ABSTRACT

Title of Dissertation:	**Framing Influence:** How Russian Trolls Use Morality and Emotion to Mimic U.S. Political Discourse
	Mathew Aaron Feehan, Doctor of Philosophy, 2026
Dissertation directed by:	Dr. Carol Graham, School of Public Policy

Efforts to counter foreign influence on social media increasingly assume that harmful activity can be identified and mitigated through automated analysis of language. This dissertation shows why that assumption is flawed. Analyses of Russian Internet Research Agency activity on Twitter demonstrate that the rhetorical features most strongly associated with foreign influence substantially overlap with legitimate domestic political speech, making content-based detection analytically informative but operationally unsafe for enforcement.

Based on these findings, this dissertation argues that policy responses should not rely on content-only detection or automated moderation to counter foreign influence. Instead, effective responses should prioritize transparency, attribution, and structural platform interventions that reduce the reach and impact of coordinated foreign influence behavior without requiring fine-grained judgments about the permissibility of individual messages. These recommendations reflect the high false-positive risks associated with content-based enforcement under realistic prevalence conditions and the corresponding normative and legal costs of suppressing lawful political expression.

Russia's information strategy integrates cyber operations with psychological influence to shape foreign publics and political processes. Against this backdrop, this dissertation examines the rhetoric of Russian influence campaigns using content-only features and interprets the results through a comparative U.S.–EU policy lens.

Study 1 analyzes engagement in Russian-attributed tweets and estimates the independent and joint effects of Moral Foundations Theory language, partisan word use, and emotions. Moral framing increases engagement beyond partisanship and emotion, with Authority and Care showing the strongest positive associations, while partisan intensity and discrete emotions such as anger and fear provide additional but smaller effects. These results establish that moral rhetoric is not just correlated with, but incrementally predictive of, higher engagement in Russian messaging.

Study 2 asks what Russian messaging most closely resembles at the tweet level. Tweet-level classifiers trained on a balanced U.S. sample using Moral Foundations, emotions, and partisan intensity achieve about 51 percent accuracy and indicate that Russian tweets are most often classified as non-elite content, resembling politically engaged or random users more than politicians. A tweet-level TF–IDF model using a 10,000-term unigram vocabulary (the 10,000 most frequent words in the U.S. training corpus, each represented as a separate feature) performs substantially better on the held-out U.S. test set for the same three-class author-type task (accuracy ≈ 67 percent, macro-F1 ≈ 0.67). When this model is applied to Russian tweets, it assigns roughly 46 percent to the random-user class, 49 percent to the engaged-user class, and 5 percent to the politician class. Together, these tweet-level results indicate that Russian messages occupy a broad, non-elite band of civic discourse: they

exhibit moral framing and institutional vocabulary associated with domestic political conversation, yet they do not consistently separate from highly engaged U.S. users on content alone.

Study 3 aggregates the same content-only features to the account level and evaluates whether they can support prevalence-aware intervention thresholds, that is, thresholds that take into account how rare Russian accounts are in the broader population and therefore how easily detection errors can overwhelm true positives. Account-level rhetorical models using percent-present Moral Foundations Theory and emotions, alongside the mean partisan word count per tweet, achieve accuracy around 77–78 percent with macro-F1 ≈ 0.80 for U.S. accounts. An account-level TF–IDF model, a representation that converts each account's combined tweet text into a vector of term weights based on how distinctive each word is within the overall data, performs slightly better, reaching roughly 83 percent accuracy and macro-F1 ≈ 0.87. Across both models, Russian accounts consistently appear non-elite: the rhetorical model predicts approximately 55 percent random, 43 percent engaged, and 2 percent politician, while the TF–IDF model classifies about 86 percent of Russian accounts as random and 14 percent as engaged, with almost none labeled as politician. Feature comparisons show that partisan and institutional terms concentrate most heavily in elite accounts; Russian accounts use these terms less frequently than politicians but at levels that overlap with engaged users. Moral Foundations profiles also differ modestly, with Russian accounts showing slightly lower Fairness and Loyalty and somewhat higher Sanctity- and Disgust-related activation than domestic users.

In the final part of Study 3, I simplify the classifier's account-level predictions to make them suitable for policy evaluation. The model assigns each account a probability of belonging to the politician, engaged user, or random classes. To translate these into more interpretable measures, I create two composite scores. The political-style score combines the probabilities of the politician and engaged user classes and reflects how similar an account is to typical political actors. The random-style score is the model's predicted probability that an account resembles a random user. Using these measures, I evaluate policy risk under realistic base-rate conditions by estimating how many U.S. accounts would be mistakenly flagged if platforms attempted to detect Russian accounts using content-only signals.

The results show that content-only rhetorical/lexical models (as constrained here) are insufficient for enforcement thresholds. Setting thresholds to capture 70–90 percent of Russian accounts produces extremely high false-positive rates among U.S. users. Under political-style scoring, thresholds in this range flag nearly all U.S. politicians and engaged users, along with more than 70 percent of ordinary accounts. When realistic Russian prevalences are introduced at 0.1, 0.5, or 1.0 percent, the positive predictive value remains below 1 percent, meaning that hundreds of U.S. accounts would be incorrectly flagged for every Russian account correctly identified. Random-style scoring performs slightly better for political elites but still labels most ordinary and engaged users as Russian-like and yields comparably low positive predictive value. Together, these prevalence-aware results show that even reasonably accurate content-only models cannot, on their own, produce operational thresholds that avoid sweeping up large amounts of legitimate domestic speech.

The concluding comparative case study interprets these patterns under U.S. constitutional constraints and the European Union's regulatory approach. In the EU, the Digital Services Act framework emphasizes systemic risk assessment (ongoing evaluation of platform-wide risks such as disinformation), transparency obligations (requirements that platforms disclose how algorithms work and how content is moderated), and researcher access (mandated data access for external auditors). These tools focus on platform processes rather than individual pieces of content. Feasible adaptations for the U.S. context include using rhetorical models for offline analysis and investigative triage and combining content-based features with structural or behavioral indicators, such as coordination patterns, bot-like posting rhythms, provenance metadata, or network ties, that provide more reliable signals than text alone. By contrast, direct enforcement based solely on rhetorical style presents a high risk of over-capture, meaning that legitimate domestic accounts, particularly politically active or morally expressive users, would be swept up alongside the very small number of Russian accounts. This risk is amplified by low base rates for foreign actors and by First Amendment protections for the right to receive information, which make content-based restrictions especially difficult to justify.

In sum, these findings show how moral, emotional, and partisan rhetoric shape engagement in Russian influence campaigns; how Russian tweets and accounts resemble the communication patterns of non-elite domestic users; and why effective policy responses must pair rhetorical insights with additional signals and institutional safeguards to mitigate harm without suppressing lawful political expression. All analysis code for this study, including data-processing pipelines, modeling scripts, and figure generation, is available in a GitHub repository located at https://github.com/mfeehan/PhD-Code.

Framing Influence: How Russian Trolls Use Morality and Emotion to Mimic U.S. Political Discourse

by

Mathew Aaron Feehan

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2026

Advisory Committee:
College Park Professor, Dr. Carol Graham, Chair
Associate Professor, Dr. Cody Buntain
Associate Research Professor, Dr. Charles Harry
Associate Professor, Dr. Niambi Carter
Associate Professor, Dr. Susannah B.F. Paletz, Dean's Representative

Acknowledgements

First and foremost is my wife, Kathy, who never gave up on her belief that I could finish this degree. She endured countless hours of complaints, theories, and being my sounding board for most of the last decade. Her confidence in me and limitless love made the long hours and difficult stretches possible. I am also grateful to my chair Dr. Carol Graham for her thoughtful feedback and the clarity she brought to both the theoretical and empirical dimensions of this work. I sincerely appreciate the advice and support from my full dissertation committee, Dean's Rep Dr. Susannah B.F. Paletz, Dr. Charlie Harry, and Dr. Niambi Carter for their time, attention, and constructive engagement with this project. I would like to especially express gratitude to my unofficial advisor, Dr. Cody Buntain, without whom I'd have never been able to pull this off. His mentorship and guidance put me back on track. I am further indebted to Dr. Robert Sprinkle, whose early confidence in my potential and personal advocacy were instrumental in my admission to the program. I also wish to recognize Dr. Tiffany Ford and Dr. Clifford Cottrell, my cohort classmates, for motivating me and just being so kind while keeping me moving forward. I would be remiss if I did not also acknowledge my current and former employers in academia. I am only here because in 2015 my Dean at National Defense University, Dr. Mary McCully, told me to "go get my doctorate". Further the School of Public Policy here at UMD hired me extensively as an adjunct, solidifying my desire to be a professional college teacher. And finally, my bosses at my current posting at National Intelligence University who have been so supportive. There is no way I could have done this alone, luckily, I didn't have to make the journey solo.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Russia's Concept of Information Confrontation (Informatsionnoye Protivoborstvo)

Russian military strategy places significant emphasis on *information confrontation*, or "informatsionnoye protivoborstvo" (Galeotti, 2016) (IPb), which integrates both technical cyber operations and psychological warfare. This method of conflict, particularly relevant in the context of hybrid warfare, enables Russia to influence perceptions and behaviors across political, economic, military, and cultural domains. IPb focuses on two primary components: *informational-technical* operations, akin to cyberattacks, and *informational-psychological* efforts designed to manipulate public opinion and the psychological state of both domestic and foreign audiences (Galeotti, 2016).

On the technical side, Russia's cyber warfare tactics include network defense, offensive cyberattacks, and information exploitation. State-sponsored or proxy groups often carry out these operations. A notable example is the cyberattack on Estonia in 2007, one of the earliest demonstrations of state-level cyber warfare (Connell & Vogler, 2017). These cyber operations typically encompass Distributed Denial of Service (DDoS) attacks that overwhelm a network with repeated and spurious connection requests, cyber espionage that attempts to steal US government and private information, and the manipulation of information for strategic purposes through targeted influence campaigns. The primary objective here is not just disruption but shaping the information landscape in favor of Russia's geopolitical aims (Rid, 2012).

Complementing these technical operations are psychological measures that manipulate public discourse. Russia has utilized bots, trolls, and state-controlled media like RT and Sputnik to spread disinformation, promote false narratives, and destabilize the informational environment.

These efforts are aimed at creating confusion, sowing discord, and discrediting adversaries, often focusing on weakening societal trust in institutions (Giles, 2016). For example, during the 2016 U.S. presidential election, Russian operatives were believed to be responsible for releasing hacked materials to create political scandals, a tactic known as cyber-enabled psychological operations (Darczewska, 2014).

The Russian approach to information confrontation is deeply rooted in Soviet-era psychological operations but has adapted to the digital age. Russia views information warfare as a critical element of hybrid warfare, allowing it to weaken adversaries without engaging in direct military conflict. Unlike traditional warfare, IPb is employed consistently, even during peacetime, as part of Russia's long-term strategy to gain strategic advantage and influence international political outcomes (Galeotti, 2016).

Several tools are central to Russia’s information confrontation strategy, blending technical cyber capabilities with psychological operations:

- **Hacktivists**: Co-opted by Russian intelligence agencies, these groups offer plausible deniability, masking state involvement in operations that are presented as grassroots cyber activism (Connell & Vogler, 2017).
- **CyberBerkut**: One example of a group used for cyberattacks, CyberBerkut has been linked to operations targeting Ukrainian governmental systems, blending technical attacks with psychological propaganda efforts aimed at undermining Ukrainian stability (Darczewska, 2014).
- **Troll Farms**: Managed by entities like the Internet Research Agency, troll farms flood social media platforms with content that promotes Russian narratives. These operations are designed to overwhelm the informational environment, making it difficult for opposing views to gain traction (Giles, 2016).
- **Bots**: Automated social media accounts, or bots, are used to amplify specific messages, drive particular narratives, and drown out dissenting opinions. These bots are strategically deployed to promote confusion and spread disinformation (Thornton, 2015).

Russia's strategy of information confrontation demonstrates a sophisticated use of both cyber and psychological tactics to achieve its geopolitical objectives without direct military engagement. By blending technical and psychological methods, Russia can exert significant influence over the global information landscape, shaping perceptions and manipulating public opinion while maintaining plausible deniability. Understanding this dual-faceted approach to information warfare is crucial for policymakers tasked with countering the influence of Russian cyber operations and disinformation campaigns.

Although this dissertation's empirical analyses focus on Russian influence operations, the broader strategic logic of cyber-enabled influence is not uniquely Russian. Public threat intelligence and U.S. government assessments document that other adversarial states, including China, Iran, and North Korea, also combine cyber activity with information effects to shape narratives, exploit polarization, and erode institutional trust. A major inflection point in U.S. policy awareness was Mandiant's APT1 report, which highlighted the strategic role of state-linked Chinese cyber operations against U.S. targets (Mandiant, 2013). ODNI has likewise assessed that state-aligned messaging and coordinated influence activity have targeted issues including Taiwan, COVID-19, and American democratic processes (U.S. ODNI, 2021). Public enforcement actions and incident reporting further document Iranian and North Korean information operations that amplify social division and discredit institutions (Department of Justice, 2020; FireEye, 2020). This wider threat environment reinforces why Russia's "information confrontation" provides a useful empirical case for studying how rhetorical mechanisms function at scale and how governance constraints shape feasible responses.

1.2 The Evolution of Information Confrontation in the Digital Age

The Internet, and most social media apps, and websites are free. Generally, we are not charged money for the websites that we visit or the search engines we use like Google, Yahoo!, and Bing; but in fact, they are not free. We pay with our data. Search engines know what we search for, and they sell that information to advertisers who in turn sell it to websites. Consider that this is done through software immediately and requires no human interaction. Several companies, including the largest social media companies, build profiles of their users for advertising purposes (Viktoratos & Tsadiras, 2021; Chen & Yang, 2021).

The advent of algorithmic, machine-learning–driven systems with the capacity not only to predict the actions, intents, and beliefs of humans but to do so almost instantaneously and then monetize its predictions by feeding them immediately by prearrangement to merchants, and others, has facilitated, even foreordained, the balkanization of society's thoughts and beliefs, making us more tribal by reflecting ideas and attitudes for which we have already shown preferences (Viktoratos & Tsadiras, 2021; Henderson et al., 2021).

Specifically, because this same technology that allows search engines to sell individual preferences for commercial goods is also employed in social media applications. Social media feeds us what we want to see as well (Henderson et al., 2021). The ability of machines to analyze and accurately predict human behavior is becoming more efficient and more widely used in evermore Internet applications. As in the case of Cambridge Analytica, this technology was used for wider and wider activities that diverged from what Internet pioneers had promised—the free and open exchange of ideas—onto a narrowing path toward self-interested manipulation of

political and moral thought (Viktoratos & Tsadiras, 2021; Turner Lee, 2018). Not only do these tools accurately predict preferences, but they can also be used to influence people in the same way that TV commercials have been doing for decades. The use of popular media stars to sell cars, beer, or anything has long been a staple in the industry. We know from a psychological standpoint that positive feelings created by one thing can be transferred to another thing. By understanding likes and dislikes an application can present ideas, images, or concepts likely to be received favorably while slowly introducing other less familiar concepts (Fennis & Stroebe, 2021). This process is the same for selling soft drinks and for shaping social media feeds.

The 2016 election highlighted what would become an example of this phenomenon moving forward. A research analytic company entitled Cambridge Analytica purchased from Facebook the account profiles of some 50 million U.S. citizens. These accounts and specifically their profiles were then used by the Trump Campaign to use Facebook to influence voter behavior. All of this was done without the users' knowledge or permission. This event caused a great public outcry, and Facebook CEO Mark Zuckerberg was called to testify before Congress, which was threatening to enact legislation to monitor how user data is collected and exploited by Facebook and other social media companies. Subsequently, Congress took no action (Turner Lee, 2018).

Russia's strategy of information confrontation demonstrates a sophisticated integration of cyber and psychological tactics aimed at manipulating perceptions, exploiting emotional triggers, and reshaping political narratives globally. The effectiveness of these operations largely depends on their ability to resonate emotionally and morally with targeted audiences. This reliance on emotional and moral resonance underscores the relevance of Moral Foundations Theory as a framework for analyzing Russian influence efforts.

1.3 Moral Foundations Theory (MFT) and Information Influence

Graham, Haidt, Koleva, Motyl, Iyer, Wojcik, and Ditto (2013) present an extensive exploration of Moral Foundations Theory (MFT) and its relevance in understanding moral pluralism. Their work delves into how human morality is structured around multiple foundational values that evolved in response to adaptive challenges. This multifaceted framework challenges the traditional belief that moral judgment is a rational, deliberative process. Instead, it argues that moral intuitions, shaped by both evolutionary and cultural influences, play a pivotal role in our moral reasoning.

The authors take readers through the historical landscape of moral psychology, critiquing earlier theories such as utilitarianism and deontology for focusing too narrowly on harm and fairness as the sole dimensions of moral judgment. In contrast, MFT introduces a broader range of moral concerns, encompassing care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, sanctity/degradation, and liberty/oppression. This inclusive approach reflects how diverse cultures navigate moral landscapes, utilizing these foundational values to resolve complex issues.

The evolutionary underpinnings of these moral foundations suggest that each arose as a response to recurrent social challenges. For example, the care foundation is seen as a response to the need to nurture and protect offspring, while loyalty evolved to promote group cohesion in the face of external threats. By connecting these foundations to evolutionary processes, the authors argue for their universality, though they acknowledge cultural differences in how these values are emphasized.

One of the most compelling insights of this theory is how political ideology influences moral priorities. Empirical evidence demonstrates that liberals typically prioritize care and fairness,

while conservatives tend to draw on a wider range of foundations, including loyalty, authority, and sanctity. This difference in moral focus helps explain the deep ideological divides observed in modern societies, with each side appealing to distinct sets of moral intuitions.

MFT as a moral theory is not without controversy. Critics take two primary positions. First that the evolutionary basis for MFT is flawed or more precisely that MFT is not innate but rather learned as a part of adolescence (Bretl & Hansen, 2022). Further, that MFT is largely a cultural phenomenon as evidenced by varying degrees of the prevalence or adherence to MFT across different cultures (Xu, Cropp, & Cameron, 2023). Secondly, that MFT as a moral theory lacks coherence with better competing moral theories (Curry, 2019). While these criticisms are valid it is not the goal of this research to argue for the veracity of MFT as a coherent moral theory but rather to use the theory to explore the relationship between the predictive power of emotion versus morality in influence campaigns. There is no question that MFT stands as a moral theory even though there may be better theories or the origins of the moral foundations may be more cultural than innate (Suhler & Churchland, 2011). MFT is chosen for this study because of its implications for political discourse. MFT can serve as a valuable tool for understanding political communication, bridging ideological divides, and resolving moral conflicts. Ultimately, the authors of the theory argue for the pragmatic validity of moral pluralism, recognizing the legitimacy of multiple moral perspectives in real-world decision-making. This aspect of the theory makes it applicable to the study of influence. If Russians and other actors intend to have an impact their messages must be well received and packaged in a familiar format. Further, MFT argues that liberals and conservatives value certain moral foundations differently and tend to characterize political and social issues through the lens of MFT. These aspects make the theory an excellent method to influence social media messaging. Understanding how moral

values differ between political ideologies also provides insight into the broader landscape of political partisanship. Russian influencers deliberately exploit U.S. political partisanship to amplify societal divisions, undermining trust in democratic institutions (Posard et al., 2020)

1.4 Political Partisanship and Polarization in the United States

Mann and Ornstein (2012) argue that the dysfunction in American politics arises from heightened polarization, which they attribute predominantly to an increasingly tribal and ideologically rigid Republican Party. The authors describe how this political extremism has led to repeated crises, such as the national debt standoffs and government shutdowns. They emphasize that the Republican Party's refusal to compromise has undermined traditional norms of governance, destabilizing the American political system. In his work, Rauch (2021) highlights the role of declining trust in traditional knowledge institutions as a driver of polarization. He notes the rise of partisan media and platforms, such as Fox News, that cater to ideological echo chambers, contributing to misinformation and the erosion of a shared factual foundation in public discourse. Both Mann and Ornstein (2012) and Rauch (2021) discuss the implications of these dynamics for governance. They suggest that the decline in bipartisan cooperation and the rise of ideological rigidity has rendered the United States increasingly incapable of addressing complex challenges. Mann and Ornstein specifically criticize the asymmetric polarization within the two-party system, while Rauch focuses on the societal and epistemic consequences of information silos and intellectual intolerance. Understanding the dynamics of U.S. political partisanship and how polarization can be exploited by external actors highlights the critical importance of strategic communication and public diplomacy efforts. Historically, the United States recognized the need to proactively address ideological threats.

1.5 U.S. Efforts to Counter Russian Influence and Disinformation

The United States has a long history of treating foreign propaganda as a national security problem while remaining deeply reluctant to centralize control over information. During the Cold War, the United States Information Agency (USIA) functioned as the information arm of U.S. foreign policy, coordinating broadcasting, cultural diplomacy, and public diplomacy campaigns aimed at countering Soviet influence. In *Inventing Public Diplomacy: The Story of the U.S. Information Agency*, Wilson Dizard traces how USIA used tools such as Voice of America, cultural exchanges, and international broadcasting to present U.S. policies and values abroad and to contest Soviet narratives, particularly in Europe, the developing world, and contested ideological spaces (Dizard, 2022). VOA's multilingual broadcasts provided relatively uncensored news to audiences behind the Iron Curtain and elsewhere, offering an alternative to state-controlled media and illustrating how information policy was woven into broader foreign policy. The agency's dissolution in 1999 reflected a post–Cold War judgment that such centralized public diplomacy was no longer critical. Dizard and others have argued that this decision left a structural gap in U.S. foreign information capabilities precisely as new forms of information warfare like cyber operations, social media manipulation, and covert funding networks were emerging.

Post-2016 responses to Russian information operations have been more fragmented. The Department of Homeland Security's Disinformation Governance Board, stood up in 2022 inside DHS, was intended to coordinate how the department understood and responded to foreign disinformation threats, including Russian interference and border-related narratives. The board never reached operational maturity. It was immediately framed by critics as an Orwellian "Ministry of Truth," with controversy centering on its director, Nina Jankowicz, and confusion

about whether the board would police speech rather than coordinate internal risk assessments. Under intense political pressure, DHS paused the board's work within weeks and formally disbanded it in August 2022, acknowledging that the initiative had become untenable (Swan & Lippman, 2022). The episode illustrates how easily an effort to coordinate federal understanding of disinformation can be turned into a partisan symbol and dismantled.

Congressional action has followed a similar pattern of concern without coherent institutional design. The RESTRICT Act (S.686), introduced in March 2023, would authorize the Secretary of Commerce to review and prohibit certain transactions involving foreign-owned information and communications technologies, with apps such as TikTok as the most visible targets (Congress.gov, 2023). The bill's scope is technology- and transaction-focused rather than message-focused: it aims at ownership, data flows, and systemic ICT risks, not at particular narratives or accounts. Civil-liberties groups and some lawmakers criticized the proposal as overbroad and potentially speech-chilling, arguing that its vague language could sweep in circumvention tools like VPNs and leave too much discretion to the executive branch. The RESTRICT Act ultimately did not advance out of committee in the 118th Congress, but political pressure over TikTok and other foreign-owned platforms contributed to subsequent, more targeted legislation such as the Protecting Americans from Foreign Adversary Controlled Applications Act, which focuses specifically on adversary-controlled apps. These initiatives share a common feature: they address foreign influence through ownership and infrastructure, not through content-based rules.

Against this background, there is no single federal agency tasked with managing disinformation writ large. The United States remains constrained by the legacy of Smith–Mundt and its

modernization. The original Smith–Mundt Act of 1948 limited the domestic dissemination of materials produced by U.S. foreign information programs, reflecting a concern about federal "propaganda" being turned inward (U.S. Congress, 1948). The Smith–Mundt Modernization Act of 2012 relaxed those constraints by allowing materials produced for foreign audiences to be made available domestically upon request and through the U.S. Agency for Global Media (USAGM), but it did not create a domestic disinformation agency or authorize direct influence operations against U.S. audiences (U.S. Congress, 2012). In practice, U.S. efforts to counter Russian and other foreign influence remain dispersed: elements of State, DHS, DOJ, the intelligence community, and DoD each hold pieces of the problem, and the gap left by USIA has not been filled by any comparable institution.

1.6 First Amendment Challenges to Countering Disinformation

Any attempt to build a more robust U.S. response to foreign disinformation runs immediately into First Amendment doctrine. Under current Supreme Court precedents, the First Amendment protects not only the right to speak but also the **right to receive** information and ideas. In cases such as *Martin v. City of Struthers* (1943), *Lamont v. Postmaster General* (1965), *Stanley v. Georgia* (1969), and *Virginia State Board of Pharmacy v. Virginia Citizens Consumer Council* (1976), the Court has held that government efforts to block or filter lawful information can violate the rights of listeners and readers, even when the content is controversial, offensive, or politically sensitive. Applied to the context of foreign influence, this line of cases means that the federal government has very limited authority to prevent U.S. citizens from accessing lawful foreign content solely because it is foreign or propagandistic.

Within this framework, the main statutory tool for addressing organized foreign influence is the Foreign Agents Registration Act (FARA) of 1938. FARA requires individuals and entities acting as "agents of foreign principals" in the United States and engaging in certain political or public-relations activities to register and disclose their relationship, activities, and funding. It does not prohibit speech; it imposes a transparency obligation on those who speak or act on behalf of foreign principals. A recent example is the 2024 Tenet Media case. The Department of Justice indicted two employees of the Russian state media outlet RT, Kostiantyn Kalashnikov and Elena Afanasyeva, alleging that they covertly funneled approximately $10 million through a U.S. media company to pay prominent social media influencers to disseminate pro-Russian and anti-Ukrainian narratives without disclosing the Russian origin of the funding (U.S. Department of Justice, 2024; "2024 Tenet Media Investigation," 2024). The indictment focuses on money laundering and FARA violations by the RT employees and their shell entities; the U.S. influencers who were paid to produce content have not been charged, because accepting money and speaking, without knowledge of the foreign principal, does not itself violate FARA or other speech laws.

The Tenet Media case illustrates both the strength and the limits of the U.S. legal response to foreign information operations. The government can bring criminal charges against covert foreign agents who launder money and fail to register under FARA, seize assets, and expose networks. It cannot easily punish domestic speakers who unknowingly amplify foreign narratives, nor can it broadly prohibit access to foreign propaganda without colliding with the right to receive information. As a result, the federal toolkit against Russian and other state-sponsored disinformation is heavily skewed toward ex post enforcement against undisclosed foreign agents, not toward ex ante filtering or blocking of content. This tension between a

constitutional order that protects access to information and a security environment in which foreign actors exploit that openness is a central theme for the policy analysis in Chapter 6.

1.7 Key Terms and Definitions

This section defines recurring terms used throughout the dissertation so that readers can interpret the study overview in section 1.8 and the empirical chapters consistently.

Term	Definition
Moral Foundations Theory (MFT)	A framework positing that human moral reasoning is structured around several innate foundations—Care, Fairness, Loyalty, Authority, Sanctity, and Liberty—that shape how individuals interpret political and social issues.
Authority	An MFT category capturing respect for hierarchy, tradition, and legitimate leadership; reflects language emphasizing rules, order, and social structure.
Loyalty	An MFT category indicating group commitment, coalition identity, and allegiance; signals include language about teams, allies, betrayal, and unity.
Sanctity	An MFT category tied to purity, contamination, taboo, and moral elevation; often appears in discourse about immorality or disgust.
Care	An MFT foundation reflecting compassion, concern for harm, and protection of vulnerable individuals; appears frequently in empathetic rhetoric.
Fairness	An MFT foundation associated with justice, equality, proportionality, and rights-based reasoning; central to debates about equity and social responsibility.
NRC Emotion Lexicon	A lexicon-based tool that categorizes words into eight emotions—anger, fear, sadness, joy, trust, disgust, surprise, anticipation—to quantify emotional tone in text.
TF–IDF (Term Frequency–Inverse Document Frequency)	A numerical representation of text in which each term is weighted by how often it appears in a document (TF) and how rare it is across the corpus (IDF). Highlights distinctive vocabulary and reduces noise from common words.
Unigram Vocabulary	A representation that treats each individual word as a feature, without accounting for word order or grammar; used in TF–IDF vectorization.
Vectorization	The process of converting raw text into numerical feature vectors so machine-learning models can process it; TF–IDF is one example.

Term	Definition
XGBoost	A gradient-boosted decision-tree algorithm optimized for speed and performance, well suited for high-dimensional text classification tasks.
Baseline Model	The simplest or reference specification against which alternative models are compared to assess performance improvements.
Ablation Model	A model variant in which specific features (e.g., partisan words) are removed to test their independent contribution to predictive accuracy.
Confusion Matrix	A table summarizing classification performance by comparing predicted versus actual labels; identifies where the model succeeds and where classes are confused.
Precision	A performance metric indicating the proportion of predicted positives that are true positives; reflects how often the model is correct when it flags a class.
Recall	A performance metric indicating the proportion of true positives the model successfully identifies; measures sensitivity.
F1 Score	The harmonic mean of precision and recall; useful when dealing with imbalanced datasets.
Macro-F1	The unweighted average of F1 scores across all classes; treats each class equally regardless of size.
Weighted-F1	An average of F1 scores weighted by class size; reflects the performance of larger classes more strongly.
Feature Importance	A measure (common in tree-based models) indicating how much each input feature contributes to a model's predictions. Higher values indicate stronger influence.
Account-Level Aggregation	Combining all tweets from an account into a single feature profile (e.g., mean partisan word count or percent-present MFT categories) to classify behavior at the account level rather than at the tweet level.
KWIC (Key Word in Context)	A text-analysis method that extracts a word along with its surrounding context (e.g., ±3 words) to examine how terms are used semantically; helps interpret top TF–IDF features.
Partisan Word Count	A measure of how many politically charged or ideologically relevant terms appear in a tweet or account; used to assess explicit political signaling.
Intervention Threshold	A score or probability cutoff above which an account or message would be flagged for further review in a detection system; lower thresholds maximize recall, higher thresholds maximize precision.
Prevalence-Aware Evaluation	A model evaluation approach that accounts for how rare a target class is in real-world populations; emphasizes PPV (precision) because false positives dominate when prevalence is low.

Term	Definition
Lexical Register	The characteristic vocabulary associated with a particular rhetorical style or discourse community (e.g., elite political speech, civic discourse). Useful for distinguishing between account types.
High-Capacity Classifier	A model capable of capturing complex, high-dimensional patterns due to its flexibility and number of parameters; examples include TF–IDF paired with XGBoost.
Aggregated Features	Features that summarize multiple observations—such as average emotion scores across all tweets—rather than analyzing individual messages; used primarily in account-level models.

1.8 Dissertation Structure and Study Overview

Russia's doctrine of "information confrontation" integrates technical cyber activity with psychological influence to shape foreign publics and political processes (Galeotti, 2016). On social media, these operations often blend into ordinary political conversation while advancing strategic aims. This dissertation examines that rhetoric empirically and then interprets the results through a comparative U.S.–EU law and policy lens, focusing on content-only signals rather than structural, network, or audience-size features.

The first study analyzes engagement mechanisms in Russian influence tweets and addresses Research Question 1 (RQ1) and Hypothesis 1 (H1). RQ1 asks how Moral Foundations Theory (MFT) language, emotion categories, and partisan vocabulary are associated with engagement in Russian IRA tweets. H1 proposes that tweets using MFT-based moral language will be associated with higher engagement than tweets relying primarily on partisan rhetoric or emotionally charged speech. Using a consistent preprocessing pipeline that extracts Moral Foundations scores (via the Mformer classifier), emotion categories, and partisan language measures from each tweet, Study 1 estimates how these rhetorical signals relate to engagement outcomes using regression models.

The second study asks what Russian messaging most resembles at the tweet level and addresses Research Questions 2–4 (RQ2–RQ4) and Hypotheses 2–4 (H2–H4). RQ2 asks whether Russian IRA tweets more closely resemble the rhetorical style of U.S. politicians, politically engaged users, or random users when relying on content-only tweet-level features. H2 predicts that Russian tweets will generally resemble non-elite, politically engaged users more than political elites and will avoid the densest partisan signals. RQ3 asks which rhetorical feature families—Moral Foundations, emotions, and partisan vocabulary—contribute most to distinguishing these groups. H3 predicts that partisan intensity will be the dominant discriminating signal, followed by moral foundations and then emotion. RQ4 asks whether a theory-agnostic lexical baseline (TF–IDF), trained on the same U.S. tweet-level classification task, produces the same substantive placement of Russian tweets across the three U.S. classes as the engineered rhetorical models. H4 predicts that TF–IDF will corroborate the engineered-feature results by recovering the same overall pattern of Russian tweet placement, indicating that the core resemblance finding is not an artifact of curated feature selection. To answer these questions, Study 2 trains a sequence of interpretable tweet-level classifiers on U.S. messages, including engineered rhetorical models and a theory-agnostic TF–IDF lexical baseline, and then applies them to Russian tweets to infer alignment with each U.S. user type.

The third study moves to the account level, revisiting RQ2–RQ4 under aggregation and directly addressing Research Question 5 (RQ5) and Hypothesis 5 (H5). RQ5 asks whether any content-only account-level profiling, using either aggregated engineered rhetorical features or aggregated TF–IDF lexical representations, can support prevalence-aware intervention thresholds given the rarity of Russian influence accounts, without substantial over-capture of legitimate U.S. political speech. H5 predicts that aggregation to the account level will improve stability and classification

performance relative to tweet-level models but that prevalence-aware evaluation will still show that content-only profiling yields low positive predictive value at operationally useful recall, making it insufficient as a standalone basis for enforcement decisions such as removal. To do so, Study 3 aggregates the same moral, emotional, and partisan measures across account histories, compares engineered-feature models to TF–IDF at the account level, and then evaluates screening thresholds under realistic base rates using false-positive rates and positive predictive value. This design distinguishes two questions that are often conflated: whether content-only models can characterize where Russian accounts sit within domestic discourse, and whether those same signals remain precise enough under realistic prevalences to support intervention.

While the empirical studies focus on Russian IRA messaging, the comparative legal-policy analysis generalizes to other state actors that employ cyber-enabled influence tactics. The concluding case study interprets the findings of Studies 1–3 against U.S. doctrine and European Union practice. In the United States, strong speech protections and fragmented platform governance constrain content-based interventions; in the European Union, instruments such as the Digital Services Act enable systemic risk assessments, transparency obligations, and process-based remedies. Using the answers to RQ1–RQ5 and H1–H5, the case study evaluates which measures are feasible, rights-preserving, and likely to reduce harm under each regime, and where interventions risk over-capture of legitimate political speech. Together, the three empirical studies and the comparative analysis provide an evidence-based foundation for targeted, legally viable responses to foreign influence in democratic information ecosystems.

Chapter 2: Related Work

2.1 Moral Foundations and Digital Influence Campaigns

Moral Foundations Theory (MFT) provides a framework for understanding how moral appeals shape engagement by linking moral language to the intuitive judgments that structure political and social evaluation. Developed most prominently by Haidt and colleagues, MFT argues that moral reasoning is often driven by fast, affect-laden intuitions, with post hoc justification following rather than preceding moral judgment. Within this approach, moral disagreement is not simply disagreement over facts, but disagreement over which moral concerns are being activated and prioritized. Haidt and colleagues (2013) identify five primary moral dimensions, care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation that capture recurring patterns in how individuals evaluate moral situations and social order. These foundations are commonly grouped into "individualizing" foundations (Care and Fairness), which emphasize protection of individuals from harm and injustice, and "binding" foundations (Loyalty, Authority, and Sanctity), which emphasize group cohesion, respect for legitimate hierarchy, and protection of sacred norms. Haidt and colleagues argue these foundations are especially relevant because moral language is a reliable mechanism for mobilization: it compresses complex disputes into intuitive evaluations, signal's identity, and group membership, and encourages sharing and reinforcement within like-minded communities.

Although Moral Foundations Theory was developed and validated primarily in Western contexts, later research has shown that MFT constructs can meaningfully relate to political and social attitudes in non-Western samples. Sychev and Protasova (2020), for example, examine MFT in a Russian youth sample and show that moral foundations are systematically associated with attitudes toward economic inequality and related social beliefs. In their results, support for

equality is most strongly associated with the individualizing foundations, especially Care and Fairness, and these associations are partly mediated by broader "jungle world" beliefs about competitive social relations. Sychev and Protasova's study provides evidence that MFT measures are not limited to Western populations and can operate as meaningful attitudinal correlates in a Russian context.

Amjadi and John (2024) applied MFT to Twitter discourse during the Ukraine-Russia conflict, categorizing tweets by their moral content to explore how public sentiment evolved in response to the war. Their study found that care/harm was the most dominant foundation early in the conflict, as global audiences reacted to civilian casualties, while fairness and liberty concerns shaped later discussions about Ukraine's sovereignty. Notably, they found little evidence of sanctity-based rhetoric, suggesting that certain moral appeals may be less salient in crisis contexts.

This dissertation expands on Amjadi and John's work by applying MFT in a different setting—Russian influence operations targeting U.S. audiences. While their study focused on organic public reactions, this research examines how state-sponsored actors strategically craft moral appeals to maximize engagement. Unlike the Ukraine-Russia conflict, where users organically expressed moral concerns, my study assesses whether moral rhetoric is deliberately used as a persuasion tactic. Additionally, this dissertation explores how moral framing interacts with political ideology to assess whether certain moral appeals are more effective.

2.2 Moral Foundations in Artificial Intelligence and Media Manipulation

Abdulhai et al. (2023) explored how large language models (LLMs), such as OpenAI's GPT, reflect moral values in their outputs. Their findings suggest that LLM-generated text tends to

emphasize care/harm and fairness/cheating over loyalty, authority, and sanctity, reflecting biases in their training data. The study raises concerns that AI-generated content may disproportionately reinforce certain moral narratives while suppressing others, creating an imbalance in digital moral discourse.

This dissertation builds on Abdulhai et al.'s research by analyzing whether Russian influence campaigns exploit these same asymmetries by using emotional language. A crucial extension of this approach involves the inclusion of specific political partisan words as an additional variable to assess whether MFT-based appeals are deployed independently or in conjunction with explicit partisan rhetoric. Prior studies have demonstrated that moralized language spreads effectively in political discourse, but the relationship between partisan signaling and moral framing remains underexplored. This dissertation assesses whether MFT-based appeals alone are sufficient to drive engagement or if their effectiveness is contingent on explicit political framing. This will help determine whether moral narratives in Russian influence campaigns function as a standalone persuasive mechanism or if they gain additional traction when intertwined with partisan identity cues.

2.3 MFT in Political and Strategic Communication

Beyond social media influence campaigns, research has explored the role of MFT in political public relations and strategic messaging. Mikhaylova and Abramov (2023) examined how political public relations (PR) professionals navigate ethical dilemmas using MFT. They found that loyalty and authority were dominant moral values in political PR, often at the expense of fairness or honesty. This is particularly relevant to my research, as it suggests that political

actors, including foreign influence campaigns, may prioritize strategic loyalty over factual accuracy when crafting persuasive messages.

This dissertation extends their findings by directly testing whether Russian influence messaging mirrors the moral strategies used in PR and political campaigns. By analyzing a dataset of Russian-affiliated tweets, I assess whether their moral framing follows traditional political persuasion tactics. If Russian disinformation aligns with standard political PR methods, it suggests that Russian disinformation is an extension of conventional strategic communication, rather than a fundamentally unique form of persuasion.

2.4 European Union Counter Influence Frameworks

The European Union (EU) has developed significant regulatory efforts to counter disinformation and improve digital platform accountability, primarily through the EU Code of Practice on Disinformation and the Digital Services Act (DSA). These frameworks aim to increase transparency, regulate content moderation, and curb the spread of foreign influence operations online. While these initiatives represent some of the most comprehensive legal frameworks for addressing disinformation, they also raise questions regarding effectiveness and applicability outside the European context.

The EU Code of Practice on Disinformation, introduced in 2018, is a voluntary agreement between major platforms (e.g., Facebook, Google, X (formerly Twitter)) and the EU to combat online disinformation through increased advertising transparency, bot detection, and media literacy initiatives (European Commission, 2018). However, because compliance remains voluntary, critics argue that enforcement mechanisms are weak, making it difficult to hold platforms accountable (Wardle & Derakhshan, 2017).

The Digital Services Act (DSA), proposed in 2020, builds on these efforts by introducing legally binding obligations for online platforms, mandating stricter content moderation policies, algorithmic transparency, and risk assessments for disinformation (European Parliament, 2021). Unlike the Code of Practice, the DSA holds platforms directly responsible for addressing harmful content. Despite these advances, scholars debate whether such regulations might inadvertently curtail free speech or create enforcement inconsistencies across EU member states (Flew, 2021).

This dissertation contributes to this discussion by evaluating the applicability of the EU's regulatory approach to U.S. policy. Given that the First Amendment places strong limitations on content moderation by the government, a direct adoption of the DSA-style regulations in the U.S. is unlikely. My research examines, as a case study, what aspects of the EU's approach to countering disinformation could be adapted for use in the United States and what legal and policy changes would be necessary for such an adoption.

2.5 Influence Operation Detection

A substantial body of research has developed and evaluated detection approaches for coordinated influence campaigns, spanning text-based signals, coordination and network-based signals, and hybrid systems that combine multiple data sources. This dissertation does not attempt to advance state-of-the-art detection performance or to benchmark against leading operational systems. Instead, it evaluates a deliberately constrained, content-only approach based on rhetorical and lexical features derived from Moral Foundations Theory, emotion measures, and partisan language. The goal is to test whether MFT features add meaningful explanatory value for influence-message engagement and whether these content-only features improve performance in the dissertation's classification tasks, while using prevalence-aware evaluation (including

positive predictive value, PPV) to assess what model outputs would imply if translated into policy-relevant or operational thresholds. Accordingly, the purpose of reviewing influence operation detection methods here is to clarify what the field already knows, where measurement and governance constraints remain unresolved, and how this dissertation is positioned within that landscape. The literature makes clear that many effective detection approaches rely on behavioral and network signals, such as coordination patterns across accounts, that can be stronger than content alone for identifying organized campaigns. This dissertation therefore treats content-only rhetoric as an intentional constraint and evaluates feasibility using prevalence-aware risk, including PPV, to make explicit how false positives scale under realistic base rates given the feature families used in this work.

The remainder of this section reviews representative work across major detection families and identifies the specific gaps this dissertation addresses. I incorporate several key influence detection papers to anchor the discussion, and then connect those approaches to the empirical design of Studies 1 through 3 using citations already introduced elsewhere in the dissertation. The objective is to frame the dissertation's research questions, modeling choices, and conclusions as contributions to disinformation measurement and governance, rather than as contributions to psycholinguistics or algorithm development for its own sake.

2.5.1 Content-based Detection Under Platform-agnostic Constraints

A useful anchor for understanding what content-only signals can support is the content-based detection work in *Science Advances* by Alizadeh et al. (2020). Rather than treating detection as a single static classification problem, Alizadeh et al. (2020) evaluate whether "industrialized"

influence content can be distinguished from organic activity using public features that are designed to generalize across time and settings. Their design is notable for the way it structures detection as multiple prediction tasks that vary difficulty and realism, including within-period detection, forward prediction to later periods, detection of previously unseen accounts, and cross-platform tests. The core implication is that influence activity can leave measurable content-linked signals that persist across time and across accounts, and that the difficulty of detection varies by campaign and context, not just by model choice. This is methodologically important for later studies because it treats detectability as an empirical property that can fluctuate as operators adapt and as political contexts shift, rather than as a one-time estimate of performance.

Alizadeh et al. (2020) also help clarify what "content-based" often means in practice. In many applied detection pipelines, content-based systems do not rely solely on rhetorical style or lexicon counts. They frequently incorporate publicly visible behaviors adjacent to content, such as linking patterns, mention behavior, and other metadata-like traces that are visible in the content stream. Their work therefore provides a concrete baseline for the broader detection literature: content-only approaches can be operationally useful under some conditions, but their performance is data-regime dependent and context sensitive, and they exist alongside stronger coordinated-action approaches that exploit network structure when available. For a dissertation that explicitly constrains itself to rhetorical signals, this study is valuable not because it sets a benchmark to beat, but because it establishes that content-based detectability can be systematically studied and that detection claims must be contextualized within a broader methodological landscape (Alizadeh et al., 2020).

A central stream of influence-operation detection research uses content and other publicly observable traces to characterize state-linked accounts and distinguish them from baseline users. This literature is important for two reasons. First, it shows that influence activity can leave systematic, measurable signals that allow researchers to compare state-linked actors to ordinary users, but it also makes clear that "content-based" detection frequently incorporates behavioral or platform-adjacent traces, such as linking and interaction patterns, rather than relying solely on rhetorical style (Zannettou et al., 2019). Second, supervised text classification work demonstrates that interpretable baselines can separate influence-linked content from organic content and can detect ideological stratification within troll output, even when the objective is characterization rather than operational deployment (Chun et al., 2019). Recent work also advances this supervised tradition by explicitly prioritizing explainability as a design goal for identifying state-sponsored agents, emphasizing outputs that are legible for human review rather than purely optimized for accuracy (Tian et al., 2025). Related work on suspended or enforcement-identified accounts further suggests that influence-aligned activity may extend beyond canonical "known IRA" lists, underscoring that the relevant actor ecology can be broader than the accounts publicly attributed to a single campaign (Serafino et al., 2024).

This content-based literature also supports shifting the unit of analysis from individual posts to persistent actors. Lewinski and Hasan (2021), for example, frame detection as an account classification problem using aggregated content across platforms, arguing that stable linguistic and behavioral regularities become more visible once messages are summarized across an account's history. This account-level framing is consistent with platform-agnostic reasoning because the primary signal remains content-derived, but it leverages persistence across many posts rather than treating any one message as decisive (Lewinski & Hasan, 2021).

Complementing this move, Mathew et al. (2024) treat trolling as a multi-level classification problem rather than a binary label, illustrating how supervised models can operationalize typologies that distinguish among multiple adversarial communication styles. That orientation helps normalize multi-class designs that separate heterogeneous domestic baselines before comparing them to foreign influence content (Mathew et al., 2024).

Collectively, these studies motivate the design of Study 2 in two ways. First, they establish that influence-linked accounts can be characterized and, in some settings, distinguished from baseline users using content-derived signals, but they also show that much of the strongest prior work relies on traces beyond rhetorical language. Study 2 therefore adopts a deliberately narrower test by isolating tweet-level rhetorical features to evaluate what moral language, emotion cues, and partisan language contribute when structural and coordination signals are excluded (Zannettou et al., 2019). Second, because prior work often frames detection as a binary "troll/not troll" task, Study 2 emphasizes interpretability by mapping Russian tweets into a role-typed U.S. discourse space defined by politicians, politically engaged users, and random users. This design provides a direct answer to what Russian messaging most closely resembles under a content-only constraint and sets up the later prevalence-aware feasibility analysis (Chun et al., 2019; Serafino et al., 2024; Tian et al., 2025; Lewinski & Hasan, 2021; Mathew et al., 2024).

2.5.2 Russian IRA Messaging Tactics and the Detection Relevance of "camouflage"

Complementing predictive detection studies, descriptive research provides detection-relevant baselines about how influence operators communicate. Linvill et al. (2019) examine IRA Twitter tactics during the final month prior to the 2016 U.S. presidential election and develop a qualitative typology of tweet functions that can be mapped onto operational goals. Their analysis

is particularly relevant for detection because it emphasizes that influence operations are not composed solely of overt propaganda. Linvill et al. (2019) identifies a "camouflage" category that comprises a large share of IRA messaging in their sample and describe this content as designed to build credibility, appear authentic, and cultivate audiences rather than to push explicit political persuasion.

This descriptive point matters for detection because it highlights a structural risk of content-only systems that focus on overt political keywords or obvious ideological markers: a large fraction of operational output may be intentionally non-political, locally flavored, or culturally generic in ways that blur into ordinary social media discourse (Linvill et al., 2019). Linvill et al. (2019) also identify IRA content categories oriented around undermining institutions and attacking the media, showing that distrust-focused themes can appear across ideological targets and can be integrated with more mundane content streams. For influence detection, this implies that rhetorical signals may be distributed across multiple functional modes, and that a model's sensitivity to overt political content may not capture the broader strategic repertoire of an influence campaign.

2.5.3 Cross-platform Propaganda and Link-mediated Influence Ecosystems

Influence operations do not stay on one platform. Accounts often use Twitter (or X) mainly as a distribution channel to push people to content hosted elsewhere, so understanding an operation sometimes requires looking at where posts send users, not just the words in the posts. Golovchenko et al. (2020) analyze the hyperlinks embedded in IRA tweets to assess competing interpretations of IRA strategy, including whether the IRA mainly supported a single candidate or instead tried to increase polarization by amplifying multiple sides. They find that IRA

accounts promoted links across the ideological spectrum, and they show that the destinations of those links are informative: which sites and content types an account repeatedly directs users toward can reveal strategic patterns, with some destinations and formats disproportionately tied to particular ideological segments.

A central contribution of Golovchenko et al. (2020) is the emphasis on YouTube as a major linked destination and the argument that cross-platform pathways shape how propaganda circulates. This is detection-relevant for two reasons. First, link behavior can provide a structured trace of campaign strategy that is less sensitive to the ambiguity of short-text rhetoric. Second, cross-platform linking underscores that influence detection may require attention to where content is being routed, not only what is being said in a single post (Golovchenko et al., 2020). In the detection literature, these results reinforce the idea that influence operations can be studied through cross-platform flows, which can yield signals that are distinct from, and sometimes stronger than, message text alone.

This dissertation does not directly test link-routing or cross-platform destination signals. The empirical focus is intentionally narrower: it evaluates what can be learned from content-only rhetorical features within the tweet text itself (moral language, emotion cues, and partisan language) and how those signals perform when used for engagement analysis, classification, and prevalence-aware risk assessment (PPV).

2.5.4 Audience-building, Network Infrastructure, and the Conditions for Influence

Where many detection approaches focus on identifying suspicious content or coordinated bursts, other work emphasizes that influence requires an infrastructure of attention and audience. Zhang et al. (2021) analyze how successful English-speaking IRA accounts built their followings

between 2015 and 2017, treating follower networks as both an ego network and an audience that enables diffusion. Their study highlights that influence operations can be detectable through the processes by which accounts gain visibility and legitimacy, and through their embedding in partisan media and information ecosystems. Rather than isolating influence to what trolls said, Zhang et al. (2021) emphasize the broader communication system that makes amplification possible, including the interaction between partisan networks, media coverage, and platform dynamics.

This focus on audience-building is relevant to detection because it reframes the problem from identifying individual posts to identifying the structural and social conditions under which influence becomes effective. If influence depends on routinized amplification and stable audience relationships, then detection strategies that incorporate network and media-ecosystem signals may be better suited for campaign identification than purely rhetorical classifiers (Zhang et al., 2021). At the same time, this literature also sharpens the policy relevance of content-only approaches: governance interventions often require content-facing rationales even when the most discriminative signals are behavioral or network-based. The detection literature therefore motivates careful attention to what content-only features can contribute, where they fail, and what risk is introduced when content models are used to justify threshold-based enforcement decisions.

This dissertation does not evaluate network-based audience-building signals or follower-growth dynamics as detection features. The empirical work is intentionally constrained to content-only rhetorical measures. However, the policy discussion does draw on the broader detection literature's emphasis on structural conditions by arguing that, when adversarial states

systematically exploit Western information systems, purely content-based countermeasures can be insufficient or too risky to deploy at scale. As a result, the dissertation includes policy recommendations oriented toward structural mitigation, including reducing or blocking adversarial access to Western information systems under clearly defined legal and governance frameworks, rather than relying exclusively on content-level thresholds to manage influence activity.

2.5.5 Coordination and Network-Based Detection

A major branch of influence-operation detection focuses on coordination and network signals rather than on message text alone. These approaches treat influence campaigns as collective phenomena that become visible through patterns of synchronized behavior across accounts, such as repeated co-retweeting, co-hashtagging, shared URL promotion, or temporally clustered activity that is unlikely to arise independently. The core advantage of coordination-based detection is that it can identify organized action even when individual messages appear rhetorically ordinary, politically ambiguous, or deliberately "camouflaged." In other words, where content-only methods are vulnerable to strategic mimicry, coordination methods often detect the operational structure of the campaign rather than the surface semantics of any single post.

Pacheco et al. (2021) provide a representative framework for uncovering coordinated networks on social media, emphasizing methods that identify groups of accounts whose behaviors are unusually aligned across time and interaction patterns. Their work is often used as a methodological reference point because it formalizes coordination detection as a network discovery problem rather than a conventional text classification task. Cinelli et al. (2022)

similarly examine coordinated inauthentic behavior as a diffusion and synchronization phenomenon, arguing that coordinated actors can be detected through systematic patterns of information spreading that differ from organic social dynamics. Together, these studies reinforce a key point in the detection literature: for campaign identification, network-level signals can be more discriminative than post-level linguistic features, especially when adversarial actors intentionally avoid explicit partisan language or inflammatory rhetoric.

More recent work also illustrates how behavioral and network signals can be integrated into machine-learning pipelines for influence detection. Ezzeddine et al. (2022) propose a behavioral-based approach to detecting state-sponsored trolls that emphasizes account activity traces and coordination-relevant features, reflecting the broader shift toward detection paradigms that do not depend solely on content semantics. Across this literature, the recurring implication is that the strongest detection systems are often those that combine multiple signal types, content, behavior, and network structure, while also recognizing that operational feasibility depends on data access, governance constraints, and the base-rate problem that affects any large-scale screening tool.

For the purposes of this dissertation, coordination and network-based detection research serves primarily as context: it clarifies what "state of the art" often looks like in operational settings and why content-only models face inherent limits when influence messaging is designed to resemble ordinary domestic discourse. Accordingly, the empirical chapters that follow treat content-only rhetorical signals as an intentional constraint and evaluate what moral language, emotion cues, and partisan language can support under that constraint. At the same time, this coordination literature informs the policy discussion and conclusions, where the dissertation argues that

certain countermeasures may be more appropriately pursued at the structural or network level, including measures that reduce or block adversarial access to Western information systems through clearly defined legal and governance frameworks, rather than relying exclusively on content-level thresholds to manage influence activity at scale.

2.6 Contributions of This Research

Motivated by Moral Foundations Theory and by the expectation that Russian influence operators tailor messages to multiple ideological audiences, this dissertation tests whether Russian influence rhetoric exhibits systematic moral framing patterns and whether those patterns provide interpretable, content-only signals for comparative classification and policy-relevant risk assessment and makes four contributions. First, it provides a rigorous test of whether Russian influence campaigns deploy Moral Foundations language systematically, not merely mirroring organic moral discourse around salient events, by measuring foundation use alongside partisan and emotional cues in large-scale data. Second, it estimates the incremental contribution of moral framing to engagement beyond partisanship and emotion, identifying when and how moral appeals add explanatory power for diffusion. Third, it situates Russian messaging within the broader U.S. communication ecosystem by comparing tweets and accounts to politicians, politically engaged users, and random users using interpretable, content-only classifiers validated against TF-IDF baselines; it also quantifies policy risk with prevalence-aware metrics (capture thresholds, false-positive rates, and positive predictive value) within in the framework of these models and findings. Fourth, it translates these empirical results into governance options through a comparative U.S.–EU lens, assessing which EU-style platform accountability mechanisms (e.g., DSA-type transparency and systemic-risk processes) could be adapted within U.S.

constitutional constraints to counter foreign influence while minimizing over-capture of legitimate political speech.

Chapter 3 Study 1: Engagement Mechanisms in Russian Influence Tweets

3.1 Overview and Research Questions

This study aims to investigate if Russian influence campaigns use Moral Foundations Theory (MFT) in social media messaging and if doing so increases engagement, measured through retweets as compared to emotionality and use of politically partisan speech. Prior work shows that moral language is systematically associated with political ideology and moral emphasis in

U.S. discourse (Graham et al., 2013), and recent work suggests these same constructs can function as meaningful attitudinal correlates in Russian samples as well (Sychev & Protasova, 2020). At the same time, lexicon-based emotion measures in social-media contexts face known limitations that can attenuate estimated affect effects, making it important to evaluate emotion alongside alternative rhetorical signal families (Paletz, Golonka, et al., 2024). What remains unclear is whether moral framing provides **incremental explanatory power** for engagement in Russian-attributed influence messaging once partisan intensity, discrete emotions, and audience-size differences are accounted for. This gap motivates RQ1 and H1.

Accordingly, research question 1 (RQ1) asks: To what extent does the presence of MFT-based moral language in Russian IRA tweets explain variation in engagement, relative to emotional tone and partisan vocabulary? The associated hypothesis (H1) is that messages utilizing MFT-based language will be associated with higher engagement than messages relying primarily on partisan rhetoric or emotionally charged speech. This approach is rooted in the idea that moral language resonates more deeply with audiences, driving greater interaction and engagement on social media platforms. This research contributes to the understanding of how moral appeals in influence operations can drive engagement, particularly in the context of Russian disinformation campaigns that seek to deepen political divides and sway public opinion on social media platforms.

3.2 Study Design

This study utilizes the Mformer model (Nguyen et al., 2019), a specialized large language model, to detect Moral Foundations Theory (MFT)-based language in tweets. Mformer tokenizes text to generate confidence scores for five moral foundations: authority, fairness, care, sanctity, and

loyalty. While the model also provides scores indicating the absence of these foundations ("not authority," "not care," etc.), these are excluded from analysis because they are not necessary for the study's objectives and complicate interpretation of moral-language presence.

The Mformer model was originally developed to analyze moral scenarios, making it well-suited for identifying moral language. However, it does not include liberty/oppression as a moral foundation. Although the omission of this foundation is a limitation, Nguyen (2023) suggests that liberty/oppression correlates with authority and fairness, which are included in the model. Future iterations of this research may incorporate liberty/oppression to provide a more comprehensive analysis of MFT.

To support the relevance of MFT constructs beyond Western samples, Sychev and Protasova (2020) examine MFT in a Russian youth sample and show that moral foundations are systematically associated with attitudes toward economic inequality and related social beliefs. In their results, support for equality is most strongly associated with the individualizing foundations, especially Care and Fairness, and these associations are partly mediated by "jungle world" beliefs about competitive social relations. While this study is not about influence operations, it provides evidence that MFT measures can operate as meaningful attitudinal correlates in a Russian context. Consistent with prior work showing that political ideology is associated with different moral emphases (Graham et al., 2013), this provides additional justification for treating MFT as a plausible feature family for analyzing Russian influence messaging.

This study also employs NRCLex (Bailey, 2019) to extract discrete emotional affect from tweets. NRCLex provides eight emotion categories of fear, anger, anticipation, trust, surprise, sadness,

disgust, and joy, and also provides a positive/negative sentiment label. The analysis retains the discrete emotion variables but excludes anticipation because it does not align cleanly with the study's theoretical focus on moral, ideological, and partisan framing. The NRCLex sentiment label is also excluded in order to keep the emotion specification focused on discrete affect rather than a redundant aggregate.

Recent critiques of lexicon-based emotion tools, including concerns raised about NRCLex in social media contexts (Paletz, Golonka, et al., 2024), note that slang, irony, and platform-specific rhetoric can reduce detection sensitivity. As a result, emotionally salient tweets may be undercounted, which can attenuate estimated emotion effects. This study therefore treats NRCLex-based emotion measures as informative but imperfect indicators and interprets them alongside moral and partisan language features rather than as definitive measures of affect.

Finally, the study identifies partisan terms (e.g., "liberal," "conservative") to measure partisan word count and individual partisan indicators. Partisan language can signal ideological intensity and, in practice, may carry affective and moral connotations even when lexicon tools miss emotional cues. These partisan features are included as a complementary rhetorical signal family rather than as a substitute for emotion measurement.

Preliminary models also included VADER sentiment scores (Hutto & Gilbert, 2014) as an overall polarity measure. VADER was subsequently excluded because it is conceptually redundant with the emotion specification and offers limited additional explanatory value beyond discrete affect measures, while also risking masking distinct effects of individual emotions. Accordingly, the final models retain the discrete NRCLex emotion categories (excluding anticipation) and exclude aggregate sentiment measures.

To investigate the relationship between moral, emotional, and partisan linguistic features and engagement, a series of nested regression models were employed. These models assess the incremental explanatory power of successive feature families on key engagement metrics, with both Ordinary Least Squares (OLS) and Generalized Linear Models (GLM) used for estimation.

The dependent variables in this study were engagement measures derived from user interactions with tweets. For the OLS models, the dependent variable was the log-transformed retweet count, represented mathematically for the full regression as:

$$\mathbf{log2(Retweets+1)=\beta0+\beta1(Emotion\ Variables)+\beta2(MFT\ Variables)+\beta3(Partisan\ Word\ Count)+\beta4(Partisan\ Word\ Indicators)+\beta5(log2(Followers+1))+\varepsilon}$$

where engagement was measured in terms of tweet amplification. For the GLM models, the dependent variable was the raw like count, modeled using a Negative Binomial distribution with a log link function to account for over-dispersion in the count data, represented mathematically for the full regression as:

$$\mathbf{E(Likes)=exp(\gamma0+\gamma1(Emotion\ Variables)+\gamma2(MFT\ Variables)+\gamma3(Partisan\ Word\ Count)+\gamma4(Partisan\ Word\ Indicators)+\gamma5(log2(Followers+1)))}$$

The independent variables were added sequentially across the models except for the first two that regressed follower count against emotion and MFT only, respectively. The first model regressed the log-transformed follower count only to establish a baseline for further multivariable regressions, each successive model included the variable: log2(Followers+1)

The second model was follower count and MFT alone. The third model included variables capturing the emotional content of tweets, derived from the NRC emotion lexicon. These variables represented the normalized frequency of eight basic emotions: fear, anger, trust, surprise, sadness, disgust, joy, and anticipation, with anticipation excluded from the final analysis. The fourth model extended this by incorporating variables reflecting the likelihood of tweets aligning with categories from Moral Foundations Theory (MFT). These included scores for loyalty, sanctity, care, fairness, and authority, computed using supervised machine learning models trained on MFT-specific classification tasks (Nguyen et al., 2019). The fifth model adds the count of partisan words as the final variable. The sixth model consisted of a multivariable regression of individual partisan words found in the tweet text and includes log-transformed follower count. These specific words were chosen as being politically sensitive for the timeframe of the data set, 2009-2018. See table 1 for a full list.

To assess whether the 96 partisan keywords reflected language that was salient in U.S. public discourse during the study period, Google Trends was used as an external validation source. Twitter's trending-topic data are not publicly archived and cannot be retrieved retrospectively, making Google Trends the only continuous, accessible record of public attention spanning 2009–2018.

Each term in the partisan lexicon was queried in the United States using Google Trends, which reports normalized search interest on a 0–100 scale. In this scale, 100 represents the week of peak search activity for that term within the selected window, and all other weeks are expressed as percentages of that maximum. The resulting averages therefore indicate how much search

interest a term maintained relative to its own historical high point rather than an absolute share of total searches.

Typical values for major political figures such as *Trump* or *Obama* ranged from roughly 6 to 15, signifying that searches for these terms averaged about 6–15 % of their own peak-week intensity across the twelve-year period. Mid-range terms, including *Fox News*, *China*, and *Clinton*, exhibited average values around 4–10, while movement-specific slogans such as *MAGA* or *Lock Her Up* often registered near zero because they trended primarily inside social-media platforms rather than in search behavior.

These normalized values provide a practical measure of temporal salience: higher averages identify terms that sustained broad public visibility across time, and lower averages mark terms that appeared mainly in short, high-intensity bursts. This validation confirmed that the lexicon captures language that was widely recognizable and politically relevant within the broader U.S. information environment. See Annex E for the full Google Trends data.

In this study, the dependent variables were carefully chosen to capture two distinct aspects of engagement. The log-transformed retweet count served as the primary dependent variable in the Ordinary Least Squares (OLS) regression models, addressing issues of skewness and heteroscedasticity in the raw count data. For the Generalized Linear Models (GLM), the raw like count was used, modeled with a Negative Binomial distribution and a log link function to handle the over-dispersion commonly found in count data. This dual approach ensured that both the spread of information (retweets) and audience appreciation (likes) were effectively modeled.

The independent variables progressively included three sets of features. Emotion variables, derived from the NRC emotion lexicon, captured the presence of eight basic emotions—fear, anger, trust, surprise, sadness, disgust, joy, and anticipation—based on the normalized frequency of emotion words in each tweet. Moral Foundations Theory (MFT) classification variables, generated using supervised machine learning models, provided probability scores for five moral foundations: loyalty, sanctity, care, fairness, and authority. The partisan word count variable reflected the presence of politically salient terms, selected to represent ideological or political partisanship in U.S. politics. Additionally, follower count was log-transformed to control for the influence of audience size on engagement metrics.

The choice of OLS and GLM methods allowed for a nuanced exploration of the data. OLS regression, which minimizes the sum of squared differences between observed and predicted values, was appropriate for modeling the log-transformed retweet count. This transformation addressed skewness and made the data more suitable for OLS assumptions, such as linearity and homoscedasticity. GLM, on the other hand, extended the linear modeling framework to handle the raw like count, using a Negative Binomial distribution to account for over-dispersion and a log link function to relate the linear predictors to the dependent variable. This flexibility made GLM an ideal choice for analyzing count data with variability exceeding the mean.

The methodological differences between OLS and GLM reflect their complementary strengths. OLS assumes that the dependent variable is normally distributed and that independent variables relate to the outcome in a linear fashion. GLM, by contrast, allows for greater flexibility by incorporating alternative distributional assumptions and link functions, enabling the modeling of non-normal, over dispersed data. In this study, applying both approaches allowed retweet activity

to be treated as a continuous, log-transformed outcome and likes as raw counts, improving the overall interpretability and robustness of the findings.

This approach allowed for a systematic assessment of how emotional, ideological, and partisan language influenced user engagement with tweets, while accounting for the influence of user-specific factors like audience size. The nested structure of the models provided insight into the incremental explanatory power of each set of variables, highlighting their respective contributions to explaining variation in engagement.

To evaluate the independence of predictors used in Study 1, a comprehensive analysis of Pearson and Spearman correlation matrices was completed across all six model configurations. Across all models, **no pair of independent variables exceeded a correlation threshold of 0.6**, a common benchmark for multicollinearity (Dormann et al., 2013; Tabachnick & Fidell, 2019).

The highest observed Pearson correlation among independent variables (excluding trivial lexical duplicates like democrat/democrats) was 0.265 between joy and fear. The highest Spearman correlation, more appropriate for bounded or skewed distributions, was 0.416 between the partisan terms hillary and clinton, which, although moderately correlated, were retained in Study 1 regressions due to their distinct ideological signaling and statistical significance.

Correlations between predictors and engagement outcomes were modest, as expected in social media studies. The control variable follower count was consistently the strongest predictor of engagement, with a Pearson correlation of +0.308 with retweets and a Spearman correlation of +0.403. Among content variables, fairness, authority, and care from the MFT framework showed the most consistent positive correlations with engagement, typically ranging from +0.06 to

+0.11. Emotion variables such as fear, anger, and joy showed weak or near-zero correlations, suggesting they do not meaningfully co-vary with engagement when treated as independent predictors.

These results confirm that the independent variables used in Study 1 are not collinear and represent distinct dimensions of political and moral expression. This supports the structural integrity of the regression models and affirms that no variable is acting as a surrogate for another. By demonstrating low X-to-X correlations across both linear and rank-based correlation methods, the analysis validates the inclusion of each variable in the modeling framework. See Annex A.2 for complete correlation tables.

Table 1 Partisan Terminology

Category	Terms
General Terms	liberal, conservative, republican, democrat, right-wing, left-wing, progressive, alt-right
Politics	Obama, Biden, Clinton, Trump, Pence, Sanders, Hillary, McConnell, Pelosi, DNC, GOP, Republicans, Democrats, libertarian, independent
Obama Presidency (2009–2012)	Obamacare, Tea Party, birth certificate, socialist, czar
2016 Election	Crooked Hillary, Lock her up, Make America Great Again, MAGA, Build the wall, Trump Train, email server, Benghazi
2018 Midterms and Trump Presidency	Trump Derangement Syndrome, TDS, covfefe, alternative facts, deep state
Social Movements	Hands up don't shoot, I can't breathe, Kaepernick, take a knee, racial divide, Black Lives Matter, BLM, all lives matter, Blue Lives Matter, antifa, white supremacy, race war, social justice, racism
Climate and Environment	Paris Agreement, climate change hoax, EPA, fracking
International Relations	Russia, Ukraine, Crimea, NATO, Putin, Syria, China, Iran, ISIS, terrorism, collusion, Russia hoax, WikiLeaks, Podesta emails, al-Baghdadi, radical Islam, caliphate, trade war, tariffs, intellectual property theft
Media and Pop Culture	CNN sucks, Fox News, Hannity, Maddow, SJW, NPC, red pill, blue pill
2018 Events	Parkland shooting, March for Our Lives, Kavanaugh, Me Too, #MeToo

Table 1 summarizes the partisan terminology used to construct the partisan language features, providing the basis for the partisan indicators and counts used throughout the empirical chapters.

3.3 Data

This research relies on datasets provided by the former Twitter Academic Research Consortium, which includes millions of tweets verified by Twitter as originating from Russian accounts from 2009 through 2018 (Twitter, 2018). These accounts are believed to represent foreign influence operations targeting the United States and other countries. The verification process involved analyzing account behavior, patterns of activity, content, and network associations by Twitter.

For this study, only English-language tweets were considered, resulting in a sample of 3,101,744 tweets. This approach reduced processing time and excluded non-English tweets that would require translation, which was beyond the available resources. While this focus on English tweets may overlook Russian influence in non-English-speaking populations, it aligns with the study's aim of assessing the impact on the U.S. audience, where English dominates online discourse. Despite English not being the official language of the United States, the high proportion of English tweets in the dataset ensures validity for this research's purpose.

However, the OLS models only observed roughly 1,765,496 tweets due to the parameter of excluding any tweets that had zero scores for emotion. This sample size is comparable to similar studies, such as Brady et al. (n=286,255) and approximates Rathje et al. (n=2.7 million), enhancing the robustness of the analysis. Lastly, a full set of regressions was run that did not exclude rows with no emotion score and the full 3,101,744 tweets for comparison.

3.4 Summary Results

Complete regression tables, detailed model analysis and correlation tables are provided in Annex A.1. In the baseline OLS model predicting retweet count, log-transformed follower count explained **9.5 percent** of the variance ($R^2 = 0.095$), confirming its role as a meaningful but limited structural predictor of visibility. Adding NRC emotion variables increased explained variance only slightly ($R^2 = 0.098$). Within this family, sadness ($\beta \approx 0.16$) and joy ($\beta \approx 0.16$) were positively associated with engagement, while anger ($\beta \approx -0.38$) and surprise ($\beta \approx -0.13$) were negatively related. These heterogeneous effects indicate that emotional tone is not a uniform driver of retweet activity.

Substantially larger gains emerged when **Moral Foundations Theory (MFT)** variables were introduced. Models including the five MFT scores explained **11.2–11.6 percent** of the variance, and these effects were stable across both OLS and GLM specifications. Fairness ($\beta \approx 0.35$), Loyalty ($\beta \approx 0.32$), and Care ($\beta \approx 0.21$) consistently appeared as the strongest positive predictors of engagement, with Sanctity and Authority contributing at smaller but still significant magnitudes. This represents a clear improvement over emotion-only models and suggests that moral rhetoric provides a more consistent content-based signal of engagement, directly addressing RQ1 and supporting H1's expectation that MFT-based language would be more strongly associated with engagement than purely emotional speech.

Partisan vocabulary improved model performance only modestly. Aggregate partisan word count increased explained variance to **approximately 11.9 percent**, and individual partisan terms exhibited substantial variation in effect size. High-salience cultural or political terms—such as *#MeToo*, *kaepernick*, *take a knee*, and *parkland shooting*—were among the strongest predictors

of engagement, whereas institutional or procedural political terms (e.g., *mcconnell*, *tea party*) were generally weak or even negatively associated. This pattern suggests that partisan words are most influential when they encode moral meaning or identity relevance, rather than ideological extremity alone.

Robustness checks using the full dataset of 3.1 million tweets showed slightly lower explanatory power overall (e.g., $R^2 = 0.114$ in the full OLS model), but the relative ranking of variables remained unchanged. Moral foundations retained stable and significant effects, emotional variables continued to exhibit mixed associations, and partisan word count weakened when tweets with no emotional content were included. Across all specifications, moral framing consistently outperformed both affective and partisan predictors in explanatory strength.

3.5 Discussion

The findings of this study demonstrate that moral language, emotional expression, and partisan cues each shape engagement with Russian influence tweets, but they do so with markedly different levels of consistency and strength. The results indicate that moral framing, as captured through Moral Foundations Theory (MFT), provides a particularly durable rhetorical mechanism. Messages invoking fairness, loyalty, and care consistently generated higher engagement than those emphasizing emotional tone or partisan identity, suggesting that audiences respond more strongly to value-laden content than to affective or ideological cues alone. Within the limits of this dataset and design, this pattern directly addresses RQ1 and supports H1's expectation that MFT-based language would be more strongly associated with engagement than purely partisan or emotional cues. This finding contributes to ongoing work on moral contagion and political communication, reinforcing research showing that moralized content spreads more effectively

because it activates shared principles and identity-relevant concerns (Brady et al., 2017; Haidt, 2012).

By contrast, emotional expression showed more modest and inconsistent effects. Joy and sadness appeared to resonate with users, while anger and surprise were associated with lower engagement, an unexpected pattern given prior findings that anger often increases virality (Brady et al., 2017). One plausible explanation is that anger overlaps with other rhetorical signals in the model, particularly moral language related to harm and fairness. Correlation analyses show that anger tends to co-occur with other negative emotions and with moral language tied to these foundations, meaning that some of its engagement-relevant signal may already be captured by other features. As a result, the regression coefficient for anger may reflect the residual emotional signal once morally framed grievances are accounted for. This residual form of anger, less clearly anchored in moral justification, may be less effective, or even counterproductive, for generating engagement, helping explain the negative association observed here. In this sense, anger may operate less as a stand-alone emotional cue and more as part of a broader pattern of moralized grievance reflected across multiple measures. This divergence also highlights the limitations of lexicon-based emotion models such as NRCLex, which can underperform on Twitter-style text and cannot distinguish between expressing anger, describing anger, or provoking anger in an audience (Paletz, Golonka, et al., 2024). These distinctions matter because a tweet expressing righteous anger about injustice may mobilize aligned users while discouraging broader amplification. Public engagement on social media is reputationally visible, and users may avoid sharing angry or hostile content to limit perceived aggression or negativity within their networks (Marwick & boyd, 2011). The strategic composition of the IRA tweet sample may further amplify this effect, as engagement-oriented messaging that foregrounds

grievance without overt hostility may be more effective than explicitly angry appeals. Platform governance and moderation incentives may also discourage the spread of overt hostility in ways not directly observable in tweet text.

Partisan vocabulary played an even more selective role. Highly salient partisan terms tied to moments of cultural and political significance, such as #metoo, kaepernick, take a knee, and parkland shooting, were associated with engagement because they blended emotional intensity with moral significance and identity relevance. These terms functioned as more than ideological markers; they activated broader narratives around harm, injustice, and collective identity. In contrast, institutional or procedural political references (e.g., mcconnell, tea party, covfefe) lacked clear emotional or moral resonance and tended to correlate weakly or negatively with engagement, reflecting the lower salience of procedural politics relative to moralized social issues (Pew Research Center, 2023). This distinction underscores that partisan cues are most impactful when they intersect with moral narratives rather than when they stand alone as ideological signals.

The broader pattern across models suggests that moral framing is a central mechanism by which influence campaigns may embed themselves within mainstream discourse. Value-laden content appears to offer influence operators a way to shape attention and amplify resonance without relying on overt polarization, emotional extremity, or explicit ideological positioning. Unlike anger or partisan rhetoric, moral language communicates why the content matters and to whom, making it more interpretable and broadly resonant. This helps explain why Russian messaging often resembles the rhetorical style of ordinary or moderately engaged U.S. users rather than political elites or ideological activists.

These findings connect Study 1's engagement results to the dissertation's subsequent classification models. If moral framing reliably increases engagement, then moral foundations represent a plausible mechanism for how influence messaging can scale within mainstream political discourse. Study 2 tests whether these content-level tendencies place Russian tweets closer to U.S. politicians, politically engaged users, or random users when evaluated at the tweet level using the same moral, emotional, and partisan features. Study 3 then aggregates these same measures across full account histories to evaluate whether persistent rhetorical profiles provide clearer separation under realistic base rates and prevalence-aware metrics, including positive predictive value. Study 1 therefore answers RQ1 and supports H1: moral foundations provide the most consistent and informative content-based signal of engagement, with emotion and partisan language contributing secondary and context-dependent effects.

Chapter 4 Study 2: Tweet-Level Classification of Russian Influence Rhetoric in U.S. Discourse Space

4.1 Overview and Research Questions

This study tests whether Russian influence tweets are distinguishable from U.S. political speech using content-only, tweet-level features that capture rhetorical style rather than audience scale. It

uses Moral Foundations Theory scores, NRC-based emotion measures, and partisan language features developed in Study 1; structural variables such as follower count are excluded. Prior work shows that state-sponsored accounts can emulate participatory political discourse while still differing systematically from ordinary users, and that supervised text classifiers offer interpretable baselines for making such comparisons (Alizadeh et al., 2020; Chun et al., 2019; Serafino et al., 2024; Tian et al., 2025; Zannettou et al., 2019). Complementary research emphasizes behavioral and coordination signals as an alternative detection family; in this study, those signals are treated as out of scope so that the analysis isolates rhetorical mechanisms under a strict content-only constraint (Cinelli et al., 2022; Ezzeddine et al., 2022; Pacheco et al., 2021).

The first research question asks whether Russian influence tweets more closely resemble the rhetorical style of U.S. politicians, politically engaged users, or random users (RQ2), with the expectation that Russian IRA tweets will resemble non-elite users (engaged or random) and will avoid overtly partisan language in order to appear non-partisan (H2). A second question asks which rhetorical features such as partisan, moral, or emotional most strongly drive classification accuracy (RQ3); here I hypothesize that partisan language will be the strongest predictor, followed by moral foundations, then emotion at both the tweet and account level (H3). A third question tests whether the stylistic ambiguity of Russian messaging is detectable across both theory-driven and theory-agnostic lexical baseline models (RQ4), with the expectation that a TF–IDF model using raw lexical features will reproduce the engaged-versus-random classification pattern, confirming rhetorical moderation at both the tweet and account level (H4). A final research question asks whether content-only rhetorical characterizations separate Russian content from domestic political speech strongly enough to enable targeted interventions without substantial over-capture of legitimate speech from U.S. politicians and engaged users (RQ5); the

associated hypothesis anticipates that high Russian recall will entail substantial false positives, and that usable detection windows, if they exist at all, emerge only once rhetoric is aggregated to the account level (H5).

In this chapter I address RQ2 through RQ4 at the tweet level and use those results to motivate the account-level and policy analysis in Study 3, where RQ5 and H5 are taken up explicitly with realistic prevalences and positive predictive value. The tweet-level sequence is intentionally simple and report-driven. First, a Baseline Tweet Model is trained on all tweets using seven emotions, five MFT scores (without z-scoring), and a single partisan feature. Second, a No-Partisan Ablation removes the partisan signal to test whether explicit partisan language materially changes class placement relative to moral and emotional framing. Third, an Enhanced-Partisan variant augments the baseline with the top ten positively and top ten negatively impactful partisan indicators coded as binary flags, allowing a localized test of whether specific partisan terms materially sharpen class distinctions. Fourth, a pair of TF–IDF (term frequency–inverse document frequency) specifications replaces the engineered features with bag-of-words representations. The primary TF–IDF model provides a theory-agnostic lexical baseline for distinguishing the politician, engaged, and random classes, and a substantively equivalent secondary model is used to confirm stability and to extract the top-weighted TF–IDF terms for a KWIC (key word in context) diagnostic. TF–IDF weights words by how distinctive they are in the corpus, allowing the model to rely on lexical patterns rather than predefined rhetorical features. The KWIC analysis then examines a short context window, typically the three words before and after each top TF–IDF term, to determine whether these ostensibly civic or institutional keywords are being used in partisan ways or instead reflect a broader public-affairs register that Russian accounts mimic while avoiding explicit domestic partisanship. Results for

each tweet-level model, including accuracy, per-class precision and recall, feature importance, and the placement of Russian tweets within the U.S. discourse map, are reported in the following sections and then used in Study 3 as the basis for account-level scoring and policy-risk assessment.

4.2 Study Design

Training data consist of three labeled U.S. datasets namely politicians (n = 3,822,905), politically engaged users (n = 13,615,549), and random users (n = 13,334,998). To isolate content-based separability rather than prevalence, I draw a class-balanced sample of 3.1 million tweets per class for model fitting, yielding 9.3 million training instances. Evaluation uses an 80/20 stratified split on U.S. tweets so that test performance reflects the same within-class composition the model sees during training while preserving independence between train and test. After fitting, I apply each model to an unlabeled Russian corpus (n = 3,104,107) to infer alignment distributions across the three U.S. rhetorical types.

All tweets are processed with the Study 1 pipeline for consistency. For each tweet I compute Moral Foundations categories (Care, Fairness, Loyalty, Authority, Sanctity) by parsing classifier outputs to numeric scores, NRC emotions (fear, anger, trust, surprise, sadness, disgust, joy) as numeric scores, and an optional partisan signal. The supervised label is author type with three classes: politician, engaged user, and random. Structural variables such as follower count and engagement metrics are excluded to focus the task on rhetorical style rather than audience scale or connectivity.

The model sequence is designed to test RQ2 through RQ4 and to set up the policy evaluation for RQ5. The Baseline Tweet Model includes seven emotions, five MFT features, and a numeric partisan count, and it provides the reference alignment pattern used to evaluate H2. A no-partisan ablation removes the partisan feature to test mechanism and to quantify its marginal contribution to performance for RQ3 by comparing macro-F1 and accuracy on the held-out U.S. test set. An enhanced partisan variant augments the baseline with the ten most positively associated and the ten most negatively associated partisan indicators from Study 1 encoded as binary flags. This tests whether a compact but salient lexicon improves separability beyond overall partisan intensity without changing the underlying theoretical mechanism. A TF-IDF baseline replaces the engineered features with bag-of-words vectors to assess RQ4 by asking whether the observed non-elite over elite alignment persists under a theory-agnostic lexical baseline representation. A TF-IDF contextual diagnostic adds a plus-minus three-word window around focal partisan terms to illustrate how co-occurrence patterns convey partisan meaning; this diagnostic is used for interpretation rather than benchmarking. For RQ5, taken up in Study 3, I convert model probabilities to political-style equal to P(engaged) plus P(politician) and random-style equal to P(random). I set operating thresholds by Russian political-style score percentiles targeting approximately 70, 80, and 90 percent capture, then evaluate false positive rates for politicians and engaged users on held-out U.S. tweets and compute positive predictive value at assumed prevalences of 0.1 percent, 0.5 percent, and 1 percent.

Models.

1. **Balanced Baseline.** Three-class classifier trained on class-equalized U.S. samples using seven NRC emotions, five MFT scores (unstandardized), and a numeric partisan word count.

2. No-Partisan Ablation. Same as the Baseline but excludes partisan word count to test the mechanism and estimate the marginal contribution of partisan intensity.
3. **Enhanced Partisan.** The Baseline model plus the top ten positive and top ten negative partisan indicators from Study 1, each encoded as a binary feature, to test whether a small but salient set of partisan terms improves separability beyond overall partisan intensity.
4. **TF-IDF Baseline.** Replaces theory-driven features with bag-of-words TF-IDF vectors to assess whether the alignment pattern among politicians, engaged users, and random users persists with theory-agnostic lexical baseline representations.
5. **TF-IDF (Contextual KWIC).** Uses the fitted TF-IDF Baseline to select the most important unigrams and computes ±3-word "key word in context" windows around those terms in U.S. and Russian tweets to examine whether ostensibly civic and institutional features are used inside partisan frames. This diagnostic clarifies mechanism and interpretability but does not constitute a separate predictive model.

4.3 Data

The Russian dataset is the same from the former Twitter Academic Research Consortium (Twitter, 2018), which includes millions of tweets verified by Twitter as originating from Russian accounts used in study 1. The same constraints were applied for English-only tweets. The political tweets that are from 2007 to 2020 come from elected members of Congress. The third and fourth datasets come from tweets of politically engaged users and random Twitter users come from the Harvard Dataverse (Alizadeh, 2020) and cover 2006 through 2019. Politically engaged users are defined by following at least five political accounts, and both datasets were restricted to users who posted at least 100 times per year.

4.4 Summary Results

Complete classification reports, confusion matrices, and feature-importance plots for all Study 2 models are provided in Annex B.1. Across tweet-level specifications, the baseline rhetorical classifier using Moral Foundations Theory (MFT) scores, NRC emotions, and partisan word count achieved **about 51 percent accuracy** with **macro-F1 ≈ 0.45** on held-out U.S. tweets. The model reliably distinguished random users from political actors but showed substantial overlap between politicians and politically engaged users, indicating that elites and highly engaged non-elites share much of the same rhetorical register at the tweet level.

Ablation tests show that removing partisan vocabulary produces only modest declines in performance. The no-partisan model preserves essentially the same accuracy and macro-F1, suggesting that partisan intensity contributes incremental value but is not the primary discriminating feature once moral framing and basic emotional tone are available. By contrast, an enhanced-partisan variant that incorporates the ten most positively and ten most negatively impactful partisan indicators yields small but consistent numerical gains, improving macro-F1 by roughly one point and slightly sharpening separation of random users. This pattern reinforces that partisan cues are most informative when tied to high-salience or moralized events rather than appearing as a diffuse aggregate signal.

Across engineered-feature models, **Authority, Care, and Fairness** consistently rank among the highest-importance variables, with partisan word count and selected NRC emotions (such as trust, disgust, and joy) occupying secondary positions. Random users tend to exhibit lower overall MFT activation, while politicians and engaged users show more frequent use of moral and civic vocabulary. The classifier therefore distinguishes user types partly through how they

deploy moral framing and institutional language, with partisan terms providing only localized refinement.

The TF–IDF specification performs **substantially better** than the engineered-feature models. Using a 10,000-term unigram vocabulary, the TF–IDF classifier achieves **roughly 67 percent accuracy and macro-F1 ≈ 0.67** on U.S. tweets, with the largest improvements appearing in recall and F1 for engaged users while maintaining strong performance for politicians and random users. When applied to Russian tweets, it classifies them at approximately **46 percent random, 49 percent engaged, and 5 percent politician**, placing the vast majority of Russian content in the non-elite region of the discourse space. Its top-weighted features consist largely of civic, procedural, and institutional vocabulary rather than overtly partisan terms, indicating that political communication online is structured around a broadly shared public-affairs lexicon that Russian accounts emulate while selectively underusing explicit partisan branding.

To better understand how Russian accounts use this lexical register, a KWIC (key word in context) diagnostic was applied to the highest-importance TF–IDF terms. Examining the ±3-word context window around each term showed that these civic and institutional keywords typically appear in neutral or informational public-affairs framing rather than embedded in partisan slogans. For Russian tweets in particular, these high-weight terms most often occur in the same issue-oriented, non-elite register used by politically engaged U.S. users, suggesting that Russian accounts consciously mimic a moderated civic discourse style while avoiding the dense partisan and culture-war signals characteristic of many domestic elites.

Overall, the Study 2 results show that tweet-level models can reliably distinguish random users from political actors but cannot cleanly separate politicians from politically engaged users,

regardless of whether engineered rhetorical features or raw lexical structure are used. Russian tweets overwhelmingly fall into the non-elite rhetorical region, aligning most closely with engaged and random users. These findings motivate the account-level analyses in Study 3, which test whether aggregating features across an account's full posting history strengthens these distinctions and supports more realistic assessments of detection risk.

4.5 Discussion

Study 2 asks a constrained question: within this Twitter dataset, and using only tweet-level rhetorical features, specifically Moral Foundations Theory scores, NRC emotion categories, and partisan word indicators, how well can we distinguish Russian IRA tweets from messages posted by U.S. politicians, politically engaged users, and random users? The models are deliberately limited to single-message text on one platform and are intended as interpretable content-only baselines for comparison, consistent with contemporary detection work that uses supervised text classifiers to characterize influence content while holding constant audience-scale variables (Alizadeh et al., 2020; Chun et al., 2019; Serafino et al., 2024; Tian et al., 2025; Zannettou et al., 2019). Within those bounds, the results identify a stable rhetorical structure that situates Russian messaging inside the U.S. discourse space in a way that is empirically legible and comparable across model families (Zannettou et al., 2019; Serafino et al., 2024).

Across specifications, moral framing emerges as the most consistent and informative signal. Authority, Care, and Fairness reliably separate political elites from ordinary users, while emotional tone and partisan vocabulary provide incremental lift. The resulting geometry is substantively meaningful for RQ2: Russian tweets fall predominantly in the non-elite region of the space rather than clustering with officeholder rhetoric, indicating that Russian influence

content aligns more closely with participatory civic discourse than with elite institutional messaging, as prior work would predict for campaigns designed to blend into everyday political talk (Zannettou et al., 2019; Tian et al., 2025).

The baseline rhetorical model reinforces this picture. It separates random users from political actors reasonably well and produces a consistent class hierarchy in which officeholder rhetoric is most distinctive. Feature importance is dominated by Authority, Care, and Fairness, with emotions and partisan word count contributing smaller effects. The sparsity diagnostics provide a clear descriptive account of why these features matter in this corpus: moral language occurs often enough, especially in politician tweets, to serve as a reliable anchor for classification, while partisan language appears in a minority of tweets across groups (14.7% of politician tweets, 12.1% of Russian tweets, 9.4% of engaged users, and 2.9% of random users). This distribution clarifies the role partisan terms play in tweet-level separability: they offer localized signal where present, but the dominant structure comes from moral framing.

The no-partisan ablation confirms that interpretation. Removing partisan word count yields essentially the same overall performance profile and preserves the same class hierarchy, indicating that moral framing and emotional tone account for most of the separability available under a strict content-only constraint. This result is useful for RQ3 because it clarifies which rhetorical dimensions are most consistently informative at the tweet level in this dataset.

Adding the top twenty individual partisan indicators provides a targeted test of whether a compact lexicon meaningfully shifts performance. The enhanced-partisan model yields small numerical gains and modestly improves engaged recall but does not materially change the feature-importance structure: MFT dimensions remain most influential, while partisan word

count and the new indicators occupy mid-tier positions. Russian predictions remain stable in aggregate. Taken together, the partisan variants indicate that curated partisan features contribute where they appear, while the broader separability derives from moral framing and a public-affairs register, consistent with the broader supervised detection literature's emphasis on stable stylistic signals over any single lexicon (Serafino et al., 2024; Alizadeh et al., 2020).

The TF-IDF baseline addresses RQ4 by replacing engineered features with 10,000 unigram TF-IDF terms and asking whether a theory-agnostic lexical baseline representation recovers similar structure. This model performs substantially better overall, particularly by improving recall for engaged users while maintaining strong performance for politicians and random users. When applied to Russian tweets, it assigns the majority of messages to the same non-elite region, with a larger share in engaged relative to the engineered-feature models. Feature importance analysis further strengthens the convergence claim: the highest-weighted TF-IDF terms are primarily civic and institutional tokens, aligning with the Authority/Care emphasis surfaced by the rhetorical models. This agreement across representations supports the inference that the observed structure is not an artifact of feature engineering and that the civic–institutional register is a robust dimension of tweet-level separability (Chun et al., 2019; Zannettou et al., 2019; Tian et al., 2025).

The TF-IDF + KWIC diagnostic adds local context by examining ±3-word windows around top TF-IDF terms in U.S. and Russian tweets. Across the top unigrams, co-occurrence with partisan language is generally low, and the highest-importance terms most often appear in neutral public-affairs contexts rather than partisan slogans. Where co-occurrences rise, they concentrate around institutional tokens such as housegop, housedemocrats, senate, scotus, sotu, and trumpcare,

primarily in elite and highly engaged U.S. tweets. Russian usages, when present, follow the same civic–institutional pattern. These KWIC results reinforce the interpretation that the most predictive lexical features reflect an institutional public-affairs register that Russian accounts adopt in a manner consistent with participatory discourse (Zannettou et al., 2019; Serafino et al., 2024).

In sum, Study 2 provides a coherent content-only answer to RQ2–RQ4. First, tweet-level rhetorical signals place Russian messages predominantly in the non-elite region of U.S. discourse rather than in elite institutional rhetoric. Second, moral foundations, especially Authority and Care, are the most central and stable drivers of separability in this dataset, with emotion and partisan language providing incremental improvements. Third, the convergence between engineered rhetorical features and a TF-IDF representation, reinforced by the KWIC diagnostic, indicates that the detected structure generalizes across theory-driven and theory-agnostic lexical baseline modeling choices. These results motivate Study 3's account-level aggregation as a natural extension: the same rhetorical dimensions can be evaluated at the level of persistent speaker profiles and assessed under realistic prevalences and operational metrics (Tian et al., 2025; Serafino et al., 2024).

These findings also clarify the limits of tweet-level detection. Message-level models necessarily abstract away from sequencing, network structure, and account-level behavior, and their performance relies on balanced samples rather than real-world prevalences. The ambiguity between politicians and engaged users, and the tendency of Russian tweets to resemble engaged or random users, suggest that practical interventions would require information beyond single tweets. Study 3 builds directly on this insight by aggregating the same moral, emotional, and

partisan signals across full account histories and evaluating detection under realistic base rates, enabling an assessment of false-positive risks and positive predictive value in operational settings.

4.6 Policy Implications

Study 2 offers a measured contribution to ongoing policy debates about detecting and constraining foreign influence on social media. At the tweet level, the models show that Russian IRA messages are not easily separable from domestic political speech when we restrict ourselves to content-only features. Moral framing and civic vocabulary distinguish political actors from random users, but the boundary between elites and highly engaged users remains weak, and most Russian tweets fall on the non-elite side of that line. From a policy perspective, this combination of partial separability and substantial overlap suggests both a constraint and a modest opportunity.

The clearest constraint lies in partisan keyword approaches. Across models, explicit partisan vocabulary plays only a limited role in classification. The partisan word count feature is sparse and heavy-tailed, individual partisan indicators provide only localized lift, and the TF-IDF plus KWIC analysis demonstrates that even highly weighted civic terms rarely appear within dense partisan windows. The KWIC results reinforce this point: when Russian tweets use high-importance civic or institutional terms, those terms seldom co-occur with partisan tokens in their immediate context, indicating that Russian messaging is designed to blend into non-elite civic discourse rather than rely on explicit partisan branding. This implies that regulatory or platform responses based primarily on partisan or slogan lists are unlikely to reliably identify Russian messaging. Such lists will miss a large fraction of influence content that uses moderated,

institutional language while still sweeping up legitimate domestic speech that happens to use the same vocabulary. As a result, keyword-based screening is better understood as a crude triage tool than as a defensible basis for sanctions or removal.

At the same time, the models do detect a consistent rhetorical signature that has potential policy value. Russian tweets tend to adopt a public-affairs register, discussing communities, security, budgets, veterans, the economy, and similar topics, while using explicit partisan labels less densely than politicians and at rates that overlap with highly engaged domestic users. In aggregate, roughly 12 percent of Russian tweets contain at least one partisan term, compared with about 15 percent for politicians, 9 percent for engaged users, and 3 percent for random users. Moral Foundations scores, particularly Authority and Care, are elevated relative to purely random users but do not reach the levels observed in elite communication. This suggests that Russian content tends to inhabit a rhetorical style between elites and highly engaged citizens—issue-oriented, civic in tone, and only selectively partisan. For platforms and regulators, this implies that Russian campaigns may be more likely to blend into the engaged layer of public discourse than to impersonate elite voices. Any intervention targeting this rhetorical band will therefore intersect inevitably with ordinary civic participation.

Methodologically, Study 2 underscores the limits of message-level enforcement. Even under favorable conditions like balanced datasets, curated features, and clean labels, tweet-level classifiers cannot cleanly separate politicians from engaged users nor reliably isolate Russian tweets without substantial confusion with domestic content. This does not imply that content features are irrelevant for detection, but it does suggest that individual tweets are an ill-suited unit of enforcement. In practice, it would be difficult to justify removing or downranking tweets

solely because they use a moral and institutional style common to both foreign influence efforts and legitimate domestic activism. These results highlight why platform governance frameworks increasingly emphasize broader systemic assessments rather than per-post labeling.

The policy implications therefore point toward account-level and contextual strategies rather than tweet-level filters. The core insight from Study 2 is that linguistic signals of Russian influence are real but subtle. They reflect a moderated, public-affairs-oriented style rather than overt propaganda or extreme partisanship, and they are shared with wide swaths of domestic political speech. Any detection system built solely on single-tweet rhetoric risks over-capture of legitimate speech, especially under low base rates. Study 3 addresses these challenges by aggregating signals across entire accounts, evaluating whether stable rhetorical profiles can distinguish Russian from U.S. users more reliably and whether content-only models can support operational thresholds with acceptable false-positive rates under realistic conditions.

In short, Study 2 suggests that linguistic signals of Russian influence are real but subtle. They reflect a moderated, public affairs oriented style rather than overt propaganda or extreme partisanship, and they are shared with large swaths of domestic political speech. This makes conservative policy use of tweet-level models more appropriate than aggressive screening. The role of Study 2 is therefore to map the rhetorical landscape and to show where Russian messages sit within it. Study 3 asks whether those same features, when accumulated over time and combined with account-level information, can support more targeted and transparent interventions that mitigate foreign influence while limiting collateral effects on legitimate political expression.

Chapter 5 Study 3: Account-Level Rhetorical Profiling and Prevalence-Aware Policy Risk

5.1 Overview and Research Questions

Study 3 extends the tweet-level analysis of Study 2 to the account level and addresses the same research questions with a more stable unit of analysis. Instead of asking whether single tweets can distinguish Russian influence from U.S. political speech, it examines whether the same rhetorical features form account-level profiles that separate Russian accounts from U.S. politicians, politically engaged users, and random users, and whether those profiles can support realistic, prevalence-aware interventions. Study 2 showed that moral framing, especially Authority and Care, is more central than partisan intensity for tweet-level separability and that most Russian tweets resemble non-elite U.S. discourse. Study 3 builds directly on that result. It tests, first, whether Russian influence accounts more closely resemble U.S. politicians, politically engaged users, or random users once rhetoric is aggregated over many tweets (RQ2), with the expectation that Russian accounts will most resemble random users and will avoid overtly partisan language in order to appear non-partisan (H2). Second, it evaluates which rhetorical feature families—partisan, moral, or emotional—contribute most strongly to account-level classification accuracy (RQ3), with the hypothesis that partisan language will be the strongest predictor, followed by moral foundations, then emotion (H3). Third, it asks whether the stylistic ambiguity observed at the tweet level remains detectable when comparing theory-driven aggregated features to an account-level TF-IDF representation (RQ4), with the expectation that a TF-IDF model over concatenated account text will reproduce the engaged-versus-random distinction and confirm rhetorical moderation at the account level (H4). Finally, Study 3 directly tests whether content-only, account-level rhetorical characterizations separate Russian influence accounts from domestic political speech strongly enough to support targeted interventions

without substantial over-capture of legitimate U.S. politicians and engaged users (RQ5). The associated hypothesis is that thresholds tuned to capture a large share of Russian accounts will generate high false-positive rates and low positive predictive value at realistic base rates, making political-style scoring unsafe at useful recall and leaving, at best, only narrow triage-oriented operating points rather than defensible enforcement thresholds (H5).

To answer these questions, Study 3 keeps the feature families and model logic of Study 2 but aggregates features from tweets to accounts. Each account is represented by the proportion of its tweets that exhibit each Moral Foundations category and NRC emotion, together with measures of partisan intensity and binary indicators for salient partisan terms. A sequence of account-level models, baseline rhetorical, no-partisan ablation, and TF-IDF parallel the tweet-level models from Study 2. This is followed by a policy-risk analysis (PPV) that converts model probabilities into political-style and random-style scores for prevalence-aware thresholding.

5.2 Study Design

Each account is represented as a single observation constructed from its complete tweet history. I aggregate Moral Foundations by the percentage of tweets exhibiting Care, Fairness, Loyalty, Authority, and Sanctity, and NRC emotions by the percentage of tweets exhibiting anger, fear, trust, surprise, sadness, disgust, and joy. Partisan usage is summarized by the sum of partisan word count across the account's tweets and the mean partisan word count per tweet. Features are used in their natural scales (percentages and counts), matching the tweet-level design in Study 2. Structural variables such as follower count, network structure, and engagement metrics are not used in order to maintain a content-only design. This aggregation follows prior work showing that supervised models can separate coordinated or troll accounts from organic users and that

such signals persist beyond individual posts (Zannettou et al., 2019; Serafino et al., 2024; Lewinski and Hasan, 2021).

A three-class gradient-boosted tree classifier is trained on the full datasets of U.S. accounts labeled as politician, engaged user, or random, with strict split hygiene by screen name (account holder) so that all tweets from a given account reside in a single split. Evaluation on a stratified held-out U.S. account set reports accuracy and macro-F1 and per-class F1. I then apply the fitted models to Russian accounts to infer alignment distributions across the three U.S. account types, addressing RQ2 and H2 at the account level.

To estimate feature-group contributions and test the mechanisms implied by partisan-language avoidance and moral framing (RQ3 and H3), I run a no-partisan ablation that removes the partisan sum and mean while holding all other features constant and report changes in macro-F1 as the primary performance metric and changes in accuracy as a secondary metric relative to the baseline. Permutation importance provides a complementary estimate of the relative contribution of Moral Foundations, emotions, and partisan features at the account scale on the held-out U.S. test set.

For representation robustness (RQ4 and H4), I concatenate each account's tweets into a single document and vectorize with bag-of-words TF-IDF using a fixed vocabulary and document-frequency constraints matched as closely as possible to the tweet-level TF-IDF setup (Alizadeh et al., 2020; Chun et al., 2019). This enables a direct comparison of engineered rhetorical features and raw lexical representations at account scope and tests whether the stylistic ambiguity observed in Study 2 persists under a theory-agnostic lexical baseline representation.

Finally, to evaluate policy risk at a more realistic unit of analysis (RQ5 and H5), I transform the model's predicted class probabilities into political-style and random-style scores. The political-style score is defined as the model's predicted probability of "politician" plus its predicted probability of "engaged user," and the random-style score is simply the predicted probability of "random." Thresholds are set using Russian-account score percentiles that target approximately 70, 80, and 90 percent capture of Russian accounts. I then apply these thresholds to held-out U.S. accounts to estimate false-positive rates for politicians and engaged users and to compute positive predictive value (PPV) at assumed Russian prevalences of 0.1 percent, 0.5 percent, and 1 percent. Let p_R denote the prevalence of Russian accounts, TPR the recall (true positive rate) at a given threshold, and FPR the false positive rate on non-Russian accounts; PPV can then be written as:

$$PPV = (p_R * TPR) / (p_R * TPR + (1 – p_R) * FPR)$$

This expression makes explicit how low base rates cause false positives on domestic users to dominate even when recall is high and the false positive rate is modest. This prevalence-aware evaluation tests whether a random-style score provides any usable operating window while political-style scoring remains unsafe at useful recall, linking the avoidance of overt partisanship and the use of moderated civic rhetoric to concrete trade-offs between detection and collateral impact (Mathew et al., 2024; Tian et al., 2025).

Models (summary).

1. **Rhetorical baseline (account).** Account-level model using percent-present Moral Foundations (5) and NRC emotions (7) plus partisan sum and mean per account trained and evaluated on a three-class U.S. account sample, then applied to Russian accounts. Purpose: account-level alignment baseline for RQ2 and reference for feature contribution under RQ3.

2. **No-partisan ablation (account).** Same as baseline, but drops all partisan features. Purpose: mechanism test for H3 and group contribution to performance/alignment (RQ3).
3. **TF-IDF account baseline.** Concatenate all tweets per account → bag-of-words TF-IDF; same splits, then apply to Russian accounts. Purpose: representation robustness (RQ4).
4. **Policy-risk scoring (account).** Transform to political-style = P(engaged)+P(politician) and random-style = P(random); set thresholds by Russian-score percentiles (~70/80/90% capture); report FPR on held-out U.S. accounts and PPV at 0.1/0.5/1% prevalence. Purpose: operational risk/utility (RQ5).

5.3 Data

This model uses the same sources as earlier analyses but shifts the unit of analysis from tweets to accounts. The labeled U.S. sets comprise 655 politicians**,** 5,102 politically engaged users**,** and 5,886 random users. Russian IRA accounts (for unlabeled scoring) total 3,168 after filtering to English-language content from active profiles (Twitter, 2018). Politically engaged and random users are drawn from the Harvard Dataverse corpus (Alizadeh et al., 2020) using sustained political interest/activity and low-political-followership/high-volume criteria, respectively. A minimum activity threshold per account ensures stable feature estimates. All account-level models are trained and evaluated under the observed class proportions (no balancing). For the TF-IDF baseline, tweets for each account are concatenated into a single document and vectorized under the same splits and evaluation protocol.

5.4 Summary Results

Complete classification reports, confusion matrices, feature-importance plots, and prevalence-aware evaluation tables for all Study 3 models are provided in Annex C.1. The baseline account-level classifier provides the first test of whether aggregated rhetorical features can distinguish Russian influence accounts from U.S. politicians, politically engaged users, and random users, directly addressing RQ2 and RQ3, and supporting H2-H3. Trained on all 11,641 U.S. accounts and evaluated on an 80/20 stratified split, it achieves about 0.78 accuracy with macro-F1 ≈ 0.80.

Errors concentrate almost entirely along the engaged–random boundary: politicians are consistently well separated from non-elites, and misclassifications involving politicians almost always flow into the engaged user category rather than random. In contrast, engaged and random accounts show reciprocal confusion, indicating a more diffuse rhetorical region among non-elites. Feature-importance rankings show Authority as the dominant predictor, followed by mean partisan word intensity, trust, and Loyalty, with disgust, Care, Sanctity, and anger forming a second tier. This pattern indicates that account-level separability relies primarily on moral framing, with partisan intensity contributing as a secondary, accumulated signal.

When this baseline model is applied to the 3,168 Russian accounts, it predicts roughly 56 percent as random, 43 percent as engaged user, and only about 2 percent as politician. Relative to the tweet-level classifiers in Study 2, which sometimes labeled more Russian content as politician-like, the account-level baseline is more conservative: very few Russian accounts exhibit the sustained moral and partisan profile characteristic of U.S. elites, and most cluster in the engaged–random region of U.S. rhetorical space. These placement patterns further support H3's expectation that account-level aggregation would sharpen the non-elite resemblance observed at the tweet level.

The no-partisan ablation model removes partisan word count mean while retaining aggregated Moral Foundations and NRC emotions. Its performance declines only slightly (accuracy ≈ 0.76, macro-F1 ≈ 0.79), and the class hierarchy remains unchanged. Authority, trust, and Loyalty remain the most important features, with disgust, Fairness, Care, and Sanctity providing additional separation. Russian placement shifts somewhat further toward random, at about 62 percent random, 36 percent engaged, and under 2 percent politician, indicating that partisan

intensity refines but does not drive the placement of Russian accounts. Moral framing alone is sufficient to keep Russian accounts in the non-elite region, reinforcing the answer to RQ3 and confirming that H3's emphasis on moral structure over partisan intensity holds at the account level.

The TF–IDF account model replaces engineered features with a 10,000-term unigram TF–IDF representation over each account's concatenated tweet history. This theory-agnostic lexical baseline classifier performs substantially better than the rhetorical models, attaining about 0.83 accuracy with macro-F1 ≈ 0.87. Politicians are classified with very high fidelity, and both engaged and random users see clear gains in precision and recall. Feature-importance analysis shows that the TF–IDF model relies heavily on overtly political and institutional vocabulary, such as references to covid-19, Congress, the GOP, the Senate, SCOTUS, voters, and major media outlets, which appear frequently in elite and highly engaged U.S. accounts, addressing RQ4 and H4's claim that a more flexible lexical representation would highlight the dense partisan and institutional register of elite speech.

Crucially, the TF–IDF model does not move Russian accounts closer to elites; it pushes them further into the non-elite region. When applied to Russian accounts, it predicts approximately 86 percent random, 14 percent engaged user, and 0 percent politician. Compared with the baseline and no-partisan models, this sharply increases the random share and eliminates the already small politician fraction. In other words, the more explicitly political and institutional the representation becomes, the less Russian accounts resemble U.S. elites or highly engaged domestic users. Both the rhetorical and TF–IDF models therefore agree that Russian accounts are "under-political" in their lexical patterns relative to domestic elites, providing additional support

for H4's prediction that theory-agnostic lexical baseline models would converge with the theory-driven specifications on Russian moderation.

Finally, the prevalence-aware analysis evaluates whether any of these account-level classifiers can support operational thresholds for detecting Russian accounts under realistic base rates, directly addressing RQ5 and H5. Using the baseline rhetorical model, political-style scoring, defined as the probability of politician plus engaged user, produces unusably high false-positive rates on U.S. accounts: thresholds that capture 70 to 90 percent of Russian accounts flag nearly all U.S. politicians and engaged users and a large fraction of random users. Under assumed Russian prevalences between 0.1 and 1.0 percent, positive predictive value remains below 1 percent at all tested thresholds. Random-style scoring, which emphasizes resemblance to random users, performs slightly better for elites but still labels the majority of ordinary and engaged users as Russian-like, again with PPV near or below 1 percent. In sum, these results indicate that although account-level aggregation produces clearer and more stable rhetorical profiles, content-only signals cannot support operational detection thresholds without sweeping in large volumes of legitimate domestic political speech, confirming H5's prediction that rhetoric alone cannot yield usable enforcement thresholds at realistic base rates.

5.5 Discussion

Study 3 asked whether the rhetorical patterns observed at the tweet level persist when features are aggregated to accounts, and whether those account-level profiles can plausibly support intervention thresholds on a single platform. The models are deliberately constrained to content-based features drawn from Moral Foundations Theory (MFT), NRC emotions, partisan-word measures, and TF–IDF over tweet text; they do not incorporate network structure, coordination,

timing, non-textual content, or off-platform intelligence. Within those limits, the account-level results sharpen the picture from Study 2 rather than overturn it. (Alizadeh et al., 2020; Zannettou et al., 2019; Serafino et al., 2024; Lewinski & Hasan, 2021).

First, aggregating rhetoric to accounts confirms that Russian IRA profiles sit firmly in the non-elite region of U.S. discourse. All three account-level classifiers locate most Russian accounts among random and politically engaged users rather than among politicians. The baseline rhetorical model cleanly separates elites from non-elites and places Russian accounts on the non-elite side of that divide, with only a small fringe ever classified as politician-like. In this feature space, Russian operators do not sustain the dense moral and partisan style that characterizes U.S. officeholders and high-profile activists.

Second, the models clarify what is doing the work. Across specifications, Authority and related moral foundations, along with a subset of emotions, provide the primary separation between elites and everyone else. Partisan intensity certainly contributes—especially when aggregated across an account's history—but the ablation models show that moral structure alone is sufficient to keep Russian accounts in the non-elite band. Rather than exhibiting a unique or exotic partisan vocabulary, Russian accounts appear as elite-adjacent civic actors: they use institutional and partisan language more than random users, but less intensely and less consistently than domestic elites.

Third, the TF–IDF model shows what happens when the classifier sees the full domestic political lexicon instead of hand-engineered rhetorical features. Once given a 10,000-term view of each account's text, the model leans heavily on explicit political and institutional words—covid-19, Congress, GOP, Senate, SCOTUS, media outlets, and culture-war terms—and becomes even

more confident about who is an elite. Under this richer representation, Russian accounts move further away from politicians rather than closer to them. The more clearly the model sees the language of domestic partisanship, the more Russian accounts look like non-elite, moderately engaged users who are participating in public affairs without fully adopting the intense partisan register of U.S. political elites. (Alizadeh et al., 2020; Tian et al., 2025).

Finally, the prevalence-aware analysis exposes the operational limits of content-only detection. Once the account-level scores are turned into "political-style" and "random-style" indicators and evaluated under realistic base rates, even reasonably accurate models break down as enforcement tools. Thresholds that capture a large share of Russian accounts simultaneously flag vast numbers of U.S. politicians and engaged users, and positive predictive value remains extremely low. In practice, a platform acting directly on these scores would primarily sanction domestic accounts while still missing many Russian operators. (Mathew et al., 2024; Tian et al., 2025).

Taken together, Study 3 shows that account-level rhetorical models are useful for mapping and interpreting the position of Russian IRA accounts within U.S. discourse. Consistent with content-based detection work that treats "detectability" as an empirical property that varies by context and signal design, the account-level models recover a stable non-elite placement for Russian accounts and clarify the feature families driving that geometry (Alizadeh et al., 2020; Zannettou et al., 2019; Serafino et al., 2024; Lewinski & Hasan, 2021). The same convergence holds under a theory-agnostic lexical baseline: when the classifier is given a broad TF–IDF view of account histories, it becomes more confident in identifying elite institutional language and moves Russian accounts further from politicians rather than closer to them, reinforcing the interpretation that Russian operators sustain a moderated civic register across time (Tian et al.,

2025). Finally, translating these scores into prevalence-aware operating points makes explicit how base rates constrain precision in realistic settings, even when separation on balanced evaluation data appears strong (Mathew et al., 2024; Tian et al., 2025). In this sense, content-only account-level models are best understood as analytic instruments that support comparative characterization and risk estimation, rather than as stand-alone levers for sanctions, downranking, or account removal.

5.6 Policy Implications

Study 3 clarifies both the potential and the limits of using rhetorical features to manage foreign influence on social media. At the account level, the models show that Russian IRA accounts sit squarely in the non-elite region of U.S. discourse: they rarely resemble politicians, and even under generous assumptions they look more like ordinary or moderately engaged users than like domestic elites. The TF-IDF models sharpen this contrast by revealing how heavily U.S. political accounts rely on explicit partisan and institutional language such as covid-19, congress, gop, senate, scotus, voters, and similar terms while Russian accounts use that lexicon less densely, at levels that are closer to engaged users than to elites. This supports the theoretical claim that Russian operations pursue a strategy of moderated, elite-adjacent rhetoric that does not fully embrace the dense partisan signals that define much domestic elite communication. From an analytic perspective, these models are valuable for mapping where influence accounts sit within the broader ecosystem and for quantifying how far their style deviates from elites and hyper-partisans.

The PPV analysis, however, shows that this analytic value does not translate into a workable enforcement rule. When the baseline account model is converted into political-style and random-

style scores and evaluated under realistic prevalences, any threshold that captures a substantial share of Russian accounts produces intolerably high false-positive rates. Political-style thresholds that capture 70–90 percent of Russians flag essentially all U.S. politicians and nearly all engaged users, with overall false-positive rates above 70 percent and PPV below one percent even when Russian accounts are assumed to comprise one percent of all users. Random-style thresholds perform slightly better for elites but still label the majority of ordinary U.S. accounts as "Russian-like," again yielding PPV on the order of fractions of a percent. In practical terms, these models cannot be used as stand-alone detectors without systematically over-capturing domestic speech and generating large volumes of "hits" that are overwhelmingly false positives. This directly supports H5: rhetorically defined political- or random-style scores are unsafe at useful recall levels for operational interventions.

For platforms and regulators, the implication is that rhetorical style should be treated as an analytic and diagnostic input, not as a primary enforcement trigger. These models are well suited to offline tasks: identifying clusters of accounts whose rhetoric sits in suspicious regions of the U.S. discourse space, characterizing how influence operations mimic or diverge from domestic actors, and informing qualitative investigation. They are poorly suited to automated removal, downranking, or labeling at scale, because the same stylistic features that distinguish Russian accounts from elites are heavily shared with ordinary and engaged domestic users. Any system that uses these scores as hard thresholds would primarily act on U.S. citizens and elected officials, not on foreign operators.

A more defensible policy approach would treat rhetorical signals as one layer in a multi-dimensional risk assessment that also incorporates structural and behavioral evidence: network

ties to known operations, coordination patterns, cross-platform reuse, IP and infrastructure indicators, and account provenance. In such a setting, moral and lexical features could be used to prioritize accounts for further review or to weight other signals, rather than to independently justify sanctions. Transparency is also critical: because style-based models are naturally biased toward high-volume political communication, any deployment should disclose their limitations, especially the low PPV and the potential for disproportionate impact on activists and elites.

Overall, Study 3 suggests that content-only models are powerful tools for understanding where Russian influence operations sit within U.S. political discourse and for theorizing their rhetorical strategies, but they cannot bear the weight of policy decisions on their own. Effective and legitimate responses to foreign influence will require combining these rhetorical insights with structural intelligence and procedural safeguards, ensuring that efforts to counter adversarial campaigns do not become broad instruments against ordinary political participation.

Chapter 6: Policy Analysis and Conclusion

6.1 The U.S. Policy Landscape

The United States approaches foreign disinformation with a mixture of old statutes, fragmented institutions, and a constitutional design that is better at limiting government than at managing information. The federal system was built to protect civil liberties and restrain centralized power, not to prevent citizens from encountering harmful or misleading speech. In that design, the First Amendment is not an inconvenient obstacle but the core constraint that must be preserved. There is no single law or central agency tasked with managing the information domain. Instead, authority is splintered across criminal law, foreign intelligence, and regulatory powers, each with its own limits and oversight mechanisms.

This fragmentation reflects a deeper philosophical split between foreign and domestic authority. The United States has a strong tradition of allowing the federal government broad latitude to defend the country from external threats and to prosecute individuals who violate the social contract through crime. It has a much weaker tradition of allowing that same government to shape or police the content of ordinary political speech. Foreign intelligence and domestic law enforcement are sharply separated in statute and practice. Disinformation sits awkwardly across that divide. It is often orchestrated abroad, but it operates through the everyday speech of people inside the United States. Any attempt to build a centralized “information agency” would cut against these constitutional and institutional norms, which is why the American approach has evolved as a loose federation of offices and programs scattered across the government rather than a single command structure for the information environment.

On the legal side, the Foreign Agents Registration Act (FARA) remains the primary tool for forcing disclosure of foreign influence. Enacted in 1938 and amended repeatedly since, FARA requires agents of foreign principals engaged in political activities or public-relations work "within the United States" to register and disclose their activities (U.S. Department of Justice, n.d.). The statute is built around physical presence and agency relationships in U.S. territory; it makes no clear provision for foreign actors who run influence campaigns entirely from abroad through social media, targeting U.S. audiences without using registered agents or organizations on U.S. soil. In practice, enforcement has been sporadic and often triggered by high-profile cases rather than systematic coverage. At the same time, First Amendment doctrine has been interpreted by the Supreme Court to protect not only the right to speak but also the right to receive information and ideas, making it difficult for the government to restrict access to lawful content even when it is misleading or foreign in origin (Lamont v. Postmaster General, 381 U.S. 301 (1965)). Together, these features leave the United States with a disclosure regime that is poorly matched to borderless online influence and a constitutional framework that sharply limits direct state control over what information citizens can see.

The Smith–Mundt Act of 1948, and its 2012 modernization, primarily regulate U.S. government information programs directed at foreign audiences and the domestic availability of those materials, clarifying boundaries for public diplomacy but doing little to address foreign manipulation on private platforms (Sager, 2015). More recent proposals such as the RESTRICT Act and the Protecting Americans from Foreign Adversary Controlled Applications Act focus on information and communications technology products and foreign-owned applications, authorizing the Department of Commerce to review and potentially prohibit transactions that pose national security risks, but they are largely concerned with ownership and data flows rather

than the day-to-day practice of running influence campaigns through ordinary-looking social media accounts (S. 686, 2023; Warner, 2023).

Institutionally, responsibility for foreign disinformation has never sat in a single "information agency." It has been spread across the Cybersecurity and Infrastructure Security Agency (CISA), the State Department's Global Engagement Center (GEC) and its successor office for foreign information manipulation, elements of the intelligence community, and the FBI (U.S. Department of State, n.d.). CISA tried to frame disinformation as a risk to critical infrastructure and election security. The GEC was created to coordinate U.S. efforts to recognize, understand, expose, and counter foreign propaganda and disinformation. The intelligence community could attribute and report on foreign activity but was constrained in what it could do domestically. All of these actors depended on voluntary cooperation from platforms, and none of them ever had a clear charter or sustained resources to manage the broader information ecosystem.

As Section 6.1.2 describes in more detail, the current administration has since dismantled or sharply downgraded many of these arrangements, allowing the GEC's authority to lapse, closing its successor office, disbanding the FBI's foreign influence task force, and hollowing out advisory and review structures around CISA. What was once a thin, fragmented architecture for tracking foreign information operations has therefore become even more attenuated, leaving a looser federation of agencies, state and local officials, and private actors to confront increasingly sophisticated influence campaigns.

Layered on top of this is the role of private platforms and Section 230 of the Communications Decency Act, which generally shields online intermediaries from liability for user-generated content while allowing them wide latitude in moderation (Congressional Research Service, 2024;

Electronic Frontier Foundation, n.d.). In practice, this has meant that platform policies about disinformation, state-linked media, and coordinated inauthentic behavior are set internally, influenced by public pressure, media cycles, and occasional congressional hearings. The government can nudge and complain. It cannot easily direct. That constraint is often frustrating for policymakers, but it is also one of the main safeguards against direct state control of political speech.

The TikTok divestiture saga illustrates how these legal and constitutional tensions play out in practice. In 2024, Congress enacted the Protecting Americans from Foreign Adversary Controlled Applications Act, requiring ByteDance to divest TikTok's U.S. operations within 270 days, with a single 90-day extension, or face an effective ban; the Supreme Court upheld the statute against a First Amendment challenge in January 2025 (Congressional Research Service, 2024, 2025; TikTok Inc. v. Garland, 2025). When the initial deadline passed and TikTok briefly shut down in the United States, the incoming Trump administration issued a series of executive orders delaying enforcement well beyond the one-time extension contemplated by the statute and directing the Department of Justice not to impose penalties for prior noncompliance (The Verge, 2025). Commentators have argued that these serial delays and non-enforcement orders sit in significant tension with the law's text and effectively nullify Congress's timeline for divestiture (Congressional Research Service, 2025; Sprigman, 2025). The resulting deal channels majority control of TikTok's U.S. business to a consortium of American investors closely aligned with the president, including Oracle, private-equity firm Silver Lake, and media and technology figures such as Larry Ellison, Rupert Murdoch, and Michael Dell, while allowing ByteDance to retain roughly a 20 percent stake and to license its recommendation algorithm to the new entity (Reuters, 2025; Washington Post, 2025). The federal government will receive a multibillion-

dollar one-time payment from the investor group paid into the Treasury, a structure that critics have described as a "shakedown" or form of crony capitalism rather than a neutral national-security remedy (Allyn, 2025; LAist, 2025). For present purposes, the TikTok case underscores the central dilemma of U.S. information policy: when Congress grants broad authority to restrict access to a foreign-owned platform, enforcement quickly becomes entangled with partisan incentives, favored investors, and opaque deals, reinforcing fears that any centralized disinformation regime would be hard to separate from ordinary struggles over control of the media environment.

6.1.1 Post-2016 Architecture

After the 2016 election, and the public confirmation of Russian interference, the federal government began to build an ad hoc architecture for addressing foreign information operations. Congress authorized the GEC to coordinate counter-disinformation efforts at the State Department (Global Engagement Center, 2024; Johnson, 2024). The FBI created a Foreign Influence Task Force to bring together counterintelligence, cyber, and criminal tools against foreign influence campaigns targeting U.S. elections. CISA emerged as a central hub for election security, including disinformation-related threat information sharing with state and local election officials (Karnowski & Smyth, 2025). At the Department of Justice, Task Force KleptoCapture was set up in March 2022 to enforce Russia-related sanctions and pursue oligarch assets after the invasion of Ukraine, targeting some of the financial networks behind Kremlin-linked operations (U.S. Department of Justice, 2022; Bernstein et al., 2022).

Election infrastructure itself was drawn more explicitly into the national security frame. In January 2017, the Department of Homeland Security designated election systems as a component of U.S. critical infrastructure, a move later reflected in DHS and Election Assistance Commission guidance on securing election systems (Johnson, 2017; U.S. Election Assistance Commission, 2017; Congressional Research Service, 2019). CISA and its partners supported the growth of the Elections Infrastructure Information Sharing and Analysis Center (EI-ISAC), which provided a dedicated channel for cyber and disinformation-related threat sharing with state, local, tribal, and territorial election officials (Center for Internet Security, n.d.; CISA, 2019, 2024). These steps did not federalize election administration, but they did create a standing mechanism for technical assistance and coordinated warning about foreign cyber and information operations targeting election infrastructure.

On the military side, the Department of Defense reframed offensive and defensive cyber operations around the concepts of "defend forward" and "persistent engagement." The 2018 DoD Cyber Strategy declared that the United States would "defend forward" by disrupting malicious cyber activity at or near its source, including activity below the level of armed conflict (U.S. Cyber Command, 2022). Congress reinforced this posture in the Fiscal Year 2019 National Defense Authorization Act: Section 1632 affirmed that cyberspace operations, including clandestine operations short of hostilities, constitute a "traditional military activity," and Section 1642 explicitly authorized U.S. Cyber Command to conduct operations against foreign adversaries' information systems to disrupt ongoing attacks or preparations (Congressional Research Service, 2019). These new authorities underwrote a series of preemptive actions, including a 2018 operation in which U.S. Cyber Command reportedly disrupted internet access

for Russia's Internet Research Agency during the midterm elections in order to degrade its ability to run social-media influence campaigns in real time (Nakashima, 2019; Thomas, 2019).

Law-enforcement and intelligence agencies also began to use traditional tools against foreign information actors. Special Counsel Robert Mueller's office secured a 2018 indictment of the Internet Research Agency, its financier, and associated individuals for conspiracy to defraud the United States and related offenses, explicitly linking Russian social-media operations to violations of U.S. election and fraud laws (United States v. Internet Research Agency LLC, 2018; U.S. Department of Justice, 2018a, 2018b). A separate 2018 indictment charged twelve GRU officers with hacking-related offenses aimed at interfering in the 2016 election (U.S. Department of Justice, 2018c). The Office of the Director of National Intelligence, for its part, began issuing declassified assessments of foreign threats to U.S. elections, summarizing intelligence community judgments about Russian, Chinese, Iranian, and other activities and providing a public baseline for discussions of foreign malign influence (Office of the Director of National Intelligence, 2021, 2024).

These pieces did not add up to a coherent "information strategy." They grew out of different statutory authorities and political impulses. Together, however, they constituted a set of guardrails. Foreign influence campaigns could be monitored, investigated, and sometimes exposed. State and local election officials had a federal partner for cyber and information-threat reporting and contingency planning. Platforms had at least some institutional counterparts in government for structured communication about foreign operations.

6.1.2 The Second Trump Administration's Retrenchment from Counter–Foreign Influence

That architecture has now been systematically thinned out. On the diplomatic side, Congress allowed the GEC's legislative authority to expire in December 2024; the office ceased operations as a stand-alone entity when its mandate lapsed (Johnson, 2024). The State Department briefly reorganized its functions into a smaller Counter Foreign Information Manipulation and Interference office, but in April 2025 Secretary of State Marco Rubio announced that he was closing that office as well, criticizing it as a vehicle for "censorship" and misused taxpayer funds (Psaledakis, 2025). Current and former officials described the move as dismantling one of the few units dedicated to systematically mapping Russian, Chinese, and Iranian disinformation campaigns (Scherling, 2025).

At the Department of Justice, Attorney General Pam Bondi announced in February 2025 that the FBI's Foreign Influence Task Force would be disbanded and that FARA enforcement would be narrowed to conduct similar to traditional espionage (Johnson, 2025). The same administration terminated Task Force KleptoCapture on 5 February 2025, signaling a shift away from aggressive enforcement of Russia sanctions against oligarchs and related networks (Reuters, 2025a). These decisions remove centralized focal points inside DOJ for investigating foreign influence and targeting financial enablers of Kremlin-linked operations.

Within the Department of Homeland Security, the administration has dismantled advisory and review structures that previously served as guardrails for cyber and information-security policy. In January 2025, DHS terminated all memberships on its advisory committees, including the Cyber Safety Review Board, abruptly ending a major investigation into Chinese-linked intrusions into U.S. telecommunications networks and eliminating a key mechanism for cross-

sector learning (CBS News, 2025). CISA, which played a central role in defending the 2018 and 2020 election cycles, has seen significant staff reductions, shifting priorities, and reduced engagement with state and local partners; election officials now report that the agency has been largely absent from recent planning and express concern about its role heading into the 2026 midterms (Karnowski & Smyth, 2025). A separate assessment describes vacancy rates approaching 40 percent at CISA and a broader weakening of federal cyber defense capabilities just as state adversaries are exploiting AI tools to automate sophisticated intrusions and information operations (Harwell & Nakashima, 2025).

The current trajectory also has roots in the previous Trump administration's confrontation with CISA's leadership after the 2020 election. Christopher Krebs, the first director of CISA and a Trump appointee, oversaw the federal effort to secure the 2018 and 2020 election cycles. In November 2020, following a joint statement by CISA and state and local partners describing the 2020 election as "the most secure in American history," Krebs publicly rejected claims that the election had been stolen (CISA, 2020). President Trump responded by dismissing Krebs via Twitter on 17 November 2020, accusing him of making a "highly inaccurate" statement (Baker, 2020). The firing was widely interpreted as punishment for CISA's refusal to endorse election-fraud allegations and as an early signal that expertise-based guardrails on election and information security could be removed when they conflicted with political narratives. The subsequent staffing cuts and strategic downgrading of CISA in 2025 continue that pattern on a larger scale, transforming what had been a relatively robust, expert-driven hub for election and cyber defense into a hollowed-out agency with diminished capacity to counter foreign information threats.

The Pentagon has followed a similar retrenchment. In early 2025, multiple outlets reported that Defense Secretary Pete Hegseth ordered U.S. Cyber Command to halt or pause planning and offensive cyber and information operations targeting Russia, rolling back aspects of the "defend forward" posture that had been developed after 2018 (CBS News, 2025). The Pentagon initially denied that any stand-down order had been given, but in May 2025 Representative Don Bacon, chair of the House Armed Services cyber subcommittee, confirmed in open testimony that Cyber Command had paused offensive operations against Russia as part of negotiation tactics during Ukrainian President Volodymyr Zelenskyy's visit to Washington (Politico, 2025). Analytical commentary characterized the episode as a symbolic reversal of the previous decade's efforts to push back against Russian cyber and information operations, with critics arguing that suspending Cyber Command's activity, even briefly, undercut persistent engagement and signaled a willingness to trade away cyber pressure in pursuit of a fragile diplomatic opening (PRIF Blog, 2025).

These moves can be described as policy shifts: rebalancing enforcement priorities, trimming budgets, reining in perceived "censorship," or recalibrating military risk. They amount to a deliberate downgrading of U.S. capabilities to systematically track, attribute, and expose foreign information campaigns and to disrupt them at their source. Journalistic and analytical accounts now talk about disappearing guardrails against disinformation and "unilateral disarmament" in the information space in Trump's first 100 days (Scherling, 2025). The empirical chapters of this dissertation show Russian operators converging on a strategy of rhetorical camouflage that is precisely the kind of pattern that requires sustained analytic attention, institutional memory, and cooperation across agencies and with platforms to detect. The current policy trajectory moves in the opposite direction, leaving a more diffuse, uncoordinated set of actors such as individual

election officials, civil-society groups, and internal platform teams to confront increasingly sophisticated influence operations without a comparably resourced federal counterpart.

6.2 Summary of Findings

Across all three studies, this dissertation asked a concrete question: how do Russian influence accounts actually use language, and what does that mean for policy. First, it examined whether Russian tweets rely on moral framing, emotional appeals, and partisan terms, and whether those features increase engagement. Second, it compared Russian rhetoric to tweets from U.S. politicians, politically engaged users, and random users to see which domestic voices they most closely mimic. Finally, it used those findings to assess what kinds of interventions against foreign influence are feasible in a system that must still protect ordinary political speech. Research Question 1 (RQ1) asked how Moral Foundations language, NRC emotions, and partisan vocabulary are associated with engagement in Russian IRA tweets, and Hypothesis 1 (H1) predicted that MFT-based moral language would be more strongly associated with engagement than partisan or purely emotional cues. Research Question 2 (RQ2) asked whether Russian influence tweets resemble the rhetorical style of U.S. politicians, politically engaged users, or random users, with Hypothesis 2 (H2) predicting that Russian content would more closely resemble non-elite, engaged users than political elites. Research Question 3 (RQ3) asked which rhetorical features, Moral Foundations, NRC emotions, and partisan vocabulary, most strongly differentiate these groups, and Hypothesis 3 (H3) predicted that partisan intensity would emerge as the dominant signal, followed by moral foundations and then emotion. Research Question 4 (RQ4) asked whether more flexible lexical representations such as TF–IDF, at both the tweet and account levels, would alter or reinforce this picture of Russian moderation, with Hypothesis 4 (H4) predicting that unconstrained text models would recover the same basic

structure. Research Question 5 (RQ5) asked whether any of these content-only features, especially when aggregated to accounts, could support operational detection thresholds under realistic base rates, and Hypothesis 5 (H5) predicted that prevalence-aware performance would be too weak to justify using rhetoric alone as an enforcement trigger.

Study 1 focused on engagement. Using Moral Foundations Theory scores and NRC emotion categories, it showed that moral and emotional appeals still matter a great deal in the modern information environment. Tweets that invoked moral foundations or carried stronger emotional affect were more likely to be retweeted and shared, even when the source was a foreign influence account rather than a domestic actor. The familiar observation that moralized, emotional rhetoric spreads further did not vanish in the move from pamphlets to social media; it scaled up. The study also showed that partisan words were present but unevenly distributed. They contributed to engagement, but in a way that depended on context and was far from uniform across users.

Study 2 shifted the question from "what gets retweeted" to "who looks like whom." The unit of analysis was the individual tweet, and the models compared Russian content to tweets from U.S. politicians, politically engaged users, and random users. Extreme partisan signals were rare, and the moral and emotional profile of Russian content fell into the same broad band as ordinary civic discourse. From the vantage point of rhetoric, Russian operators were not shouting fringe slogans; they were imitating the middle of the conversation.

Study 3 aggregated these features to the account level and asked a different question: if we step back from individual messages and look at the overall rhetorical pattern of an account, can we reliably separate Russian operators from U.S. users. The baseline rhetorical model, using percent-present moral foundations, NRC emotions, and mean partisan word usage, achieved

solid performance on U.S. accounts, cleanly separating politicians from non-elites while confusion remained concentrated on the engaged-versus-random boundary. Authority, followed by trust, Loyalty, and disgust, emerged as the dominant predictors; mean partisan word count was important but secondary to moral framing. Applied to Russian accounts, this baseline model labeled the vast majority as non-elite, primarily random and engaged users, with only a very small share resembling politicians. Removing partisan features barely degraded performance and shifted Russian accounts slightly further toward random users, confirming that moral structure, especially Authority, carries most of the account-level signal and that partisan intensity refines but does not define Russian placement.

An account-level TF–IDF model provided a complementary perspective. By representing each account with a high-dimensional TF–IDF vector built from its concatenated tweets, that model attained somewhat higher accuracy and macro–F1 on U.S. accounts than the rhetorical model and leaned heavily on explicitly political and institutional vocabulary, terms such as covid-19, congress, GOP, senate, SCOTUS, and similar partisan and culture-war signifiers. Yet, despite its improved performance on U.S. accounts, the TF–IDF model pushed Russian accounts even further into the non-elite region. The account-level TF–IDF model labeled 86.4 percent of Russian accounts as random and almost all of the rest as engaged, with virtually none classified as politicians.

The richer the representation of domestic political language, the less Russian accounts resembled U.S. elites. In the aggregate, partisan and institutional terms are densest in elite accounts, with Russian accounts using this lexicon less intensively than politicians but at levels that overlap with engaged users rather than avoiding it entirely. That pattern reinforces the interpretation that

Russian operators occupy an intermediate, elite-adjacent rhetorical register rather than sharing the full partisan saturation that characterizes many domestic political actors.

These findings show that Russian influence campaigns rely less on transparent propaganda and more on mimicry. Their success depends on sounding enough like ordinary political conversation to blend in while quietly adjusting emphasis and framing. That rhetorical camouflage frustrates naïve content filters and simple "disinformation" labels. It also complicates policy responses that treat foreign influence as something that will always look obviously foreign.

6.3 Extending the Framework to China, Iran, and North Korea

Although this dissertation focuses on Russian activity, the account-level frameworks developed here are not specific to Moscow. China, Iran, and North Korea all run information operations against the United States and its allies, though with different objectives. The question is whether the Russian pattern of rhetorical mimicry, intermediate partisan intensity, and account-level camouflage among non-elites is unique to Russian influence or a broader template foreign influence actors follow. The account-level framework developed here offers a way to test that, by treating rhetoric as a measurable habit over time rather than a series of isolated posts. The People's Republic of China operates the broadest and most resource-intensive information strategy among U.S. adversaries. A 2023 special report by the State Department's Global Engagement Center describes how Beijing seeks to reshape the global information environment through a blend of state media, covert content placement, co-opted local outlets, and coordinated online campaigns (U.S. Department of State, 2023). The report documents tactics that include acquiring stakes in foreign media companies, pressuring international organizations and

newsrooms to avoid criticism of the Chinese Communist Party (CCP), deploying fake personae and inauthentic accounts to amplify preferred narratives, and using domestic repression and censorship to silence dissent on Chinese platforms (U.S. Department of State, 2023). Subsequent analyses by Microsoft, Meta, and independent researchers show that these efforts increasingly rely on cross-platform networks, such as the "Spamouflage" operation, using thousands of coordinated accounts across Facebook, X, YouTube, TikTok, Reddit, and dozens of smaller forums to push pro-PRC messages and attack critics (Meta, 2023; Mandiant, 2021).

Rhetorically, much of China's external messaging emphasizes themes of stability, prosperity, national rejuvenation, and technocratic competence, contrasted with alleged Western hypocrisy, racism, and dysfunction (U.S. Department of State, 2023). When PRC-linked networks target specific foreign debates, for example U.S. arguments over race, public health, or foreign policy, they borrow familiar polarizing tropes and attempt to inject them into existing conversations rather than inventing entirely new narratives. Mandiant and Google have documented pro-China COVID-19 campaigns that spread conspiracy theories about the virus's origins, praised Chinese vaccines, and tried to mobilize protests in the United States, using English-language slogans and imagery designed to resemble grassroots activism (Mandiant, 2021; German Marshall Fund, 2021).

These tactics are not limited to the U.S. information space. Investigations in Europe and the Indo-Pacific show PRC-linked actors running localized influence operations that mimic domestic voices. A 2025 Reuters investigation into Chinese information operations in the Philippines found that a Chinese-funded marketing firm created fake social-media accounts posing as Filipino users, launched a pro-Beijing online outlet that mimicked local news, and paid local

elites to promote narratives that undermined the U.S.–Philippines alliance and lauded Chinese vaccines and investments (Reuters, 2025). Recent reports from Taiwan's government describe a 60 percent year-on-year increase in PRC disinformation incidents targeting the island's democracy, with fake accounts, AI-generated videos, and comment manipulation used to erode trust in elections and U.S. security commitments (Taiwan National Security Bureau, 2024).

China's operators are also experimenting with generative AI in ways that echo the dynamics observed in this dissertation. Microsoft's Threat Analysis Center reports that PRC-linked actors increasingly use AI to generate articles, memes, and deepfake-style images tailored to wedge issues in U.S. politics, including race, policing, and pandemic response, as well as to sustain long-running persona accounts on X and other platforms (Microsoft Threat Analysis Center, 2024). OpenAI has likewise reported small but growing PRC-linked campaigns using ChatGPT to produce polarized English-language content criticizing Taiwan-centric media, commenting on U.S. tariffs, and seeding misleading narratives into social-media threads, alongside AI-assisted tooling for cyber operations (OpenAI, 2025).

From the perspective of this study's framework, PRC information operations share three features with Russian IRA activity. First, they rely heavily on mimicry: fake personae and co-opted outlets adopt the tone and vocabulary of local civic and policy debate, rather than overt CCP rhetoric (U.S. Department of State, 2023; Mandiant Threat Intelligence, 2021; Freedom House, 2022). Second, they increasingly use moderated partisan cues and institutional language talking about public health, infrastructure, and security in ways that resemble domestic bureaucratic or advocacy speech rather than consistently deploying extreme slogans (U.S. Department of State, 2023; Microsoft Threat Analysis Center, 2024). Third, they are starting to exploit AI-assisted

content generation to scale that mimicry across languages and platforms (Microsoft Threat Analysis Center, 2024; OpenAI, 2025). An account-level rhetorical model similar to the one used here could therefore be adapted to PRC campaigns by incorporating PRC-specific moral and ideological markers (for example appeals to harmony, social order, anti-imperial justice, and national rejuvenation) alongside generic moral, emotional, and partisan features, and then asking whether PRC-linked accounts occupy a distinct engaged-like band in foreign discourse, as Russian accounts do in the U.S. Twitter space.

Iranian operations illustrate a different configuration of resources and rhetoric. The U.S. Department of Justice has documented a series of cyber-enabled disinformation efforts linked to Iranian actors, including the 2020 voter-intimidation campaign in which emails spoofed to appear from the Proud Boys threatened registered Democrats with physical harm if they did not vote for Donald Trump (U.S. Department of Justice, 2021; U.S. Attorney's Office, 2021). The indictment describes an operation that combined hacking to obtain voter information, intimidation through threatening emails and videos, and the dissemination of misleading claims about election integrity. Outside the United States, Iranian state media and proxy outlets promote narratives that present Iran as a defender of oppressed Muslims, frame the United States and its allies as corrupt and hypocritical, and cast regional rivals as puppets of foreign powers.

Compared with Russian content aimed at U.S. audiences, Iranian messaging often features more explicit ideological and religious content: moral condemnation of injustice and corruption, emotive appeals to humiliation and resistance, and frequent references to martyrdom and oppression. An account-level model for Iranian campaigns would therefore require a different moral vocabulary attuned to Shi'a religious references and anti-imperialist rhetoric, but the

underlying logic is the same. In the aggregate, do accounts associated with Iranian operations display distinctive combinations of moral and emotional language and partisan or ideological markers relative to the baseline conversation among the same diasporic or regional audiences. Answering that question empirically would clarify whether Iran leans more toward overt ideological mobilization or toward the kind of moderated mimicry identified in Russian operations.

North Korea occupies yet another corner of this space. Historically, the DPRK's propaganda was directed inward, with stylized films and broadcasts extolling the Kim family and vilifying external enemies. In the last decade, Pyongyang has experimented with outward-facing "soft" content: YouTube channels and short-form videos featuring young women showing daily life in Pyongyang, eating ice cream, shopping, or discussing pop culture, as well as lifestyle-oriented content on Chinese and Western platforms (Loh, 2023; Reuters, 2023). Analyses of channels such as *New DPRK* argue that they are run by state-linked media organizations and designed to present a sanitized, normal-looking North Korea to foreign viewers (Goldman, 2024; Radio Free Asia, 2023). In parallel, North Korean cyber operations have included efforts to plant pro-regime stories on foreign websites and to use fake profiles to spread disinformation about sanctions, nuclear tests, or military exercises (Reuters, 2023).

Here, the rhetorical camouflage problem takes a slightly different form. Many overtly propagandistic materials are easy to identify. The more subtle challenge lies in understanding how lifestyle and vlog content functions as a gateway, normalizing aspects of North Korean life while downplaying repression and militarization (Loh, 2023; Radio Free Asia, 2023; Goldman, 2024). Applying an account-level rhetorical model could help distinguish between state-linked

soft-power channels and genuinely independent content. For example, one could examine whether these accounts show unusually consistent affective profiles, with high levels of joy and pride and limited expressions of anger or sadness; whether references to the regime and its leadership are framed in systematically positive moral terms; and how this compares to ordinary travel or lifestyle influencers covering other countries.

Across these three states, one pattern is already clear in public reporting: they face the same operational problem Russia confronted in the Twitter data analyzed here. They need to operate inside someone else's information environment without triggering immediate defenses. Beijing spends heavily on online and offline media influence, using a mix of overt and covert tactics to normalize its narratives (U.S. Department of State, 2023). Iranian actors blend intimidation, ideological appeals, and opportunistic exploitation of existing U.S. tensions (U.S. Department of Justice, 2021; U.S. Attorney's Office, 2021). North Korea experiments with unusually glossy representations of everyday life to soften its image while maintaining more traditional hard-edged propaganda elsewhere (Goldman, 2024; Loh, 2023; Reuters, 2023). Extending the account-level framework to these cases would allow researchers to test whether the Russian pattern that uses intermediate partisan intensity, engaged-like rhetoric, and less dense explicit partisan markers than domestic elites is unique to the IRA or a broader feature of contemporary foreign influence techniques.

6.4 International Regulatory Models: The EU and Beyond

The European Union has taken a more explicit regulatory path in managing the information environment. It operates in a legal context that protects expression but does not have an

equivalent to the U.S. First Amendment. As a result, the EU has been able to develop a set of tools that sit between self-regulation and full command-and-control of the information domain.

In 2018, following concerns about Russian interference in U.S. and European elections and the Cambridge Analytica scandal, the European Commission convened major platforms and advertising firms to develop a voluntary Code of Practice on Disinformation (European Commission, 2018). The Code defined disinformation, set out commitments for demonetization of disinformation, transparency in political advertising, reduction of fake accounts and bots, and support for media literacy, and required signatories to submit annual self-assessment reports (European Commission, 2018; EU Code of Practice on Disinformation, 2024). A 2022 "strengthened" version expanded the number and specificity of commitments and was explicitly designed to be recognized as a code of conduct under the Digital Services Act (DSA), but it remains a self-regulatory instrument whose impact depends on platform cooperation and uneven reporting (Mündges, 2024).

Building on that experience, the EU adopted the DSA, a binding regulatory framework for online intermediaries and very large online platforms. The DSA applies in principle to all intermediary services in the EU but imposes its heaviest obligations on "Very Large Online Platforms" and "Very Large Online Search Engines" with more than 45 million monthly active users, including many of the global social-media and search services that host contemporary political debate (European Commission, 2022; Digital Services Act, 2022). These very large services must assess systemic risks, including those related to disinformation and electoral integrity, implement mitigation measures, and provide meaningful data access to vetted researchers, along with detailed transparency reporting about moderation practices and recommender systems (European

Commission, 2022, 2025). A 2025 delegated act further specifies how these very large platforms must grant data access to qualified researchers to study systemic risks and mitigation efforts (European Commission, 2025; Weizenbaum Institute, 2025).

The EU model is not about the state micromanaging content. It is about imposing process duties. Platforms must assess, document, and mitigate systemic risks, explain how their algorithms work in broad terms, and allow external researchers and regulators to examine whether mitigation measures are effective. Enforcement happens through EU institutions and national digital services coordinators rather than ad hoc political pressure, and the DSA gives the Commission tools ranging from audits and corrective orders to fines and, in extreme cases, bans from the single market (European Commission, 2022; Protiviti, 2023).

At the same time, the European approach has important limitations that matter for this dissertation's concerns. First, the most demanding obligations, including systemic risk assessments, independent audits, and researcher data access, apply only to very large services; smaller platforms, niche networks, and many messaging applications fall under lighter, more generic rules and can remain largely outside the disinformation-focused parts of the regime (Digital Services Act, 2022; Halil, 2025). Second, neither the Code of Practice nor the DSA create a comprehensive system for tracking the country of origin of accounts or messages. They focus on platform processes and political advertising transparency, not on systematically labeling whether content originates with foreign or domestic actors, and they do not by themselves solve the attribution and provenance problems highlighted in this dissertation (European Commission, 2018; Wolters, 2025). Third, the strengthened Code remains voluntary and compliance has been uneven, with analyses finding gaps and inconsistencies in how major signatories such as Google,

Meta, and TikTok report and implement their commitments (Mündges, 2024; EU DisinfoLab, 2022). Data-access provisions under the DSA are still in early stages and researchers report friction and partial responses to requests, underscoring the gap between formal rights and operational access (Weizenbaum Institute, 2025; Stiković, 2024). Finally, even under the EU's relatively ambitious transparency and data-access rules, researchers cannot easily obtain current, actor-linked datasets comparable to the Russian IRA corpus used here. As a result, much of the empirical literature and many policy discussions rely on historical data and partial disclosures. This gap between formal transparency obligations and usable research access is itself a structural limitation on evidence-based regulation of foreign influence.

It is also unclear why the EU has not been more aggressive in exploiting its regional vantage point for provenance and infrastructure-level tracking. Later sections discuss how the mechanics of the internet complicate attribution, but even with the limitations of IPv4 and the widespread use of VPNs and proxies, infrastructure-level patterns are not meaningless. Foreign influence operations and commercial influence farms repeatedly reuse the same VPN providers, hosting ranges, and access patterns, and these "tactics, techniques, and procedures" (TTPs) can be aggregated into probabilistic provenance signals. A serious regulatory regime under frameworks like the DSA could require large platforms to monitor and report on these recurring infrastructure clusters, whether they originate from Russian access points or from paid influence firms in third countries, and to combine that information with behavioral and rhetorical indicators. The problem is not that such signals are unattainable; it is that existing transparency and enforcement mechanisms stop well short of exploiting them systematically.

From a U.S. perspective, the EU experience offers two clear lessons and one red line. First, governments can demand transparency, access, and basic accountability for how platforms manage information without dictating substantive viewpoints. Second, this kind of regulatory scaffold is most effective when it is treated as a long-term governance regime rather than a short-term political weapon. The red line is that the EU model assumes a degree of state authority over platform practices, including systemic-risk definitions and mitigation duties, that would be much more difficult to justify under U.S. constitutional doctrine. An American approach can borrow the procedural elements, such as transparency, researcher access, and risk assessments, but it cannot easily import European-style obligations to "combat disinformation" in ways that are likely to be interpreted as viewpoint-based, and it would still face the same attribution and base-rate problems documented in this dissertation.

6.5 The Role of the Private Sector in the Information Environment

The infrastructure that carries modern political speech is overwhelmingly private. The long-haul lines, local access networks, and data centers that move packets across the network are owned by telecommunications and cloud companies. The dominant venues for public discourse are advertising-funded platforms. These businesses do not operate as public utilities; their legal obligations are limited, and their primary incentives are commercial.

At the network layer, longstanding reliance on IPv4 makes even basic origin-tracking structurally difficult. A 32-bit address space supports roughly 4.3 billion unique IPv4 addresses, but tens of billions of devices are now online, a gap that is managed through extensive use of network address translation (NAT) and carrier-grade NAT (CGNAT), in which dozens or even hundreds of subscribers share a single public address (Cloudflare, 2025; NFware, 2023). In

practice, an IP address identifies an access network or shared gateway rather than an individual speaker; a single address may front a household, a coffee shop, a university dorm, or a commercial proxy service. Dynamic addressing, mobile networks, and VPNs further blur the picture: commercial IP-geolocation providers report country-level accuracy well above 95–99 percent but substantially lower precision at the region and city level, especially for mobile and rural connections (MaxMind, n.d.; NetAcuity, n.d.;). Even a complete transition to IPv6 would not eliminate these constraints. A larger address space reduces pressure for NAT, but privacy extensions, temporary addresses, and continued use of intermediaries still mean that addresses primarily identify networks and devices, not reliably distinct individuals.

Foreign operators can also route traffic through domestic infrastructure, commercial VPNs, cloud-hosting providers, or residential proxy services, so that their activity appears to originate from U.S. IP space regardless of physical location. Under these conditions, IP information is at best a weak and easily manipulated provenance signal. Any attempt to police foreign influence purely through IP space faces a rough trade-off: either rely on coarse blocks that risk substantial collateral damage or maintain detailed logs linking individual sessions to customers and ports, with corresponding privacy, compliance, and liability costs. Making IP-based attribution precise enough to support account-level enforcement would require pervasive logging and long-term retention that push toward mass data-retention regimes, raising significant privacy and constitutional concerns, while still failing to distinguish sophisticated foreign campaigns from ordinary domestic users.

From a commercial perspective, there is limited upside in investing heavily in this kind of fine-grained, auditable provenance. Advertisers generally need approximate location and rich

behavioral profiles rather than legally robust origin tracing, and existing real-time bidding markets already support highly targeted advertising without that additional expense (Electronic Frontier Foundation [EFF], 2025; LeadDigital, 2020;). The business model is straightforward. Users are not primarily charged for the amount of data they consume. Instead, platforms collect data on their behavior, including search queries, clicks, likes, follows, and dwell time, assemble behavioral profiles, and sell targeted access to those profiles to advertisers and political campaigns. The Cambridge Analytica scandal made this visible: investigative reporting showed that data from tens of millions of Facebook profiles were harvested without consent and used to build psychological profiles that could be targeted with personalized political advertisements (Cadwalladr & Graham-Harrison, 2018a, 2018b). The firm's work, linked to Steve Bannon and others, demonstrated how granular behavioral data could be weaponized for political microtargeting.

Shoshana Zuboff describes this broader pattern as "surveillance capitalism," an economic order that treats human experience as raw material for behavioral prediction markets (Zuboff, 2019). In this model, engagement is not a by-product; it is the product. Platforms engineer ranking systems and interfaces to maximize time on site, clicks, and reactions because those metrics drive advertising revenue. Content that provokes strong emotional responses, fuels conflict, or keeps users scrolling tends to be rewarded, regardless of whether it is accurate or beneficial for democratic discourse.

The empirical findings in this dissertation intersect directly with that incentive structure. Study 1 showed that moral and emotional language reliably increases engagement. Study 2 and Study 3 demonstrated that Russian operators have learned to exploit that fact without resorting to overtly

partisan or obviously foreign rhetoric. They sound like ordinary politically engaged users, use moral and emotional framing at levels that fit the ambient conversation, and avoid the most extreme partisan markers. In an engagement-driven ecosystem, this is a rational strategy. Accounts that mimic the tone and emotional profile of mainstream discourse are more likely to be rewarded by algorithms that care about interaction rather than origin.

Platforms are not neutral conduits. They write and enforce community standards that go far beyond First Amendment limits on state action, ban or demote some content, and boost other content. In principle, those tools can be used to blunt foreign influence operations. In practice, they are deployed in ways that reflect a blend of public pressure, brand management, and political risk, with limited transparency and weak external oversight. Recent moves by large providers underscore both the potential and the limits of private-sector intervention: in 2025, Google's vice president for threat intelligence, Sandra Joyce, described a planned "cyber disruption unit" that would proactively identify and legally disrupt threat campaigns, arguing that defenders "have to get from a reactive position to a proactive one" if they hope to make a difference (Uren, 2025). Such initiatives hint at a more assertive role for platforms in dismantling hostile operations, but they remain ad hoc, legally constrained, and focused on clearly illegal activity rather than the ambiguous political speech examined in this dissertation.

In late November 2025, X (formerly Twitter) rolled out an "about this account" feature that showed the country or broader region where an account is based and, in some cases, additional metadata such as the account's creation date and username history (Business Insider, 2025; NDTV, 2025). The initial implementation also displayed the location where an account was created, but widespread reports of inaccurate data, especially for older accounts and VPN users,

led X to remove the creation-location field within days (Robertson, 2025; NDTV, 2025). The rollout triggered immediate backlash: critics described the update as a form of "forced doxxing," particularly for users in authoritarian states, and X responded by adding warnings that VPNs can affect accuracy, allowing users in high-risk environments to generalize their region, and noting that location data may be updated with delays to protect privacy (Business Insider, 2025; Privacy Guides, 2025; TechRadar, 2025).

Despite these problems, the feature produced immediate and politically salient information. Journalists and researchers used it to examine U.S. political accounts and found that many prominent "Make America Great Again" and right-wing accounts presenting themselves as American patriots were in fact based in countries such as Russia, Nigeria, India, Thailand, and others (Business Insider, 2025; The Guardian, 2025). The same tool exposed foreign-run accounts posing as Gaza residents soliciting donations, as well as mis-labeled accounts whose locations were corrected after public pushback (Robertson, 2025). A single provenance feature thus made visible both genuine foreign participation in U.S. political discourse and the difficulty of building accurate labeling systems at scale.

From the perspective of this dissertation, this example matters for two reasons. It is a concrete implementation of the kind of provenance-oriented intervention recommended here: it shifts attention from the surface content of a tweet to the conditions of its production, including likely geographic origin. It also illustrates the need for independent evaluation and public oversight. Location labels can help identify foreign influence, but only if users have reason to trust that they are applied with reasonable accuracy and not selectively disabled when they produce politically inconvenient results.

The private sector is therefore not simply a vector for foreign influence. It is the terrain on which influence campaigns operate. Business models built on surveillance and engagement create an information environment that foreign and domestic actors can exploit. Government can respond by setting minimum standards through law, for example, by imposing transparency and research-access requirements akin to the Digital Services Act, by strengthening data-protection and competition rules, and by reviving net-neutrality-style constraints on infrastructure providers, and by building public–private cooperation on terms that recognize platform incentives and constraints. The empirical tools developed in this dissertation offer one way to structure such cooperation: not as a mechanism for automated censorship, but as a diagnostic instrument for mapping how foreign operations sit relative to domestic rhetorical baselines and for evaluating the likely side effects of interventions.

In sum, these dynamics suggest that the private sector's role in countering foreign influence extends well beyond individual content decisions. Infrastructure providers, cloud hosts, payment intermediaries, and social platforms all control structural and network-level levers like account creation, coordination detection, amplification, monetization, and friction in sharing, that can significantly shape the reach of foreign campaigns without relying on simple "leave up or take down" judgments. At the same time, these actors operate under powerful commercial, legal, and political constraints: they face free-rider problems in cross-platform coordination, litigation, and reputational risks when enforcement appears partisan, and persistent asymmetries between the rich data they hold and the limited visibility available to researchers and regulators. Before turning to specific policy proposals, it is therefore necessary to revisit the ethical and constitutional assumptions that have guided U.S. thinking about speech and foreign influence since the end of the Cold War. If Russian operations are best understood as part of a renewed

long-term contest in the information domain, then a purely domestic, content-neutral model of platform governance is unlikely to be adequate. The next sections examine how liberal-democratic commitments to free expression, audience autonomy, and the "right to receive" information might be reconciled with a more security-conscious approach to foreign political operations.

6.6 Ethical and Constitutional Considerations

Any serious U.S. policy on disinformation sits inside a narrow channel. On one bank is the national security impulse to stop foreign manipulation. On the other is the constitutional commitment to free expression and the historical suspicion of state control over political speech. It is easy to say that both can be honored. It is much harder to design actual institutions and laws that do so. After the attacks of September 11, 2001, the balance between liberty and security in the physical domain shifted visibly: airport screening, intelligence authorities, and surveillance practices were rapidly expanded in the name of preventing another catastrophic attack. In the information domain, by contrast, the United States still largely operates with a pre–social media conception of speech rights, even as foreign states treat the online environment as a continuous battlespace. The core challenge of this chapter is not whether to act, but how far liberal democracies can, and should, recalibrate their ethical and constitutional reasoning about political communication so that foreign adversaries cannot freely exploit U.S. information infrastructure while domestic political speech remains protected.

The European Union can ask platforms to "combat disinformation" with relatively little constitutional friction (European Commission, 2018, 2022). That is not because the European Union lacks rights protections but because its legal and political tradition balances individual

expression against other values in a way that differs from the United States. Many member states spent significant parts of the twentieth century under fascist or communist rule, and postwar constitutional settlements therefore place heavy emphasis on human dignity, group protection, and the prevention of extremist mobilization. Not to mention virtually all EU member states have a long history as monarchies. Within that framework, speech is understood as one fundamental right among others, to be weighed against equality, privacy, and public order, rather than as an almost absolute constraint on state action. Regulatory language that invites platforms to reduce harmful or misleading content therefore sits more comfortably in the European context than in a U.S. system built around a strong First Amendment tradition and a deeper suspicion of state involvement in political discourse.

These differences are sharpened by the way U.S. law structures the role of private intermediaries. Section 230 of the Communications Decency Act provides that providers of "interactive computer services" shall not be treated as the publisher or speaker of information supplied by users and also shields good-faith efforts to restrict content they consider objectionable (47 U.S.C. § 230; Congressional Research Service, 2024; Douek, 2023). This combination gives large platforms a hybrid position that no traditional utility enjoyed: they are broadly immune from liability for most user-generated content, yet they retain wide discretion to curate, rank, and remove that content according to their own policies. At the infrastructure layer, broadband providers are not generally regulated as public utilities either. The Federal Communications Commission briefly classified broadband as a Title II telecommunications service in the 2015 Open Internet Order and imposed net neutrality rules, but those rules were later repealed in the Restoring Internet Freedom Order and recent litigation has constrained the Commission's ability to treat internet service providers as common carriers under existing statutes (FCC, 2015;

Restoring Internet Freedom Order, 2018; U.S. Telecom Ass'n v. FCC, 2016; Sixth Circuit net-neutrality decisions, 2025). The result is an information infrastructure that is privately owned, enjoys strong statutory protections, and is not subject to the kind of neutrality and public-interest obligations that historically constrained common carriers and utilities. That arrangement was tolerable when foreign information operations were a marginal concern. In a renewed long-term contest with authoritarian powers, it amounts to granting foreign states de facto access to critical infrastructure that is neither regulated as a utility nor held to clear security obligations.

In the United States, the same "combat disinformation" language that the European Union can deploy, if enforced by the state, immediately raises questions about which speech is being targeted and why. Courts are skeptical of regulations that are content based and even more skeptical of regulations that appear aimed at particular viewpoints. The constitutional risks are not hypothetical. In Missouri v. Biden, several states and individual plaintiffs sued the Biden administration over efforts to "flag" misinformation for major platforms, arguing that behind-the-scenes pressure to remove COVID-19 and election-related content turned private moderation into state censorship. On July 4, 2023, a federal district court agreed and issued a sweeping preliminary injunction barring a wide range of executive-branch officials from urging platforms to remove or downgrade protected speech (Missouri v. Biden, 2023). The Fifth Circuit later narrowed the order but still concluded that the White House, the Surgeon General's office, the CDC, and the FBI had crossed the line into unconstitutional coercion by becoming "significantly entangled" in platforms' content-moderation decisions (Missouri v. Biden, 2023). Although the Supreme Court ultimately vacated the injunction on standing grounds in Murthy v. Missouri, it did not resolve the merits, and three justices described the record as documenting a "serious threat" to the First Amendment (Murthy v. Missouri, 2024). For present purposes, the key point

is that even relatively modest, informal government efforts to shape platforms' responses to "disinformation" have already triggered major litigation and produced findings of First Amendment violations in the lower courts. Those decisions do not mean that the government is powerless to respond to foreign information operations; they mean that it cannot simply outsource broad, content-based censorship of domestic speech to private platforms under the banner of "combating disinformation." In fact the user agreements of social media companies to adhere to their stated values are the only content moderation beyond explicitly illegal or directly threatening content.

Ethically and constitutionally, the lesson is that **domestic content control is the wrong lever**, not that foreign adversaries must be given equal access to U.S. information infrastructure. Legitimacy in this space depends on procedural neutrality and on a sharp distinction between domestic speakers and foreign principals. Any sustainable regime must be framed and implemented in ways that apply regardless of which party or ideology is currently harmed or helped by foreign narratives. A system that targets "disinformation" only when it undermines one side's candidates or causes will not survive legal scrutiny and does not deserve to. The empirical work in this dissertation reinforces that point. Russian operators do not rely on a distinct ideological vocabulary that can be easily banned or filtered. They borrow from mainstream political and civic language and adjust emphasis, context, and targeting. Any content rule aimed at their rhetoric will inevitably sweep up legitimate domestic speech as well. The appropriate target of regulation is therefore not the **content style** of individual messages but the **structural position** of foreign, state-linked networks within U.S. infrastructures that were never designed for adversarial use.

The models used in Studies 2 and 3 also highlight a related ethical issue: base rates and error costs. Even accurate classifiers produce large numbers of false positives when the underlying event is rare. Study 3's account-level positive predictive value analysis makes this concrete. When the baseline rhetorical model's class probabilities are converted into political-style and random-style scores and evaluated under plausible Russian prevalences on the order of 0.1–1 percent of all accounts, any threshold that captures even 70 percent of Russian accounts produces false-positive rates that approach or exceed 70 percent across U.S. users, with positive predictive values that remain below one percent. Political-style thresholds that capture 70–90 percent of Russians flag essentially all U.S. politicians and nearly all engaged users; random-style thresholds perform slightly better for elites but still label the majority of ordinary U.S. accounts as "Russian-like." In practical terms, hundreds of domestic accounts would be flagged for every Russian account correctly identified.

These results, though limited to my selected features, directly support the conclusion that rhetorically defined scores are unsafe as stand-alone enforcement criteria at any level of recall that would matter for national security. At best, such models can help describe aggregate patterns, map where foreign accounts sit relative to domestic actors, and support human analysis and triage. They are a poor basis for automated sanctions or account-level penalties. The legal and empirical constraints in this chapter rule out a simple model in which the U.S. government deputizes platforms to remove or suppress "Russian-like" speech. They do not, however, preclude a far more assertive approach that distinguishes foreign operations from domestic speakers and focuses state power on the former through ownership restrictions, utility-style obligations for key intermediaries, and targeted disruption of foreign networks, while leaving the vast bulk of domestic political expression beyond the reach of automated censorship.

The ethical path forward is therefore to focus on **who controls the infrastructure** and **how that control is exercised**, not on policing the semantic content of political messages. The central liberal commitments at stake are not only speakers' rights to express themselves but also audiences' rights to receive information and evaluate it for themselves. In a renewed long-term information contest with foreign powers, the goal should not be to insulate citizens from contested or disturbing speech, but to deny foreign adversaries' effortless ownership and exploitation of the systems through which that speech travels, and to give citizens reliable information about who is speaking and how that speech is being amplified. That suggests a regulatory emphasis on provenance, process, structural design choices, and foreign ownership restrictions, rather than on viewpoint-based content judgments. In 1985 the idea that the Soviet Union should operate and broadcast a television station within the United States was ludicrous.

In practical terms, this means encouraging or requiring platforms to disclose when content originates from state-linked entities, when patterns of covert coordination are detected, and how their moderation and recommendation systems operate. It also means drawing clear constitutional lines around state involvement. The government has legitimate interests in identifying and countering foreign political operations, but it should be far more cautious about endorsing or participating in systems that remove or suppress political content because it matches a statistical pattern of "disinformation." Where the state does act, whether through sanctions, designations, ownership limits, or legal action against foreign principals, it should do so on the basis of provenance, funding, and covert conduct, not on the basis of whether particular messages are judged true, false, or politically desirable. The next section takes that normative starting point and asks what concrete, process-based obligations might be imposed on platforms

and public institutions to expose and constrain foreign operations while preserving a meaningful sphere of political freedom for domestic speakers.

6.7 Policy Recommendations

The empirical results and comparative analysis in this chapter point toward a set of policy directions that treat foreign information operations as a security problem embedded in private infrastructure, rather than as a matter of domestic truth arbitration. Given the legal and ethical constraints outlined above, effective policy must focus on provenance, process, and structural incentives, not on broad, content-based control of political speech by the state.

First, the legal framework should explicitly distinguish foreign political operations from domestic speech and adjust infrastructure obligations accordingly. Congress should update the Telecommunications Act and related statutes so that broadband internet access and the major services that ride on it, specifically social media platforms, search engines, and large content-hosting providers, are **classified and regulated as common carriers under Title II**, with modernized public-interest obligations that include the security and integrity of democratic communication. In practical terms, this would mean treating the internet and its dominant services as utilities: privately owned, but licensed and subject to clear duties of nondiscrimination among domestic users and to heightened responsibilities when dealing with foreign, state-linked operations. The model is not Soviet-style state media; it is closer to how the United States already treats nuclear power plants and other high-risk utilities, where private operators are bound by stringent design, reporting, and safety requirements overseen by a mix of regulatory agencies and technical advisory bodies. For information infrastructure, those responsibilities should not take the form of direct government control over individual posts.

Instead, they should be implemented through public–private oversight boards with a statutory mandate to set process standards, review risk assessments, audit provenance and labeling systems, and recommend targeted restrictions on ownership, access, and amplification for foreign principals, much as the financial sector carries special obligations with respect to money laundering and sanctions evasion.

These reforms would not require the United States to invent a new doctrine from scratch. They would instead revive and adapt a long-standing practice of limiting foreign control over core communications media. Since the 1930s, section 310 of the Communications Act has restricted broadcast licenses to U.S. citizens and imposed a 20–25 percent benchmark on foreign ownership of licensees and their parent companies, justified explicitly in terms of wartime homeland-security risks and the danger that hostile governments could disrupt ship-to-shore and governmental communications or use broadcast outlets to disseminate propaganda to the American public (47 U.S.C. § 310; Federal Communications Commission [FCC], 2013a, 2013b; Agarwal, 2013). For most of the Cold War, that system reflected a bipartisan consensus that foreign states should not own or control U.S. radio and television stations. Only in the 2010s did the FCC begin to treat the 25 percent benchmark as a flexible ceiling rather than a hard cap, concluding in 2013 that the original national security concerns had been "overtaken by marketplace and technological developments" and announcing a case-by-case approach to approving higher levels of foreign investment (FCC, 2013a, 2016). That shift was a policy judgment about the threat environment, not a constitutional mandate. Nothing in the First Amendment prevents Congress and the Commission from revisiting it in light of a renewed information contest with foreign adversaries.

The same logic is already reappearing in other domains. In response to national security concerns about Chinese and other adversary-linked entities purchasing land near sensitive military sites, Congress and state legislatures have begun to restrict foreign ownership of farmland and real property around military bases and other sensitive locations in national security. Recent Congressional Research Service work notes that between 2023 and 2024 at least twenty-two states enacted legislation regulating foreign ownership of real property, with several laws specifically targeting land near military installations and other critical infrastructure (Congressional Research Service, 2023). Federal proposals such as the Promoting Agriculture Safeguards and Security (PASS) Act would bar entities controlled by China, Russia, Iran, or North Korea from purchasing agricultural land and agribusinesses near U.S. military facilities (Rounds & Cortez Masto, 2025). Treasury has simultaneously moved to expand the jurisdiction of the Committee on Foreign Investment in the United States (CFIUS) over foreign real estate purchases around sensitive sites (U.S. Department of the Treasury, 2024). These measures reflect a simple principle: foreign adversaries have no inherent right to own sensitive physical infrastructure in the United States when that ownership creates an unacceptable security risk. The same principle can be applied to control of key information infrastructure. A revitalized foreign-ownership regime for broadcast, broadband, and large-scale social media platforms focused on state-linked or adversary-controlled entities would not prevent Americans from accessing foreign content if they chose to seek it out, but it would deny hostile governments an easy path to owning the pipes and megaphones through which U.S. public life is conducted.

Within that framework, it is important to clarify how these recommendations interact with Supreme Court decisions on the “right to receive” information. Cases such as Lamont v. Postmaster General and Kleindienst v. Mandel recognize that U.S. citizens have a First

Amendment interest in accessing foreign ideas, but they do not create a right to government-provided amplification or carriage on private infrastructure (Kleindienst v. Mandel, 1972; Lamont v. Postmaster General, 1965). The state may not criminalize an individual who chooses to visit a foreign website, read a foreign newspaper, or seek out foreign broadcasts. It does not follow that foreign principals have a constitutional right to operate large, covert influence campaigns on U.S.-based platforms or to secure nondiscriminatory access to privately owned distribution systems. The goal of the reforms proposed here is to preserve individuals' ability to seek out foreign content if they wish, while allowing internet providers as regulated utilities and major platforms to deny carriage, amplification, or paid promotion to foreign state-linked operations as part of a clearly defined security and provenance regime.

Second, FARA should be updated for the environment that actually exists. Foreign influence in this century does not look like a registered lobbyist handing talking points to a senator. It looks like networks of semi-anonymous accounts, shell organizations, data-brokered targeting, and proxy media outlets. Modernized disclosure rules should therefore focus primarily on **organizations** that run or enable online political campaigns on behalf of foreign principals, rather than on chasing every individual operator (U.S. Department of Justice, n.d.). Any company, nonprofit, media outlet, advertising firm, data broker, or front organization that purchases political advertising, manages high-reach accounts or pages, or provides amplification and targeting services in the United States **at the direction, control, or substantial funding of a foreign principal** should be required to register as a foreign online political operator and disclose its beneficial ownership, funding sources, and campaign clients. Registration would take place in a publicly accessible, machine-readable database. Platforms and internet providers, as regulated utilities, would in turn have a duty to: (1) verify FARA registration for foreign-linked

entities that seek political ad access, large-scale promotional services, or privileged content-distribution status; (2) label registered operators clearly to users; and (3) deny those services to foreign-linked entities that appear to meet the statutory criteria but refuse to register. Registration will not stop covert activity. But by concentrating obligations on identifiable organizations, rather than on individual accounts, it would bring a significant portion of foreign-directed online campaigning into the open and provide a clearer legal basis for denying carriage, cutting off monetization, imposing civil and criminal penalties, and, where appropriate, applying sanctions when actors are caught. The target is undisclosed, state-directed online campaigns, not ordinary contact with foreign media or ideas.

Third, the interagency structure for dealing with information threats needs to be more than a loose collection of working groups. A permanent coordinating body inside the National Security Council system or an equivalent venue should have the mandate to integrate intelligence, cyber, diplomatic, and domestic risk perspectives on foreign influence. Its job would not be to decide which posts to remove or which narratives to endorse. Its function would be to develop shared threat pictures, maintain structured relationships with platforms and infrastructure providers, and support state and local partners, especially around elections. In parallel, the Director of National Intelligence should be directed by statute to establish a dedicated interagency intelligence center focused specifically on identifying, tracking, and reporting on state-sponsored information operations. That center should have clear authorities to fuse collection and analysis across agencies, and it should be subject to mandatory, regular reporting requirements to Congress so that its work cannot be quietly sidelined or dismantled by an administration that finds its assessments politically inconvenient. Where appropriate, the NSC body and the ODNI center

together should be able to trigger targeted actions under sanctions law, FARA, or updated telecommunications authorities against clearly identified foreign operations and their technical infrastructure, while keeping domestic content decisions at arm's length.

Fourth, the United States should adopt the procedural core of the Digital Services Act without the parts that sit uneasily with U.S. law. Large platforms that operate at national scale should be required to conduct regular systemic risk assessments, including risks from foreign influence, coordinated inauthentic behavior, and the concentration of highly moralized and emotional rhetoric documented in Studies 1–3, and to publish high-level findings (European Commission, 2018, 2022, 2025). For domestic users, these assessments and any resulting interventions should operate at the level of system design and provenance rather than at the level of viewpoint-specific takedowns, consistent with common-carrier obligations to carry lawful speech neutrally. Platforms should also be required to provide meaningful transparency reports that describe moderation practices, enforcement metrics, and significant changes to recommender systems in a way that is comprehensible to the public and to policymakers. Oversight of these duties should be assigned to **the same independent, statutorily created public–private board described above**, with a professional staff drawn from civil society, technical communities, national security agencies, and the major platforms. That board's mandate would be to review risk assessments and mitigation plans, audit provenance and labeling systems, and report publicly on the state of foreign information operations and platform responses, while operating under strict constraints against directing the removal or downranking of lawful domestic content based on its political viewpoint.

Fifth, these reforms should be embedded in a transatlantic strategy rather than pursued in isolation. The European Union is already experimenting with the DSA, the Code of Practice on

Disinformation, and coordinated responses to Russian and other state-backed campaigns. The United States and the EU face largely overlapping threats from Russian, Chinese, Iranian, and other authoritarian information operations, and they share a common interest in protecting open societies. Policy should therefore aim at interoperable provenance labels, compatible transparency and data-access regimes, and routine sharing of indicators and sanctions designations across the Atlantic. A joint U.S.–EU information integrity forum, linked to existing NATO and G7 structures, could help align standards for risk assessments, coordinate responses to major campaigns, and avoid a fragmented regulatory landscape that foreign operators can exploit.

Sixth, there must be a secure, legally grounded framework for data access by independent researchers and designated public bodies. The work in this dissertation depended on a historical dataset that is increasingly difficult to replicate as platforms restrict access. Serious oversight is impossible if the only people who can see how the system operates are the companies that profit from it. A U.S. model could follow the DSA's lead by allowing vetted researchers to work with pseudonymized or aggregated data under strict safeguards, with clear limits on how it can be used and disclosed (EU Digital Services Act, 2025). Regulators and the proposed oversight board should have structured access to more detailed operational data under confidentiality rules, so that foreign campaigns and platform responses can be evaluated independently rather than on the basis of voluntary disclosures. These risk assessments should not focus solely on volumes of 'false' content, but also on the concentration and targeting of highly moralized and emotional rhetoric documented in Study 1, since that is the substrate foreign and domestic actors exploit to drive engagement.

Seventh, policymakers should look at the identity and infrastructure layers, not only at individual messages. Proposals for stronger digital identity online are often framed in maximalist terms, such as national digital IDs or blanket real-name requirements, which would collide quickly with U.S. protections for anonymous speech and with legitimate privacy concerns. A narrower, more defensible approach is to tie higher levels of identity assurance to higher levels of reach and privilege. If an account wants to purchase political advertising, manage a large network of public pages, or present itself as a governmental or institutional actor, it is reasonable to require stronger proof of identity, location, and organizational control. At the network level, the transition from IPv4 to IPv6 and the deployment of improved routing security, logging, and authentication can help with attribution and forensics. These tools should be used to make it more difficult and costly for foreign actors to operate large, covert networks at scale, and to support targeted disruption of known assets, not as a blunt instrument for real-time blocking of political content. Sophisticated operators will still rely on mainstream cloud providers, VPNs, satellites, and compromised systems that are indistinguishable at the network level from ordinary traffic. Network measures are therefore best suited to supporting investigations, sanctions, and denial of service to identified foreign operations rather than to adjudicating what domestic users may see.

Finally, no legal or technical measure will substitute for a public that is less easily manipulated. Long-term resilience depends on citizens who understand that foreign actors are present in their feeds, that domestic actors will also manipulate, and that moralized, emotional language is both a genuine part of politics and a tool for influence. Digital-literacy and civic-education efforts are often treated as soft or secondary. In a liberal system, they are the closest thing to a sustainable defense, and they are the only tools that work equally against foreign and domestic manipulative

campaigns. Public investments should therefore support sustained programs in schools, community institutions, and media organizations that teach how foreign operations work, how algorithms shape information exposure, and how to interpret provenance cues and platform labels

Study 3's results, particularly the positive predictive value analysis, imply that rhetorical features should enter these policy workflows only as one layer in a multi-signal risk model, combined with structural, network, provenance, and coordination evidence, rather than as stand-alone enforcement criteria. The quantitative models developed here are best understood as diagnostic instruments that can help map where foreign operations sit relative to domestic actors and estimate the likely collateral effects of interventions. They should inform the design of process-based obligations and technical controls, not serve as automated gatekeepers for political speech.

6.8 Limitations and Future Research

This dissertation is necessarily bounded by the data and methods available. The analysis relied on Twitter as a single source. Twitter was useful because historical data were available at sufficient scale and granularity and because the platform matched the research question: text-based political messages circulating in a largely public network. But Twitter, now X, is not the entire social media ecosystem. Other platforms such as Facebook, YouTube, TikTok, Telegram, and encrypted or semi-closed messaging applications have different affordances, audience structures, and content formats. Visual and audiovisual content, recommendation-heavy feeds, and closed-group dynamics all play a larger role elsewhere than they do in this text-focused dataset. The account- and tweet-level findings here therefore describe one important slice of the information environment rather than a complete picture of global influence operations.

The temporal scope, roughly 2009 to 2020, also matters. It captures the period in which Russian IRA activity was most visible to researchers and in which Twitter's API was comparatively open, but it does not fully encompass the rise of large-scale AI-generated text, synthetic media, and subsequent generations of influence techniques. Many of the models here implicitly assume that operators must choose content carefully and laboriously: selecting themes, crafting messages, and tuning rhetoric by hand. That assumption may not hold if generative systems can produce hundreds of plausible variations with minimal human oversight and can adapt interactively to engagement metrics or platform responses. At the same time, the feature space used here as in moral foundations, emotions, partisan language, and structural indicators could be applied to the problem of distinguishing foreign AI-assisted campaigns from ordinary domestic speech. Future work should test whether the rhetorical and account-level patterns identified in this dissertation remain stable when foreign actors begin to rely systematically on generative models, and whether those patterns can be used to detect such activity without collapsing into broad suspicion of all AI-mediated expression.

Methodologically, the studies rely on supervised learning and lexicon-based measures. Moral Foundations Theory and the NRC emotion lexicon provide interpretable structure, but they are still abstractions laid over the complexity of language. Labeling decisions, sampling strategies, and feature engineering all shape the results. The classifiers are only as good as the training data and do not capture the full range of sarcasm, multimodal cues, or evolving slang. While care was taken to avoid obvious forms of leakage and contamination, no model is perfect. The dissertation also focused on Russian activity. Other state and non-state actors, including China, Iran, North Korea, and domestic communities, may use different rhetorical strategies, organizational structures, and platform mixes that would require recalibrated models and possibly different

feature sets. The results should therefore be read as a detailed map of one adversary's behavior in one environment, not as a universal template.

Future research should move in several directions. One priority is to extend rhetorical and account-level classification to other actors and platforms, with particular attention to cross-platform coordination and multimodal content. The State Department's analysis of PRC information operations and public reporting on Iranian and North Korean campaigns demonstrate that these actors blend overt propaganda with more subtle attempts to sit inside foreign conversations, often across multiple services at once (Loh, 2023; Reuters, 2023; U.S. Department of Justice, 2021; U.S. Department of State, 2023). Applying the account-level framework to these cases would test whether the Russian pattern that occupy an intermediate rhetorical band between elites and ordinary users is idiosyncratic or part of a general doctrine of blended foreign influence.

Second, there is a need for more work at the intersection of law, policy, and computer science on what acceptable interventions look like in a liberal democracy when foreign ownership and control are explicitly on the table. The DSA, data-protection rules, and net-neutrality frameworks offer examples of how legislatures can shape the operating environment of private actors without dictating content (European Commission, 2018, 2022, 2025). The empirical patterns documented here, especially the difficulty of distinguishing foreign from domestic speech on the basis of rhetoric alone and the poor positive predictive value of content-based classifiers at realistic base rates, should inform debates about foreign ownership bans, critical-infrastructure designations, and process-based duties for major platforms. Researchers should ask not only whether a given technical or legal intervention is effective in disrupting foreign operations, but also whether it

can be implemented in ways that are procedurally neutral, transparent, and reversible, and whether it keeps the focus on foreign principals rather than drifting into domestic viewpoint management.

Third, as platforms and governments roll out new transparency and provenance tools, researchers will need to evaluate whether these measures actually change behavior or simply add labels and legal categories to an already noisy environment. It is one thing to expose that a "patriotic" account is operated from abroad or that a major platform is ultimately controlled by a foreign adversary. It is another to show that users register those facts and treat the content differently, that advertisers, civil-society organizations, and public officials adjust their behavior in response, and that policymakers and journalists use such information in a disciplined way rather than as another weapon in partisan conflict. The next phase of research will need to move beyond demonstrating that structural and rhetorical interventions are possible and focus on how they are received, contested, and repurposed in practice, both by the citizens they are supposed to protect and by the foreign actors they are intended to constrain.

Finally, future work should examine how shifts in moral foundations and emotional framing at the platform or ecosystem level correlate with the deployment of structural interventions, so that policymakers can track not only whether foreign operations are disrupted but also whether the moral tone of domestic discourse is being unintentionally distorted.

6.9 Conclusion

This dissertation began as an attempt to see whether Russian influence campaigns used the moral vocabulary described by Moral Foundations Theory and, if so, what effect that had on

engagement. The models showed that they do: Russian accounts routinely frame their messages in ways that map onto MFT categories and pair those frames with distinct emotional tones. They also revealed a second, equally important pattern. Russian operators make extensive use of moral and emotional language while deliberately avoiding dense partisan branding, positioning their content so that, on many features, it blends with the rhetoric of ordinary, politically engaged U.S. users rather than with elite partisan messaging.

These results show that moral and emotional rhetoric remains a powerful force online. In the hands of foreign actors, it can be used to raise the temperature of debates that are already hot, to amplify content that was already polarizing, and to keep attention focused on fault lines that were already there. The models in this dissertation demonstrated that it is possible to detect some of these patterns, especially when we look at accounts over time and treat rhetoric as a habit rather than a single choice. They also underscored the limits of purely content-based approaches. The closer foreign operators get to the center of normal political discourse, the harder they are to distinguish from the citizens they are trying to influence, and the easier it is for any content filter built on these features to sweep up legitimate domestic speech. This dissertation also points to several questions that remain open. Do the same rhetorical patterns and base-rate constraints hold on other platforms and in multimodal environments, and do they persist in the post-2020 era of large-scale AI-generated content? Beyond the Russian case, do other state and non-state influence actors show similar “intermediate” profiles, and can provenance and transparency interventions change user and advertiser behavior without distorting legitimate domestic discourse?

The policy challenge for the United States is to respond without abandoning the values that make it a target in the first place. The country cannot and should not copy European regulatory models wholesale, but it can adopt their emphasis on transparency, research access, and systemic risk assessment. It can modernize existing tools such as FARA, clarify the line between foreign principals and domestic speakers, and build more coherent institutional arrangements for tracking and reporting on state-sponsored operations. It can also recognize that platforms and networks now function as critical public infrastructure and regulate them accordingly: as utilities that must carry lawful domestic speech neutrally, but that also bear clear obligations to manage foreign state-linked operations with the same seriousness applied to other forms of critical infrastructure risk.

In the end, securing the information space in a democracy is not about silencing adversaries. It is about making it harder for them to masquerade as us and easier for citizens to recognize when that is happening. That requires better data, better models, and better policy, but it also requires a basic commitment to keep the conversation open even when it is messy. The work here is a small step toward that goal: an attempt to describe, with some precision, how foreign influence actually operates and what kinds of interventions might blunt its effects without sacrificing the open, argumentative, and sometimes chaotic public sphere that a democracy needs to survive.

Annex A.1 Study 1 Results and Full Regression Tables

The first regression simply performed the baseline coefficient of log-transformed follower count. In the OLS model focused on retweet count, follower count was statistically significantly and positively related to engagement ($\beta = 0.119$, $p < .001$). However, the model explained only 9.5% of the variance in retweet activity ($R^2 = 0.095$), suggesting that audience size contributes meaningfully to visibility but leaves most of the variation unexplained. These results are consistent with the theoretical understanding that reach enables, but does not guarantee, diffusion. See table 2.

In contrast, the GLM model focused on like count, estimated using a Negative Binomial distribution due to overdispersion, showed a much stronger model fit. The pseudo-R^2 was substantially higher (pseudo $R^2 = 0.865$), and follower count was statistically significant and positively associated with like count ($\beta = 1.071$, $p < .001$). This finding suggests that likes show

a stronger statistical association with account size than retweets and are more tightly coupled to structural visibility. In other words, users are more likely to "like" content from accounts with large followings, but retweet behavior likely hinges more on content resonance or personal alignment. See table 3.

These baseline results confirm what is often assumed in social media analysis: accounts with larger followings enjoy greater visibility and are more likely to receive engagement. However, the relatively low explanatory power of follower count in the OLS model also underscores that audience size is a necessary but not sufficient condition for virality. Much of the variation in engagement remains unexplained by follower count alone, reinforcing the value of this research and underscoring the need to explore content-based variables.

Table 2 OLS Regression Follower Count Only

Dep. Variable: retweet_count			R-squared:	0.095		
Model: OLS			Adj. R-squared:	0.095		
Method: Least Squares			F-statistic:	1.845e+05		
DF Model: 1			Prob (F-statistic):	0.00		
Covariance Type: nonrobust			Log-Likelihood:	-2.8162e+06		
No. Observations: 1765496			AIC:	5.632e+06		
Df Residuals: 1765494			BIC:	5.632e+06		
Variable	Coef	Std Err	t (or z)	P>|t|	[0.025	0.975]
const	-0.9590	0.003	-294.994	0	-0.965	-0.953
follower count	0.1190	0.000	429.520	0	0.118	.0120

Table 3 GLM Regression Follower Count Only

Dep. Variable: like_count	No. Observations:	1765496
Model: GLM	Df Residuals:	1765494
Model Family: Negative Binomial	Df Model:	1

Method: IRLS			Log-Likelihood:		-2.3416e+06	
Covariance Type: HC2			Deviance		3.2074e+06	
Iterations: 15			Pearson chi2:		1.54e+09	
			Pseudo R-squared		0.8653	
Variable	Coef	Std Err	t (or z)	P>|t|	[0.025	0.975]
const	-12.7228	0.143	-88.780	0	-13.004	-12.442
follower count	1.0711	0.012	89.825	0	1.048	1.094

The next set of models examined the relationship between emotional variables (e.g., fear, anger, trust, surprise, sadness, disgust, and joy) and tweet engagement—specifically, retweet count and like count—while controlling for account popularity via follower count. In the OLS regression the model showed some improved performance, with the model now explaining approximately 9.8% of the variance ($R^2 = 0.098$). Within this model, fear ($\beta = 0.066$, $p < 0.001$) and sadness ($\beta = 0.161$, $p < 0.001$) were statistically significantly positively associated with retweet counts, whereas anger ($\beta = –0.379$, $p < 0.001$) and surprise ($\beta = –0.128$, $p < 0.001$) were statistically significantly negatively associated. In addition, both disgust ($\beta = 0.242$, $p < 0.001$) and joy ($\beta = 0.158$, $p < 0.001$) showed statistically significant positive associations. Follower count remained a strong and statistically significant predictor ($\beta = 0.123$, $p < .001$), reinforcing the established effect of account visibility on engagement. See Table 4.

Table 4 OLS Regression Emotion + Follower Count

Dep. Variable: retweet_count			R-squared:		0.098	
Model: OLS			Adj. R-squared:		0.098	
Method: Least Squares			F-statistic:		2.410e+04	
DF Model: 8			Prob (F-statistic):		0.00	
Covariance Type: nonrobust			Log-Likelihood:		-2.8125e+06	
No. Observations: 1765496			AIC:		5.625e+06	
Df Residuals: 1765488			BIC:		5.625e+06	
Variable	Coef	Std Err	t (or z)	P>|t|	[0.025	0.975]

const	-1.0039	0.004	-265.291	0	-1.011	-0.997
fear	0.0663	0.006	11.037	0	0.055	0.078
anger	-0.3787	0.005	-69.145	0	-0.389	-0.368
trust	0.0143	0.005	2.916	0.004	0.005	0.024
surprise	-0.1284	0.008	-15.467	0	-0.145	-0.112
sadness	0.1606	0.008	20.241	0	0.145	0.176
disgust	0.2422	0.011	22.910	0	0.221	0.263
joy	0.1576	0.008	18.874	0	0.141	0.174
follower count	0.1232	0.000	434.581	0	0.123	0.124

The GLM regression for like_count—modeled with a Negative Binomial specification using a log link—produced a similar pattern. Here, anger ($\beta = -2.825$, $p < 0.001$) and surprise ($\beta = -1.278$, $p < 0.001$) were significantly negatively associated with likes, while disgust ($\beta = 1.903$, $p < 0.001$) and joy ($\beta = 2.146$, $p < 0.001$) were statistically significant positive predictors. Although fear ($\beta = -0.264$, $p = 0.407$) did not reach significance, and trust ($\beta = 0.079$, $p = 0.703$) was not statistically significant, follower count again showed statistically significant association ($\beta = 1.088$, $p < 0.001$) with like counts. These findings suggest that even when accounting for the considerable influence of account popularity (follower count), emotional content remains a meaningful predictor of engagement. The inclusion of follower count not only improves the explanatory power of the models but also helps clarify the specific contributions of each emotional variable to retweet and like counts. See table 5.

Table 5 GLM Regression Emotion + Follower Count

Dep. Variable: like_count	No. Observations:	1765496
Model: GLM	Df Residuals:	1765487
Model Family: Negative Binomial	Df Model:	8
Method: IRLS	Log-Likelihood:	-2.2767e+06
Covariance Type: HC2	Deviance	3.0775e+06
Iterations: 20	Pearson chi2:	1.12e+09
	Pseudo R-squared	0.8749

Variable	Coef	Std Err	t (or z)	P>|t|	[0.025	0.975]
const	-12.9468	0.164	-78.781	0.000	-13.269	-12.625
fear	-0.2641	0.319	-0.829	0.407	-0.889	0.361
anger	-2.8246	0.190	-14.875	0.000	-3.197	-2.452
trust	0.0785	0.206	0.381	0.703	-0.325	0.482
surprise	-1.2778	0.322	-3.973	0.000	-1.908	-0.647
sadness	0.6702	0.363	1.846	0.065	-0.041	1.382
disgust	1.9033	0.546	3.485	0.000	0.833	2.974
joy	2.1462	0.496	4.330	0.000	1.172	3.118
Follower count	1.0878	0.012	92.789	0.000	1.065	1.111

The next regression which tested the central research question of does MFT better explain engagement than emotion or partisan extremity consisted of Moral Foundations Theory (MFT) classification variables—loyalty, sanctity, care, fairness, and authority—alongside follower count as a control variable. In the OLS model predicting $\log_2$(retweet_count), the inclusion of MFT features only increased the explained variance to 11.2% ($R^2 = 0.112$), compared to 9.5% with follower count alone, and 9.8% of emotion and follower count.

Among the MFT predictors, fairness ($\beta = 0.354$, $p < 0.001$), loyalty ($\beta = 0.319$, $p < 0.001$), and care ($\beta = 0.208$, $p < 0.001$) were statistically significant and showed the strongest associations with increased retweet activity. Sanctity ($\beta = 0.240$, $p < 0.001$) and authority ($\beta = 0.177$, $p < 0.001$) also exhibited statistically significant positive coefficients, albeit with slightly smaller effects. Follower count retained its expected statistical significance and influence ($\beta = 0.117$, $p < 0.001$), underscoring the foundational role of audience size in driving visibility and engagement.

These results suggest that moral language, independent of emotional or partisan cues, is a salient predictor of engagement. Even in the absence of emotional variables, MFT-based appeals carry explanatory weight, indicating that moral resonance may operate through a distinct mechanism in the amplification of influence-oriented messages. See tables 6 and 7.

Table 6 OLS Regression Follower Count+MFT

Dep. Variable: retweet_count			R-squared:	0.112		
Model: OLS			Adj. R-squared:	0.112		
Method: Least Squares			F-statistic:	3.712e+04		
DF Model: 6			Prob (F-statistic):	0.00		
Covariance Type: nonrobust			Log-Likelihood:	-2.7991e+06		
No. Observations: 1765496			AIC:	5.598e+06		
Df Residuals: 1765489			BIC:	5.598e+06		
Variable	Coef	Std Err	t (or z)	P>|t|	[0.025	0.975]
const	-1.0506	0.003	-319.097	0	-1.057	-1.044
Loyalty	0.3191	0.005	59.155	0	0.309	0.33
Sanctity	0.24	0.005	44.598	0	0.229	0.251
Care	0.2075	0.003	77.858	0	0.202	0.213
Fairness	0.3536	0.003	108.151	0	0.347	0.36
Authority	0.1768	0.004	48.785	0	0.17	0.184
Follower Count	0.1169	0	420.391	0	0.116	0.117

Table 7 GLM Regression Emotion + MFT

Dep. Variable: like_count			No. Observations:	1765496		
Model: GLM			Df Residuals:	1765489		
Model Family: NegativeBinomial			Df Model:	6		
Method: IRLS			Log-Likelihood:	-2.2744e+06		
Covariance Type: HC2			Deviance	3.0729e+06		
Iterations: 19			Pearson chi2:	1.94e+09		
			Pseudo R-squared.	0.8752		
Variable	Coef	Std Err	t (or z)	P>|t|	[0.025	0.975]
const	-13.3686	0.176	-76.149	0	-13.713	-13.025
Loyalty	1.3301	0.142	9.356	0	1.051	1.609
Sanctity	0.7311	0.162	4.514	0	0.414	1.049

Care	0.6444	0.198	3.253	0.001	0.256	1.033
Fairness	0.9058	0.114	7.954	0	0.683	1.129
Authority	0.0736	0.113	0.651	0.515	-0.148	0.296
Follower Count	1.0856	0.013	83.843	0	1.06	1.111

Incorporating the full set of Moral Foundations Theory (MFT) classification variables—loyalty, sanctity, care, fairness, and authority—into the regression models (along with follower count as a control) modestly improved explanatory power. In the OLS model for $\log_2$(retweet_count), the inclusion of Moral Foundations Theory (MFT) variables increased the explained variance to 11.6% ($R^2 = 0.116$). Among these measures, loyalty ($\beta = 0.306$, $p < 0.001$), fairness ($\beta = 0.362$, $p < 0.001$), and care ($\beta = 0.212$, $p < 0.001$) were statistically significant and most strongly associated with higher retweet counts, while sanctity ($\beta = 0.211$, $p < 0.001$) and authority ($\beta = 0.206$, $p < 0.001$) also showed statistically significant positive associations. The inclusion of MFT reduced the magnitude of some emotional associations—specifically, anger ($\beta = -0.427$, $p < 0.001$) and surprise ($\beta = -0.073$, $p < 0.001$)—though other emotional variables, such as joy ($\beta = 0.178$, $p < 0.001$), remained significant. Follower count continued to maintain a robust positive effect ($\beta = 0.121$, $p < 0.001$), highlighting again the central role of account popularity in driving retweet engagement. See table 8.

Table 8 OLS Regression Emotion+MFT

Dep. Variable: retweet_count			R-squared:	0.116		
Model: OLS			Adj. R-squared:	0.116		
Method: Least Squares			F-statistic:	1.786e+04		
DF Model: 13			Prob (F-statistic):	0.00		
Covariance Type: nonrobust			Log-Likelihood:	-2.7986e+06		
No. Observations: 1765496			AIC:	5.590e+06		
Df Residuals: 1765482			BIC:	5.590e+06		
Variable	Coef	Std Err	t (or z)	P>|t|	[0.025	0.975]

const	-1.0792	0.004	-285.139	0	-1.087	-1.072
fear	-0.0181	0.006	-2.983	0.003	-0.03	-0.006
anger	-0.4269	0.005	-78.514	0	-0.438	-0.416
trust	-0.0314	0.005	-6.47	0	-0.041	-0.022
surprise	-0.0728	0.008	-8.848	0	-0.089	-0.057
sadness	0.116	0.008	14.66	0	0.1	0.131
disgust	0.0946	0.011	8.984	0	0.074	0.115
joy	0.1784	0.008	21.415	0	0.162	0.195
MFT Loyalty	0.3064	0.005	56.808	0	0.296	0.317
MFT Sanctity	0.2108	0.005	38.957	0	0.2	0.221
MFT Care	0.2124	0.003	77.766	0	0.207	0.218
MFT Fairness	0.3616	0.003	110.253	0	0.355	0.368
MFT Authority	0.2055	0.004	56.345	0	0.198	0.213
follower count	0.1214	0	429.042	0	0.121	0.122

In the GLM model for like_count similar trends emerge. Loyalty ($\beta = 1.326$, $p < 0.001$) and fairness ($\beta = 1.169$, $p < 0.001$) remain statistically significant and the strongest association with likes, while care ($\beta = 0.770$, $p < 0.001$) and sanctity ($\beta = 0.499$, $p < 0.001$) also positively associate substantially. These results suggest that tweets containing stronger expressions of these moral foundations tend to receive more likes, even when controlling for emotional content. See table 9.

Table 9 GLM Regression Emotion + MFT

Dep. Variable: like_count			No. Observations:	1765496		
Model: GLM			Df Residuals:	1765482		
Model Family: NegativeBinomial			Df Model:	13		
Method: IRLS			Log-Likelihood:	-2.1912e+06		
Covariance Type: HC2			Deviance	2.9065e+06		
Iterations: 23			Pearson chi2:	1.24e+09		
			Pseudo R-squared.	0.8864		
Variable	Coef	Std Err	t (or z)	P>|t|	[0.025	0.975]
const	-13.5696	0.171	-79.175	0	-13.905	-13.234
fear	-0.9752	0.252	-3.871	0	-1.469	-0.481
anger	-3.1726	0.179	-17.708	0	-3.524	-2.821

trust	-0.0757	0.238	-0.318	0.75	-0.542	0.390
surprise	-1.3191	0.287	-4.591	0	-1.882	-0.756
sadness	0.2932	0.361	0.813	0.416	-0.414	1.000
disgust	0.5	0.408	1.226	0.22	-0.299	1.299
joy	2.2344	0.535	4.176	0	1.186	3.283
MFT Loyalty	1.3262	0.152	8.754	0	1.029	1.623
MFT Sanctity	0.4988	0.122	4.087	0	0.26	0.738
MFT Care	0.77	0.15	5.118	0	0.475	1.065
MFT Fairness	1.1694	0.108	10.802	0	0.957	1.382
MFT Authority	0.3315	0.103	3.216	0.001	0.13	0.534
follower count	1.106	0.012	89.34	0	1.082	1.130

Adding the partisan word count variable provided a modest improvement in model performance. The OLS model explained 11.9% of the variance ($R^2 = 0.119$), with partisan word count showing a statistically significant positive association with retweet count ($\beta = 0.117$, $p < 0.001$). This suggests that tweets containing more partisan language—measured by the count of matched ideological terms—were more likely to be shared. Notably, this association remained robust even when accounting for emotional tone, moral foundation language, and follower count. See table 10.

Table 10 OLS Regression Emotion+MFT+Partisan Words Aggregate Count

Dep. Variable: retweet_count			R-squared:	0.119		
Model: OLS			Adj. R-squared:	0.119		
Method: Least Squares			F-statistic:	1.704e+04		
DF Model: 14			Prob (F-statistic):	0.00		
Covariance Type: nonrobust			Log-Likelihood:	-2.7921e+06		
No. Observations: 1765496			AIC:	5.584e+06		
Df Residuals: 1765481			BIC:	5.584e+06		
Variable	Coef	Std Err	t (or z)	P>|t|	[0.025	0.975]
const	-1.0947	0.004	-289.249	0	-1.102	-1.087
fear	-0.0014	0.006	-0.232	0.817	-0.013	0.010
anger	-0.4579	0.005	-84.109	0	-0.469	-0.447
trust	-0.0313	0.005	-6.456	0	-0.041	-0.022
surprise	-0.0753	0.008	-9.162	0	-0.091	-0.059

sadness	0.1218	0.008	15.418	0	0.106	0.137
disgust	0.1011	0.011	9.612	0	0.080	0.122
joy	0.2069	0.008	24.845	0	0.191	0.223
MFT Loyalty	0.2637	0.005	48.694	0	0.253	0.274
MFT Sanctity	0.2083	0.005	38.557	0	0.198	0.219
MFT Care	0.2215	0.003	81.149	0	0.216	0.227
MFT Fairness	0.3264	0.003	98.654	0	0.320	0.333
MFT Authority	0.1558	0.004	42.087	0	0.149	0.163
follower count	0.1210	0	427.982	0	0.120	0.122
partisan word count	0.1172	0.002	77.449	0	0.114	0.120

In the GLM model partisan word count also emerged as a significant contributor ($\beta = 0.344$, $p < .001$), with the model maintaining a strong pseudo R^2 of 0.8877. These findings suggest that not only do partisan cues contribute to diffusion (retweets), but they also play a meaningful role in affective approval (likes). The inclusion of this variable did not diminish the explanatory power of other core content features. Key moral foundations such as loyalty ($\beta = 1.281$, $p < .001$) and fairness ($\beta = 1.093$, $p < .001$) remained statistically significant and showed strong positive associations. These results suggest that partisan language operates as an additive force in political messaging, enhancing engagement outcomes when layered alongside emotional tone and moral framing. See table 11.

Table 11 GLM Regression Emotion + MFT+Partisan Words Aggregate Count

Dep. Variable: like_count			No. Observations:	1765496		
Model: GLM			Df Residuals:	1765481		
Model Family: NegativeBinomial			Df Model:	14		
Method: IRLS			Log-Likelihood:	-2.1811e+06		
Covariance Type: HC2			Deviance	2.88630e+06		
Iterations: 25			Pearson chi2:	1.29e+09		
			Pseudo R-squared.	0.8877		
Variable	Coef	Std Err	t (or z)	P>|t|	[0.025	0.975]
const	-13.6550	0.171	-79.905	0	-13.990	-13.320
fear	-1.0002	0.250	-3.997	0	-1.491	-0.510

anger	-3.2879	0.174	-18.903	0	-3.629	-2.947
trust	-0.1146	0.246	-0.465	0.642	-0.597	0.368
surprise	-1.3170	0.303	-4.351	0	-1.910	-0.724
sadness	0.2570	0.364	0.705	0.481	-0.457	0.971
disgust	0.5441	0.416	1.309	0.191	-0.271	1.359
joy	2.3889	0.561	4.258	0	1.289	3.489
MFT Loyalty	1.2808	0.164	7.829	0	0.960	1.601
MFT Sanctity	0.4087	0.120	3.418	0.001	0.174	0.643
MFT Care	0.7800	0.143	5.471	0	0.501	1.060
MFT Fairness	1.0932	0.107	10.240	0	0.884	1.302
MFT Authority	0.1640	0.101	1.619	0.106	0.035	0.362
follower count	1.1068	0.012	90.882	0	1.083	1.131
partisan word count	0.3435	0.057	5.987	0	0.231	0.456

Next, OLS and GLM regressions were conducted using binary indicators for the presence of partisan keywords. Each term was modeled as a dummy variable equal to 1 if it appeared in a tweet and 0 otherwise, enabling estimation of its independent association with engagement while controlling for follower count. The OLS model associating retweet count yielded a statistically significant result with an R^2 of 0.111. While follower count contributed statistically significantly ($\beta = 0.117$, $p < .001$), specific partisan terms also demonstrated substantial independent associations with retweet volume.

The strongest positive effects in the OLS model were linked to emotionally and morally charged cultural flashpoints. These included parkland shooting ($\beta = 4.27$, $p < .001$), march for our lives ($\beta = 3.46$, $p < .001$), cnn sucks ($\beta = 3.42$, $p < .001$), #metoo ($\beta = 2.65$, $p < .001$), and take a knee ($\beta = 2.19$, $p < .001$). Identity-based and protest-associated terms like kaepernick ($\beta = 1.17$, $p < .001$), blm ($\beta = 1.23$, $p < .001$), blue lives matter ($\beta = 1.37$, $p < .001$), antifa ($\beta = 0.69$, $p < .001$), and radical islam ($\beta = 0.86$, $p < .001$) were also statistically significant and strongly associated with elevated retweet engagement. These findings align with Moral Foundations Theory (Haidt,

2012), suggesting that language invoking moral outrage, identity, or authority violations is more likely to spread.

Conversely, terms referring to policy domains, elite figures, or recycled political slogans were more likely to show neutral or negative associations with diffusion. Several of these terms—mcconnell ($\beta = -0.20$, $p < .001$), biden ($\beta = -0.24$, $p < .001$), tea party ($\beta = -0.34$, $p < .001$), and hannity ($\beta = -0.26$, $p < .001$)—were statistically significantly negatively associated with retweet counts. Others, such as covfefe ($\beta = -0.20$, $p < .01$), showed weaker but still statistically significant effects, while intellectual property theft ($\beta = -0.75$, $p = .52$) was not statistically significant. These patterns suggest that partisan terms lacking a strong moral-emotional frame or that are perceived as outdated may be less likely to be amplified, particularly if they are employed in sarcastic or critical contexts.

The GLM model predicting like count showed a much stronger fit (pseudo $R^2 = 0.8723$), with a similar structure but some notable differences in magnitude and direction. Likes were most strongly associated with terms expressing identity, protest, and moral outrage. These included take a knee ($\beta = 3.94$, $p < .001$), march for our lives ($\beta = 3.25$, $p < .001$), parkland shooting ($\beta = 2.62$, $p < .001$), #metoo ($\beta = 2.53$, $p < .001$), kaepernick ($\beta = 2.09$, $p < .001$), and lock her up ($\beta = 2.03$, $p < .01$). Also positively associated were terms tied to group identity or justice discourse, such as social justice ($\beta = 1.79$, $p < .05$), blm ($\beta = 1.30$, $p < .001$), all lives matter ($\beta = 1.44$, $p < .001$), and racism ($\beta = 1.26$, $p < .001$). Unlike in the OLS model, follower count showed a substantially stronger effect on like count ($\beta = 1.07$, $p < .001$), suggesting that popularity and visibility exert greater influence on passive approval than on retweeting behavior.

The strongest negative predictors in the GLM model included several terms associated with ideological fatigue, fringe discourse, or declining moral salience. These included red pill ($\beta = -39.49$, $p < .001$), climate change hoax ($\beta = -38.99$, $p < .001$), and kavanaugh ($\beta = -37.61$, $p < .001$), all of which may suffer from overuse, backlash, or toxicity. Other negatively associated terms included ukraine ($\beta = -2.40$, $p < .001$), tea party ($\beta = -2.69$, $p < .001$), mcconnell ($\beta = -1.87$, $p < .001$), and obamacare ($\beta = -1.20$, $p < .001$). These results suggest that institutional or policy-heavy language receives less affective approval, potentially due to user fatigue or declining emotional resonance.

This dynamic may reflect broader trends in political engagement. According to a 2023 Pew Research Center survey, 65% of U.S. adults report feeling "exhausted" by politics, particularly among heavy news consumers. This study did not model temporal effects (e.g., when terms emerged, peaked, or declined), but given the 11-year span of data overlapping with the rise of Gen Z/iGen (Twenge, 2017), it is likely that some terms diminished in impact over time due to saturation or generational shifts in discourse.

In sum, these findings demonstrate that partisan language increases engagement most reliably when embedded within narratives of protest, moral urgency, or identity. The presence of polarizing terms alone is insufficient; what matters is how they map onto emotional salience and moral framing. This is consistent with prior work on moral-emotional contagion and affective polarization (Brady et al., 2017), and highlights the rhetorical scaffolding that makes certain political messages "go viral" while others fade into noise. See tables 12 and 13 for the top

fifteen most positively and negatively associated terms for each regression. See table 1 for the complete list of terms regressed.

Table 12 OLS Regression Partisan Words

Dep. Variable: retweet_count			R-squared:	0.111		
Model: OLS			Adj. R-squared:	0.110		
Method: Least Squares			F-statistic:	2238		
DF Model: 98			Prob (F-statistic):	0.00		
Covariance Type: nonrobust			Log-Likelihood:	-2.8006e+06		
No. Observations: 1765496			AIC:	5.601e+06		
Df Residuals: 1765397			BIC:	5.603e+06		
const	-0.9789	0.003	-298.658	0.000	-0.985	-0.972
parkland shooting	4.273	0.264	16.164	0	3.755	4.791
march for our lives	3.4592	0.836	4.138	0	1.821	5.098
cnn sucks	3.4201	0.341	10.02	0	2.751	4.089
#metoo	2.6546	0.171	15.557	0	2.32	2.989
take a knee	2.1943	0.167	13.116	0	1.866	2.522
al-baghdadi	1.1582	0.178	6.493	0	0.809	1.508
kaepernick	1.1673	0.036	32.115	0	1.096	1.239
blue lives matter	1.3683	0.134	10.219	0	1.106	1.631
all lives matter	1.3536	0.097	13.887	0	1.163	1.545
build the wall	1.0103	0.097	10.394	0	0.82	1.201
russia hoax	1.0099	0.357	2.832	0.005	0.311	1.709
radical islam	0.8617	0.051	16.991	0	0.762	0.961
blm	1.2316	0.021	60.024	0	1.191	1.272
hillary	0.5833	0.008	72.187	0	0.567	0.599
racism	0.6247	0.016	38.289	0	0.593	0.657
intellectual property theft	-0.7525	1.182	-0.637	0.524	-3.07	1.565
blue pill	-0.5387	0.271	-1.986	0.047	-1.07	-0.007
crooked hillary	-0.5087	0.073	-7.005	0	-0.651	-0.366
tea party	-0.3431	0.088	-3.893	0	-0.516	-0.17
red pill	-0.3388	0.241	-1.404	0.16	-0.812	0.134
hannity	-0.2627	0.025	-10.599	0	-0.311	-0.214
email server	-0.262	0.1	-2.628	0.009	-0.457	-0.067
czar	-0.2663	0.094	-2.84	0.005	-0.45	-0.083
china	-0.2402	0.015	-16.168	0	-0.269	-0.211
biden	-0.2423	0.028	-8.745	0	-0.297	-0.188
covfefe	-0.2043	0.073	-2.801	0.005	-0.347	-0.061
mcconnell	-0.2019	0.044	-4.597	0	-0.288	-0.116

gop	-0.2007	0.009	-23.504	0	-0.217	-0.184
putin	-0.1913	0.021	-9.13	0	-0.232	-0.15
deep state	-0.1882	0.075	-2.51	0.012	-0.335	-0.041

Table 13 GLM Regression Partisan Words

Dep. Variable: like_count			No. Observations:	1765496		
Model: GLM			Df Residuals:	1765397		
Model Family: NegativeBinomial			Df Model:	98		
Method: IRLS			Log-Likelihood:	1.0000		
Covariance Type: HC2			Deviance	3.1133e+06		
Iterations: 100			Pearson chi2:	1.78e+09		
			Pseudo R-squared.	0.8723		
Variable	Coef	Std Err	t (or z)	P>|t|	[0.025	0.975]
const	-12.9087	0.115	-112.651	0.000	-13.133	-12.684
take a knee	3.9445	0.738	5.345	0	2.498	5.391
march for our lives	3.2542	0.711	4.578	0	1.861	4.648
parkland shooting	2.6249	0.277	9.478	0	2.082	3.168
#metoo	2.5317	0.295	8.573	0	1.953	3.11
kaepernick	2.0879	0.228	9.148	0	1.641	2.535
lock her up	2.0347	0.807	2.522	0.012	0.453	3.616
racial divide	2.9153	0.927	3.146	0.002	1.099	4.732
antifa	1.5611	0.226	6.897	0	1.117	2.005
all lives matter	1.4381	0.227	6.333	0	0.993	1.883
blm	1.3043	0.207	6.29	0	0.898	1.711
racism	1.2555	0.153	8.189	0	0.955	1.556
paris agreement	1.2668	0.88	1.439	0.15	-0.459	2.993
liberal	0.9658	0.155	6.213	0	0.661	1.27
hillary	0.8702	0.088	9.921	0	0.698	1.042
white supremacy	1.2191	0.27	4.512	0	0.69	1.749
red pill	-39.4942	0.051	-777.324	0	-39.594	-39.395
climate change hoax	-38.999	0.055	-704.732	0	-39.107	-38.891
kavanaugh	-37.6052	0.08	-468.477	0	-37.763	-37.448
i can't breathe	-4.6719	1.107	-4.221	0	-6.841	-2.503
czar	-3.403	0.631	-5.394	0	-4.64	-2.167
tea party	-2.6919	0.346	-7.769	0	-3.371	-2.013
ukraine	-2.3985	0.281	-8.545	0	-2.949	-1.848

intellectual property theft	-1.8522	0.16	-11.611	0	-2.165	-1.539
mcconnell	-1.8711	0.347	-5.398	0	-2.55	-1.192
obamacare	-1.2021	0.293	-4.104	0	-1.776	-0.628
biden	-1.0415	0.44	-2.367	0.018	-1.904	-0.179
email server	-1.0585	0.563	-1.878	0.06	-2.163	0.046
me too	-0.8927	0.427	-2.092	0.036	-1.729	-0.056
deep state	-0.8938	0.285	-3.135	0.002	-1.453	-0.335
gop	-0.7326	0.132	-5.535	0	-0.992	-0.473

Annex A.2: Study 1 Correlation Tables

Table 14 Spearman Correlation Tables: Follower Count vs. Retweets

	follower count	y
follower count	1	0.403
y	0.403	1

Table 15 Spearman Correlation Tables: Follower Count vs. Likes

	follower count	y
follower count	1	0.389
y	0.389	1

Table 16 Spearman Correlation Table: Emotion + Follower Count vs. Retweets

	fear	anger	trust	surprise	sadness	disgust	joy	follower count	y
fear	1	0.317	-0.232	0.016	0.374	0.21	-0.276	0.079	0.064
anger	0.317	1	-0.256	0.012	0.216	0.319	-0.244	0.142	0.046
trust	-0.232	-0.256	1	0.03	-0.25	-0.165	0.275	0.003	0.012
surprise	0.016	0.012	0.03	1	0.09	0.059	0.257	-0.033	-0.01
sadness	0.374	0.216	-0.25	0.09	1	0.31	-0.167	-0.001	0.023
disgust	0.21	0.319	-0.165	0.059	0.31	1	-0.124	-0.042	-0.004
joy	-0.276	-0.244	0.275	0.257	-0.167	-0.124	1	-0.147	-0.049
follower count	0.079	0.142	0.003	-0.033	-0.001	-0.042	-0.147	1	0.403
y	0.064	0.046	0.012	-0.01	0.023	-0.004	-0.049	0.403	1

Table 17 Spearman Correlation Table: MFT + Follower Count vs. Likes

	Loyalty	Sanctity	Care	Fairness	Authority	follower count	y
Loyalty	1	0.061	0.063	0.115	0.221	-0.021	0.029
Sanctity	0.061	1	0.101	0.012	0.013	-0.116	-0.017
Care	0.063	0.101	1	0.059	0.036	0.035	0.044
Fairness	0.115	0.012	0.059	1	0.196	0.055	0.063
Authority	0.221	0.013	0.036	0.196	1	0.089	0.07
follower count	-0.021	-0.116	0.035	0.055	0.089	1	0.389
y	0.029	-0.017	0.044	0.063	0.07	0.389	1

Table 18 Spearman Correlation Table: Emotion + MFT + Follower Count vs. Retweets

	fear	anger	trust	surprise	sadness	disgust	joy	follower count	Loyalty	Sanctity	Care	Fairness	Authority	y

fear	1	0.317	-0.232	0.016	0.374	0.21	-0.276	0.079	-0.02	-0.03	0.283	0.031	0.091	0.064
anger	0.317	1	-0.256	0.012	0.216	0.319	-0.244	0.142	-0.001	-0.039	0.159	0.104	0.097	0.046
trust	-0.232	-0.256	1	0.03	-0.25	-0.165	0.275	0.003	0.06	0.021	-0.048	0.051	0.063	0.012
surprise	0.016	0.012	0.03	1	0.09	0.059	0.257	-0.033	-0.022	-0.002	0.052	-0.052	-0.044	-0.01
sadness	0.374	0.216	-0.25	0.09	1	0.31	-0.167	-0.001	-0.02	0.001	0.199	0.034	-0.024	0.023
disgust	0.21	0.319	-0.165	0.059	0.31	1	-0.124	-0.042	0.008	0.073	0.104	0.1	0.014	-0.004
joy	-0.276	-0.244	0.275	0.257	-0.167	-0.124	1	-0.147	0.023	0.099	-0.032	-0.06	-0.103	-0.049
follower count	0.079	0.142	0.003	-0.033	-0.001	-0.042	-0.147	1	-0.021	-0.116	0.035	0.055	0.089	0.403
Loyalty	-0.02	-0.001	0.06	-0.022	-0.02	0.008	0.023	-0.021	1	0.061	0.063	0.115	0.221	0.022
Sanctity	-0.03	-0.039	0.021	-0.002	0.001	0.073	0.099	-0.116	0.061	1	0.101	0.012	0.013	-0.02
Care	0.283	0.159	-0.048	0.052	0.199	0.104	-0.032	0.035	0.063	0.101	1	0.059	0.036	0.077
Fairness	0.031	0.104	0.051	-0.052	0.034	0.1	-0.06	0.055	0.115	0.012	0.059	1	0.196	0.07
Authority	0.091	0.097	0.063	-0.044	-0.024	0.014	-0.103	0.089	0.221	0.013	0.036	0.196	1	0.074
y	0.064	0.046	0.012	-0.01	0.023	-0.004	-0.049	0.403	0.022	-0.02	0.077	0.07	0.074	1

Table 19 Spearman Correlation Table: Emotion + MFT + Follower Count vs. Likes

fear	1	0.317	-0.232	0.016	0.374	0.21	-0.276	0.079	-0.02	-0.03	0.283	0.031	0.091	0.029
anger	0.317	1	-0.256	0.012	0.216	0.319	-0.244	0.142	-0.001	-0.039	0.159	0.104	0.097	0.027
trust	-0.232	-0.256	1	0.03	-0.25	-0.165	0.275	0.003	0.06	0.021	-0.048	0.051	0.063	0.015
surprise	0.016	0.012	0.03	1	0.09	0.059	0.257	-0.033	-0.022	-0.002	0.052	-0.052	-0.044	-0.013
sadness	0.374	0.216	-0.25	0.09	1	0.31	-0.167	-0.001	-0.02	0.001	0.199	0.034	-0.024	0.005
disgust	0.21	0.319	-0.165	0.059	0.31	1	-0.124	-0.042	0.008	0.073	0.104	0.1	0.014	-0.008
joy	-0.276	-0.244	0.275	0.257	-0.167	-0.124	1	-0.147	0.023	0.099	-0.032	-0.06	-0.103	-0.034
follower count	0.079	0.142	0.003	-0.033	-0.001	-0.042	-0.147	1	-0.021	-0.116	0.035	0.055	0.089	0.389

Loyalty	-0.02	-0.001	0.06	-0.022	-0.02	0.008	0.023	-0.021	1	0.061	0.063	0.115	0.221	0.029
Sanctity	-0.03	-0.039	0.021	-0.002	0.001	0.073	0.099	-0.116	0.061	1	0.101	0.012	0.013	-0.017
Care	0.283	0.159	-0.048	0.052	0.199	0.104	-0.032	0.035	0.063	0.101	1	0.059	0.036	0.044
Fairness	0.031	0.104	0.051	-0.052	0.034	0.1	-0.06	0.055	0.115	0.012	0.059	1	0.196	0.063
Authority	0.091	0.097	0.063	-0.044	-0.024	0.014	-0.103	0.089	0.221	0.013	0.036	0.196	1	0.07
y	0.029	0.027	0.015	-0.013	0.005	-0.008	-0.034	0.389	0.029	-0.017	0.044	0.063	0.07	1

Table 20 Spearman Correlation Table: Emotion + MFT + Partisan Word Count + Follower Count vs. Retweets

	fear	anger	trust	surprise	sadness	disgust	joy	follower count	Loyalty	Sanctity	Care	Fairness	Authority	partisan word count	y
fear	1	0.317	-0.232	0.016	0.374	0.21	-0.276	0.079	-0.02	-0.03	0.283	0.031	0.091	-0.024	0.064
anger	0.317	1	-0.256	0.012	0.216	0.319	-0.244	0.142	-0.001	-0.039	0.159	0.104	0.097	0.081	0.046
trust	-0.232	-0.256	1	0.03	-0.25	-0.165	0.275	0.003	0.06	0.021	-0.048	0.051	0.063	0.003	0.012
surprise	0.016	0.012	0.03	1	0.09	0.059	0.257	-0.033	-0.022	-0.002	0.052	-0.052	-0.044	-0.039	-0.01
sadness	0.374	0.216	-0.25	0.09	1	0.31	-0.167	-0.001	-0.02	0.001	0.199	0.034	-0.024	-0.025	0.023
disgust	0.21	0.319	-0.165	0.059	0.31	1	-0.124	-0.042	0.008	0.073	0.104	0.1	0.014	0.013	-0.004
joy	-0.276	-0.244	0.275	0.257	-0.167	-0.124	1	-0.147	0.023	0.099	-0.032	-0.06	-0.103	-0.087	-0.049
follower count	0.079	0.142	0.003	-0.033	-0.001	-0.042	-0.147	1	-0.021	-0.116	0.035	0.055	0.089	0.058	0.403
Loyalty	-0.02	-0.001	0.06	-0.022	-0.02	0.008	0.023	-0.021	1	0.061	0.063	0.115	0.221	0.156	0.022
Sanctity	-0.03	-0.039	0.021	-0.002	0.001	0.073	0.099	-0.116	0.061	1	0.101	0.012	0.013	0.001	-0.02
Care	0.283	0.159	-0.048	0.052	0.199	0.104	-0.032	0.035	0.063	0.101	1	0.059	0.036	-0.02	0.077
Fairness	0.031	0.104	0.051	-0.052	0.034	0.1	-0.06	0.055	0.115	0.012	0.059	1	0.196	0.194	0.07
Authority	0.091	0.097	0.063	-0.044	-0.024	0.014	-0.103	0.089	0.221	0.013	0.036	0.196	1	0.25	0.074
partisan word count	-0.024	0.081	0.003	-0.039	-0.025	0.013	-0.087	0.058	0.156	0.001	-0.02	0.194	0.25	1	0.053

y	0.064	0.046	0.012	-0.01	0.023	-0.004	-0.049	0.403	0.022	-0.02	0.077	0.07	0.074	0.053	1

Table 21 Spearman Correlation Table: Emotion + MFT + Partisan Word Count + Follower Count vs. Likes

	fear	anger	trust	surprise	sadness	disgust	joy	follower count	Loyalty	Sanctity	Care	Fairness	Authority	partisan word count	y
fear	1	0.317	-0.232	0.016	0.374	0.21	-0.276	0.079	-0.02	-0.03	0.283	0.031	0.091	-0.024	0.029
anger	0.317	1	-0.256	0.012	0.216	0.319	-0.244	0.142	-0.001	-0.039	0.159	0.104	0.097	0.081	0.027
trust	-0.232	-0.256	1	0.03	-0.25	-0.165	0.275	0.003	0.06	0.021	-0.048	0.051	0.063	0.003	0.015
surprise	0.016	0.012	0.03	1	0.09	0.059	0.257	-0.033	-0.022	-0.002	0.052	-0.052	-0.044	-0.039	-0.013
sadness	0.374	0.216	-0.25	0.09	1	0.31	-0.167	-0.001	-0.02	0.001	0.199	0.034	-0.024	-0.025	0.005
disgust	0.21	0.319	-0.165	0.059	0.31	1	-0.124	-0.042	0.008	0.073	0.104	0.1	0.014	0.013	-0.008
joy	-0.276	-0.244	0.275	0.257	-0.167	-0.124	1	-0.147	0.023	0.099	-0.032	-0.06	-0.103	-0.087	-0.034
follower count	0.079	0.142	0.003	-0.033	-0.001	-0.042	-0.147	1	-0.021	-0.116	0.035	0.055	0.089	0.058	0.389
Loyalty	-0.02	-0.001	0.06	-0.022	-0.02	0.008	0.023	-0.021	1	0.061	0.063	0.115	0.221	0.156	0.029
Sanctity	-0.03	-0.039	0.021	-0.002	0.001	0.073	0.099	-0.116	0.061	1	0.101	0.012	0.013	0.001	-0.017
Care	0.283	0.159	-0.048	0.052	0.199	0.104	-0.032	0.035	0.063	0.101	1	0.059	0.036	-0.02	0.044
Fairness	0.031	0.104	0.051	-0.052	0.034	0.1	-0.06	0.055	0.115	0.012	0.059	1	0.196	0.194	0.063
Authority	0.091	0.097	0.063	-0.044	-0.024	0.014	-0.103	0.089	0.221	0.013	0.036	0.196	1	0.25	0.07
partisan word count	-0.024	0.081	0.003	-0.039	-0.025	0.013	-0.087	0.058	0.156	0.001	-0.02	0.194	0.25	1	0.059
y	0.029	0.027	0.015	-0.013	0.005	-0.008	-0.034	0.389	0.029	-0.017	0.044	0.063	0.07	0.059	1

Table 22 Spearman Correlation Table: Partisan Terms vs. Retweets (Abridged to the highest 10 positive coefficients)

	parkland shooting	march for our lives	cnn sucks	#metoo	take a knee	kaepernick	blm	blue lives matter	all lives matter	build the wall	follower count	y
parkland shooting	1	0	0	0	0	0	0	0	0	0	0.004	0.006
march for our lives	0	1	0	0	0	0	0	0	0	0	0.001	0.001
cnn sucks	0	0	1	0	0	0	0	0	0	0	0.003	0.003
#metoo	0	0	0	1	0	0	0	0	0	0	0.003	0.008
take a knee	0	0	0	0	1	0.035	0	0	0	0	0.002	0.007
kaepernick	0	0	0	0	0.035	1	0.007	0	0	0	0.01	0.013
blm	0	0	0	0	0	0.007	1	0.002	0.008	0	0.021	0.035
blue lives matter	0	0	0	0	0	0	0.002	1	0.009	0	0.002	0.005
all lives matter	0	0	0	0	0	0	0.008	0.009	1	0	0.005	0.007
build the wall	0	0	0	0	0	0	0	0	0	1	0.003	0.005
follower count	0.004	0.001	0.003	0.003	0.002	0.01	0.021	0.002	0.005	0.003	1	0.403
y	0.006	0.001	0.003	0.008	0.007	0.013	0.035	0.005	0.007	0.005	0.403	1

Table 23 Spearman Correlation Table: Partisan Terms vs. Likes (Abridged to the highest 10 positive coefficients)

	parkland shooting	march for our lives	cnn sucks	#metoo	take a knee	kaepernick	blm	blue lives matter	all lives matter	build the wall	follower count	y
parkland shooting	1	0	0	0	0	0	0	0	0	0	0.004	0.007
march for our lives	0	1	0	0	0	0	0	0	0	0	0.001	0.001
cnn sucks	0	0	1	0	0	0	0	0	0	0	0.003	0.003
#metoo	0	0	0	1	0	0	0	0	0	0	0.003	0.009
take a knee	0	0	0	0	1	0.035	0	0	0	0	0.002	0.007
kaepernick	0	0	0	0	0.035	1	0.007	0	0	0	0.01	0.013
blm	0	0	0	0	0	0.007	1	0.002	0.008	0	0.021	0.035
blue lives matter	0	0	0	0	0	0	0.002	1	0.009	0	0.002	0.004
all lives matter	0	0	0	0	0	0	0.008	0.009	1	0	0.005	0.008
build the wall	0	0	0	0	0	0	0	0	0	1	0.003	0.006
follower count	0.004	0.001	0.003	0.003	0.002	0.01	0.021	0.002	0.005	0.003	1	0.389
y	0.007	0.001	0.003	0.009	0.007	0.013	0.035	0.004	0.008	0.006	0.389	1

Table 24 Pearson Correlation Tables: Follower Count vs. Retweets

	follower count	y

follower count	1	0.308
y	0.308	1

Table 25 Pearson Correlation Tables: Follower Count vs. Likes

	follower count	y
follower count	1	0.029
y	0.029	1

Table 26 Pearson Correlation Table: Emotion + Follower Count vs. Retweets

	fear	anger	trust	surprise	sadness	disgust	joy	follower count	y
fear	1	0.317	-0.232	0.016	0.374	0.21	-0.276	0.079	0.064
anger	0.317	1	-0.256	0.012	0.216	0.319	-0.244	0.142	0.046
trust	-0.232	-0.256	1	0.03	-0.25	-0.165	0.275	0.003	0.012
surprise	0.016	0.012	0.03	1	0.09	0.059	0.257	-0.033	-0.01
sadness	0.374	0.216	-0.25	0.09	1	0.31	-0.167	-0.001	0.023
disgust	0.21	0.319	-0.165	0.059	0.31	1	-0.124	-0.042	-0.004
joy	-0.276	-0.244	0.275	0.257	-0.167	-0.124	1	-0.147	-0.049
follower count	0.079	0.142	0.003	-0.033	-0.001	-0.042	-0.147	1	0.403
y	0.064	0.046	0.012	-0.01	0.023	-0.004	-0.049	0.403	1

Table 27 Pearson Correlation Table: MFT + Follower Count vs. Likes

	Loyalty	Sanctity	Care	Fairness	Authority	follower count	y
Loyalty	1	0.061	0.063	0.115	0.221	-0.021	0.029
Sanctity	0.061	1	0.101	0.012	0.013	-0.116	-0.017
Care	0.063	0.101	1	0.059	0.036	0.035	0.044
Fairness	0.115	0.012	0.059	1	0.196	0.055	0.063
Authority	0.221	0.013	0.036	0.196	1	0.089	0.07
follower count	-0.021	-0.116	0.035	0.055	0.089	1	0.389
y	0.029	-0.017	0.044	0.063	0.07	0.389	1

Table 28 Pearson Correlation Table: Emotion + MFT + Follower Count vs. Retweets

	fear	anger	trust	surprise	sadness	disgust	joy	follower count	Loyalty	Sanctity	Care	Fairness	Authority	y
fear	1	0.317	-0.232	0.016	0.374	0.21	-0.276	0.079	-0.02	-0.03	0.283	0.031	0.091	0.064
anger	0.317	1	-0.256	0.012	0.216	0.319	-0.244	0.142	-0.001	-0.039	0.159	0.104	0.097	0.046
trust	-0.232	-0.256	1	0.03	-0.25	-0.165	0.275	0.003	0.06	0.021	-0.048	0.051	0.063	0.012
surprise	0.016	0.012	0.03	1	0.09	0.059	0.257	-0.033	-0.022	-0.002	0.052	-0.052	-0.044	-0.01
sadness	0.374	0.216	-0.25	0.09	1	0.31	-0.167	-0.001	-0.02	0.001	0.199	0.034	-0.024	0.023
disgust	0.21	0.319	-0.165	0.059	0.31	1	-0.124	-0.042	0.008	0.073	0.104	0.1	0.014	-0.004
joy	-0.276	-0.244	0.275	0.257	-0.167	-0.124	1	-0.147	0.023	0.099	-0.032	-0.06	-0.103	-0.049
follower count	0.079	0.142	0.003	-0.033	-0.001	-0.042	-0.147	1	-0.021	-0.116	0.035	0.055	0.089	0.403
Loyalty	-0.02	-0.001	0.06	-0.022	-0.02	0.008	0.023	-0.021	1	0.061	0.063	0.115	0.221	0.022
Sanctity	-0.03	-0.039	0.021	-0.002	0.001	0.073	0.099	-0.116	0.061	1	0.101	0.012	0.013	-0.02
Care	0.283	0.159	-0.048	0.052	0.199	0.104	-0.032	0.035	0.063	0.101	1	0.059	0.036	0.077
Fairness	0.031	0.104	0.051	-0.052	0.034	0.1	-0.06	0.055	0.115	0.012	0.059	1	0.196	0.07
Authority	0.091	0.097	0.063	-0.044	-0.024	0.014	-0.103	0.089	0.221	0.013	0.036	0.196	1	0.074
y	0.064	0.046	0.012	-0.01	0.023	-0.004	-0.049	0.403	0.022	-0.02	0.077	0.07	0.074	1

Table 29 Pearson Correlation Table: Emotion + MFT + Follower Count vs. Likes

fear	1	0.317	-0.232	0.016	0.374	0.21	-0.276	0.079	-0.02	-0.03	0.283	0.031	0.091	0.029
anger	0.317	1	-0.256	0.012	0.216	0.319	-0.244	0.142	-0.001	-0.039	0.159	0.104	0.097	0.027
trust	-0.232	-0.256	1	0.03	-0.25	-0.165	0.275	0.003	0.06	0.021	-0.048	0.051	0.063	0.015
surprise	0.016	0.012	0.03	1	0.09	0.059	0.257	-0.033	-0.022	-0.002	0.052	-0.052	-0.044	-0.013
sadness	0.374	0.216	-0.25	0.09	1	0.31	-0.167	-0.001	-0.02	0.001	0.199	0.034	-0.024	0.005
disgust	0.21	0.319	-0.165	0.059	0.31	1	-0.124	-0.042	0.008	0.073	0.104	0.1	0.014	-0.008
joy	-0.276	-0.244	0.275	0.257	-0.167	-0.124	1	-0.147	0.023	0.099	-0.032	-0.06	-0.103	-0.034
follower count	0.079	0.142	0.003	-0.033	-0.001	-0.042	-0.147	1	-0.021	-0.116	0.035	0.055	0.089	0.389

Loyalty	-0.02	-0.001	0.06	-0.022	-0.02	0.008	0.023	-0.021	1	0.061	0.063	0.115	0.221	0.029
Sanctity	-0.03	-0.039	0.021	-0.002	0.001	0.073	0.099	-0.116	0.061	1	0.101	0.012	0.013	-0.017
Care	0.283	0.159	-0.048	0.052	0.199	0.104	-0.032	0.035	0.063	0.101	1	0.059	0.036	0.044
Fairness	0.031	0.104	0.051	-0.052	0.034	0.1	-0.06	0.055	0.115	0.012	0.059	1	0.196	0.063
Authority	0.091	0.097	0.063	-0.044	-0.024	0.014	-0.103	0.089	0.221	0.013	0.036	0.196	1	0.07
y	0.029	0.027	0.015	-0.013	0.005	-0.008	-0.034	0.389	0.029	-0.017	0.044	0.063	0.07	1

Table 30 Pearson Correlation Table: Emotion + MFT + Partisan Word Count + Follower Count vs. Retweets

	fear	anger	trust	surprise	sadness	disgust	joy	follower count	Loyalty	Sanctity	Care	Fairness	Authority	partisan word count	y
fear	1	0.317	-0.232	0.016	0.374	0.21	-0.276	0.079	-0.02	-0.03	0.283	0.031	0.091	-0.024	0.064
anger	0.317	1	-0.256	0.012	0.216	0.319	-0.244	0.142	-0.001	-0.039	0.159	0.104	0.097	0.081	0.046
trust	-0.232	-0.256	1	0.03	-0.25	-0.165	0.275	0.003	0.06	0.021	-0.048	0.051	0.063	0.003	0.012
surprise	0.016	0.012	0.03	1	0.09	0.059	0.257	-0.033	-0.022	-0.002	0.052	-0.052	-0.044	-0.039	-0.01
sadness	0.374	0.216	-0.25	0.09	1	0.31	-0.167	-0.001	-0.02	0.001	0.199	0.034	-0.024	-0.025	0.023
disgust	0.21	0.319	-0.165	0.059	0.31	1	-0.124	-0.042	0.008	0.073	0.104	0.1	0.014	0.013	-0.004
joy	-0.276	-0.244	0.275	0.257	-0.167	-0.124	1	-0.147	0.023	0.099	-0.032	-0.06	-0.103	-0.087	-0.049
follower count	0.079	0.142	0.003	-0.033	-0.001	-0.042	-0.147	1	-0.021	-0.116	0.035	0.055	0.089	0.058	0.403
Loyalty	-0.02	-0.001	0.06	-0.022	-0.02	0.008	0.023	-0.021	1	0.061	0.063	0.115	0.221	0.156	0.022
Sanctity	-0.03	-0.039	0.021	-0.002	0.001	0.073	0.099	-0.116	0.061	1	0.101	0.012	0.013	0.001	-0.02
Care	0.283	0.159	-0.048	0.052	0.199	0.104	-0.032	0.035	0.063	0.101	1	0.059	0.036	-0.02	0.077
Fairness	0.031	0.104	0.051	-0.052	0.034	0.1	-0.06	0.055	0.115	0.012	0.059	1	0.196	0.194	0.07
Authority	0.091	0.097	0.063	-0.044	-0.024	0.014	-0.103	0.089	0.221	0.013	0.036	0.196	1	0.25	0.074
partisan word count	-0.024	0.081	0.003	-0.039	-0.025	0.013	-0.087	0.058	0.156	0.001	-0.02	0.194	0.25	1	0.053
y	0.064	0.046	0.012	-0.01	0.023	-0.004	-0.049	0.403	0.022	-0.02	0.077	0.07	0.074	0.053	1

Table 31 Pearson Correlation Table: Emotion + MFT + Partisan Word Count + Follower Count vs. Likes

	fear	anger	trust	surprise	sadness	disgust	joy	follower count	Loyalty	Sanctity	Care	Fairness	Authority	partisan word count	y
fear	1	0.317	-0.232	0.016	0.374	0.21	-0.276	0.079	-0.02	-0.03	0.283	0.031	0.091	-0.024	0.029
anger	0.317	1	-0.256	0.012	0.216	0.319	-0.244	0.142	-0.001	-0.039	0.159	0.104	0.097	0.081	0.027
trust	-0.232	-0.256	1	0.03	-0.25	-0.165	0.275	0.003	0.06	0.021	-0.048	0.051	0.063	0.003	0.015
surprise	0.016	0.012	0.03	1	0.09	0.059	0.257	-0.033	-0.022	-0.002	0.052	-0.052	-0.044	-0.039	-0.013
sadness	0.374	0.216	-0.25	0.09	1	0.31	-0.167	-0.001	-0.02	0.001	0.199	0.034	-0.024	-0.025	0.005
disgust	0.21	0.319	-0.165	0.059	0.31	1	-0.124	-0.042	0.008	0.073	0.104	0.1	0.014	0.013	-0.008
joy	-0.276	-0.244	0.275	0.257	-0.167	-0.124	1	-0.147	0.023	0.099	-0.032	-0.06	-0.103	-0.087	-0.034
follower count	0.079	0.142	0.003	-0.033	-0.001	-0.042	-0.147	1	-0.021	-0.116	0.035	0.055	0.089	0.058	0.389
Loyalty	-0.02	-0.001	0.06	-0.022	-0.02	0.008	0.023	-0.021	1	0.061	0.063	0.115	0.221	0.156	0.029
Sanctity	-0.03	-0.039	0.021	-0.002	0.001	0.073	0.099	-0.116	0.061	1	0.101	0.012	0.013	0.001	-0.017
Care	0.283	0.159	-0.048	0.052	0.199	0.104	-0.032	0.035	0.063	0.101	1	0.059	0.036	-0.02	0.044
Fairness	0.031	0.104	0.051	-0.052	0.034	0.1	-0.06	0.055	0.115	0.012	0.059	1	0.196	0.194	0.063
Authority	0.091	0.097	0.063	-0.044	-0.024	0.014	-0.103	0.089	0.221	0.013	0.036	0.196	1	0.25	0.07
partisan word count	-0.024	0.081	0.003	-0.039	-0.025	0.013	-0.087	0.058	0.156	0.001	-0.02	0.194	0.25	1	0.059
y	0.029	0.027	0.015	-0.013	0.005	-0.008	-0.034	0.389	0.029	-0.017	0.044	0.063	0.07	0.059	1

Table 32 Pearson Correlation Table: Partisan Terms vs. Retweets (Abridged to the highest 10 positive coefficients)

	parkland shooting	march for our lives	cnn sucks	#metoo	take a knee	kaepernick	blm	blue lives matter	all lives matter	build the wall	follower count	y
parkland shooting	1	0	0	0	0	0	0	0	0	0	0.003	0.013
march for our lives	0	1	0	0	0	0	0	0	0	0	0.001	0.003
cnn sucks	0	0	1	0	0	0	0	0	0	0	0.003	0.008
#metoo	0	0	0	1	0	0	0	0	0	0	0.003	0.012
take a knee	0	0	0	0	1	0.035	0	0	0	0	0.002	0.011
kaepernick	0	0	0	0	0.035	1	0.007	0	0	0	0.011	0.026
blm	0	0	0	0	0	0.007	1	0.002	0.008	0	0.021	0.051
blue lives matter	0	0	0	0	0	0	0.002	1	0.009	0	0.003	0.008
all lives matter	0	0	0	0	0	0	0.008	0.009	1	0	0.005	0.013
build the wall	0	0	0	0	0	0	0	0	0	1	0.003	0.009
follower count	0.003	0.001	0.003	0.003	0.002	0.011	0.021	0.003	0.005	0.003	1	0.308
y	0.013	0.003	0.008	0.012	0.011	0.026	0.051	0.008	0.013	0.009	0.308	1

Table 33 Pearson Correlation Table: Partisan Terms vs. Likes (Abridged to the highest 10 positive coefficients)

	parkland shooting	march for our lives	cnn sucks	#metoo	take a knee	kaepernick	blm	blue lives matter	all lives matter	build the wall	follower count	y
parkland shooting	1	0	0	0	0	0	0	0	0	0	0.003	0.002
march for our lives	0	1	0	0	0	0	0	0	0	0	0.001	0.001
cnn sucks	0	0	1	0	0	0	0	0	0	0	0.003	0.002
#metoo	0	0	0	1	0	0	0	0	0	0	0.003	0.001
take a knee	0	0	0	0	1	0.035	0	0	0	0	0.002	0.001
kaepernick	0	0	0	0	0.035	1	0.007	0	0	0	0.011	0.009
blm	0	0	0	0	0	0.007	1	0.002	0.008	0	0.021	0.002
blue lives matter	0	0	0	0	0	0	0.002	1	0.009	0	0.003	0
all lives matter	0	0	0	0	0	0	0.008	0.009	1	0	0.005	0.001
build the wall	0	0	0	0	0	0	0	0	0	1	0.003	0.011
follower count	0.003	0.001	0.003	0.003	0.002	0.011	0.021	0.003	0.005	0.003	1	0.029
y	0.002	0.001	0.002	0.001	0.001	0.009	0.002	0	0.001	0.011	0.029	1

Table 34 Summary Stats: Follower Count vs. Retweets

	Mean	SD	N
follower count	11.277	3.24	1765496
y	0.383	1.253	1765496

Table 35 Summary Stats: MFT + Follower Count vs. Likes

	Mean	SD	N
follower count	11.277	3.24	1765496
y	13.016	606.433	1765496

Table 36 Summary Stats: MFT + Follower Count vs. Retweets

	Mean	SD	N
Loyalty	0.039	0.17	1765496
Sanctity	0.038	0.167	1765496
Care	0.174	0.336	1765496
Fairness	0.111	0.278	1765496
Authority	0.104	0.257	1765496
follower count	11.277	3.24	1765496
y	0.383	1.253	1765496

Table 37 Summary Stats: MFT + Follower Count vs. Likes

	Mean	SD	N
Loyalty	0.039	0.17	1765496
Sanctity	0.038	0.167	1765496
Care	0.174	0.336	1765496
Fairness	0.111	0.278	1765496
Authority	0.104	0.257	1765496
follower count	11.277	3.24	1765496
y	13.016	606.433	1765496

Table 38 Summary Stats: Emotion + MFT + Follower Count vs. Retweets

fear	0.097	0.161	1765496

anger	0.095	0.179	1765496
trust	0.127	0.207	1765496
surprise	0.042	0.11	1765496
sadness	0.068	0.12	1765496
disgust	0.035	0.088	1765496
joy	0.07	0.12	1765496
follower count	11.277	3.24	1765496
Loyalty	0.039	0.17	1765496
Sanctity	0.038		1765496
Care	0.174	0.336	1765496
Fairness	0.111	0.278	1765496
Authority	0.104	0.257	1765496
y	0.383	1.253	1765496

Table 39 Summary Stats: Emotion + MFT + Follower Count vs. Likes

	Mean	SD	N
fear	0.097	0.161	1765496
anger	0.095	0.179	1765496
trust	0.127	0.207	1765496
surprise	0.042	0.11	1765496
sadness	0.068	0.12	1765496
disgust	0.035	0.088	1765496
joy	0.07	0.12	1765496
follower count	11.277	3.24	1765496
Loyalty	0.039	0.17	1765496
Sanctity	0.038	0.167	1765496
Care	0.174	0.336	1765496
Fairness	0.111	0.278	1765496
Authority	0.104	0.257	1765496
y	13.016	606.433	1765496

Table 40 Summary Stats: Emotion + MFT + Partisan Word Count + Follower Count vs. Retweets

	Mean	SD	N
fear	0.097	0.161	1765496
anger	0.095	0.179	1765496
trust	0.127	0.207	1765496

surprise	0.042	0.11	1765496
sadness	0.068	0.12	1765496
disgust	0.035	0.088	1765496
joy	0.07	0.12	1765496
follower count	11.277	3.24	1765496
Loyalty	0.039	0.17	1765496
Sanctity	0.038	0.167	1765496
Care	0.174	0.336	1765496
Fairness	0.111	0.278	1765496
Authority	0.104	0.257	1765496
partisan word count	0.245	0.594	1765496
y	0.383	1.253	1765496

Table 41 Summary Stats: Emotion + MFT + Partisan Word Count + Follower Count vs. Likes

	Mean	SD	N
fear	0.097	0.161	1765496
anger	0.095	0.179	1765496
trust	0.127	0.207	1765496
surprise	0.042	0.11	1765496
sadness	0.068	0.12	1765496
disgust	0.035	0.088	1765496
joy	0.07	0.12	1765496
follower count	11.277	3.24	1765496
Loyalty	0.039	0.17	1765496
Sanctity	0.038	0.167	1765496
Care	0.174	0.336	1765496
Fairness	0.111	0.278	1765496
Authority	0.104	0.257	1765496
partisan word count	0.245	0.594	1765496
y	13.016	606.433	1765496

Table 42 Summary Stats: Partisan Words vs. Retweets

	Mean	SD	N
blue pill	0	0.003	1765496
tea party	0	0.01	1765496
democrats	0.004	0.062	1765496
racism	0.003	0.055	1765496
syria	0.006	0.08	1765496
tds	0	0.012	1765496
sjw	0	0.017	1765496
alternative facts	0	0.006	1765496
dnc	0.002	0.041	1765496
black lives matter	0.001	0.023	1765496
czar	0	0.009	1765496
hannity	0.001	0.036	1765496
fake news	0.001	0.034	1765496
antifa	0.001	0.024	1765496
trump	0.058	0.234	1765496
socialist	0.001	0.028	1765496
al-baghdadi	0	0.005	1765496
left-wing	0	0.009	1765496
hands up don't shoot	0	0.003	1765496
democrat	0.007	0.081	1765496
collusion	0	0.02	1765496
tariffs	0	0.005	1765496
hillary	0.015	0.122	1765496
right-wing	0	0.011	1765496
social justice	0	0.013	1765496
russia	0.009	0.096	1765496
clinton	0.014	0.119	1765496
terrorism	0.002	0.044	1765496
lock her up	0	0.006	1765496
all lives matter	0	0.009	1765496
radical islam	0	0.018	1765496
intellectual property theft	0	0.001	1765496
russia hoax	0	0.002	1765496
take a knee	0	0.005	1765496
white supremacy	0	0.021	1765496
ukraine	0.008	0.087	1765496
march for our lives	0	0.001	1765496
pence	0.002	0.045	1765496
obamacare	0.002	0.042	1765496

benghazi	0.001	0.029	1765496
msm	0.002	0.039	1765496
fox news	0.001	0.027	1765496
deep state	0	0.012	1765496
iran	0.004	0.062	1765496
epa	0.009	0.092	1765496
racial divide	0	0.005	1765496
i can't breathe	0	0.004	1765496
parkland shooting	0	0.003	1765496
covfefe	0	0.012	1765496
nato	0.003	0.057	1765496
blue lives matter	0	0.007	1765496
race war	0	0.007	1765496
gop	0.011	0.105	1765496
#metoo	0	0.005	1765496
liberal	0.005	0.069	1765496
caliphate	0	0.009	1765496
putin	0.002	0.043	1765496
email server	0	0.009	1765496
alt-right	0	0.012	1765496
kaepernick	0.001	0.024	1765496
independent	0.002	0.048	1765496
paris agreement	0	0.004	1765496
kavanaugh	0	0.002	1765496
make america great again	0	0.014	1765496
libertarian	0	0.015	1765496
biden	0.001	0.032	1765496
cnn sucks	0	0.003	1765496
crimea	0	0.009	1765496
trump derangement syndrome	0	0.003	1765496
build the wall	0	0.009	1765496
npc	0	0.013	1765496
blm	0.002	0.043	1765496
maddow	0	0.017	1765496
trade war	0	0.004	1765496
republicans	0.002	0.049	1765496
wikileaks	0.001	0.036	1765496
republican	0.005	0.07	1765496
podesta emails	0	0.005	1765496
mainstream media	0	0.016	1765496
sanders	0.004	0.063	1765496
trump train	0	0.008	1765496
obama	0.027	0.161	1765496

isis	0.009	0.095	1765496
fracking	0	0.013	1765496
birth certificate	0	0.013	1765496
progressive	0.001	0.025	1765496
pelosi	0.001	0.024	1765496
conservative	0.002	0.046	1765496
mcconnell	0	0.02	1765496
establishment	0	0.021	1765496
maga	0.005	0.07	1765496
climate change hoax	0	0.002	1765496
crooked hillary	0	0.012	1765496
china	0.004	0.06	1765496
drain the swamp	0	0.009	1765496
me too	0	0.022	1765496
red pill	0	0.004	1765496
follower count	11.277	3.24	1765496
y	0.383	1.253	1765496

Table 43 Summary Stats: Partisan Words vs. Likes

	Mean	SD	N
blue pill	0	0.003	1765496
tea party	0	0.01	1765496
democrats	0.004	0.062	1765496
racism	0.003	0.055	1765496
syria	0.006	0.08	1765496
tds	0	0.012	1765496
sjw	0	0.017	1765496
alternative facts	0	0.006	1765496
dnc	0.002	0.041	1765496
black lives matter	0.001	0.023	1765496
czar	0	0.009	1765496
hannity	0.001	0.036	1765496
fake news	0.001	0.034	1765496
antifa	0.001	0.024	1765496

trump	0.058	0.234	1765496
socialist	0.001	0.028	1765496
al-baghdadi	0	0.005	1765496
left-wing	0	0.009	1765496
hands up don't shoot	0	0.003	1765496
democrat	0.007	0.081	1765496
collusion	0	0.02	1765496
tariffs	0	0.005	1765496
hillary	0.015	0.122	1765496
right-wing	0	0.011	1765496
social justice	0	0.013	1765496
russia	0.009	0.096	1765496
clinton	0.014	0.119	1765496
terrorism	0.002	0.044	1765496
lock her up	0	0.006	1765496
all lives matter	0	0.009	1765496
radical islam	0	0.018	1765496
intellectual property theft	0	0.001	1765496
russia hoax	0	0.002	1765496
take a knee	0	0.005	1765496
white supremacy	0	0.021	1765496
ukraine	0.008	0.087	1765496
march for our lives	0	0.001	1765496
pence	0.002	0.045	1765496
obamacare	0.002	0.042	1765496
benghazi	0.001	0.029	1765496
msm	0.002	0.039	1765496
fox news	0.001	0.027	1765496
deep state	0	0.012	1765496
iran	0.004	0.062	1765496
epa	0.009	0.092	1765496
racial divide	0	0.005	1765496
i can't breathe	0	0.004	1765496
parkland shooting	0	0.003	1765496
covfefe	0	0.012	1765496
nato	0.003	0.057	1765496
blue lives matter	0	0.007	1765496
race war	0	0.007	1765496
gop	0.011	0.105	1765496
#metoo	0	0.005	1765496
liberal	0.005	0.069	1765496
caliphate	0	0.009	1765496
putin	0.002	0.043	1765496
email server	0	0.009	1765496

alt-right	0	0.012	1765496
kaepernick	0.001	0.024	1765496
independent	0.002	0.048	1765496
paris agreement	0	0.004	1765496
kavanaugh	0	0.002	1765496
make america great again	0	0.014	1765496
libertarian	0	0.015	1765496
biden	0.001	0.032	1765496
cnn sucks	0	0.003	1765496
crimea	0	0.009	1765496
trump derangement syndrome	0	0.003	1765496
build the wall	0	0.009	1765496
npc	0	0.013	1765496
blm	0.002	0.043	1765496
maddow	0	0.017	1765496
trade war	0	0.004	1765496
republicans	0.002	0.049	1765496
wikileaks	0.001	0.036	1765496
republican	0.005	0.07	1765496
podesta emails	0	0.005	1765496
mainstream media	0	0.016	1765496
sanders	0.004	0.063	1765496
trump train	0	0.008	1765496
obama	0.027	0.161	1765496
isis	0.009	0.095	1765496
fracking	0	0.013	1765496
birth certificate	0	0.013	1765496
progressive	0.001	0.025	1765496
pelosi	0.001	0.024	1765496
conservative	0.002	0.046	1765496
mcconnell	0	0.02	1765496
establishment	0	0.021	1765496
maga	0.005	0.07	1765496
climate change hoax	0	0.002	1765496
crooked hillary	0	0.012	1765496
china	0.004	0.06	1765496
drain the swamp	0	0.009	1765496
me too	0	0.022	1765496
red pill	0	0.004	1765496
follower count	11.277	3.24	1765496
y	13.016	606.433	1765496

Annex A.3: Mformer Model with NRCLex Emotion and Partisan Words

```
import os
import ssl
import pandas as pd
import numpy as np
import nltk
from langdetect import detect, LangDetectException
from nrclex import NRCLex
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from transformers import pipeline, AutoTokenizer
from concurrent.futures import ThreadPoolExecutor
import matplotlib.pyplot as plt
from matplotlib.offsetbox import AnchoredText
import time

# Start timer at the very beginning
start_time = time.time()
```

```
# SSL context fix
ssl._create_default_https_context = ssl._create_unverified_context

# Download necessary NLTK resources if not already downloaded
nltk.download('punkt', quiet=True)
nltk.download('vader_lexicon', quiet=True)
nltk.download('stopwords', quiet=True)

# File paths
base_file_path = "C:\\Users\\feeha\\iCloudDrive\\PhD
Research\\IRA\\BigDataFolder\\PoliticalUsers.csv"
final_output_path = "C:\\Users\\feeha\\iCloudDrive\\PhD
Research\\IRA\\BigDataFolder\\PoliticalUsersProcessed.csv"  # ALWAYS use a new file name
```
RECOMMEND A BIG FAT GPU FOR THIS CODE!

```
# Expanded partisan keywords list
PARTISAN_KEYWORDS = [
    # General political terms
    'liberal', 'conservative', 'republican', 'democrat', 'right-wing', 'left-wing', 'progressive', 'alt-right',
    'establishment', 'deep state', 'drain the swamp', 'fake news', 'mainstream media', 'MSM',
    'Obama', 'Biden', 'Clinton', 'Trump', 'Pence', 'Sanders', 'Hillary', 'McConnell', 'Pelosi',
    'DNC', 'GOP', 'Republicans', 'Democrats', 'libertarian', 'independent',
    # 2009–2012 (Obama Presidency)
    'Obamacare', 'Tea Party', 'birth certificate', 'socialist', 'czar',
    # 2016 Election (Trump vs. Clinton)
    'Crooked Hillary', 'Lock her up', 'Make America Great Again', 'MAGA', 'Build the wall',
    'Trump Train', 'email server', 'Benghazi',
    # 2018 (Midterms and Trump Presidency)
    'Trump Derangement Syndrome', 'TDS', 'covfefe', 'alternative facts', 'deep state',
    # Social Movements
    'Hands up don\'t shoot', 'I can\'t breathe', 'Kaepernick', 'take a knee',
    'racial divide', 'Black Lives Matter', 'BLM', 'all lives matter', 'Blue Lives Matter',
    'antifa', 'white supremacy', 'race war', 'social justice', 'racism',
    # Climate and Environment
    'Paris Agreement', 'climate change hoax', 'EPA', 'fracking',
    # International Relations
    'Russia', 'Ukraine', 'Crimea', 'NATO', 'Putin', 'Syria', 'China', 'Iran', 'ISIS', 'terrorism',
    'collusion', 'Russia hoax', 'WikiLeaks', 'Podesta emails', 'al-Baghdadi', 'radical Islam',
'caliphate',
    'trade war', 'tariffs', 'intellectual property theft',
    # Media and Pop Culture
    'CNN sucks', 'Fox News', 'Hannity', 'Maddow', 'SJW', 'NPC', 'red pill', 'blue pill',
    # 2018 Events
    'Parkland shooting', 'March for Our Lives', 'Kavanaugh', 'Me Too', '#MeToo'
]
```

```
# Initialize the Sentiment Analyzer once
sia = SentimentIntensityAnalyzer()

# Define MFT models and columns
models = {
    "joshnguyen/mformer-loyalty": "Loyaltyclassification_result",
    "joshnguyen/mformer-sanctity": "Sanctityclassification_result",
    "joshnguyen/mformer-care": "Careclassification_result",
    "joshnguyen/mformer-fairness": "Fairnessclassification_result",
    "joshnguyen/mformer-authority": "Authorityclassification_result"
}

# Function to process sentiment and engagement on CPU
def process_sentiment_and_engagement(chunk):
    chunk = chunk.copy()  # Create a copy to avoid SettingWithCopyWarning

    # Ensure tweet_text is a string
    chunk['tweet_text'] = chunk['tweet_text'].astype(str)

    # Sentiment analysis using VADER
    chunk['sentiment_score'] = chunk['tweet_text'].apply(lambda x:
sia.polarity_scores(x)['compound'])

    # Engagement calculation with log transformation
    chunk['engagement'] = chunk[['quote_count', 'reply_count', 'like_count',
'retweet_count']].sum(axis=1).apply(np.log1p)

    # Partisan word detection
    chunk['partisan_words'] = chunk['tweet_text'].str.lower().apply(
        lambda text: ', '.join([word for word in PARTISAN_KEYWORDS if word in text.split()]) or
'None'
    )

    # Affect analysis using NRCLex
    chunk['affect'] = chunk['tweet_text'].apply(lambda x: NRCLex(x).affect_frequencies)
    return chunk

# Function to calculate engagement per follower
def calculate_engagement_per_follower(chunk):
    chunk = chunk.copy()  # Create a copy to avoid SettingWithCopyWarning

    if 'follower count' in chunk.columns:
        chunk['engagement_per_follower'] = chunk.apply(
            lambda row: row['engagement'] / row['follower count'] if row['follower count'] > 0 else
np.nan,
            axis=1
```

```
        )
    else:
        print("Warning: 'follower count' column is missing.")
        chunk['engagement_per_follower'] = np.nan
    return chunk

# GPU-based classification for all MFT models with increased batch size
def classify_mft_gpu_all(chunk, models, batch_size=8192):
    chunk = chunk.copy()  # Create a copy to avoid SettingWithCopyWarning

    for model_name, output_column in models.items():
        print(f"Processing model: {model_name}")
        tokenizer = AutoTokenizer.from_pretrained(model_name)
        pipe = pipeline("text-classification", model=model_name, tokenizer=tokenizer, device=0)

        text_data = chunk['tweet_text'].tolist()
        results = []

        for i in range(0, len(text_data), batch_size):
            batch = text_data[i:i + batch_size]
            outputs = pipe(batch, truncation=True, max_length=512)
            results.extend([f"{output['label']} ({output['score']:.2f})" for output in outputs])

        chunk[output_column] = results
    return chunk

# Function to process each chunk: filtering, sentiment, engagement, and GPU-based MFT
classification
def process_and_classify_chunk(chunk, models):
    # Remove rows where 'tweet_text' is blank or NaN
    chunk = chunk[chunk['tweet_text'].notna() & (chunk['tweet_text'].str.strip() != '')]

    # Retain all original columns
    original_columns = chunk.columns.tolist()

    # Process sentiment and engagement, adding new columns
    chunk = process_sentiment_and_engagement(chunk)

    # Calculate engagement per follower
    chunk = calculate_engagement_per_follower(chunk)

    # GPU-based MFT classification
    chunk = classify_mft_gpu_all(chunk, models)

    # Ensure original columns are preserved in the output
    return chunk[original_columns + [
```

```
        'sentiment_score', 'engagement', 'partisan_words', 'affect', 'engagement_per_follower',
        "Loyaltyclassification_result", "Sanctityclassification_result", "Careclassification_result",
        "Fairnessclassification_result", "Authorityclassification_result"
    ]]


# Process the CSV file in chunks with parallel processing and collect log engagement values
log_engagement_values = []
chunk_size = 50000
with ThreadPoolExecutor() as executor:
    with pd.read_csv(base_file_path, chunksize=chunk_size) as reader:
        futures = []
        for chunk in reader:
            futures.append(executor.submit(process_and_classify_chunk, chunk, models))

        for future in futures:
            processed_chunk = future.result()
            log_engagement_values.extend(processed_chunk['engagement'].values)  # Collect log
engagement values
            processed_chunk.to_csv(final_output_path, mode='a', header=not
os.path.exists(final_output_path), index=False)

# Stop timer and print total runtime
end_time = time.time()
total_time = end_time - start_time
print(f"Data processing complete. Results saved to {final_output_path}")
print(f"Total runtime: {total_time:.2f} seconds")
```

Annex B.1: Study 2 Full Results

Model 1: Baseline Model

This model uses the balanced U.S. sample (9.3M tweets; 3.1M per class), seven NRC emotions and five MFT scores in native units, plus a raw partisan-word count. The 80/20 stratified split preserves class composition in train/test. After fitting on U.S. tweets, the model is applied to the unlabeled Russian corpus to estimate alignment shares.

Held-out performance is 0.51 accuracy (macro-F1 = 0.45). Class recalls show the familiar asymmetry: politician is learned best (recall = 0.62), random is high due to the preponderance of neutral/low-affect tweets (recall = 0.80) and engaged is weakest (recall = 0.10). Full metrics are reported in Table 14.

Table 44 Classification Report Baseline Tweet Model

Class	Precision	Recall	F1-score	Support
politician	0.61	0.62	0.61	620000
engaged user	0.47	0.10	0.17	620000
random	0.46	0.80	0.58	620000
Accuracy			0.51	1860000
Macro avg	0.51	0.51	0.45	1860000
Weighted avg	0.51	0.51	0.45	1860000

Feature importance (has) is led by Authority and Care, with Fairness close behind; partisan word count ranks in the middle of the distribution, and emotions such as trust, disgust, and joy contribute smaller but non-trivial effects. To understand why these features separate classes, I ran per-class sparsity diagnostics on the training split. For each emotion and MFT variable, I computed the share of tweets with a non-zero value (present_pct) and the mean, noting that medians are zero throughout. These checks show that non-zero rates are substantially higher for politicians: Care appears in roughly 30 percent of politician tweets versus about 11 percent for engaged users and 7 percent for random users, and trust in about 44 percent versus 27 percent and 22 percent, respectively. I then ran a partisan intensity diagnostic (mean and median counts, plus >0 prevalence), which revealed a sparse, heavy-tailed distribution, with means of

approximately 0.172 for politicians, 0.117 for engaged users, and 0.035 for random users, and median = 0 for all classes. These diagnostics explain the feature-importance pattern: MFT and emotion variables separate politicians primarily by higher activation rates across many tweets, while partisan terms provide localized lift when they appear but, given their sparsity, have limited impact on overall accuracy.

The confusion matrix (Figure 2) confirms that errors concentrate on the engaged and random boundary rather than politician and others. Applied to Russian tweets, the fitted model assigns 57.3 percent as random, 30.7 percent as politician, and 12.1 percent as engaged (Figure 3). Relative to the U.S. classes, Russian messages therefore cluster primarily with ordinary/neutral discourse, with a sizeable minority exhibiting the moral-authority cadence characteristic of elite communication. This placement is driven more by MFT-style differences than by explicit partisan terms, which are too sparse per tweet to materially shift outcomes.

Substantively, Model 1 addresses RQ1**:** content-only rhetorical features (MFT + emotions) separate politicians from other U.S. users at scale, while the engaged vs. random boundary remains difficult without lexical/topic cues. The partisan word count contributes limited incremental signal at the tweet level due to sparsity. Applied to Russian tweets, the model places a majority with random-style discourse and a sizable minority with elite-style (politician-like) tone, consistent with the interpretation that influence messages emulate moral-authority framing more than overt partisan language.

Figure 1 Feature Importance Baseline Tweet Classification

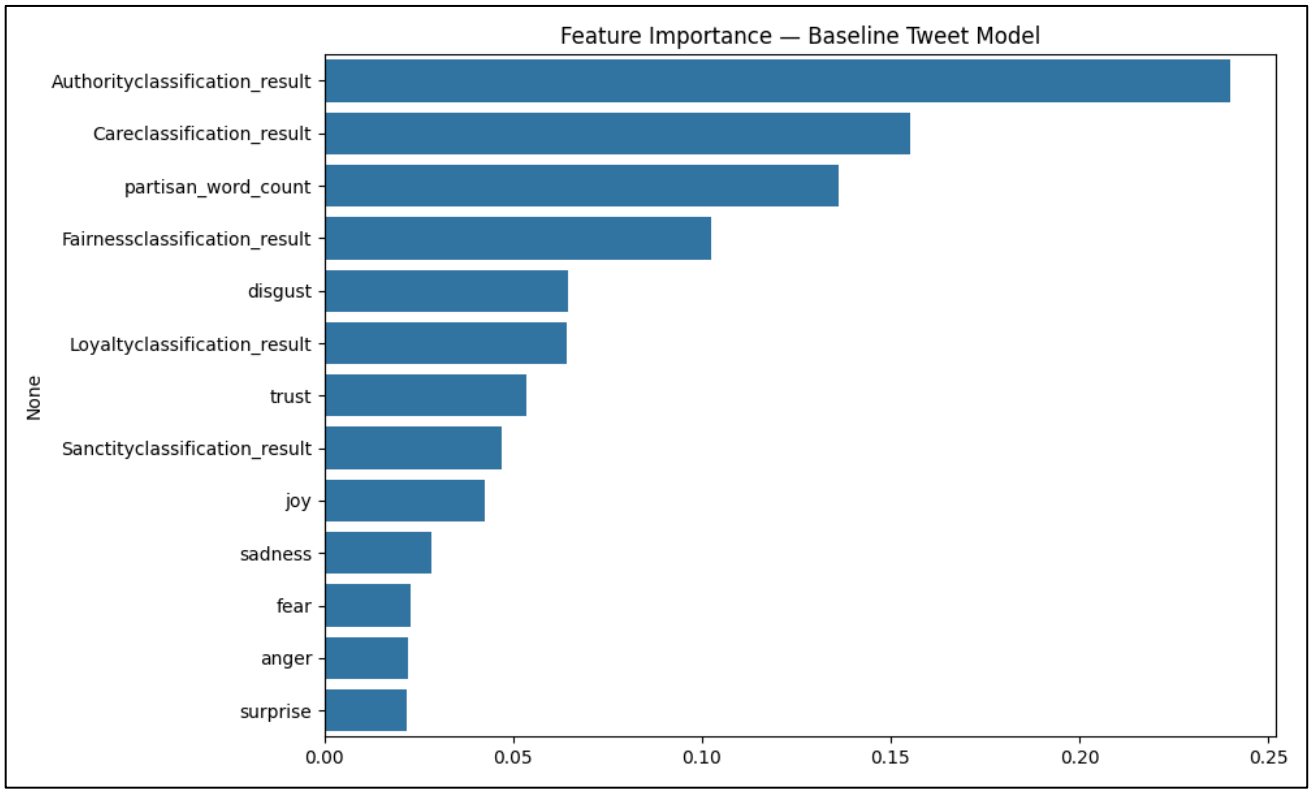


Figure 2 Confusion Matrix Baseline Tweet Model

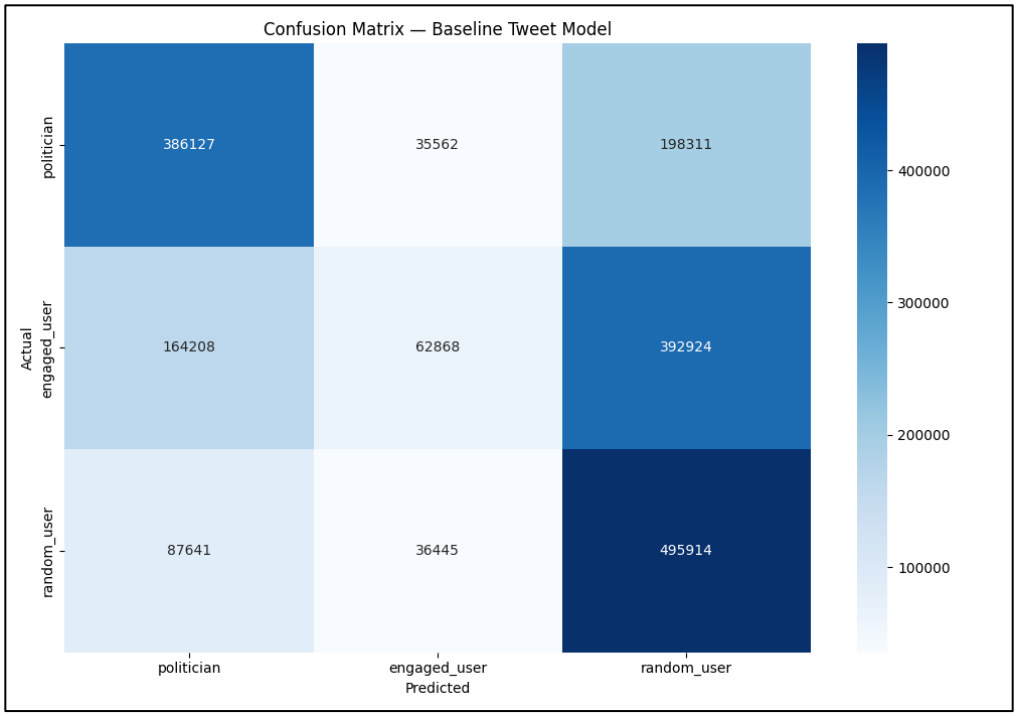
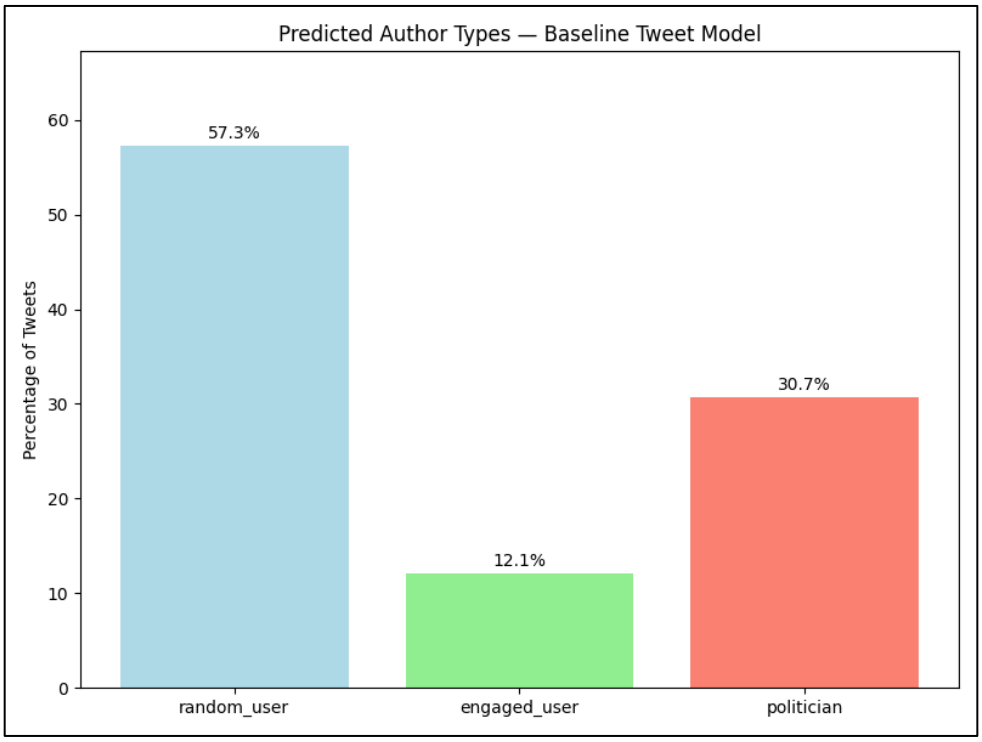


Figure 3 Predicted Author Types Baseline Tweet Classification



Model 2: No Partisan Words

This variant removes all partisan features from the baseline feature set, retaining only the seven NRC emotions and five Moral Foundations Theory (MFT) scores. The model is trained on the balanced U.S. tweet sample (9.3 million total; 3.1 million per class) with an 80/20 stratified split to preserve class composition. It is then applied to the unlabeled Russian corpus to estimate class alignment shares.

Held-out performance is 0.50 accuracy (macro-F1 = 0.46). Class-level metrics show similar asymmetry to the baseline: politician (precision = 0.61, recall = 0.60), engaged user (precision =

0.42, recall = 0.12), and random user (precision = 0.45, recall = 0.77). Full metrics are reported in Table 15.

Table 45 Classification Report No Partisan Words Tweet

Class	Precision	Recall	F1-score	Support
politician	0.61	0.60	0.61	620000
engaged user	0.42	0.12	0.19	620000
random	0.45	0.77	0.57	620000
Accuracy			0.50	1860000
Macro avg	0.49	0.50	0.46	1860000
Weighted avg	0.49	0.50	0.46	1860000

Feature importance (Figure 4) in the no-partisan model is dominated by Authority, Care, and Fairness, which together account for over half of the total importance. Loyalty follows at a mid-tier level, while emotions such as trust, disgust, and joy contribute smaller but consistent effects. Removing partisan features reveals the underlying structure of the classifier: moral framing alone is sufficient to differentiate elite from non-elite rhetoric.

Per-class sparsity diagnostics on the training split clarify these differences. Non-zero activation rates for key MFT features are far higher among politicians: Care appears in approximately 30.4 percent of politician tweets versus 11.3 percent for engaged and 7.2 percent for random; Authority in 28.7 versus 11.8 versus 4.2 percent; and Fairness in 27.1 versus 10.7 versus 4.3 percent. Trust shows the same stratification (about 44 versus 27 versus 22 percent). Other

emotions are more uniformly distributed and have weaker discriminative utility. These diagnostics confirm that higher moral-signal prevalence among politicians drives most of the separability in this model.

The confusion matrix (Figure 5) shows errors concentrated between engaged and random classes rather than between politicians and others, consistent with a rhetoric-only model that lacks topical or network context. When applied to Russian tweets, the fitted model labels 58.2 percent as random, 29.0 percent as politician, and 12.8 percent as engaged (Figure 6). Relative to U.S. patterns, Russian discourse therefore clusters mostly with neutral or low-affect tweets, while nearly a third adopt the moral-authority style characteristic of elite communication.

Substantively, this model isolates RQ1's rhetorical baseline: content features derived from moral and emotional tone can distinguish politicians from ordinary users at scale, even without explicit partisan language. The engaged-versus-random boundary remains weak because both groups exhibit similarly low moral activation and minimal emotional differentiation. Applied to Russian content, the model reproduces the random-dominant distribution observed earlier, confirming that elite-style moral framing, rather than partisan vocabulary, anchors the political resemblance in Russian influence tweets.

Figure 4 Feature Importance No Partisan Words Tweet Classification

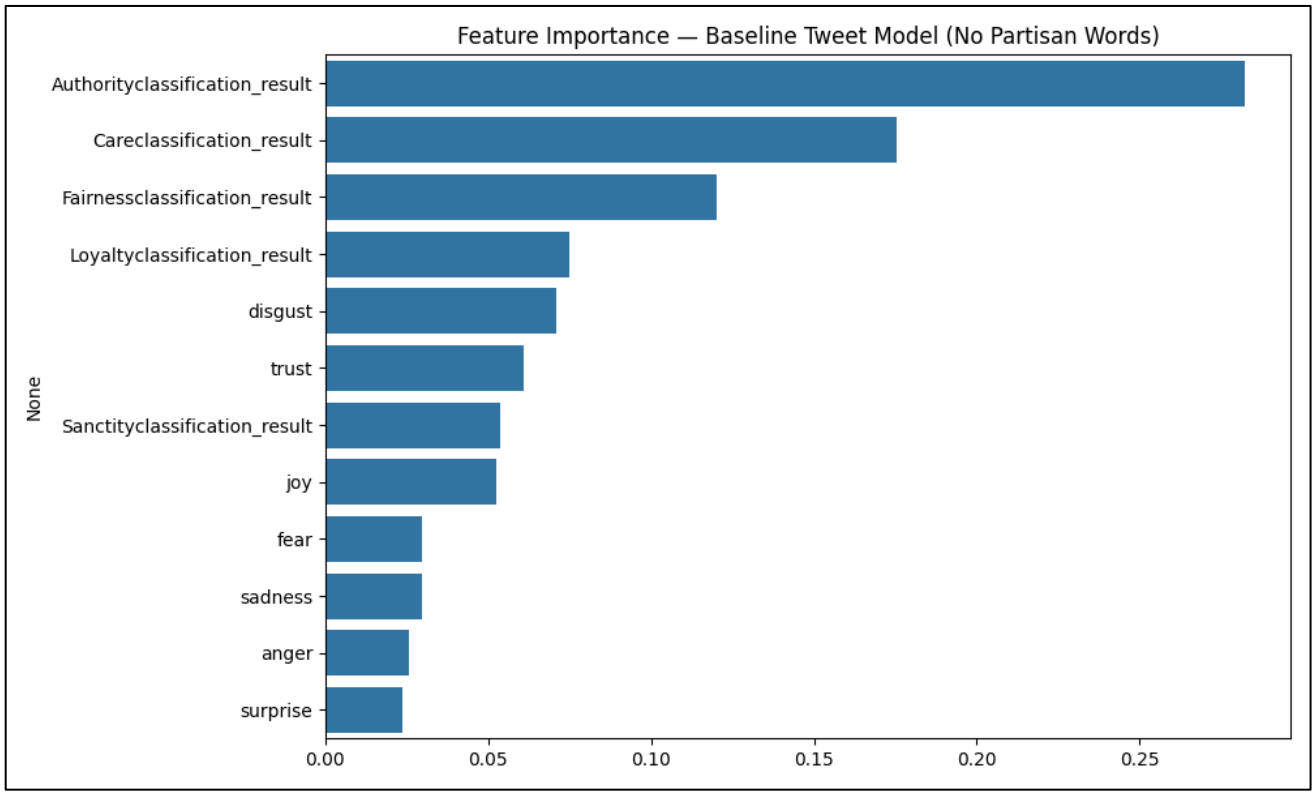


Figure 5 Confusion Matrix No Partisan Words Tweet Classification

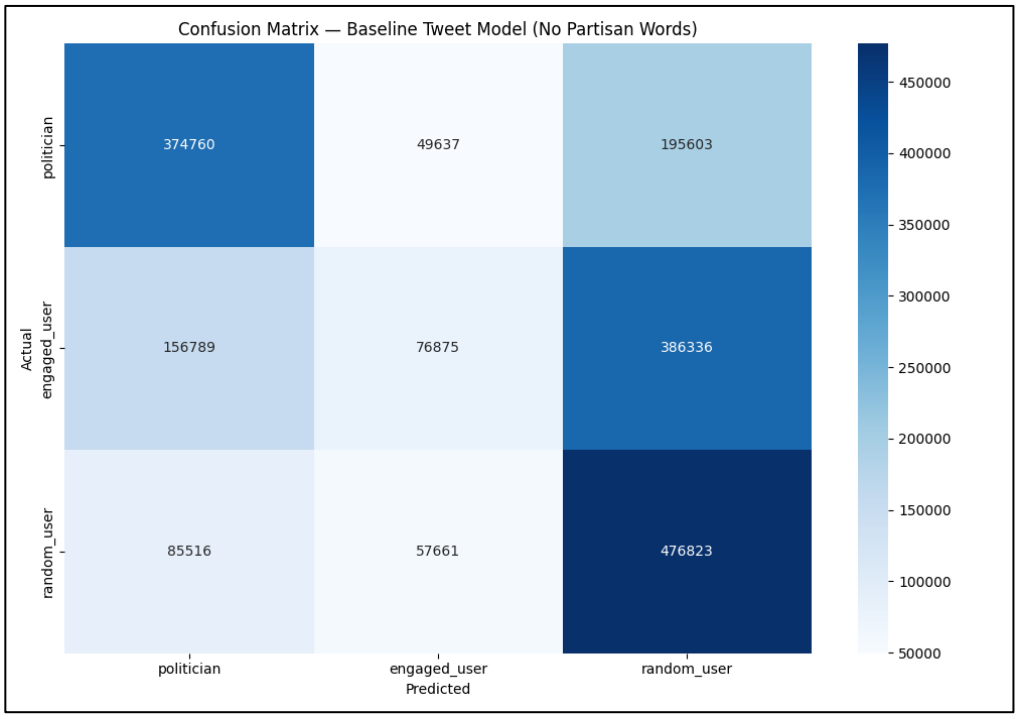
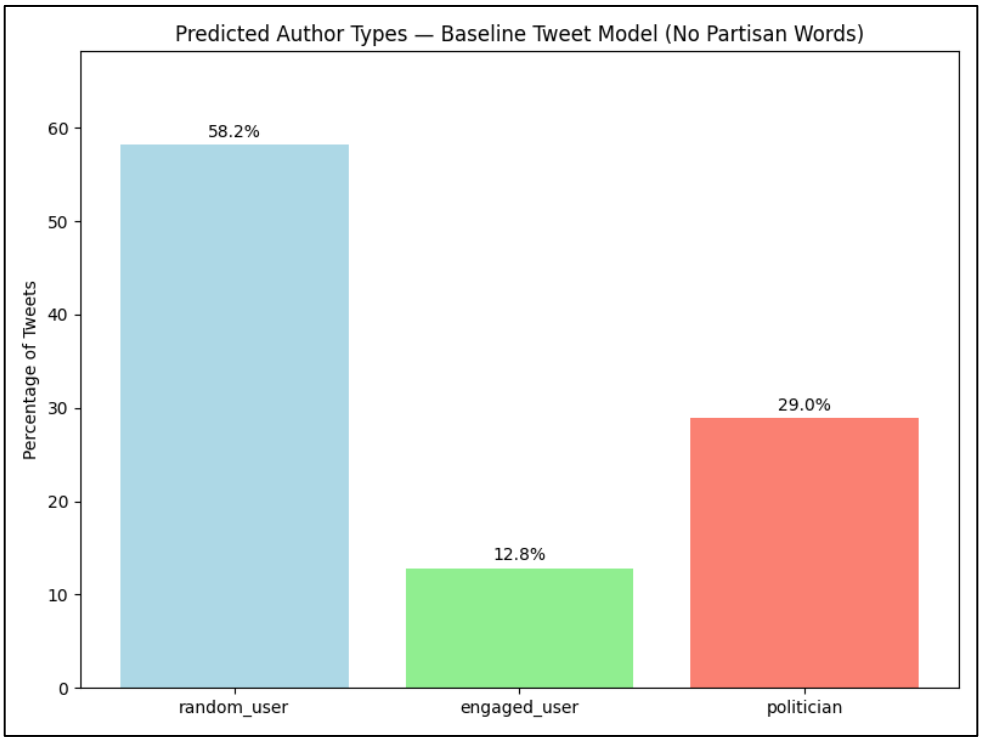


Figure 6 Predicted Author Types No Partisan Words Tweet Classification



Model 3: Enhanced Partisan Model

This variant adds twenty individual partisan indicators (binary phrase matches) to the baseline feature set of seven NRC emotions, five MFT scores, and raw partisan word count. Training uses the balanced U.S. sample (9.3 million tweets; 3.1 million per class) with an 80/20 stratified split, and the fitted model is then applied to Russian tweets to estimate alignment shares. Held-out performance is 0.51 accuracy with macro-F1 = 0.46. Class recalls retain the familiar asymmetry: recall is 0.61 for politicians, 0.80 for random users, and 0.12 for engaged users. Relative to the baseline without these indicators, changes are minimal (engaged recall increases by 0.02; macro-F1 by 0.01). The full classification report appears in Table 16.

Table 46 Classification Report Top Partisan Words Tweet

Class	Precision	Recall	F1-score	Support
politician	0.61	0.61	0.61	397050
engaged user	0.48	0.12	0.19	289214
random	0.46	0.80	0.58	260737
Accuracy			0.51	947001
Macro avg	0.52	0.51	0.46	947001
Weighted avg	0.52	0.51	0.46	947001

Feature importance (Figure 7) remains led by Authority, Care, and Fairness, with partisan word count still in the middle of the distribution. Among the new individual indicators, the largest effects are modest: terms such as obamacare, gop, mcconnell, and ukraine rank above trust and joy, while others (e.g., climate change hoax, tea party, and blue lives matter) sit near the bottom of the importance spectrum. This pattern is consistent with sparse, heavy-tailed partisan usage, where a few terms provide localized lift without displacing MFT features as the primary signals. The confusion matrix (Figure 8) again shows errors concentrated along the engaged versus random boundary rather than between politicians and others, and the additional partisan indicators do not materially reduce overlap or error mass in the engaged column.

Russian predictions (Figure 9) are correspondingly stable: 57.6 percent of Russian tweets are labeled random, 28.3 percent politician, and 14.1 percent engaged. Compared to the baseline, the engaged share increases by roughly two percentage points and the politician share declines by about two to three points, but the qualitative picture is unchanged: most Russian tweets align

with ordinary or neutral discourse, with a sizable minority exhibiting elite-style moral framing. Substantively, adding the top twenty partisan terms provides incremental signal without altering the core story. Tweet-level separability continues to hinge on moral framing via Authority, Care, and Fairness, while partisan vocabulary is too sparse at the tweet level to drive substantial accuracy gains. The Russian placement is robust across models, reinforcing the claim that Russian messaging emulates elite moral cadence more than overt partisan language.

Figure 7 Feature Importance Top Partisan Words Tweet Classification

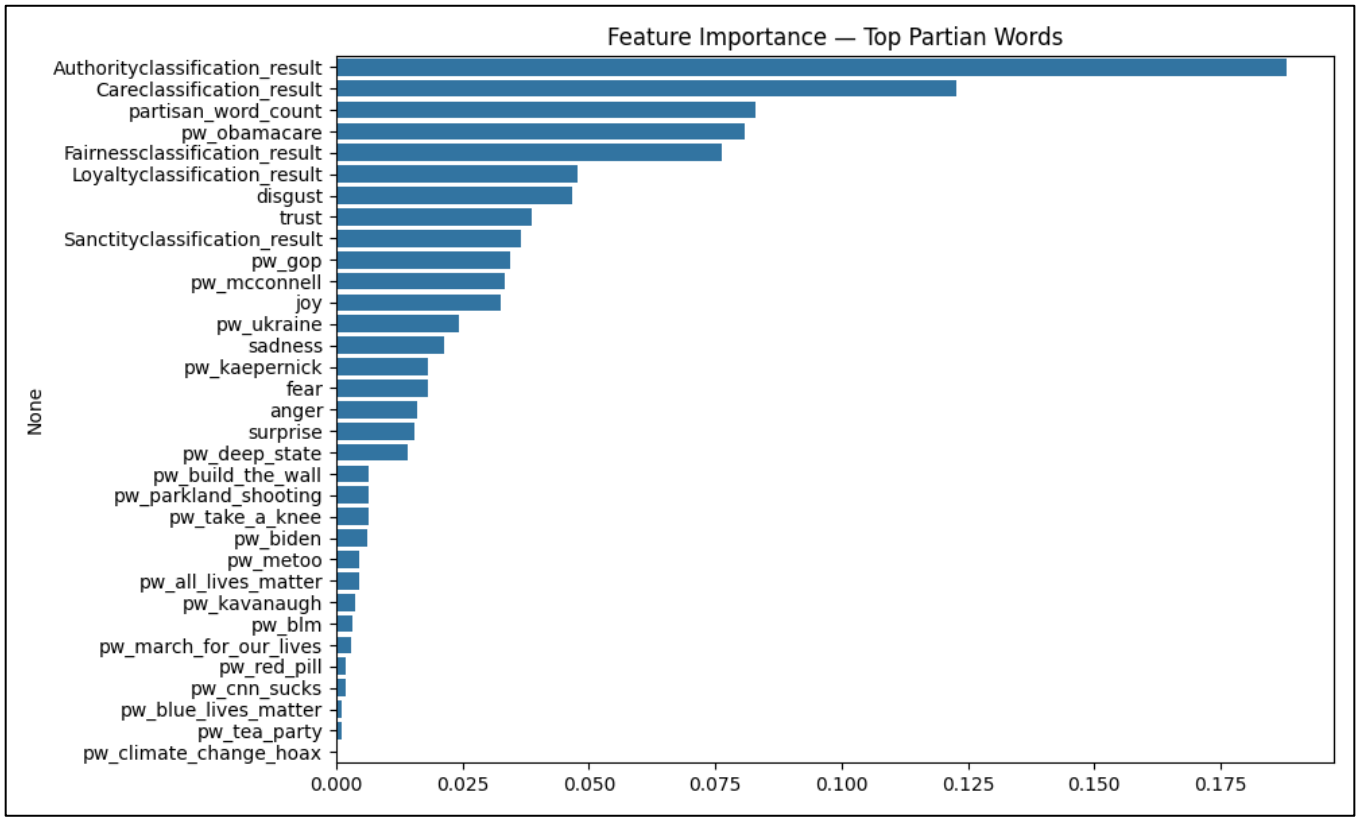


Figure 8 Confusion Matrix Top Partisan Words Tweet Classification

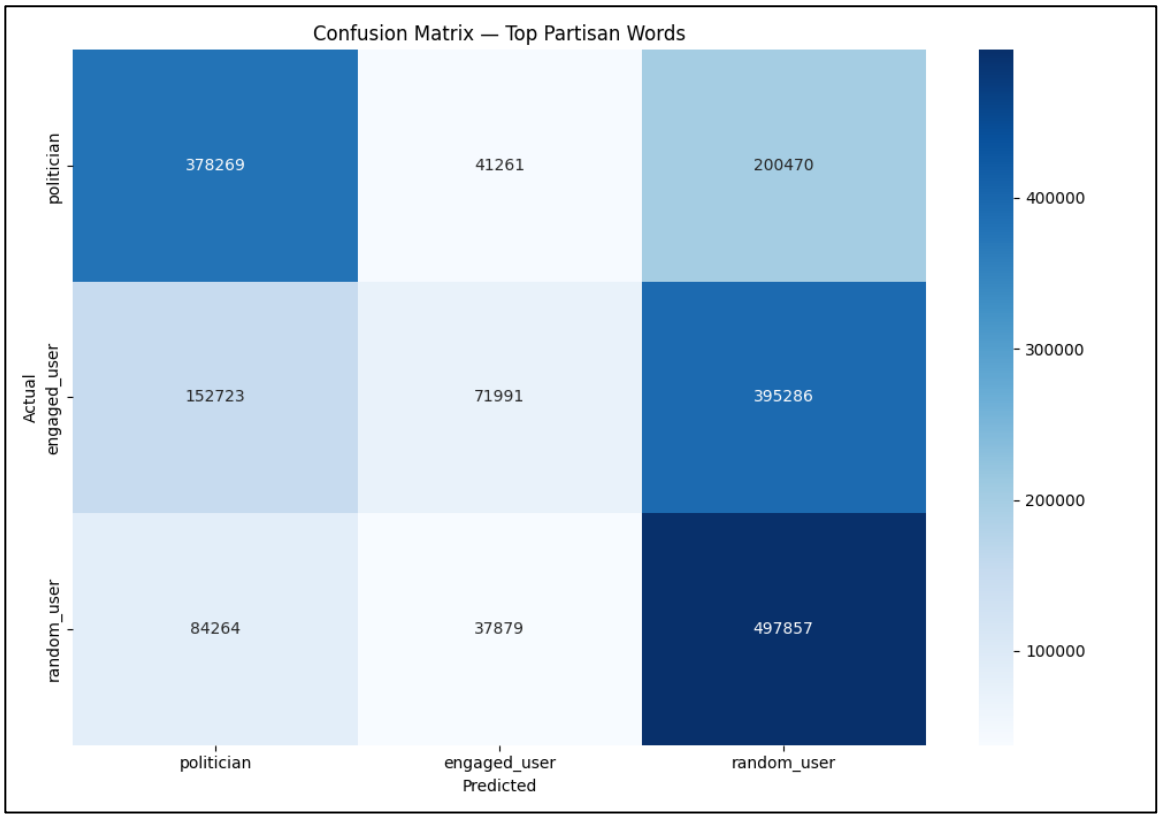
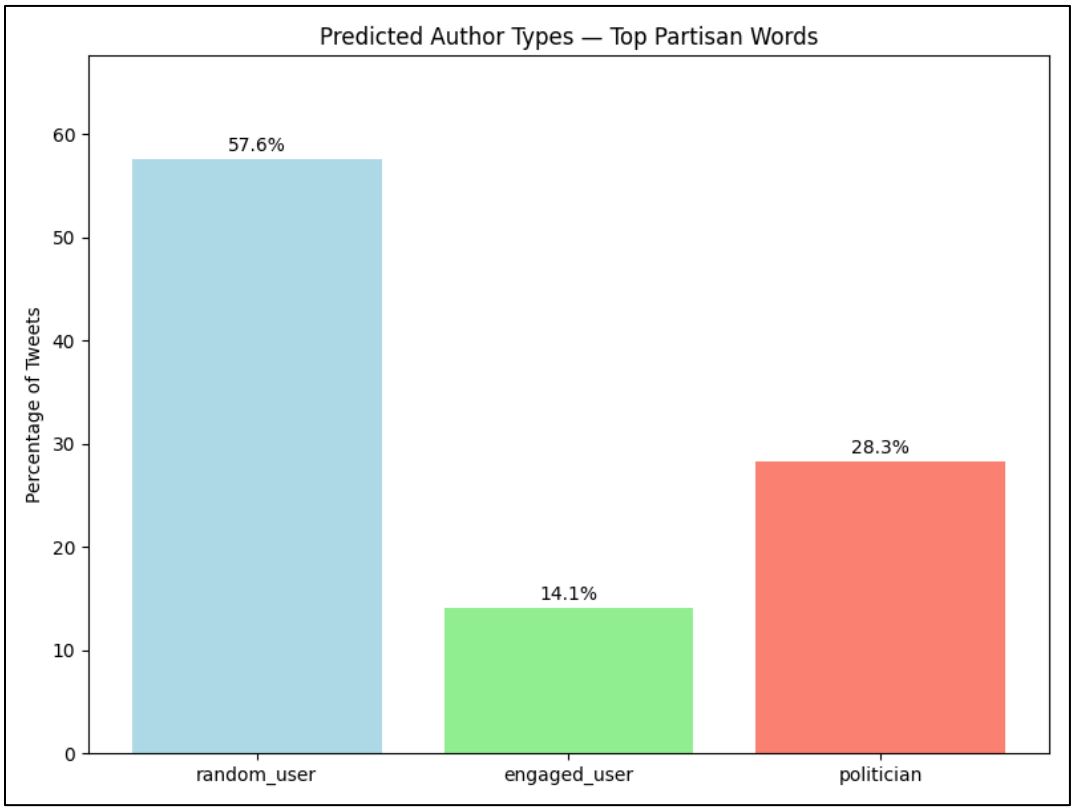


Figure 9 Predicted Author Types Top Partisan Words Tweet Classification



Model 3: Term Frequency–Inverse Document Frequency (TF-IDF)

To validate the structure of the tweet-level classifiers, I trained a theory-agnostic lexical baseline model that uses only lexical features (10,000 TF-IDF unigrams from raw text), excluding all curated signals such as Moral Foundations scores, NRC emotions, and partisan indicators. The model was trained on the same balanced U.S. sample (3.1 million tweets per class: 9.3 million total) using an XGBoost classifier.

Performance is strong relative to the rhetorical models. Overall accuracy is 0.67 with macro-F1 = 0.67. Class metrics are 0.83 precision, 0.80 recall, and 0.81 F1 for politicians; 0.55, 0.55, and 0.55 for engaged users; and 0.62, 0.65, and 0.64 for random users (Table 17). Compared with the MFT- and emotion-based baselines, TF-IDF substantially improves recall for engaged users while keeping politician and random performance high. This indicates that general lexical patterns alone capture a large share of tweet-level separability.

Table 47 Classification Report TF-IDF Tweet Model

Class	Precision	Recall	F1-score	Support
politician	0.83	0.80	0.81	620,000
engaged user	0.55	0.55	0.55	620,000
random	0.62	0.65	0.64	620,000
Accuracy			0.67	1,860,000
Macro avg	0.67	0.67	0.67	1,860,000

Weighted avg	0.67	0.67	0.67	1,860,000

The connection to the new MFT-dominant baselines is instructive. After stabilizing partisan vocabulary in the rhetorical models (binary partisan count, regularized trees), feature importance now weights MFT, especially Authority and Care, above partisan intensity. The TF-IDF feature-importance plot (Figure 10) converges on a similar signal. The highest-weight unigrams are civic and institutional terms such as continue, communities, nation, federal, veterans, families, security, economy, budget, senate, housedemocrats, housegop, and sotu, rather than overt partisan slogans. In other words, both the theory-driven and theory-agnostic lexical baseline models point to policy and governance vocabulary, conceptually aligned with Authority and Care framing, as a central dimension separating elite from non-elite tweet styles, while the engaged-versus-random boundary remains harder to resolve at the single-tweet level.

Applied to 3.1 million Russian tweets, the TF-IDF model predicts 45.7 percent random, 49.0 percent engaged, and 5.2 percent politician (Figure 11). Relative to the rhetorical baselines, this shifts Russian tweets toward engaged and away from politician, but the qualitative picture is unchanged: most Russian tweets resemble non-elite U.S. discourse, with only a small elite-like minority. Feature-importance patterns help contextualize this result. Because both politicians and highly engaged users make heavy use of civic and institutional vocabulary, a theory-agnostic lexical baseline bag-of-words model can more easily distinguish elites from ordinary users than separate elites from highly engaged non-elites, and Russian accounts tend to cluster in that latter region.

Substantively, the convergence between models that emphasize MFT over raw partisan intensity and a purely lexical TF-IDF model is consistent with the view that tweet-level separability is driven more by broad institutional and public-affairs language than by partisan keyword spikes. This supports the interpretation that the observed rhetorical structure reflects properties of the content rather than artifacts of feature engineering, and that Russian messages mostly blend into mainstream or engaged discourse rather than consistently mimicking political elites. Those TF-IDF unigrams, however, clearly inhabit a political register, and with the right neighbors they may function as partisan cues. The next section probes this directly by using the fitted TF-IDF model to select top unigrams and then examining ±3-word key-word-in-context windows around them to assess whether they tend to appear in partisan or non-partisan frames.

Figure 10 Top 30 Features TF-IDF

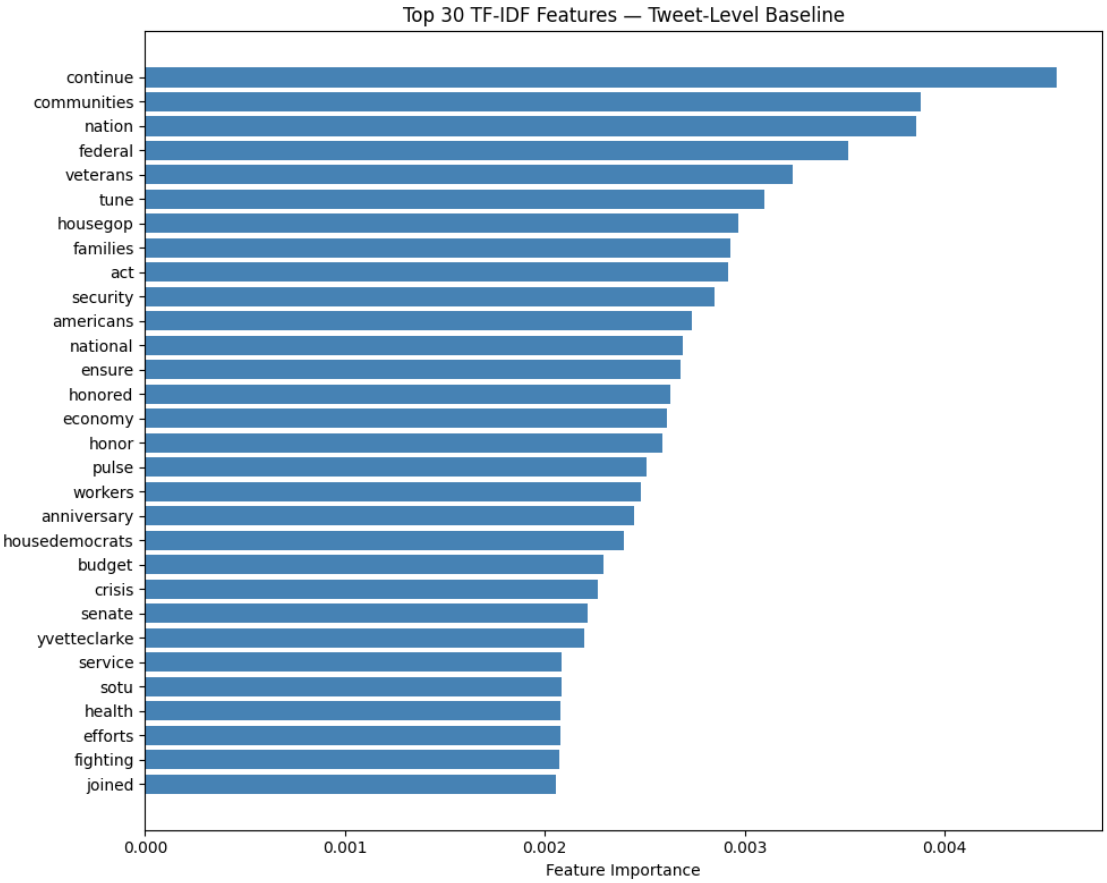


Figure 11 TF-IDF Confusion Matrix

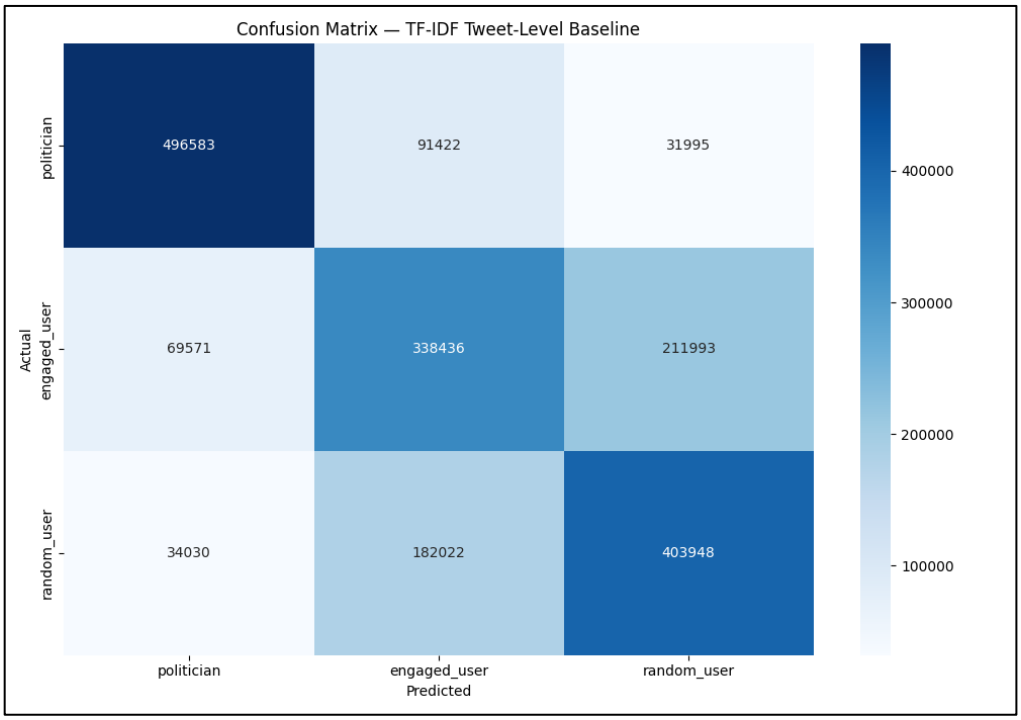
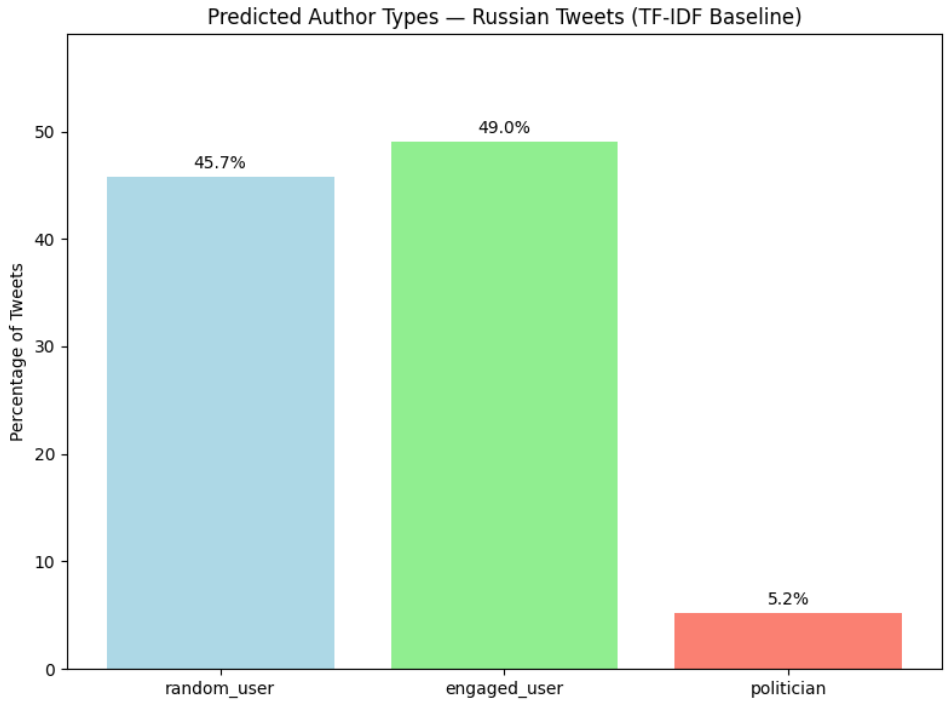


Figure 12 Predicted Author Types TF-IDF



Model 4: TF-IDF KWIC Diagnostic

To probe whether the top TF-IDF features function as covert partisan cues, I use the fitted TF-IDF tweet-level model to identify the 50 most important unigrams and then conduct a key-word-in-context (KWIC) analysis. For each focal term, I extract a ±3-word window around each occurrence in four corpora: U.S. politicians, U.S. engaged users, U.S. random users, and Russian tweets. Within each window, I record whether any partisan token from my dictionary appears and compute, for each term class combination, the total number of windows, the number containing a partisan token, and the resulting co-occurrence rate.

Quantitatively, partisan co-occurrence remains modest across the top 50 TF-IDF terms. Aggregating over terms, the median share of partisan contexts is 2.9 percent for politicians, 3.4 percent for engaged users, 3.0 percent for random users, and 8.0 percent for Russians, with mean co-occurrence rates of 3.6, 5.2, 3.6, and 9.2 percent respectively. When weighting by the number of contexts per term, the overall share of partisan windows is 4.1 percent for politicians, 5.3 percent for engaged users, 3.5 percent for random users, and 7.8 percent for Russians. Only 14 of the 200 term–class combinations exceed 15 percent partisan co-occurrence, concentrated on strongly institutional tokens such as senate, house, scotus, housegop, congressional, sotu, trumpcare, and senatortimscott. Even in these cases, partisan tokens appear in a minority of windows: most uses of these high-importance unigrams occur without any partisan term in the immediate ±3-word neighborhood.

The KWIC examples clarify how these terms are used in practice. For focal terms such as budget and communities, contexts across all U.S. classes are dominated by public-affairs and distributional concerns, “confusion in our communities,” “helping families and communities,”

“budget that protects our communities”, without explicit partisan tokens in the local window. Where partisan co-occurrence does appear, it is tightly linked to a small set of intensely institutional terms. For example, housegop appears in strings such as “gop housegop senategop do your…” or “senategop housegop gopleader…,” sotu appears alongside Trump or Obama, and senate or scotus appear in references to Democratic or Republican actions on nominations or legislation. Russian tweets show the same pattern on these focal terms: partisan windows are concentrated on a few institutional tokens and often mention Trump, the GOP, Democrats, or issue-linked phrases such as trumpcare, while the majority of occurrences of civic terms like communities, economy, or security do not contain any partisan token in the ±3-word window.

Substantively, the TF-IDF + KWIC results are consistent with the view that the high-importance unigrams identified by the TF-IDF model inhabit a civic and institutional register rather than acting primarily as dense partisan triggers. Co-occurrence rates with partisan tokens are low for most term–class combinations and only moderately higher for Russians and highly engaged users on a small set of strongly institutional terms. This aligns with the rhetorical models, which emphasize Moral Foundations, especially Authority and Care, over raw partisan intensity: both approaches suggest that tweet-level separability is driven more by broad public-affairs language than by partisan keyword spikes. Russian IRA tweets participate in this same civic vocabulary and, in a minority of cases, place it near partisan tokens, but most uses of the top TF-IDF terms occur in non-partisan local contexts. The KWIC diagnostic therefore supports the interpretation that rhetorical ambiguity in Russian content arises from moderated, elite-adjacent political style rather than from the absence of political or institutional language.

Annex B.2: Python Code Study 2 Tweet-Level Classification

```
import pandas as pd
import numpy as np
import re
from sklearn.model_selection import train_test_split
from xgboost import XGBClassifier
from sklearn.metrics import classification_report, confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns

# === File Paths (balanced 3.1M per class) ===
politicians_path = '/Users/mathewfeehan/Library/Mobile
Documents/com~apple~CloudDocs/PhD Research/Study 2
Redo/PoliticiansDownsampled3.1Mil.csv'
engaged users_path = '/Users/mathewfeehan/Library/Mobile
Documents/com~apple~CloudDocs/PhD Research/Study 2
Redo/EngagedUsersDownsampled3.1Mil.csv'
randoms_path = '/Users/mathewfeehan/Library/Mobile Documents/com~apple~CloudDocs/PhD
Research/Study 2 Redo/RandomUsersDownsampled3.1Mil.csv'
russian_path = '/Users/mathewfeehan/Library/Mobile Documents/com~apple~CloudDocs/PhD
Research/Study 2 Outputs/partisan_fixed/RussianDatasetStudy2workswithXGBoost.csv'

# === Load datasets ===
politicians   = pd.read_csv(politicians_path, low_memory=True)
engaged users = pd.read_csv(engaged users_path, low_memory=True)
randoms  = pd.read_csv(randoms_path,  low_memory=True)
russian       = pd.read_csv(russian_path,       low_memory=True)

# === Labels & combine ===
politicians['author_type']   = 'politician'
engaged users['author_type'] = 'engaged user'
randoms['author_type']  = 'random'
combined_df = pd.concat([politicians, engaged users, randoms], ignore_index=True)

# === Columns & helpers ===
EMO_COLS = ["fear","anger","trust","surprise","sadness","disgust","joy"]
MFT_COLS = [
   'Loyaltyclassification_result', 'Sanctityclassification_result',
   'Careclassification_result', 'Fairnessclassification_result',
   'Authorityclassification_result'
]

_num_pat = re.compile(r".*?\(([-+]?\d*\.?\d+(?:[eE][-+]?\d+)?)\)")
def extract_score(val):
   if isinstance(val, str):
      m = _num_pat.match(val)
```

```
        if m: return float(m.group(1))
        return 0.0
    return float(val) if pd.notnull(val) else 0.0

def ensure_emotions(df: pd.DataFrame) -> pd.DataFrame:
    emo = pd.DataFrame(index=df.index)
    for c in EMO_COLS:
        emo[c] = pd.to_numeric(df[c], errors='coerce').fillna(0.0) if c in df.columns else 0.0
    return emo[EMO_COLS]

# --- Emotions numeric (features only; NO FILTERING) ---
emotion_df = ensure_emotions(combined_df)
ru_emotions = ensure_emotions(russian)

# --- Class counts BEFORE rebalance ---
print("Counts BEFORE rebalance:")
print(combined_df['author_type'].value_counts())

# === REBALANCE (no filtering applied) ===
TARGET_PER_CLASS = 3_100_000  # cap; will drop to min available if any class has fewer
post_counts = combined_df['author_type'].value_counts()
sample_n = int(min(TARGET_PER_CLASS, post_counts.min()))
print(f"Sampling per class: {sample_n:,} each")

balanced_parts = []
for cls in ['politician','engaged user','random']:
    part = combined_df[combined_df['author_type'] == cls]
    if len(part) > sample_n:
        part = part.sample(n=sample_n, random_state=42)
    balanced_parts.append(part)
combined_df = pd.concat(balanced_parts, ignore_index=True)
emotion_df  = ensure_emotions(combined_df)  # align with sampled rows

print("Counts AFTER rebalance:")
print(combined_df['author_type'].value_counts())

# --- Robust MFT parsing to numeric ---
for col in MFT_COLS:
    combined_df[col] = combined_df[col].apply(extract_score) if col in combined_df.columns
else 0.0
    russian[col]     = russian[col].apply(extract_score)     if col in russian.columns     else 0.0

# --- Partisan feature: log1p(count) ---
for df_ in (combined_df, russian):
    df_["partisan word count"] = pd.to_numeric(df_.get("partisan word count", 0),
errors="coerce").fillna(0.0)
```

```
    df_["partisan_log1p"] = np.log1p(df_["partisan word count"])

# === Features (emotions + MFT + partisan_log1p) ===
FEATURES = EMO_COLS + MFT_COLS + ['partisan_log1p']

# --- Build X/y ---
for c in FEATURES:
    if c not in combined_df.columns:
        combined_df[c] = 0.0
    combined_df[c] = pd.to_numeric(combined_df[c], errors='coerce').fillna(0.0)

label_map = {'politician':0, 'engaged user':1, 'random':2}
combined_df['author_type_numeric'] = combined_df['author_type'].map(label_map)

X_full = pd.concat([emotion_df, combined_df[MFT_COLS + ['partisan_log1p']]], axis=1)
y_full = combined_df['author_type_numeric']

print(f"Total training rows used: {len(X_full):,}")

# --- Split (stratified) ---
X_train, X_test, y_train, y_test = train_test_split(
    X_full, y_full, test_size=0.20, stratify=y_full, random_state=42
)

# --- Z-score MFT on TRAIN ONLY; apply to test & Russian ---
mft_mean = X_train[MFT_COLS].mean()
mft_std  = X_train[MFT_COLS].std().replace(0, 1)
def apply_mft_z(df):
    df[MFT_COLS] = (df[MFT_COLS] - mft_mean) / mft_std
    return df
X_train = apply_mft_z(X_train.copy())
X_test  = apply_mft_z(X_test.copy())

# Prepare Russian features with same columns/order + scaling
for c in EMO_COLS:
    if c not in russian.columns:
        russian[c] = 0.0
    russian[c] = pd.to_numeric(russian[c], errors='coerce').fillna(0.0)
russian_feat = russian[EMO_COLS + MFT_COLS + ['partisan_log1p']].copy()
russian_feat = apply_mft_z(russian_feat)

# === Model (simple, no class weights) ===
model = XGBClassifier(
    objective='multi:softprob',
    eval_metric='mlogloss',
    num_class=3,
```

```
    random_state=42,
    tree_method="hist"
)
print("Training XGBoost (tweet baseline, rebalanced no-filter)…")
model.fit(X_train, y_train)

# === Evaluation ===
y_pred = model.predict(X_test)
print("Classification Report:")
print(classification_report(y_test, y_pred, target_names=['politician','engaged user','random']))

cm = confusion_matrix(y_test, y_pred, labels=[0,1,2])
plt.figure(figsize=(10,7))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
            xticklabels=['politician','engaged user','random'],
            yticklabels=['politician','engaged user','random'])
plt.title("Confusion Matrix — Tweet-Level Baseline (no emotion filter)")
plt.xlabel("Predicted"); plt.ylabel("Actual")
plt.tight_layout(); plt.show()

# === Feature importance ===
imp = pd.Series(model.feature_importances_, index=FEATURES).sort_values(ascending=False)
plt.figure(figsize=(10,6))
sns.barplot(x=imp.values, y=imp.index)
plt.title("Feature Importance — Tweet-Level Baseline (no emotion filter)")
plt.tight_layout(); plt.show()

# === Russian tweet-level inference & chart (fixed colors) ===
inv_map = {v:k for k,v in label_map.items()}
rus_pred = model.predict(russian_feat)
rus_labels = pd.Series([inv_map[int(p)] for p in rus_pred])

order  = ['random','engaged user','politician']  # fixed color order
colors = ['lightblue','lightgreen','salmon']

rus_counts = rus_labels.value_counts().reindex(order, fill_value=0)
rus_pct = (rus_counts / rus_counts.sum()) * 100
print("Russian prediction breakdown (%):")
print(rus_pct.round(2))

plt.figure(figsize=(8,6))
bars = plt.bar(order, rus_pct.values, color=colors)
for b, pct in zip(bars, rus_pct.values):
    plt.text(b.get_x() + b.get_width()/2, pct + 0.8, f'{pct:.1f}%', ha='center')
plt.title("Predicted Author Types — Russian Tweets (no emotion filter)")
plt.ylabel("Percentage of Tweets")
```

```
plt.ylim(0, max(rus_pct.values) + 10)
plt.tight_layout(); plt.show()

# === Tiny diagnostics: partisan log1p in TRAIN by class ===
train_df = combined_df.loc[X_train.index].copy()
train_df['author_type'] = train_df['author_type_numeric'].map(inv_map)
diag_means = train_df.groupby('author_type')['partisan_log1p'].agg(['mean','median']).round(3)
print("partisan_log1p mean/median in TRAIN by class:\n", diag_means)
```

Annex C.1 Study 3 Full Results

Model 1: Baseline Account Classification

The baseline account-level model provides a first test of whether aggregated rhetorical features can distinguish Russian influence accounts from U.S. politicians, politically engaged users, and random users. Trained on all U.S. accounts without downsampling (11,641 total) and evaluated on an 80/20 stratified split, the classifier achieves 0.78 accuracy with macro-F1 = 0.80. As shown in Figure 14, errors are concentrated almost entirely on the engaged–random boundary: engaged user accounts are correctly identified in most cases (recall = 0.79, F1 = 0.76), but 211 engaged accounts are misclassified as random. Random accounts show the mirror pattern (recall = 0.76, F1 = 0.78), with 277 misclassified as engaged. Politician accounts are cleanly separated from both groups (precision = 0.90, recall = 0.85, F1 = 0.87), and when they are misclassified they are almost always labeled as engaged user rather than random. These patterns indicate that, once tweets are aggregated to accounts, moral and partisan profiles yield a stable separation between elites and non-elites, while engaged and random users remain comparatively hard to disentangle. See table 18.

Table 48 Classification Report Baseline Account Model

Class	Precision	Recall	F1-score	Support
politician	0.90	0.85	0.87	131
engaged user	0.73	0.79	0.76	1021
random	0.81	0.76	0.78	1177
Accuracy			0.78	2329
Macro avg	0.81	0.80	0.80	2329
Weighted avg	0.78	0.78	0.78	2329

Feature importance at the account level (Figure 13) is consistent with the tweet-level findings but sharper. Authority is by far the dominant predictor, followed by partisan word count mean, trust, and Loyalty. Disgust, Care, Sanctity, and anger form a second tier, while Fairness and the remaining emotions (joy, fear, sadness, surprise) contribute smaller but non-zero effects. This ranking supports RQ2 and H2 at the account scale: moral framing, especially Authority, with contributions from Care, Loyalty, and related emotions, remains the primary signal distinguishing elites from others, and partisan intensity gains influence when allowed to accumulate across an account's history but does not replace moral structure as the main driver of separability.

When applied to 3,168 Russian accounts, the model predicts 55.5 percent as random, 42.6 percent as engaged user, and only 2.0 percent as politician (Figure 15). Relative to the tweet-level baseline, which placed a higher fraction of Russian tweets in the politician class, the account-level model is more conservative about elite resemblance: very few Russian accounts exhibit the sustained moral and partisan profile characteristic of U.S. politicians. Instead, Russian

accounts cluster overwhelmingly in the engaged–random region of the U.S. rhetorical space. These results address RQ1 and H1 at the account level by showing that Russian IRA accounts resemble non-elite users, predominantly random and engaged users, rather than political elites when rhetoric is aggregated across messages. Combined with the feature-importance pattern, this baseline supports the Study 3 premise that account-level moral and partisan profiles offer a stronger and more stable signal than individual tweets, and it provides the foundation for the subsequent ablation, TF-IDF, and policy-risk models that more directly test RQ2–RQ4 and their associated hypotheses.

Figure 13 Feature Importance Baseline Account Model

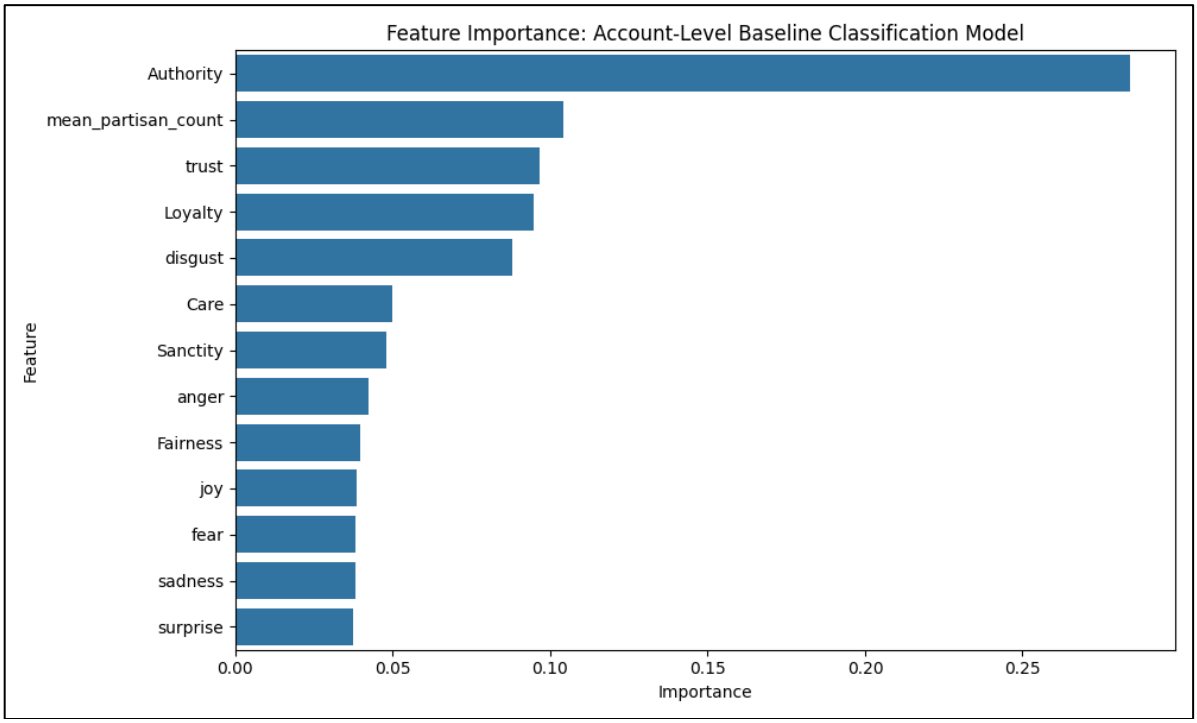


Figure 14 Confusion Matrix Baseline Account Model

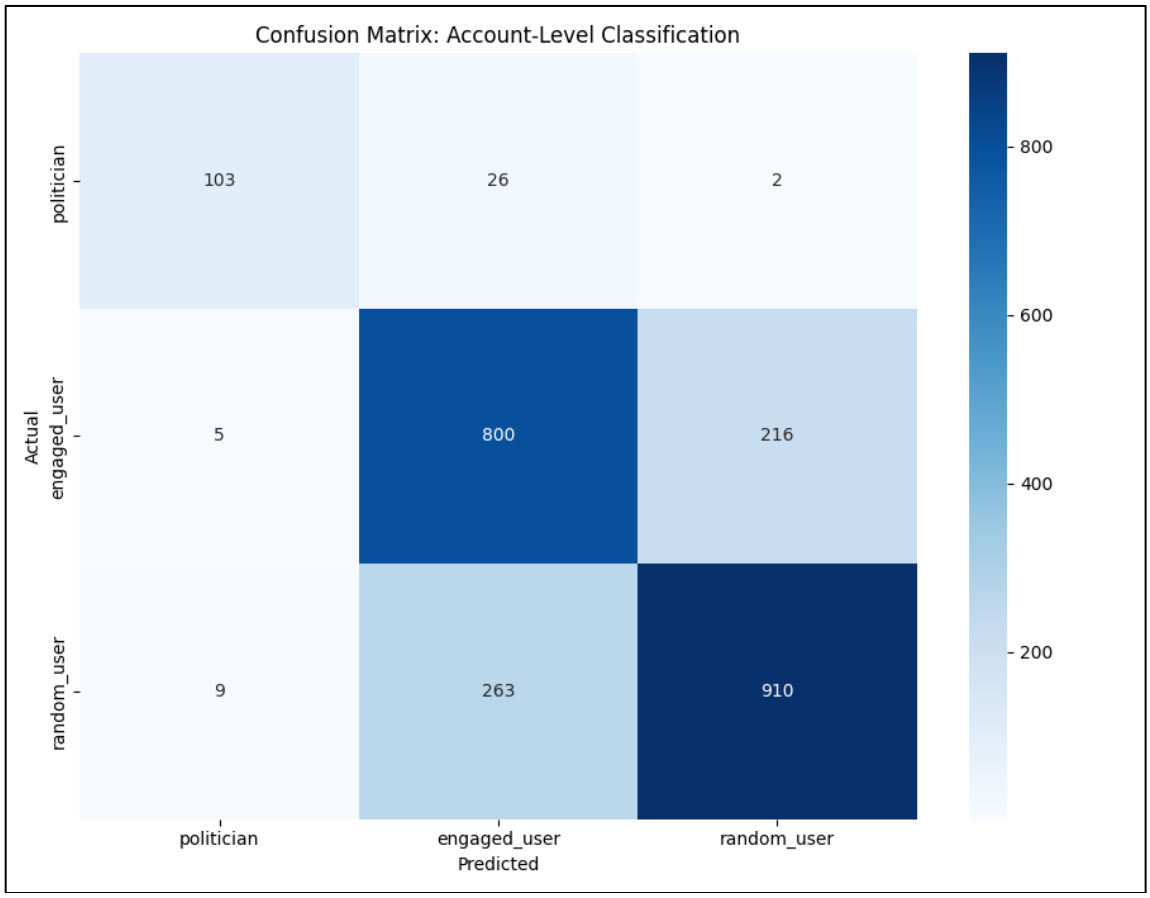
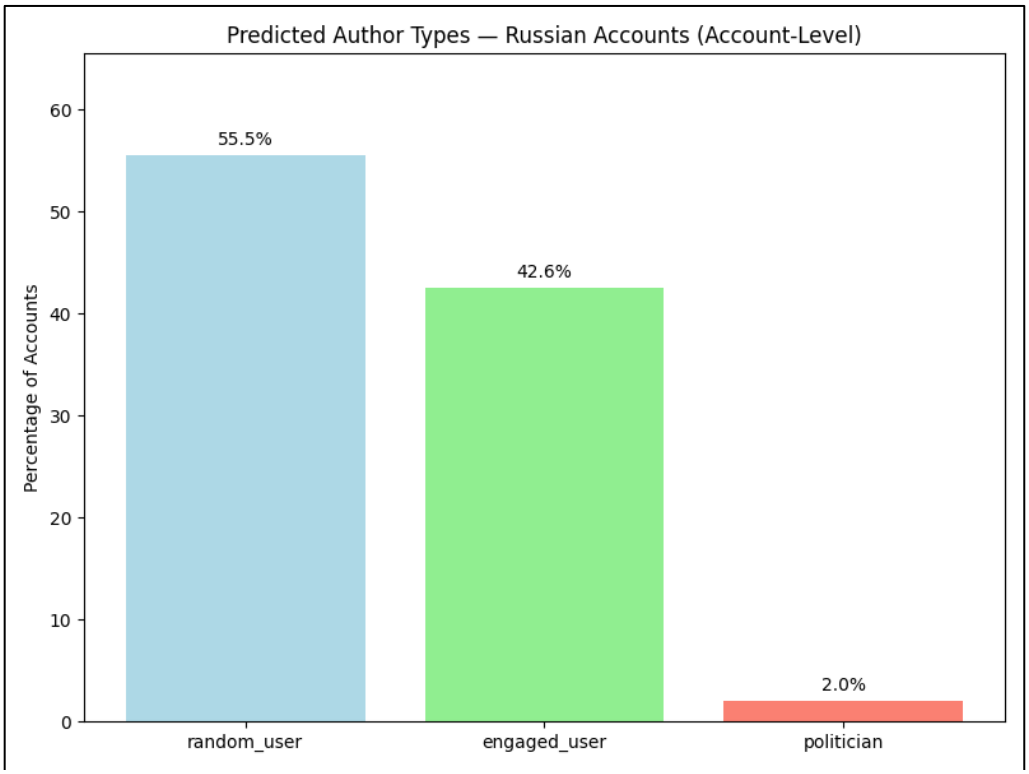


Figure 15 Predicted Author Type Baseline Account Model



Model 2: Account Classification No Partisan Words

The account-level no-partisan model removes partisan word count mean from the baseline feature set and uses only aggregated Moral Foundations and emotion signals to separate politicians, engaged users, and random users. On the held-out U.S. accounts (n = 2,329), overall accuracy is 0.76 with macro-F1 = 0.79, a small decline relative to the baseline (0.78 accuracy, macro-F1 = 0.80). Class metrics remain strong: politician accounts retain high performance (precision = 0.88, recall = 0.83, F1 = 0.85), engaged user accounts show modestly lower precision but similar recall (precision = 0.71, recall = 0.78, F1 = 0.75), and random accounts perform comparably (precision = 0.80, recall = 0.74, F1 = 0.77). See Table 19.

Table 49 No Partisan Words Account Classification Report

Class	Precision	Recall	F1-score	Support
politician	0.88	0.83	0.85	131
engaged user	0.71	0.78	0.75	1021
random	0.80	0.74	0.77	1177
Accuracy			0.76	2329
Macro avg	0.80	0.78	0.79	2329
Weighted avg	0.77	0.76	0.76	2329

Feature importance for the no-partisan model (Figure 16) underscores the centrality of moral framing. Authority is the dominant predictor, followed by trust and Loyalty, with disgust, Fairness, Care, and Sanctity forming a second tier. The remaining emotions (anger, joy, fear, surprise, sadness) contribute smaller but non-zero effects. With partisan intensity removed, the classifier relies entirely on these moral and emotional dimensions and still maintains nearly the same performance as the full baseline. This pattern supports the interpretation that Authority and related moral cues form the core "elite style" signal, while partisan intensity provides additional refinement rather than being the primary driver of account-level separability.

The confusion matrix in Figure 17 again concentrates errors on the engaged–random boundary, with 216 engaged accounts predicted as random and 301 random accounts predicted as engaged. Misclassification involving politicians remains rare and flows almost entirely into the engaged category rather than random. These changes indicate that partisan intensity improves account-

level separability but that most of the structure is preserved when partisan features are removed, consistent with RQ2 and H2.

When applied to the 3,168 Russian accounts, the no-partisan model predicts 62.4 percent as random, 35.9 percent as engaged user, and 1.7 percent as politician (Figure 18). Compared with the baseline account model (55.5 percent random, 42.6 percent engaged, 2.0 percent politician), removing partisan features shifts Russian accounts further toward random users and slightly reduces the already small politician share. Substantively, this strengthens the answer to RQ1 and H1 at the account level: even without any partisan features, Russian IRA accounts continue to resemble non-elite U.S. users, and the model becomes even more conservative about labeling them as politician-like. Moral framing alone is sufficient to keep Russian accounts in the engaged–random band, reinforcing the view that their aggregate rhetorical profile is closer to ordinary or engaged civic participation than to elite political communication.

Figure 16 Feature Importance No Partisan Words Account Model

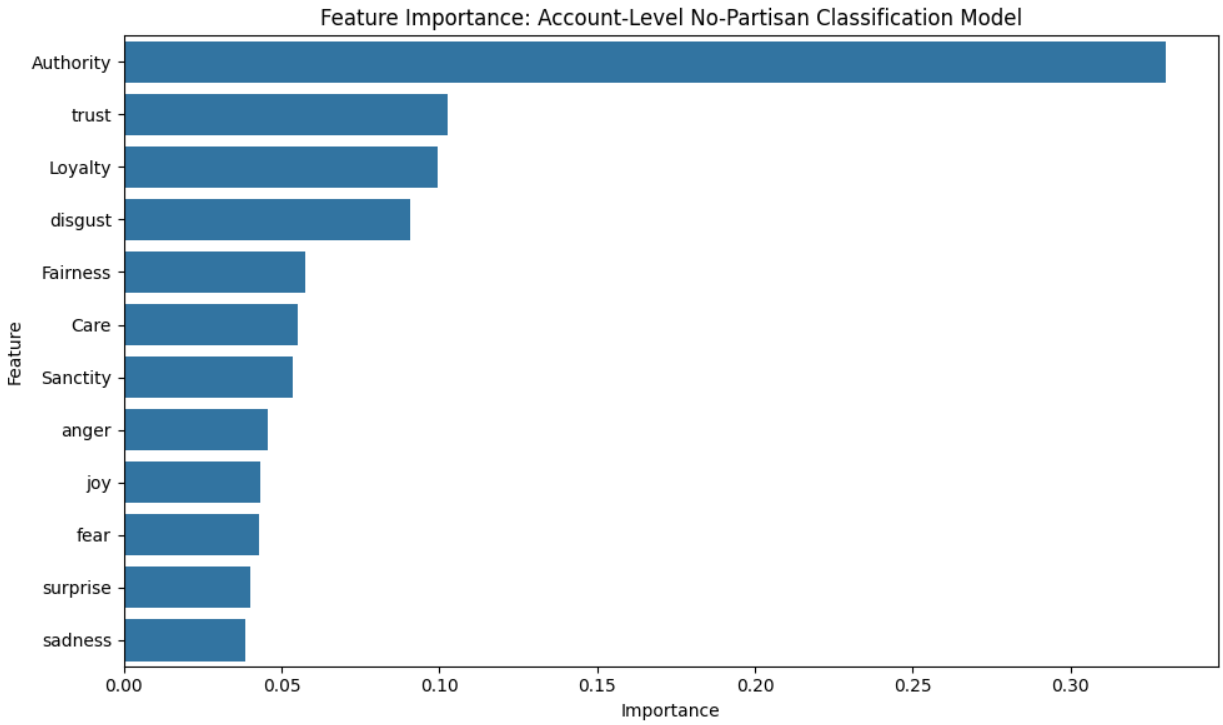


Figure 17 Confusion Matrix No Partisan Words Account Model

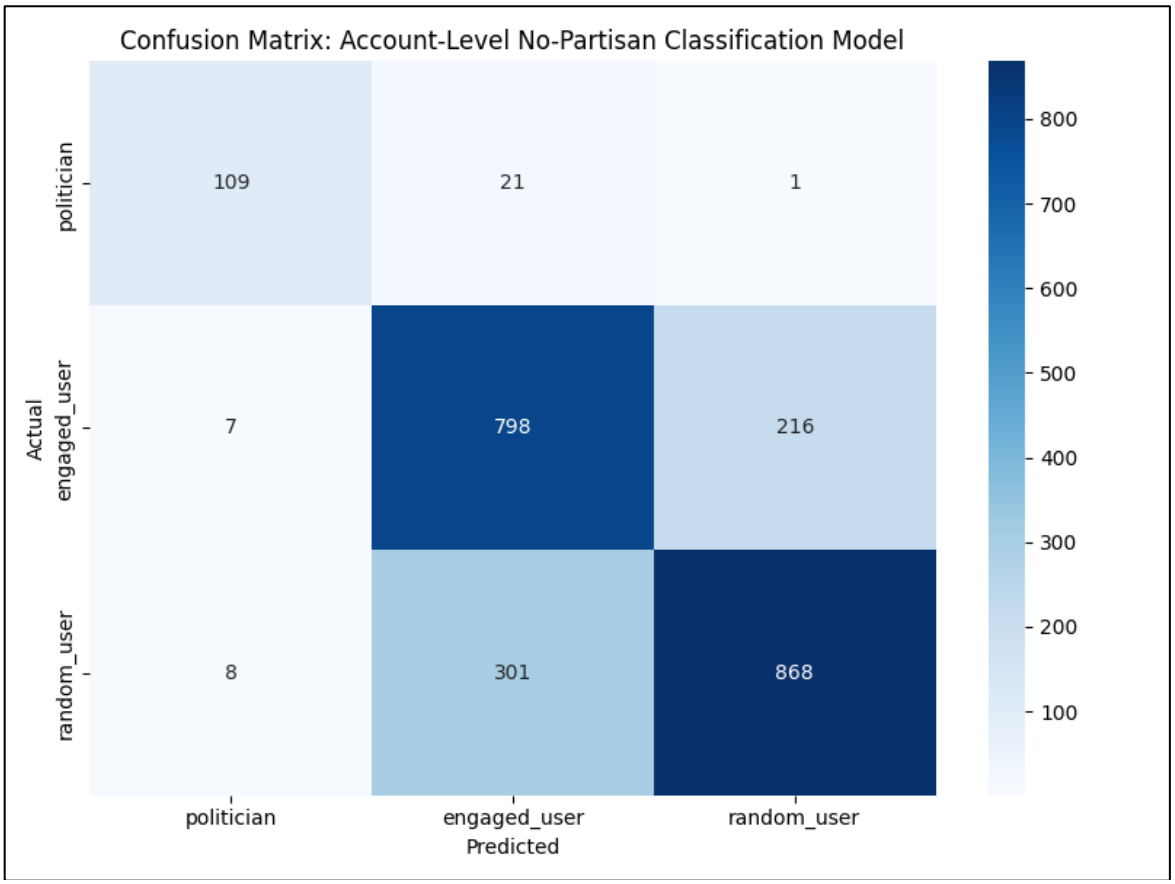
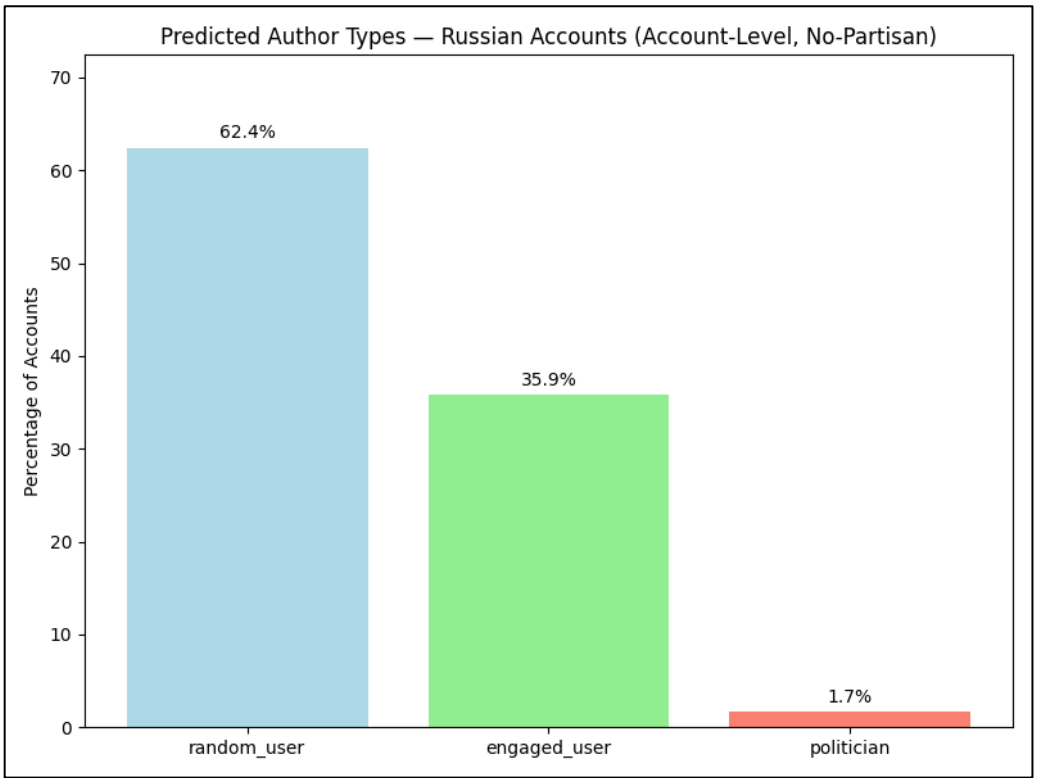


Figure 18 Predicted Author Type No Partisan Words Account Model



Model 3: TF-IDF Account Classification

The account-level TF-IDF model replaces engineered rhetorical features with a bag-of-words representation over each account's concatenated tweet history. For every U.S. account, I join all tweets into a single document, vectorize with a 10,000-term unigram TF-IDF vocabulary (matched to the tweet-level setup), and train a three-class XGBoost classifier on all U.S. accounts without downsampling. As before, an 80/20 stratified split at the account level ensures that each account appears in only one split. This model addresses RQ3 and H3 by testing whether a theory-agnostic lexical baseline lexical representation at the account scale reproduces the structure observed in the rhetorical models and whether it leads to different conclusions about Russian alignment.

Held-out performance improves substantially compared with the rhetorical account baseline. On the U.S. test set (n = 2,329), the TF-IDF model attains 0.83 accuracy with macro-F1 = 0.87, versus 0.78 and 0.80 for the rhetorical model. Politician accounts are classified with very high fidelity (precision = 0.97, recall = 0.92, F1 = 0.95), and both engaged user (precision = 0.79, recall = 0.87, F1 = 0.82) and random accounts (precision = 0.87, recall = 0.80, F1 = 0.83) see clear gains. See Table 20.

Table 50 Classification Report TF-IDF Account Model

Class	Precision	Recall	F1-score	Support
politician	0.97	0.92	0.95	131
engaged user	0.79	0.87	0.82	1021
random	0.87	0.80	0.83	1177
Accuracy			0.83	2329
Macro avg	0.87	0.86	0.87	2329
Weighted avg	0.84	0.83	0.83	2329

The feature-importance profile (Figure 19) makes clear what the TF-IDF model is exploiting. The highest-weight terms are overtly political and divisive, including covid-19, congress, gop, coronavirus, senate, scotus, voters, bipartisan, treason, cnn, and nytimes, along with slurs and culture-war language. In contrast to the more abstract moral structure of the rhetorical models, the TF-IDF model relies heavily on explicit partisan, institutional, and issue-specific vocabulary that characterizes contemporary U.S. political discourse. In terms of RQ3 and H3, this confirms

that the stronger performance arises from direct access to the partisan and issue lexicon rather than from a different latent structure: once the model can see the full political vocabulary, it separates politicians and highly engaged domestic users from random users even more cleanly.

The confusion matrix in Figure 20 shows that misclassifications continue to concentrate on the engaged–random boundary, but the error mass is smaller than in the rhetorical models and mislabeling of politicians is rare. These results demonstrate that, at the account level, theory-agnostic lexical baseline lexical representation is a strictly stronger classifier of U.S. accounts than aggregated moral and partisan features alone.

Crucially, this stronger classifier does not pull Russian accounts toward elites; it pushes them further into the non-elite region. When applied to the 3,168 Russian accounts, the TF-IDF model predicts 86.4 percent as random, 13.6 percent as engaged user, and 0 percent as politician (Figure 21). Compared with the rhetorical account baseline (55.5 percent random, 42.6 percent engaged, 2.0 percent politician) and the no-partisan variant (62.4 percent random, 35.9 percent engaged, 1.7 percent politician), the TF-IDF model sharply increases the random share and eliminates the already small politician fraction. **In other words, the more explicitly political the representation becomes, the less Russian accounts resemble U.S. elites or highly engaged domestic users.**

These outcomes reinforce the core theoretical claims from RQ1–RQ3. First, they show that domestic political actors are defined lexically by a dense partisan and institutional register, while Russian IRA accounts largely avoid that register even when operating at scale across many tweets. Second, they demonstrate that the placement of Russian accounts in the engaged–random band is not an artifact of curated features: a high-capacity TF-IDF model, free to use any lexical

cue, **reaches an even stronger conclusion that Russian accounts look "under-political" relative to U.S. elites and activists.** Finally, they support the interpretation that Russian operators pursue a strategy of rhetorical moderation and camouflage, participating in public-affairs discourse while remaining below the lexical threshold that characterizes domestic political actors. For policy purposes, this implies that account-level lexical models tuned to identify political discourse will preferentially flag U.S. politicians and activists rather than Russian influence accounts, underscoring both the detectability of domestic political style and the difficulty of targeting Russian operations without substantial collateral capture of legitimate speech. The next model addresses this question directly.

Figure 19 Feature Importance TF-IDF Account Model

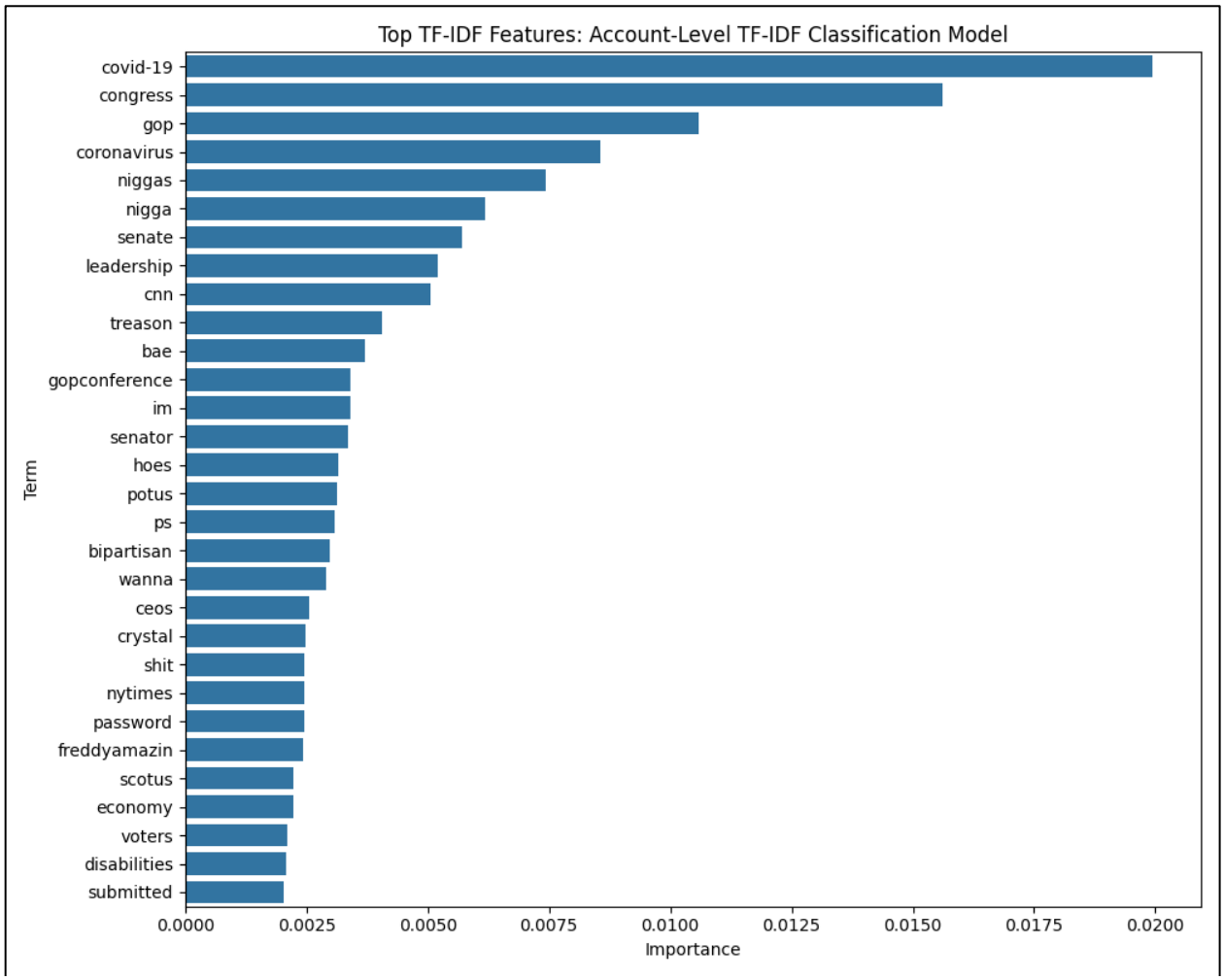


Figure 20 Confusion Matrix TF-IDF Account Model

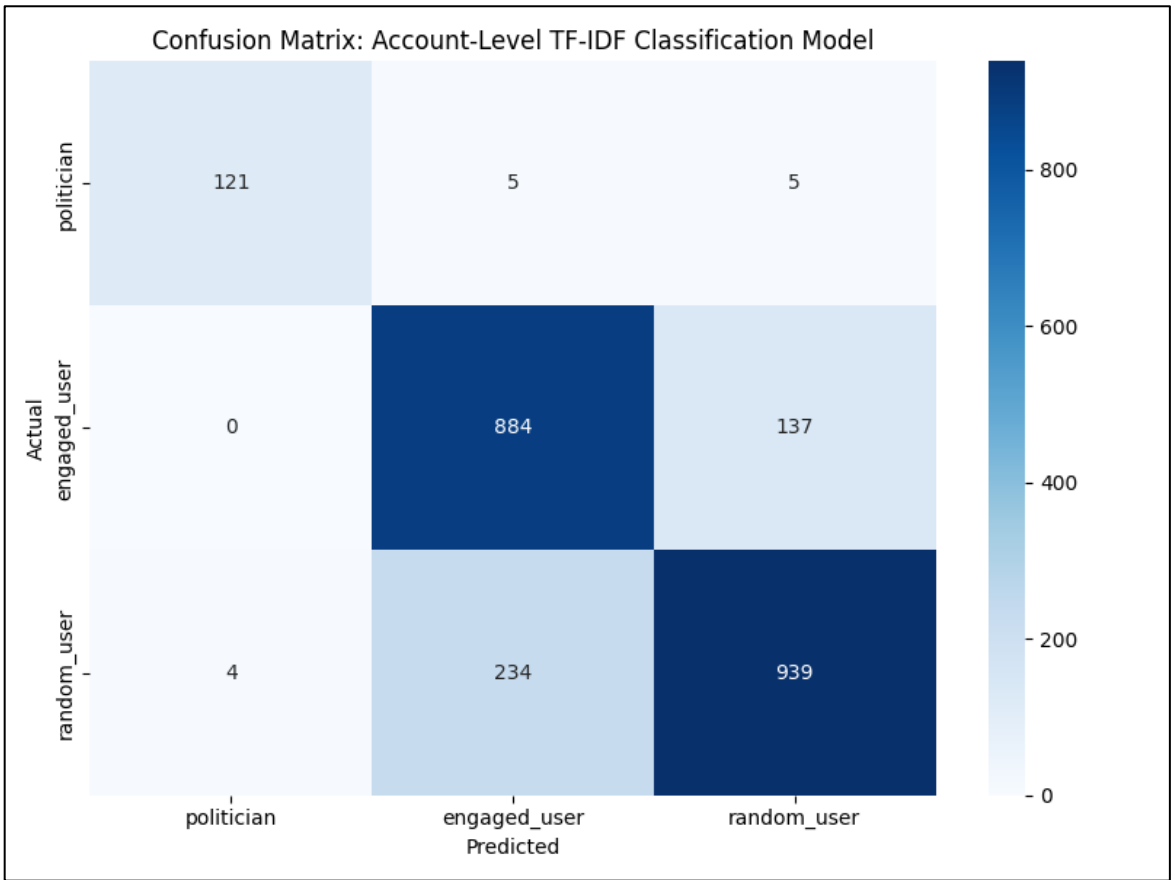
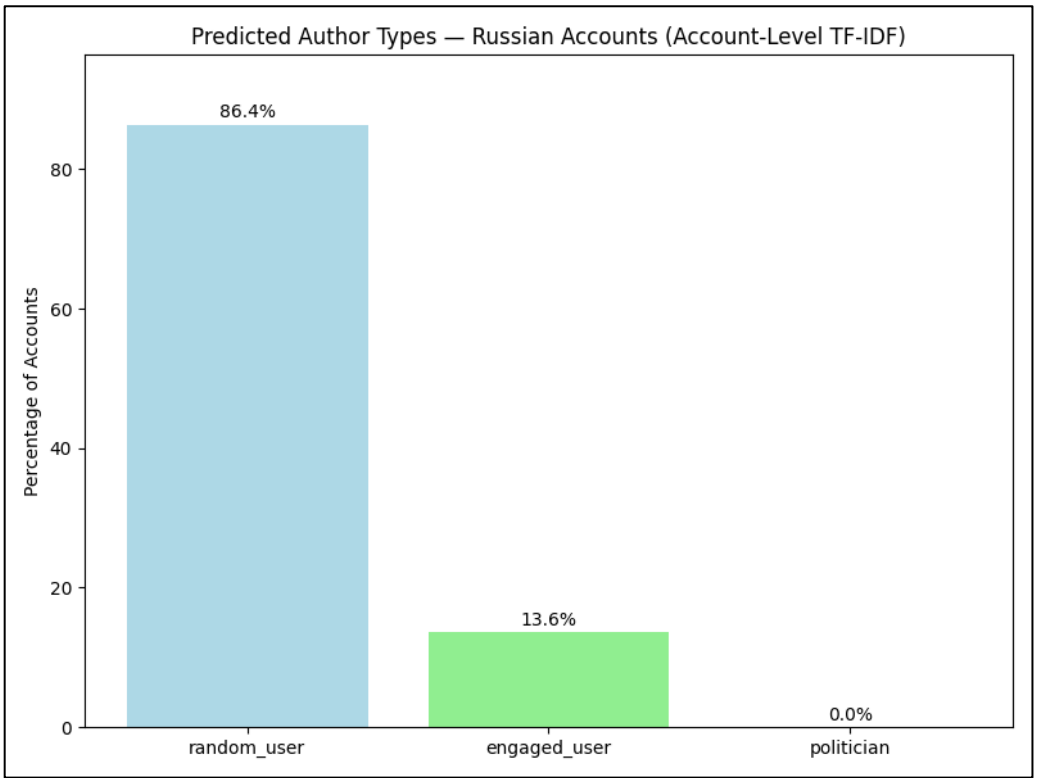


Figure 21 Predicted Author Type TF-IDF Account Model



Model 4 Account-Level Policy-Risk and PPV Analysis

The final set of analyses evaluates whether the account-level baseline model can support operational thresholds for detecting Russian influence while avoiding substantial over-capture of legitimate U.S. speech (RQ5, H5). I use the same account-level baseline classifier described above—percent-present Moral Foundations and NRC emotions plus mean partisan word count per tweet—trained on all U.S. politician, engaged user, and random accounts (N = 11,641) with an 80/20 stratified split. Performance on the held-out U.S. accounts is similar to the earlier baseline: accuracy is 0.77 with macro-F1 = 0.80 (Figure 22; Table 51). Politician accounts remain the easiest to identify (precision = 0.90, recall = 0.84, F1 = 0.87), while engaged user (precision = 0.72, recall = 0.78, F1 = 0.75) and random accounts (precision = 0.80, recall = 0.75, F1 = 0.77) show the familiar engaged–random confusion pattern (Table 51). The confusion matrix illustrates this structure directly. Of 131 politician accounts, 110 are correctly classified as politician, 19 are classified as engaged user, and 2 are classified as random. Of 1,021 engaged-user accounts, 797 are correctly classified, 218 are classified as random, and 6 are classified as politician. Of 1,177 random accounts, 882 are correctly classified as random, 289 are classified as engaged user, and 6 are classified as politician (Figure 22). This provides a realistic foundation for assessing policy risk: the classifier is reasonably accurate on U.S. accounts but far from perfect.

To turn this classifier into a detection score, I compute two derived metrics from the predicted class probabilities for each account. Political-style is defined as the model's predicted probability of "politician" plus its predicted probability of "engaged user," capturing how strongly an account resembles elites or politically engaged users; random-style is simply the predicted

probability of "random," reflecting resemblance to ordinary users. For each metric, I treat Russian accounts as "positives" and U.S. accounts as "negatives" and set thresholds so that approximately 70, 80, or 90 percent of Russian accounts fall above the threshold. I then evaluate, on the held-out U.S. accounts, the resulting false-positive rates (FPRs) for politicians, engaged users, and random users, as well as overall FPR on all U.S. accounts. Finally, assuming Russian accounts constitute 0.1, 0.5, or 1.0 percent of the total user population, I compute positive predictive value (PPV) at each threshold—using PPV = (p_R * TPR) / (p_R * TPR + (1 – p_R) * FPR), where p_R is Russian prevalence, TPR is recall, and FPR is the false positive rate—to estimate the fraction of flagged accounts that would truly be Russian under realistic prevalences.

Political-style scoring is clearly unusable as an intervention rule. At a threshold chosen to capture 70 percent of Russian accounts, every U.S. politician in the held-out sample is flagged (FPR = 100 percent), as are 95 percent of engaged users and roughly half of random users; the overall FPR on U.S. accounts is 73 percent. Pushing capture to 80 or 90 percent drives the overall FPR above 84 percent with essentially all elites and engaged users labeled as suspicious. Under these conditions, PPV remains extremely low: at a prevalence of 0.5 percent Russian accounts, fewer than one in 200 flagged accounts is actually Russian (PPV ≈ 0.5 percent), and at 1.0 percent prevalence PPV remains below 1.1 percent. In substantive terms, any political-style threshold that captures a meaningful share of Russian accounts would function as a broad net over domestic political speech, primarily flagging U.S. politicians and activists rather than foreign influence operations. These results directly support H5's claim that political-style scoring is unsafe at useful recall.

Random-style scoring performs slightly better but remains problematic. At a threshold that captures 70 percent of Russian accounts, FPR for politicians is modest (about 3 percent), but 38 percent of engaged users and 84 percent of random users are flagged, yielding an overall FPR of roughly 59 percent. Thresholds targeting 80 and 90 percent capture drive overall U.S. FPRs to approximately 70 and 83 percent, respectively, with false positives for engaged users and random users approaching saturation. PPV remains low across all thresholds: at 0.5 percent prevalence, PPV ranges from about 0.6 percent at 70 percent capture to roughly 0.6–0.7 percent at 80–90 percent capture; even at 1.0 percent prevalence, PPV stays near 1 percent. Random-style scoring therefore offers, at best, a coarse analytic signal indicating which accounts resemble the broad "random" band in which Russian operations cluster, but it does not yield a viable rule for removal, throttling, or labeling without sweeping in the majority of ordinary U.S. users.

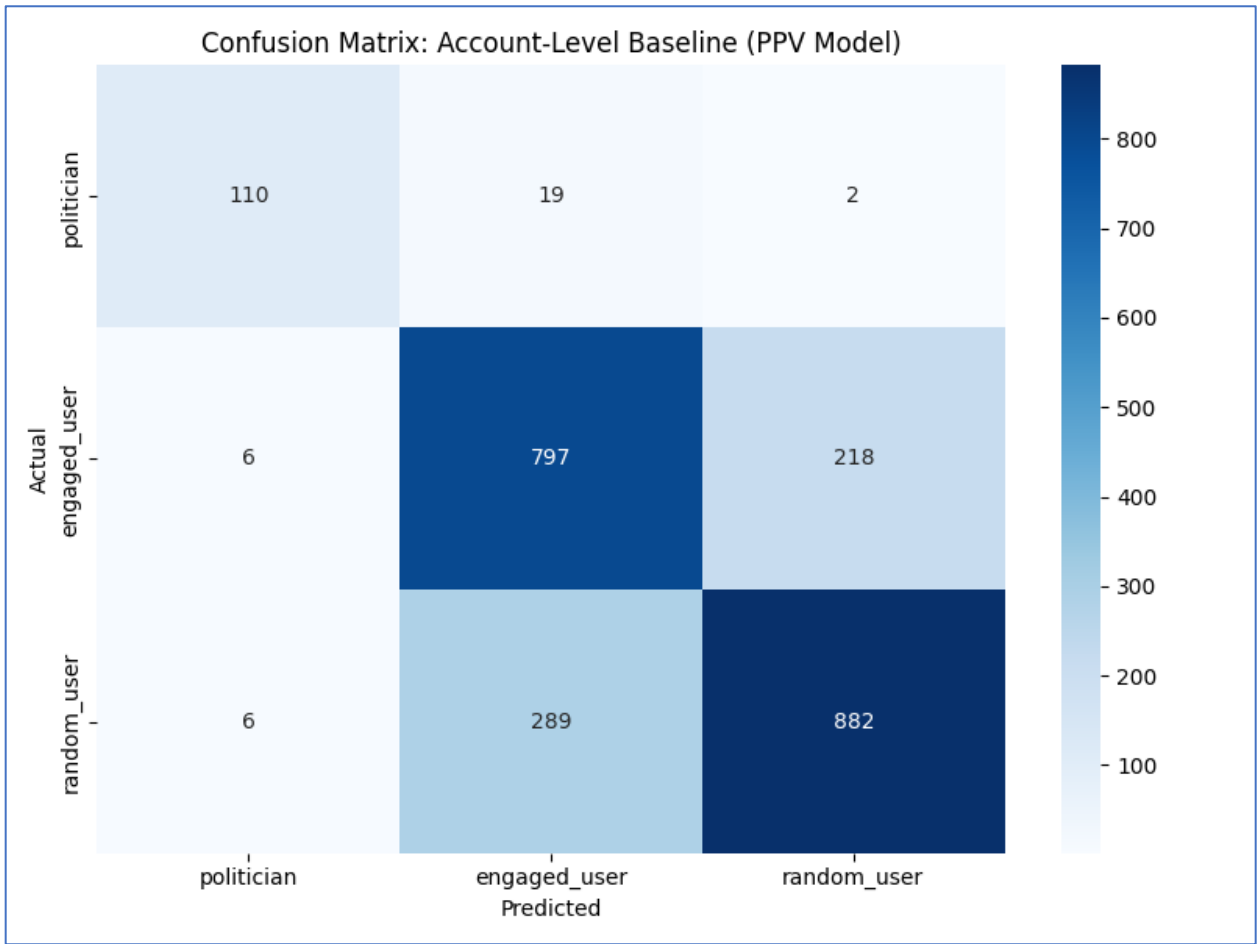
In sum, these policy-risk results provide a clear answer to RQ5 and strong support for H5. My specific rhetorical characterizations at the account level are informative for analysis: they separate politicians from non-elites, show that Russian accounts consistently resemble non-elite users, and align with the theoretical claim that Russian operators occupy an intermediate, engaged-like rhetorical register that is less saturated with partisan and elite cues than domestic politicians but overlaps with politically engaged users. However, when translated into decision thresholds under realistic base rates, the same models perform poorly as enforcement tools. Political-style scoring effectively flags domestic elites and activists rather than Russian accounts, and random-style scoring labels most ordinary users as suspicious while yielding PPV near zero. Study 3 therefore concludes that rhetorical features are valuable for mapping influence operations within the broader discourse, but they are not sufficient on their own to justify

targeted platform interventions without unacceptable collateral effects on legitimate political expression.

Table 51 Classification Report PPV Model

Class	Precision	Recall	F1-score	Support
politician	0.90	0.84	0.87	131
engaged user	0.72	0.78	0.75	1021
random	0.80	0.75	0.77	1177
Accuracy			0.80	2329
Macro avg	0.81	0.79	0.77	2329

Figure 22 Confusion Matrix PPV Model



Annex C.2: Python Code Study 3 Account-Level Classification Python Code

```
import pandas as pd
import re
from sklearn.model_selection import train_test_split
from xgboost import XGBClassifier
from sklearn.metrics import classification_report, confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns

# === File Paths ===
```

```
politicians_path = '/Users/mathewfeehan/Library/Mobile
Documents/com~apple~CloudDocs/PhD Research/Study 2
Outputs/partisan_fixed/PoliticianFullDatasetStudy2.csv'
engaged users_path = '/Users/mathewfeehan/Library/Mobile
Documents/com~apple~CloudDocs/PhD Research/Study 2
Outputs/partisan_fixed/EngagedUsersFullDataset_augmented__20251111_175631.csv'
randoms_path   = '/Users/mathewfeehan/Library/Mobile
Documents/com~apple~CloudDocs/PhD Research/Study 2
Outputs/partisan_fixed/RandomUsersFullDataset_augmented__20251111_182612.csv'
russian_path        = '/Users/mathewfeehan/Library/Mobile
Documents/com~apple~CloudDocs/PhD Research/Study 2
Outputs/partisan_fixed/RussianDatasetStudy2workswithXGBoost.csv'

# === Load tweet-level datasets ===
politicians   = pd.read_csv(politicians_path)
engaged users = pd.read_csv(engaged users_path)
randoms  = pd.read_csv(randoms_path)
russian       = pd.read_csv(russian_path)

# === Label the U.S. training data ===
politicians['author_type']   = 'politician'
engaged users['author_type'] = 'engaged user'
randoms['author_type']  = 'random'

# Combine U.S. tweets
us_tweets = pd.concat([politicians, engaged users, randoms], ignore_index=True)

# === Extract numeric MFT scores from strings like "care (0.87)" ===
def extract_score(val):
    if isinstance(val, str):
        match = re.search(r"\((\d\.\d+)\)", val)
        return float(match.group(1)) if match else 0.0
    return float(val) if pd.notnull(val) else 0.0

mft_cols = [
    'Loyaltyclassification_result',
    'Sanctityclassification_result',
    'Careclassification_result',
    'Fairnessclassification_result',
    'Authorityclassification_result'
]

emotion_cols = ['fear', 'anger', 'trust', 'surprise', 'sadness', 'disgust', 'joy']

# === Pretty labels for feature importance ===
MFT_LABEL_MAP = {
```

```
    'Careclassification_result': 'Care',
    'Fairnessclassification_result': 'Fairness',
    'Loyaltyclassification_result': 'Loyalty',
    'Authorityclassification_result': 'Authority',
    'Sanctityclassification_result': 'Sanctity'
}

def pretty_feature_name(name: str) -> str:
    """
    Clean up feature names for plotting:
    - Strip '_present_pct' suffix.
    - Map MFT columns to simple labels (Care, Authority, etc.).
    - Leave emotion names as 'fear', 'anger', etc.
    - Make partisan word count mean a bit more readable.
    """
    if name.endswith('_present_pct'):
        base = name.replace('_present_pct', '')
        # If it's one of the MFT columns, map to a clean moral label
        if base in MFT_LABEL_MAP:
            return MFT_LABEL_MAP[base]
        # Otherwise it's an emotion like 'fear', 'anger', etc.
        return base
    if name == 'partisan word count mean':
        return 'mean_partisan_count'
    return name

# Apply MFT extraction to U.S. and Russian tweets
for col in mft_cols:
    if col in us_tweets.columns:
        us_tweets[col] = us_tweets[col].apply(extract_score)
    if col in russian.columns:
        russian[col] = russian[col].apply(extract_score)

# Ensure all feature columns are numeric and fill NaNs with 0
all_feature_cols = emotion_cols + mft_cols + ['partisan word count']

for col in all_feature_cols:
    if col in us_tweets.columns:
        us_tweets[col] = pd.to_numeric(us_tweets[col], errors='coerce').fillna(0.0)
    if col in russian.columns:
        russian[col] = pd.to_numeric(russian[col], errors='coerce').fillna(0.0)

# Sanity check: screen_name must exist
if 'screen_name' not in us_tweets.columns or 'screen_name' not in russian.columns:
    raise ValueError("Expected 'screen_name' column in all datasets for account-level
aggregation.")
```

```
# === Account-level aggregation ===
# For each account: percent of tweets with MFT/emotion present (value > 0), and mean partisan word count

def aggregate_account(group: pd.DataFrame) -> pd.Series:
    agg = {}

    # Percent-present for each MFT and emotion (fraction of tweets with value > 0)
    for col in mft_cols + emotion_cols:
        col_vals = group[col]
        present_pct = (col_vals > 0).astype(float).mean() if len(group) > 0 else 0.0
        agg[col + '_present_pct'] = present_pct

    # Mean partisan word count per tweet
    if 'partisan word count' in group.columns:
        agg['partisan word count mean'] = group['partisan word count'].mean()
    else:
        agg['partisan word count mean'] = 0.0

    # Author type (U.S. only; Russian accounts will not have this)
    if 'author_type' in group.columns:
        agg['author_type'] = group['author_type'].iloc[0]

    return pd.Series(agg)

print("Aggregating U.S. tweets to account level...")
us_accounts = us_tweets.groupby('screen_name', as_index=False).apply(aggregate_account)
print(f"Total U.S. accounts: {len(us_accounts):,}")

print("Aggregating Russian tweets to account level...")
rus_accounts = russian.groupby('screen_name', as_index=False).apply(aggregate_account)
print(f"Total Russian accounts: {len(rus_accounts):,}")

# === Define account-level feature set ===
account_features = [c + '_present_pct' for c in (mft_cols + emotion_cols)] + [
    'partisan word count mean'
]

# Ensure no NaNs in aggregated feature columns
for col in account_features:
    us_accounts[col] = pd.to_numeric(us_accounts[col], errors='coerce').fillna(0.0)
    rus_accounts[col] = pd.to_numeric(rus_accounts[col], errors='coerce').fillna(0.0)

# === Prepare X, y for U.S. accounts ===
label_mapping = {'politician': 0, 'engaged user': 1, 'random': 2}
```

```
us_accounts['author_type_numeric'] = us_accounts['author_type'].map(label_mapping)

X = us_accounts[account_features]
y = us_accounts['author_type_numeric']

print(f"Total U.S. training accounts used: {len(X):,}")

# === Train/test split at the account level ===
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, stratify=y, random_state=42
)

# === Train XGBoost account-level model ===
model = XGBClassifier(
    objective='multi:softprob',
    eval_metric='mlogloss',
    num_class=3,
    random_state=42,
    tree_method='hist'
)

print("Training XGBoost (account-level baseline)...")
model.fit(X_train, y_train)

# === Evaluate on held-out U.S. accounts ===
y_pred = model.predict(X_test)
print("Classification Report (account-level):")
print(classification_report(y_test, y_pred, target_names=list(label_mapping.keys())))

conf_matrix = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(8, 6))
sns.heatmap(
    conf_matrix,
    annot=True,
    fmt='d',
    cmap='Blues',
    xticklabels=list(label_mapping.keys()),
    yticklabels=list(label_mapping.keys())
)
plt.title("Confusion Matrix: Account-Level Baseline Classification Model")
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.tight_layout()
plt.show()

# === Feature importance (account-level) ===
```

```
importances = pd.DataFrame({
    'Feature': account_features,
    'Importance': model.feature_importances_
}).sort_values(by='Importance', ascending=False)

# Add pretty labels for plotting
importances['FeatureLabel'] = importances['Feature'].apply(pretty_feature_name)

plt.figure(figsize=(10, 6))
sns.barplot(x='Importance', y='FeatureLabel', data=importances)
plt.title("Feature Importance: Account-Level Baseline Classification Model")
plt.xlabel("Importance")
plt.ylabel("Feature")
plt.tight_layout()
plt.show()


# === Predict Russian accounts ===
rus_X = rus_accounts[account_features]
rus_pred = model.predict(rus_X)

# Map numeric predictions back to labels
inv_label_map = {v: k for k, v in label_mapping.items()}
rus_accounts['predicted_author_type'] = [inv_label_map[int(p)] for p in rus_pred]

# Breakdown of Russian accounts by predicted type
rus_counts = rus_accounts['predicted_author_type'].value_counts()
rus_pct = (rus_counts / len(rus_accounts)) * 100

print("Russian account-level prediction breakdown (%):")
print(rus_pct)

# --- Bar chart with fixed colors: random=blue, engaged=green, politician=salmon ---
order  = ['random', 'engaged user', 'politician']
colors = ['lightblue',    'lightgreen',   'salmon']

rus_counts_ordered = rus_counts.reindex(order, fill_value=0)
rus_pct_ordered = (rus_counts_ordered / rus_counts_ordered.sum()) * 100

plt.figure(figsize=(8, 6))
bars = plt.bar(order, rus_pct_ordered.values, color=colors)
for b, pct in zip(bars, rus_pct_ordered.values):
    plt.text(b.get_x() + b.get_width() / 2, pct + 1, f'{pct:.1f}%', ha='center')
plt.title("Predicted Author Types — Russian Accounts (Account-Level)")
plt.ylabel("Percentage of Accounts")
plt.ylim(0, max(rus_pct_ordered.values) + 10)
```

```
plt.tight_layout()
plt.show()

# === Save Russian account-level predictions ===
save_path = '/Users/mathewfeehan/Library/Mobile Documents/com~apple~CloudDocs/PhD
Research/IRA/Study 2 and 3 Final Data/predicted_russian_accounts_account_baseline.csv'
rus_accounts.to_csv(save_path, index=False)
print(f"Account-level predictions saved to: {save_path}")
```

Annex D: Study 3 PPV Python Code

```
import pandas as pd
import re
import numpy as np
from sklearn.model_selection import train_test_split
from xgboost import XGBClassifier
from sklearn.metrics import classification_report, confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns

# === File Paths (full partisan_fixed datasets; update if needed) ===
politicians_path = '/Users/mathewfeehan/Library/Mobile
Documents/com~apple~CloudDocs/PhD Research/Study 2
Outputs/partisan_fixed/PoliticianFullDatasetStudy2.csv'
```

```
engaged users_path = '/Users/mathewfeehan/Library/Mobile
Documents/com~apple~CloudDocs/PhD Research/Study 2
Outputs/partisan_fixed/EngagedUsersFullDataset_augmented__20251111_175631.csv'
randoms_path   = '/Users/mathewfeehan/Library/Mobile
Documents/com~apple~CloudDocs/PhD Research/Study 2
Outputs/partisan_fixed/RandomUsersFullDataset_augmented__20251111_182612.csv'
russian_path       = '/Users/mathewfeehan/Library/Mobile
Documents/com~apple~CloudDocs/PhD Research/Study 2
Outputs/partisan_fixed/RussianDatasetStudy2workswithXGBoost.csv'

# Output (new file, will not overwrite prior outputs)
scores_output_path  = '/Users/mathewfeehan/Library/Mobile
Documents/com~apple~CloudDocs/PhD Research/IRA/Study 2 and 3 Final
Data/account_baseline_with_style_scores.csv'

# === Load tweet-level datasets ===
politicians   = pd.read_csv(politicians_path, low_memory=True)
engaged users = pd.read_csv(engaged users_path, low_memory=True)
randoms  = pd.read_csv(randoms_path,  low_memory=True)
russian      = pd.read_csv(russian_path,      low_memory=True)

# === Label the U.S. training data ===
politicians['author_type']   = 'politician'
engaged users['author_type'] = 'engaged user'
randoms['author_type']  = 'random'

us_tweets = pd.concat([politicians, engaged users, randoms], ignore_index=True)

# Drop missing text
us_tweets = us_tweets.dropna(subset=['tweet_text'])
russian   = russian.dropna(subset=['tweet_text'])

# Ensure screen_name exists
if 'screen_name' not in us_tweets.columns or 'screen_name' not in russian.columns:
   raise ValueError("Expected 'screen_name' column in all datasets for account-level
aggregation.")

# === Extract numeric MFT scores from strings like "care (0.87)" ===
def extract_score(val):
   if isinstance(val, str):
      match = re.search(r"\((\d\.\d+)\)", val)
      return float(match.group(1)) if match else 0.0
   return float(val) if pd.notnull(val) else 0.0

mft_cols = [
   'Loyaltyclassification_result',
```

```
    'Sanctityclassification_result',
    'Careclassification_result',
    'Fairnessclassification_result',
    'Authorityclassification_result'
]

emotion_cols = ['fear', 'anger', 'trust', 'surprise', 'sadness', 'disgust', 'joy']

# Apply MFT extraction
for col in mft_cols:
    if col in us_tweets.columns:
        us_tweets[col] = us_tweets[col].apply(extract_score)
    if col in russian.columns:
        russian[col] = russian[col].apply(extract_score)

# Force numeric, fill NaNs
all_feature_cols = emotion_cols + mft_cols + ['partisan word count']
for col in all_feature_cols:
    if col in us_tweets.columns:
        us_tweets[col] = pd.to_numeric(us_tweets[col], errors='coerce').fillna(0.0)
    if col in russian.columns:
        russian[col] = pd.to_numeric(russian[col], errors='coerce').fillna(0.0)

# === Account-level aggregation ===
def aggregate_account(group: pd.DataFrame) -> pd.Series:
    agg = {}
    # Percent-present for each MFT and emotion
    for col in mft_cols + emotion_cols:
        col_vals = group[col]
        present_pct = (col_vals > 0).astype(float).mean() if len(group) > 0 else 0.0
        agg[col + '_present_pct'] = present_pct
    # Mean partisan word count per tweet
    if 'partisan word count' in group.columns:
        agg['partisan word count mean'] = group['partisan word count'].mean()
    else:
        agg['partisan word count mean'] = 0.0
    # Author type (U.S. only)
    if 'author_type' in group.columns:
        agg['author_type'] = group['author_type'].iloc[0]
    return pd.Series(agg)

print("Aggregating U.S. tweets to account level...")
us_accounts = us_tweets.groupby('screen_name', as_index=False).apply(aggregate_account)
print(f"Total U.S. accounts: {len(us_accounts):,}")

print("Aggregating Russian tweets to account level...")
```

```
rus_accounts = russian.groupby('screen_name', as_index=False).apply(aggregate_account)
print(f"Total Russian accounts: {len(rus_accounts):,}")

# === Feature set ===
account_features = [c + '_present_pct' for c in (mft_cols + emotion_cols)] + [
    'partisan word count mean'
]

for col in account_features:
    us_accounts[col]  = pd.to_numeric(us_accounts[col], errors='coerce').fillna(0.0)
    rus_accounts[col] = pd.to_numeric(rus_accounts[col], errors='coerce').fillna(0.0)

# === Labels ===
label_mapping = {'politician': 0, 'engaged user': 1, 'random': 2}
us_accounts['author_type_numeric'] = us_accounts['author_type'].map(label_mapping)

X = us_accounts[account_features]
y = us_accounts['author_type_numeric']

print(f"Total U.S. training accounts used (baseline for PPV): {len(X):,}")

# === Train/test split ===
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, stratify=y, random_state=42
)

# === Train XGBoost baseline account model ===
model = XGBClassifier(
    objective='multi:softprob',
    eval_metric='mlogloss',
    num_class=3,
    random_state=42,
    tree_method='hist',
    n_estimators=300
)

print("Training XGBoost (account-level baseline for PPV)...")
model.fit(X_train, y_train)

# === Evaluate on held-out U.S. accounts ===
y_pred = model.predict(X_test)
print("Classification Report (account-level baseline for PPV):")
print(classification_report(y_test, y_pred, target_names=list(label_mapping.keys())))

conf_matrix = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(8, 6))
```

```
sns.heatmap(
    conf_matrix,
    annot=True,
    fmt='d',
    cmap='Blues',
    xticklabels=list(label_mapping.keys()),
    yticklabels=list(label_mapping.keys())
)
plt.title("Confusion Matrix: Account-Level Baseline (PPV Model)")
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.tight_layout()
plt.show()

# === Probabilities and style scores ===
# U.S. test set
proba_test = model.predict_proba(X_test)
# Russian accounts (all positives for PPV purposes)
proba_rus  = model.predict_proba(rus_accounts[account_features])

# Index mapping
idx_politician = label_mapping['politician']
idx_engaged    = label_mapping['engaged user']
idx_random     = label_mapping['random']

# Style scores
political_score_test = proba_test[:, idx_politician] + proba_test[:, idx_engaged]
random_score_test    = proba_test[:, idx_random]

political_score_rus = proba_rus[:, idx_politician] + proba_rus[:, idx_engaged]
random_score_rus    = proba_rus[:, idx_random]

# Add scores into dataframes for potential export
us_test_accounts = us_accounts.iloc[X_test.index].copy()
us_test_accounts['political_score'] = political_score_test
us_test_accounts['random_score']    = random_score_test

rus_accounts['political_score'] = political_score_rus
rus_accounts['random_score']    = random_score_rus

# === Thresholds and PPV computation ===
targets = [0.70, 0.80, 0.90]
prevalences = [0.001, 0.005, 0.01]  # 0.1%, 0.5%, 1%

def compute_metrics(style_name, rus_scores, test_scores, y_test):
    """
```

```
    style_name: 'political' or 'random'
    rus_scores: scores for Russian accounts (positives)
    test_scores: scores for U.S. test accounts (negatives)
    y_test: numeric labels for U.S. test accounts
    """
    print(f"\n=== {style_name.capitalize()}-style scoring ===")
    for target in targets:
        # threshold so that ~target of Russians are above it
        thr = np.quantile(rus_scores, 1 - target)
        rus_flag = rus_scores >= thr
        capture = rus_flag.mean()

        test_flag = test_scores >= thr
        # FPR per class (politician, engaged, random) on U.S. accounts
        fpr_politician = test_flag[(y_test == idx_politician)].mean()
        fpr_engaged    = test_flag[(y_test == idx_engaged)].mean()
        fpr_random     = test_flag[(y_test == idx_random)].mean()
        fpr_all        = test_flag.mean()

        print(f"\nTarget capture: {int(target*100)}% (threshold={thr:.3f})")
        print(f"  Actual Russian capture: {capture*100:.1f}%")
        print(f"  FPR politicians: {fpr_politician*100:.2f}%")
        print(f"  FPR engaged:     {fpr_engaged*100:.2f}%")
        print(f"  FPR random:      {fpr_random*100:.2f}%")
        print(f"  FPR all U.S.:    {fpr_all*100:.2f}%")

        # PPV at different prevalences
        for p in prevalences:
            tpr = capture
            fpr = fpr_all
            # PPV = (p * TPR) / (p * TPR + (1 - p) * FPR)
            numerator   = p * tpr
            denominator = p * tpr + (1 - p) * fpr
            ppv = numerator / denominator if denominator > 0 else 0.0
            print(f"    PPV at prevalence {p*100:.1f}%: {ppv*100:.2f}%")

# Political-style scoring
compute_metrics('political', political_score_rus, political_score_test, y_test.values)

# Random-style scoring
compute_metrics('random', random_score_rus, random_score_test, y_test.values)

# === Save scores for further analysis ===
full_accounts_with_scores = pd.concat(
    [us_accounts.assign(dataset='US'), rus_accounts.assign(dataset='Russian')],
    ignore_index=True
```

```
)

full_accounts_with_scores.to_csv(scores_output_path, index=False)
print(f"\nSaved account-level scores to: {scores_output_path}")
```

Annex E: Google Trends of Partisanship (2009–2020)

year	liberal	conservative	republican	democrat	right-wing	left-wing	progressive	alt-right	obama	biden	clinton
2009	1	1	1	1	0	0	2	0	13.83	0.17	2.25
2010	1	0.67	1.33	1	0	0	2	0	6.5	0.08	2.58
2011	1	0.42	1.67	1	0	0	2.08	0	6.25	0	2.92
2012	1	0.83	2.58	1.25	0	0	2.17	0	11.92	0.5	3.17
2013	1	0.33	1.08	1	0	0	2.08	0	6.5	0.08	3
2014	1	0.5	1.08	1	0	0	2.75	0	5.5	0	3.08
2015	1	0.75	2.58	1	0	0	2.75	0	5.25	0.5	4.5
2016	1.08	1	4.42	1.33	0	0	2.75	0	7.17	0.5	15
2017	1	1	1.08	1	0	0	2.83	0	5.08	0.33	4.17
2018	1	1	1.25	1.17	0	0	3	0	3.58	0.17	3.75
2019	1	1	1.17	1	0	0	3	0	2.83	1.75	3.33
2020	1	1	2.58	1.67	0	0	2.67	0	4.08	13.42	3.5

year	democrats	libertarian	independent	obamacare	tea party	birth certificate	socialist	czar	crooked hillary	lock her up	make america great again
2009	0	0	2.96	0	0.85	1	0	0	0	0	0
2010	0.08	0	2.62	0	1.44	1	0	0	0	0	0
2011	0	0	2.71	0	0.85	1.08	0	0	0	0	0
2012	0.25	0.17	2.54	0.76	0.42	1	0	0	0	0	0
2013	0	0	2.45	2.12	0.34	1	0	0	0	0	0
2014	0	0	2.29	1.44	0.08	1	0	0	0	0	0
2015	0	0	2.2	0.68	0	1	0	0	0	0	0
2016	0.42	0.34	2.2	0.42	0	1	0.17	0	0	0	0
2017	0.33	0	2.12	0.68	0	1	0	0	0	0	0
2018	0.67	0	2.12	0	0	1	0	0	0	0	0
2019	1	0	2.12	0	0	1	0.08	0	0	0	0
2020	1.08	0.17	2.37	0	0	0.92	0.17	0	0	0	0

year	deep state	hands up don't shoot	i can't breathe	kaepernick	take a knee	racial divide	black lives matter	blm	all lives matter	blue lives matter	antifa

2009	0	0	0	0	0	0	0	0	0	0	0
2010	0	0	0	0	0	0	0	0	0	0	0
2011	0	0	0	0	0	0	0	0	0	0	0
2012	0	0	0	0.17	0	0	0	0	0	0	0
2013	0	0	0	0.76	0	0	0	0	0	0	0
2014	0	0	0	0.51	0	0	0	0	0	0	0
2015	0	0	0	0.08	0	0	0.08	0	0	0	0
2016	0	0	0	0.85	0	0	0.34	0.08	0	0	0
2017	0	0	0	0.76	0	0	0	0	0	0	0.59
2018	0	0	0	0.42	0	0	0	0	0	0	0.08
2019	0	0	0	0.51	0	0	0	0	0	0	0.17
2020	0	0	0	0.17	0	0	1.35	1.27	0.08	0	1.18
year	ukraine	crimea	nato	putin	syria	china	iran	isis	terrorism	collusion	russia hoax
2009	0.08	0	0	0	0	7.62	1.27	0	0	0	0
2010	0	0	0	0	0	8.04	1.02	0	0	0	0
2011	0	0	0.08	0	0	7.96	1.1	0.17	0	0	0
2012	0.08	0	0.17	0	0.85	8.04	1.35	0.25	0	0	0
2013	0.08	0	0	0.08	1.95	7.7	1.02	0.51	0.08	0	0
2014	2.12	0.25	0.08	0.59	0.59	7.7	1.1	3.05	0	0	0
2015	0.85	0	0.08	0.51	0.76	8.29	1.44	4.15	0.17	0	0
2016	0.34	0	0.08	0.42	1.02	8.29	1.1	1.86	0	0	0
2017	0.08	0	0.17	0.59	1.1	8.38	1.02	1.1	0.08	0	0
2018	0.17	0	0.08	0.34	0.93	8.04	1.18	0.34	0	0	0
2019	0.59	0	0.08	0	0.25	8.38	1.44	0.25	0	0	0
2020	0.25	0	0	0.17	0	11.59	2.54	0	0	0	0

year	fox news	hannity	maddow	sjw	npc	red pill	blue pill	parkland shooting	march for our lives	kavanaugh	me too
2009	7.53	0.25	0	0	0	0	0	0	0	0	0
2010	6.52	0.08	0	0	0	0	0	0	0	0	0
2011	8.04	0	0	0	0	0	0	0	0	0	0
2012	9.22	0.08	0.34	0	0	0	0	0	0	0	0
2013	10.66	0	0	0	0	0	0	0	0	0	0
2014	9.39	0	0	0	0	0	0	0	0	0	0
2015	9.22	0	0	0	0	0	0	0	0	0	0
2016	9.9	0.17	0.08	0	0	0	0	0	0	0	0.34
2017	11.51	0.25	0.17	0	0	0	0	0	0	0	0.08
2018	12.61	0.08	0	0	0.08	0	0	0.08	0.17	2.7	0.34
2019	11.76	0	0	0	0	0	0	0	0	0.08	0

2020	17.1	0.25	0.08	0	0	0	0	0	0	0	0
year	al-baghdadi	radical islam	caliphate	trade war	tariffs	racism	paris agreement	climate change hoax	epa	hillary	mcconnel
2009	0	0	0	0	0	0.08	0	0	1.02	0.92	0
2010	0	0	0	0	0	0.17	0	0	1.02	0.83	0
2011	0	0	0	0	0	0	0	0	1.02	0.75	0
2012	0	0	0	0	0	0.08	0	0	1.02	0.83	0
2013	0	0	0	0	0	0	0	0	1.02	0.92	0
2014	0	0	0	0	0	0.34	0	0	0.93	1	0.08
2015	0	0	0	0	0	0.51	0	0	0.93	2	0
2016	0	0	0	0	0	0.85	0	0	0.59	11.17	0
2017	0	0	0	0	0	0.76	0.08	0	0.51	1.58	0.08
2018	0	0	0	0	0.17	0.68	0	0	0.34	1	0.08
2019	0	0	0	0	0.08	0.51	0	0	0	1	0.42
2020	0	0	0	0	0	1.1	0	0	0.08	1	0.83

year	email server	benghazi	trump derangement syndrome	tds	covfefe	alternative facts	republicans
2009	0	0	0	0	0	0	0
2010	0	0	0	0	0	0	0.08
2011	0	0	0	0	0	0	0
2012	0	0.17	0	0	0	0	0.33
2013	0	0.08	0	0	0	0	0.08
2014	0	0.08	0	0	0	0	0.08
2015	0	0.08	0	0	0	0	0
2016	0	0.51	0	0	0	0	0.42
2017	0	0	0	0	0.42	0.08	0.33
2018	0	0	0	0	0	0	0.25
2019	0	0	0	0	0	0	0.25
2020	0	0	0	0	0	0	1

Bibliography

47 U.S.C. § 230 (2018).

A10 Networks. (n.d.). *What is carrier-grade NAT (CGN/CGNAT)?* A10 Networks. https://www.a10networks.com/glossary/what-is-carrier-grade-nat-cgn-cgnat/

Abdulhai, M., Serapio-Garcia, G., Crepy, C., Valter, D., Canny, J., & Jaques, N. (2023). *Moral foundations of large language models*. arXiv preprint arXiv:2310.15337.

Adler-Bell, S., Ford, M., Nwanevu, O., Segall, P., Al-Agba, N., The New Republic, & Cohen, R. M. (2021, December 2). *The radical young intellectuals who want to take over the American right*. *The New Republic*. Retrieved from https://newrepublic.com/article/164408/young-intellectuals-illiberal-revolution-conservatism

Agence France-Presse. (2025, April 28). *US anti-disinformation guardrails fall in Trump's first 100 days*. France 24. https://www.france24.com/en/live-news/20250428-us-anti-disinformation-guardrails-fall-in-trump-s-first-100-days

Agarwal, N. (2013, December 3). FCC loosens restrictions on foreign investments in broadcasting. The Regulatory Review. https://www.theregreview.org/2013/12/03/03-agarwal-fcc-loosens-restrictions/

Aggarwal, A., Rajadesingan, A., & Kumar, S. (2018). *The follower count fallacy: Detecting Twitter users with manipulated follower count*. *arXiv preprint arXiv:1802.03625*. https://arxiv.org/abs/1802.03625

Alizadeh, M. (2020). Replication data for: Content-based features predict social media influence operations [Data set]. Harvard Dataverse. https://doi.org/10.7910/DVN/PMY0KF

Alizadeh, M., Shapiro, J. N., Buntain, C., & Tucker, J. A. (2020). Content-based features predict social media influence operations. *Science Advances, 6*(30), eabb5824. https://doi.org/10.1126/sciadv.abb5824

Alliance for Securing Democracy. (2021, September 8). *Pro-China COVID-19 misinformation campaign targets U.S.* German Marshall Fund. https://securingdemocracy.gmfus.org/incident/pro-china-covid-19-misinformation-campaign-targets-us/

Allyn, B., & Folkenflik, D. (2025, September 26). Trump's TikTok deal payment criticized as "shakedown scheme" by experts. *LAist / NPR.* Retrieved from https://laist.com/news/trumps-tiktok-deal-payment-criticized-as-shakedown-scheme-by-experts

Amjadi, E., & John, R. S. (2024). *All is fair in love and war: Moral foundations in English-language tweets during the first 36 weeks of conflict between Ukraine and Russia*. Proceedings of the 57th Hawaii International Conference on System Sciences.

Associated Press. (2025, November 23). *Big changes to the agency charged with securing elections lead to midterm worries*. AP News. https://apnews.com/article/6d18799c6c5fdd1bc001544b2dca12bf

Aragon, J. (2025, November 26). X (Twitter) now shows your account's country/region. Privacy Guides. https://www.privacyguides.org/news/2025/11/26/x-twitter-now-shows-your-accounts-country-region/

Bailey, M. (2019). *NRCLex* [Python library]. GitHub. https://github.com/metalcorebear/NRCLex

Baker, P. (2020, November 17). Trump fires head of election cybersecurity who debunked voter fraud claims. *The New York Times*. https://www.nytimes.com/2020/11/17/us/politics/trump-fires-christopher-krebs.html

BakerHostetler. (2025, February 11). *Requiem for Task Force KleptoCapture: The future of Russia sanctions enforcement*. https://www.bakerlaw.com/insights/requiem-for-task-force-kleptocapture-the-future-of-russia-sanctions-enforcement/

Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Artemova, K., Tutubalina, E., & Chowell, G. (2023). A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration [Data set]. In Epidemiologia (Version 162, Vol. 2, Number 3, pp. 315–324). Zenodo. https://doi.org/10.5281/zenodo.7834392

Benson, P. J., Brannon, V. C., & Lampe, J. R. (2024). *Regulation of TikTok under the Protecting Americans from Foreign Adversary Controlled Applications Act: Analysis of selected legal issues* (CRS Legal Sidebar LSB11127). Congressional Research Service. Retrieved from https://www.congress.gov/crs-product/LSB11127

Benson, P. J., & Brannon, V. C. (2025). *TikTok Inc. v. Garland: Supreme Court rejects challenge to TikTok divestiture law* (CRS Legal Sidebar LSB11261). Congressional Research Service. Retrieved from https://www.congress.gov/crs-product/LSB11261

Bernstein, D., Fishman, P. J., & Weiss, B. (2022, March 7). *DOJ establishes Task Force KleptoCapture to enforce Russia-related sanctions* [Client Alert]. Greenberg Traurig LLP. Retrieved from https://www.gtlaw.com/en/insights/2022/3/doj-establishes-task-force-kleptocapture-sanctions-export-restrictions-russias-invasion-ukraine

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2018). An ideological asymmetry in the diffusion of moralized content on social media among political leaders. *Proceedings of the National Academy of Sciences, 116*(7), 2274–2279. https://www.researchgate.net/publication/329951229_An_Ideological_Asymmetry_in_the_Diffusion_of_Moralized_Content_on_Social_Media_Among_Political_Leaders

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, 114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bretl, B. L., & Hansen, D. M. (2022). An exploration of the structure of moral intuitions in early adolescence. Cognitive Development, 64, 101248. https://doi.org/10.1016/j.cogdev.2022.101248

Business Insider. (2025, November 25). *X erupts after the platform reveals the locations where accounts are based*. Business Insider. https://www.businessinsider.com/x-shows-locations-users-about-this-account-2025-11

Brumfiel, G., Allyn, B., & Simon, S. (2025, September 27). Experts say Trump's TikTok deal payment is a shakedown. *KUOW / NPR.* Retrieved from https://www.kuow.org/stories/experts-say-trump-s-tiktok-deal-payment-is-a-shakedown

Business Insider. (2025, November). X erupts after the platform reveals the locations where accounts are based. Business Insider. https://www.businessinsider.com/x-shows-locations-users-about-this-account-2025-11

Cadwalladr, C., & Graham-Harrison, E. (2018a, March 17). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

Cadwalladr, C., & Graham-Harrison, E. (2018b, March 25). The Cambridge Analytica files: The story so far. *The Guardian*. https://www.theguardian.com/news/series/cambridge-analytica-files

CBS News. (2025, January 22). *DHS terminates all advisory committees, ending investigation into Chinese-linked telecom hack*. https://www.cbsnews.com/news/dhs-terminates-all-advisory-committees-ends-investigation-chinese-linked-telecom-hack-salt-typhoon/

CBS News. (2025, May 16). *GOP congressman confirms Hegseth ordered pause in cyber operations against Russia, despite Pentagon denial.* CBS News. https://www.cbsnews.com/news/gop-congressman-don-bacon-hegseth-ordered-pause-cyber-operations-against-russia/

Cambridge Core. (2021). *A New Order: The Digital Services Act and Consumer Protection*. Retrieved from https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/new-order-the-digital-services-act-and-consumer-protection

Center for Internet Security. (n.d.). *EI-ISAC.* Retrieved from https://www.cisecurity.org/ei-isac

Center for Internet Security. (2025, April 29). *The Center for Internet Security's essential guide to election security.* Retrieved from https://docs.cisecurity.org/essentialguide

Chen, S., & Yang, X. (2021). *The Rise of User Profiling in Social Media: Review, Challenges, and Future Directions*. Springer.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chesney, R. (2019, March 22). What a U.S. operation against Russian trolls predicts about escalation in cyberspace. *War on the Rocks.* Retrieved from https://warontherocks.com/2019/03/what-a-u-s-operation-against-russian-trolls-predicts-about-escalation-in-cyberspace/

Chun, S., Holowczak, R., Dharan, K., Wang, R., Basu, S., & Geller, J. (2019). Detecting political bias trolls in Twitter data. In Proceedings of the 15th International Conference on Web Information Systems and Technologies (WEBIST 2019) (pp. 334–342). SciTePress. https://doi.org/10.5220/0008350303340342

Cinelli, M., Cresci, S., Quattrociocchi, W., Tesconi, M., & Zola, P. (2022). Coordinated inauthentic behavior and information spreading on Twitter. *Decision Support Systems, 160*, 113819. https://doi.org/10.1016/j.dss.2022.113819

Common Cause. (2025, February 6). *Elections endangered by DOJ plan to close Foreign Influence Task Force (FITF)*. https://www.commoncause.org/press/elections-endangered-by-doj-plan-to-close-foreign-influence-task-force-fitf/

Congress.gov. (2023). *S.686 - RESTRICT Act*. Retrieved from https://www.congress.gov/bill/118th-congress/senate-bill/686

Congressional Research Service. (2024, January 4). *Section 230: A brief overview* (IF12584). https://crsreports.congress.gov/product/pdf/IF/IF12584

Congressional Research Service. (2024, December 26). *Termination of the State Department's Global Engagement Center (GEC)* (IN12475). https://crsreports.congress.gov/product/pdf/IN/IN12475

Congressional Research Service. (2019). *The designation of election systems as critical infrastructure* (CRS In Focus IF10677). Congressional Research Service. Retrieved from https://www.congress.gov/crs-product/IF10677

Congressional Research Service. (2024). *TikTok: Frequently asked questions and issues for Congress* (CRS Report R48023). Congressional Research Service. Retrieved from https://www.everycrsreport.com/reports/R48023.html

Congressional Research Service. (2023, July 28). State regulation of foreign ownership of U.S. land (LSB11013). https://www.congress.gov/crs-product/LSB11013

Congressional Research Service. (2025, January 22). Supreme Court rejects challenge to TikTok divestiture law (LSB11261). https://www.congress.gov/crs-product/LSB11261

Connell, M., & Vogler, S. (2017). Russia's approach to cyber warfare. Center for Naval Analyses (CNA). https://www.cna.org/analyses/2017/russias-approach-to-cyber-warfare

Cortez Masto, C., & Rounds, M. (2025, March 6). Cortez Masto leads bipartisan legislation to ban foreign adversaries from buying American farmland [Press release]. https://www.cortezmasto.senate.gov/news/press-releases/cortez-masto-leads-bipartisan-legislation-to-ban-foreign-adversaries-from-buying-american-farmland/

Cloudflare. (2022, May 12). *Magic NAT: Everywhere, unbounded, and lower cost.* Cloudflare Blog. https://blog.cloudflare.com/magic-nat/

Cloudflare. (2025, October 29). *One IP address, many users: Detecting CGN to reduce collateral damage.* Cloudflare Blog. https://blog.cloudflare.com/detecting-cgn-to-reduce-collateral-damage/

CNN. (2022, February 5). *Joe Rogan apologizes for using racial slur after India Arie shares video*. CNN. Retrieved from https://www.cnn.com/2022/02/05/media/joe-rogan-racial-slur-apology-india-arie/index.html

Curry, O. S. (2019, November 13). *What's wrong with moral foundations theory (and how to get moral psychology right)*. Behavioral Scientist. https://behavioralscientist.org/whats-wrong-with-moral-foundations-theory-and-how-to-get-moral-psychology-right/

Cybersecurity and Infrastructure Security Agency. (2020, November 12). *Joint statement from elections infrastructure government coordinating council & the election infrastructure sector coordinating executive committees on the 2020 election.* CISA. https://www.cisa.gov/news-events/news/joint-statement-elections-infrastructure-government-coordinating-council-election

Darczewska, J. (2014). *The anatomy of Russian information warfare: The Crimean operation, a case study*. Centre for Eastern Studies.

Digital Element. (n.d.). *Guide to IP geolocation.* Digital Element. https://www.digitalelement.com/resources/whitepapers/guide-to-ip-geolocation/

Digital Element. (2025). *Digital Element's state of the IP address report.* Digital Element. https://www.digitalelement.com/resources/whitepapers/state-of-the-ip-address-report/

Dizard, W. P. Jr., Brown, K. L., & Funseth, R. L. (2022). *Inventing Public Diplomacy: The Story of the U.S. Information Agency*. Boulder, CO: Lynne Rienner Publishers.

Douek, E. (2023). *The law and politics of online content moderation.* Brookings Institution. (See also Douek's broader work on systems thinking in content moderation: https://harvardlawreview.org/print/vol-136/content-moderation-as-systems-thinking/).

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ... & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Electronic Frontier Foundation. (n.d.). *Section 230 of the Communications Decency Act*. https://www.eff.org/issues/cda230

Electronic Frontier Foundation. (2025, August 4). Data brokers are ignoring privacy law. We deserve better. https://www.eff.org/deeplinks/2025/08/data-brokers-are-ignoring-privacy-law-we-deserve-better

EU Code of Practice on Disinformation. (2024). Transparency Centre (reports and documentation). https://disinfocode.eu/

EU Digital Services Act. (2025). *Data access for researchers under the Digital Services Act*. European Commission. https://digital-strategy.ec.europa.eu/en/news/commission-facilitates-data-access-researchers-under-digital-services-act

EU DisinfoLab. (2022). User-guide to the EU Digital Services Act. https://www.disinfo.eu/wp-content/uploads/2022/11/20221020_DSAUserGuide_Final.pdf

European Commission. (2018). *EU Code of Practice on Disinformation*. Retrieved from https://ec.europa.eu/digital-strategy/en/policies/code-practice-disinformation

European Commission. (2020). *Digital Services Act: Ensuring a Safe and Accountable Online Environment*. Retrieved from https://ec.europa.eu/digital-strategy/en/policies/digital-services-act

European Commission. (2022). *The Digital Services Act: Ensuring a safer and more accountable online environment.* European Commission. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act

European Commission. (2022). *The Digital Services Act package*. https://ec.europa.eu/digital-strategy/our-policies/digital-services-act_en

European Commission. (2022). *The Digital Services Act package*. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

European Commission. (2025, July 3). *FAQs: DSA data access for researchers.* Algorithmic Transparency. https://algorithmic-transparency.ec.europa.eu/news/faqs-dsa-data-access-researchers-2025-07-03_en

European Commission. (2025, September 24). *How the Digital Services Act enhances transparency online*. https://digital-strategy.ec.europa.eu/en/policies/dsa-brings-transparency

European Journal of Risk Regulation. (2021). *A new order: The Digital Services Act and consumer protection*. Cambridge University Press. Retrieved from https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/new-order-the-digital-services-act-and-consumer-protection/8E34BA8A209C61C42A1E7ADB6BB904B1

European Parliament. (2021). *Digital Services Act: A New Set of Rules for Digital Platforms*. Retrieved from https://www.europarl.europa.eu/news/en/press-room/20210119IPR96204/digital-services-act-ensuring-accountability-in-the-online-environment

European Parliament and Council of the European Union. (2022). Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act). https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng

Ezzeddine, F., Luceri, L., Ayoub, O., Sbeity, I., Nogara, G., Ferrara, E., & Giordano, S. (2022). *Exposing influence campaigns in the age of LLMs: A behavioral-based AI approach to detecting state-sponsored trolls* (arXiv:2210.08786). arXiv. https://doi.org/10.48550/arXiv.2210.08786

Feinberg, M., & Willer, R. (2015). From gulf to bridge: When do moral arguments facilitate political influence? Personality and Social Psychology Bulletin, 41(12), 1665–1681. https://doi.org/10.1177/0146167215607842

Federal Bureau of Investigation (FBI). (2016, May 18). Aldrich Ames. FBI. Retrieved from https://www.fbi.gov/history/famous-cases/aldrich-ames

Federal Bureau of Investigation (FBI). (2016, July 20). Ana Montes: Cuban spy. FBI. Retrieved from https://www.fbi.gov/history/famous-cases/ana-montes-cuba-spy

Federal Communications Commission. (2015). *Protecting and promoting the open Internet* (Report and Order on Remand, Declaratory Ruling, and Order), 30 FCC Rcd. 5601. Available at: https://www.federalregister.gov/documents/2015/04/13/2015-07841/protecting-and-promoting-the-open-internet

Federal Communications Commission. (2018). *Restoring Internet Freedom* (Declaratory Ruling, Report and Order, and Order), 33 FCC Rcd. 311. Available at: https://www.federalregister.gov/documents/2018/02/22/2018-03464/restoring-internet-freedom and FCC summary at: https://www.fcc.gov/document/fcc-releases-restoring-internet-freedom-order

Fennis, B. M., & Stroebe, W. (2021). The psychology of advertising (3rd ed.). Routledge. https://www.routledge.com/The-Psychology-of-Advertising/Fennis-Stroebe/p/book/9780367346355

FireEye. (2013). *APT1: Exposing one of China's cyber espionage units*. Report formerly available at www.fireeye.com; now archived or unavailable following company acquisition.

FireEye. (2020). *Fake News and Influence Operations: China, Iran, Russia*. Report formerly available at www.fireeye.com; now archived or unavailable following company acquisition.

Flew, T. (2021). *Regulating Platforms in the Age of Disinformation: Lessons from the EU's Digital Services Act*. Journal of Information Policy, 11(2), 35-49.

Freedom House. (2022). *Beijing's global media influence: United States.* Retrieved from https://freedomhouse.org/report/beijing-global-media-influence/2022/authoritarian-expansion-power-democratic-resilience

Galeotti, M. (2016). *Hybrid war or gibridnaya voina? Getting Russia's non-linear military challenge right*. NATO Defense College.

German Marshall Fund of the United States. (2025, October 15). *The EU's Digital Markets Act and Digital Services Act.* GMF. https://www.gmfus.org/news/eus-digital-markets-act-and-digital-services-act

Giles, K. (2016). *Handbook of Russian information warfare*. NATO Defense College.

Global Engagement Center. (2024). *Report to Congress on an assessment of the Global Engagement Center's operations* (Department of State Report 006006). U.S. Department of State. Retrieved from https://www.state.gov/wp-content/uploads/2025/08/Report-Global-Engagement-006006-Accessible-07.29.2025.pdf

Goldman, I. (2024). North Korea's "New DPRK" YouTube channel: New public diplomacy attempt or international propaganda? A case study. *Humanities and Social Sciences Communications, 11*, Article 34. https://doi.org/10.1057/s41599-024-03406-6

Golovchenko, Y., Buntain, C., Eady, G., Brown, M. A., & Tucker, J. A. (2020). Cross-platform state propaganda: Russian trolls on Twitter and YouTube during the 2016 U.S. presidential election. *The International Journal of Press/Politics, 25*(3), 357–389. https://doi.org/10.1177/1940161220912682

Government disinformation board succumbs to disinformation. (2022). The Free Speech Project, Georgetown University. Retrieved from https://freespeechproject.georgetown.edu/government-disinformation-board-succumbs-to-disinformation/

Guardian. (2025, November 23). Many prominent Maga personalities on X are based outside US, new tool reveals. *The Guardian*. https://www.theguardian.com/us-news/2025/nov/23/rightwing-influencers-outside-us-x-twitter-tool

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). *Moral Foundations Theory: The pragmatic validity of moral pluralism*. Advances in Experimental Social Psychology, 47, 55-130.

Haidt, J. (2012). The righteous mind: Why good people are divided by politics and religion. Pantheon Books.

Halil, D. (2025). Regulating pressing systemic risks – but not too soon? *Internet Policy Review, 14*(2). (preprint version via SSRN). https://policyreview.info/pdf/policyreview-2025-2-2010.pdf

Harwell, D., & Nakashima, E. (2025, November 28). The U.S. has been cutting cyber defenses as AI boosts attacks. *The Washington Post*. https://www.washingtonpost.com/technology/2025/11/28/cyberdefense-ai-cisa/

Henderson, M., et al. (2021). *Identity, Advertising, and Algorithmic Targeting: Or How (Not) to Target Your "Ideal User"*. MIT.

Holland & Knight LLP. (2025). *Trump's 2025 executive orders chart.* Holland & Knight. Retrieved from https://www.hklaw.com/en/general-pages/trumps-2025-executive-orders-chart

H.R. 5736, 112th Cong. (2012). *Smith–Mundt Modernization Act of 2012*. https://www.congress.gov/bill/112th-congress/house-bill/5736

Hultquist, J. (2021, September 8). Global Chinese influence campaign sought to spark protests in US. *Breaking Defense.* https://breakingdefense.com/2021/09/mandiant-details-global-chinese-influence-campaign-that-sought-to-spark-protests-in-us/

Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Iyengar et al. (2019): Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science, 22*, 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

Iyer, R., Koleva, S., Graham, J., Ditto, P., & Haidt, J. (2012). Understanding libertarian morality: the psychological dispositions of self-identified libertarians. *PloS one*, *7*(8), e42366. https://doi.org/10.1371/journal.pone.0042366
Martin v. City of Struthers**,** 319 U.S. 141 (1943).

Johnson, D. B. (2024, December 23). State Department's disinformation office to close after funding nixed in NDAA. *CyberScoop*. https://cyberscoop.com/state-departments-disinformation-office-to-close-after-funding-nixed-in-ndaa/

Johnson, D. B. (2025, February 6). DOJ disbands foreign influence task force, limits scope of FARA prosecutions. *CyberScoop*. https://cyberscoop.com/doj-disbands-foreign-influence-task-force/

Johnson, J. (2017, January 6). *Statement by Secretary Jeh Johnson on the designation of election infrastructure as a critical infrastructure subsector.* U.S. Department of Homeland Security. Retrieved from https://www.dhs.gov/archive/news/2017/01/06/statement-secretary-johnson-designation-election-infrastructure-critical

Johnson, N. (2024, December 26). *Termination of the State Department's Global Engagement Center and disappearing guardrails against disinformation* (CRS Insight IN12475). Congressional Research Service. Retrieved from https://www.congress.gov/crs-product/IN12475

Jones, D. (2025, January 22). DHS disbands existing advisory board memberships, raising questions about CSRB. *Cybersecurity Dive*. https://www.cybersecuritydive.com/news/dhs-disbands-advisory-board-csrb/737976/

Karnowski, S., & Smyth, J. C. (2025, November 23). Big changes to the agency charged with securing elections lead to midterm worries. *AP News*. https://apnews.com/article/election-security-cisa-2026-secretaries-state-midterms-6d18799c6c5fdd1bc001544b2dca12bf

Kleindienst v. Mandel, 408 U.S. 753 (1972).

Lamont v. Postmaster General, 381 U.S. 301 (1965).

LeadDigital. (2020, February 24). Paid media & privacy: What's changing & what it means for marketers. LeadDigital. https://www.leaddigital.com/blog/paid-media-privacy-whats-changing-what-it-means-for-marketers/

Lewinski, D., & Hasan, M. R. (2021). *Russian troll account classification with Twitter and Facebook data* (arXiv:2101.05983). arXiv. https://doi.org/10.48550/arXiv.2101.05983 (arXiv)

Lintner, B. (2025, October 6). How China waged an infowar against U.S. interests in the Philippines. *Reuters.* https://www.reuters.com/world/asia-pacific/how-china-waged-an-infowar-against-us-interests-philippines-2025-10-06/

Linvill, D. L., Boatwright, B. C., Grant, W. J., & Warren, P. L. (2019). The Russians are hacking my brain!: Investigating Russia's Internet Research Agency Twitter tactics during the 2016 United States presidential campaign. *Computers in Human Behavior, 99*, 292–300. https://doi.org/10.1016/j.chb.2019.05.027

Loh, M. (2023, February 11). *Meet North Korea's YouTubers, who go fishing, eat ice cream, and gush about "Harry Potter" to fuel Kim Jong Un's propaganda machine*. Business Insider. https://www.businessinsider.com/north-korea-youtubers-propaganda-sogwang-kim-jong-un-2023-2

Mandiant. (2013). *APT1: Exposing one of China's cyber espionage units*. Mandiant Intelligence Center Report. *(Original publication by Mandiant prior to Google.inc acquisition)*

Mandiant Threat Intelligence. (2021, September 8). *Pro-PRC influence campaign expands to dozens of social media platforms, websites, and forums.* Google Cloud Security Blog. https://cloud.google.com/blog/topics/threat-intelligence/pro-prc-influence-campaign-expands-dozens-social-media-platforms-websites-and-forums

Mann, T. E., & Ornstein, N. J. (2012). It's even worse than it looks: How the American constitutional system collided with the new politics of extremism. Basic Books.

Martin v. City of Struthers, 319 U.S. 141 (1943).

Marwick, A., & boyd, danah. (2011). To See and Be Seen: Celebrity Practice on Twitter. *Convergence*, *17*(2), 139-158. https://doi.org/10.1177/1354856510394539

Mathew, S. K., Alex, D., Deshpande, N., Sharma, R., Arya, A., & Balendra, D. P. (2024). *Multilevel troll classification of Twitter data using machine learning techniques. International Journal of Computer Theory and Engineering, 16*(1), 21–28. https://doi.org/10.7763/IJCTE.2024.V16.1350 (journal.kpu.go.id)

MaxMind. (n.d.-a). *Geolocation accuracy.* MaxMind. https://support.maxmind.com/geoip-data-correction/geolocation-accuracy

MaxMind. (n.d.-b). *GeoIP2 City accuracy comparison for locations worldwide.* MaxMind. https://dev.maxmind.com/geoip/docs/databases/city-and-country#accuracy

Media Monitors**.** (2020, March 2). *Audience demographic variations are specific to genre and even individual podcasts*. Media Monitors. Retrieved from https://www.mediamonitors.com/audience-demographic-variations-specific-to-genre/

Meta. (2023, August 29). *Raising online defenses through transparency and accountability.* Meta Newsroom. https://about.fb.com/news/2023/08/raising-online-defenses/

Meta. (2024, May 1). *Meta adversarial threat report, Q1 2024.* Meta. https://md.teyit.org/file/meta-threat-report.pdf

Microsoft Threat Analysis Center. (2024, October 23). *As the U.S. election nears, Russia, Iran and China step up influence efforts.* Microsoft On the Issues Blog. https://blogs.microsoft.com/on-the-issues/2024/10/23/as-the-u-s-election-nears-russia-iran-and-china-step-up-influence-efforts/

Mikhaylova, O., & Abramov, R. (2023). *Exploring the ethics of political PR professionals using moral foundations theory*. PLoS One, 18(6), e0286217. https://doi.org/10.1371/journal.pone.0286217

Missouri v. Biden, 640 F. Supp. 3d 754 (W.D. La. 2023), aff'd in part, vacated in part, and remanded, 83 F.4th 350 (5th Cir. 2023).

Mohammad, S., & Turney, P. (2013). Crowdsourcing a word-emotion association lexicon. Computational Intelligence, 29(3), 436–465.

Mündges, S., & Park, K. (2024). But did they really? Platforms' compliance with the Code of Practice on Disinformation in review. Internet Policy Review, 13(3). https://doi.org/10.14763/2024.3.1786

Murthy v. Missouri, 602 U.S. ___ (2024).

Mutlu, E. Ç., Oghaz, T., Tütüncüler, E., Jasser, J., & Garibay, I. (2020). Quantifying latent moral foundations in Twitter narratives: The case of the Syrian White Helmets misinformation. *arXiv preprint arXiv:2004.13142*. https://arxiv.org/abs/2004.13142

Nakashima, E. (2019, February 27). U.S. Cyber Command operation disrupted internet access of Russian troll factory on day of 2018 midterms. *The Washington Post.* Retrieved from https://www.washingtonpost.com/world/national-security/us-cyber-command-operation-disrupted-internet-access-of-russian-troll-factory-on-day-of-2018-midterms/2019/02/26/1827fc9e-36d6-11e9-af5b-b51b7ff322e9_story.html

National Law Review. (2024). *RESTRICT Act and National Security*. Retrieved from https://www.natlawreview.com/restrict-act

NDTV News Desk. (2025, November 23). X's "About this account" feature faces "accuracy" complaints by users. NDTV. https://www.ndtv.com/world-news/xs-about-this-account-feature-faces-accuracy-complaints-by-users-9685624

New York Times Co. v. United States (1971)
Supreme Court of the United States. (1971). *New York Times Co. v. United States, 403 U.S. 713*. https://supreme.justia.com/cases/federal/us/403/713/

NFWare. (2023, August 1). The role of CGNAT in dual stack networks. https://nfware.com/blog/the-role-of-cgnat-in-dual-stack-networks

Nguyen, T. D., Chen, Z., Carroll, N. G., Tran, A., Klein, C., & Xie, L. (2024). *Measuring moral dimensions in social media with Mformer*. arXiv preprint. https://arxiv.org/abs/2311.10219

Nguyen, T. D. (2023). *Data-Driven Understanding of Real-Life Moral Dilemmas via Topic Mapping and Moral Foundations* [Master's thesis, Australian National University].

Office of the Director of National Intelligence. (2021). *Foreign Threats to the 2020 U.S. Federal Elections*. https://www.dni.gov/files/ODNI/documents/assessments/ICA-declass-16MAR21.pdf

Office of the Director of National Intelligence. (2021, March 16). *Foreign threats to the 2020 U.S. federal elections* (ICA 2021-00002D). Office of the Director of National Intelligence. Retrieved from https://www.dni.gov/index.php/newsroom/reports-publications/reports-publications-2021/item/2203

Office of the Director of National Intelligence. (2024). *Annual threat assessment of the U.S. intelligence community 2024.* Office of the Director of National Intelligence. Retrieved from https://www.dni.gov/index.php/newsroom/reports-publications/reports-publications-2024

Official President Edict. (2015, December 31). *Presidential Edict 683 approving appended text of "The Russian Federation's National Security Strategy."* Official website of the Russian Federation President. Retrieved from http://www.kremlin.ru

Omondi Otieno, D., Abri, F., Siami-Namini, S., & Siami Namin, A. (2024). The accuracy of domain specific and descriptive analysis generated by large language models. *arXiv*. https://ar5iv.labs.arxiv.org/html/2405.19578v1

OpenAI. (2025, February 14). *How state-affiliated threat actors misuse our models.* OpenAI Safety & Security. https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors/

Pacheco, D., Hui, P.-M., Torres-Lugo, C., Truong, B. T., Flammini, A., & Menczer, F. (2021). Uncovering coordinated networks on social media: Methods and case studies. *Proceedings of the International AAAI Conference on Web and Social Media, 15*(1), 455–466. https://doi.org/10.1609/icwsm.v15i1.18075

Paletz, S. B. F., Golonka, E. M., Pandža, N. B., Stanton, G., Ryan, D., Adams, N., Rytting, C. A., Murauskaite, E. E., Buntain, C., Johns, M. A., & Bradley, P. (2024). Social media emotions annotation guide (SMEmo): Development and initial validity. *Behavior Research Methods, 56*(5), 4435–4485. https://doi.org/10.3758/s13428-023-02195-1

Pariser (2011): Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you.* Penguin Press.

PBS NewsHour. (2025, February 7). *Trump administration ends Biden-era task force aimed at seizing Russian oligarchs' assets*. https://www.pbs.org/newshour/politics/trump-administration-ends-biden-era-task-force-aimed-at-seizing-russian-oligarchs-assets

Politico. (2025, May 16). Hegseth briefly paused cyber ops against Russia as part of negotiations, GOP Rep. Bacon says. *Politico.* https://www.politico.com/news/2025/05/16/hegseth-cyber-operations-russia-pause-00354072

Posard, M. N., Kepe, M., Reininger, H., Marrone, J. V., & Helmus, T. C. (2020). *From consensus to conflict: Understanding foreign measures targeting U.S. elections* (RR-A704-1). RAND Corporation. https://www.rand.org/pubs/research_reports/RRA704-1.html

Pew Research Center. (2020). *The role of social media in polarization and populism in the United States: A historical perspective.* https://www.pewresearch.org/

Pew Research Center. (2023, September 19). *Americans' feelings about politics, polarization and the tone of political discourse.* https://www.pewresearch.org/politics/2023/09/19/americans-feelings-about-politics-polarization-and-the-tone-of-political-discourse/

PRIF Blog. (2025, March 13). *US halts defensive cyber activities against Russia: A digital withdrawal from Europe.* Peace Research Institute Frankfurt. https://blog.prif.org/2025/03/13/us-halts-defensive-cyber-activities-against-russia-a-digital-withdrawal-from-europe/

Protiviti. (2023, August 25). *Preparing for the EU Digital Services Act.* Protiviti. https://www.protiviti.com/nl-en/blogs/preparing-for-digital-services-act
Psaledakis, D. (2025, April 16). U.S. State Department closing office aimed at countering foreign disinformation. *Reuters.* https://www.reuters.com/business/media-telecom/us-state-department-closing-office-aimed-countering-foreign-disinformation-2025-04-16/

Robertson, A. (2025, November 25). X's messy "About This Account" rollout has caused utter chaos. *The Verge*. https://www.theverge.com/news/827279/xs-messy-about-this-account-rollout-has-caused-utter-chaos

Radio Free Asia. (2023, July 3). Soft propaganda? YouTube deletes channels promoting North Korean life. https://www.rfa.org/english/news/korea/vlogs-07032023120605.html

Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences, 118*(26), e2024292118. https://doi.org/10.1073/pnas.2024292118

Rauch, J. (2021). The Constitution of Knowledge: A defense of truth. Brookings Institution Press.

Reuters. (2023, September 21). North Korea using social media to soften its image abroad, analysts say. https://www.reuters.com/world/asia-pacific/north-korea-sees-propaganda-value-slimmer-kim-analysts-say-2021-06-28/

Reuters. (2025a, February 6). Trump administration disbands task force targeting Russian oligarchs. *Reuters.* https://www.reuters.com/world/us/trump-administration-disbands-task-force-targeting-russian-oligarchs-2025-02-06/

Reuters. (2025b, November 23). DOGE 'doesn't exist' with eight months left on its charter – official. *Reuters.* https://www.reuters.com/world/us/doge-doesnt-exist-with-eight-months-left-its-charter-2025-11-23/

Russian Federation. (2021). *National Security Strategy of the Russian Federation*. Retrieved from http://en.kremlin.ru/events/president/news/66098

Rid, T. (2012). *Cyber war will not take place*. Journal of Strategic Studies, 35(1), 5-32.

Rival IQ. (2024). *2024 social media industry benchmark report*. https://www.rivaliq.com/blog/social-media-industry-benchmark-report/

S. 686, 118th Cong. (2023). *Restricting the Emergence of Security Threats that Risk Information and Communications Technology Act (RESTRICT Act)*. https://www.congress.gov/bill/118th-congress/senate-bill/686

Sager, W. R. (2015). Apple pie propaganda? The Smith–Mundt Act before and after the repeal of the domestic dissemination ban. *Northwestern University Law Review, 109*(3), 511–564. https://scholarlycommons.law.northwestern.edu/nulr/vol109/iss2/7/

Scherling, L. (2025, April 23). The downfall of the Global Engagement Center and disappearing guardrails against disinformation. *Tech Policy Press*. https://www.techpolicy.press/the-downfall-of-the-global-engagement-center-and-disappearing-guardrails-against-disinformation/

Sixth Circuit net-neutrality decisions. (2025). In re: MCP No. 185, 25a0002p-06 (6th Cir. Jan. 2, 2025). https://www.opn.ca6.uscourts.gov/opinions.pdf/25a0002p-06.pdf

Seiling, L. K., Iglesias Keller, C., Ohme, J., Klinger, U., & de Vreese, C. H. (2025). *Data access for researchers under the Digital Services Act* (Weizenbaum Policy Paper No. 14). Weizenbaum Institute. https://doi.org/10.34669/WI.WPP/14

Serafino, M., Zhou, Z., Andrade, J. S., Jr., Bovet, A., & Makse, H. A. (2024). Suspended accounts align with the Internet Research Agency misinformation campaign to influence the 2016 US election. *EPJ Data Science, 13*, Article 29. https://doi.org/10.1140/epjds/s13688-024-00464-3

Smith, M., & Smith, R. (2021). *How the Russian influence operation on Twitter weaponized American military culture*. Proceedings of the 16th International Conference on Cyber Warfare and Security, Academic Conferences International Limited. Retrieved from https://papers.academic-conferences.org/index.php/iccws/article/download/985/982

Sprigman, C. (2025, January 17). The Supreme Court upheld the US TikTok ban. Now what? (Q&A). NYU News. https://www.nyu.edu/about/news-publications/news/2025/january/sprigman-tiktok-q---a.html

Swan, B., & Lippman, D. (2022, May 5). *Small group, big headache: Inside DHS' messy Disinformation Governance Board launch*. Politico. Retrieved from https://www.politico.com/news/2022/05/05/dhs-disinformation-governance-board-00030251

Stanford Law School. (2024). *The EU's Digital Services Act and its impact on online platforms*. Stanford-Vienna Transatlantic Technology Law Forum, EU Law Working Papers No. 85. Retrieved from https://law.stanford.edu/publications/no-85-the-eus-digital-services-act-and-its-impact-on-online-platforms/

Starbird, K., Arif, A., & Wilson, T. (2019). *Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations.* Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–26. https://doi.org/10.1145/3359229

Sychev, O. A., & Protasova, I. N. (2020). *The relationships between moral foundations, social beliefs, and attitudes towards economic inequality among Russian youth: A case study of Altai Krai*. RUDN Journal of Psychology and Pedagogics, 17(4), 705-718. https://doi.org/10.22363/2313-1683-2020-17-4-705-718

Stanley v. Georgia, 394 U.S. 557 (1969).

Stiković, S. K. (2024). The EU's Digital Services Act and its impact on online platforms (Stanford–Vienna European Union Law Working Paper No. 85). Stanford Law School. https://law.stanford.edu/wp-content/uploads/2024/02/EU-Law-WP-85-Stikovic.pdf

Suhler, C. L., & Churchland, P. S. (2011). Can innate, modular "foundations" explain morality? Challenges for Haidt's moral foundations theory. Journal of Cognitive Neuroscience, 23(9), 2103–2116. https://patriciachurchland.com/wp-content/uploads/2020/05/2011-Innate-Modular-Foundations.pdf

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson.

Taiwan National Security Bureau. (2024).
Tech Policy Press. (2024, December 5). *Non-public data access for researchers: Challenges in the draft delegated act of the Digital Services Act.* Tech Policy Press. https://techpolicy.press/non-public-data-access-for-researchers-challenges-in-the-draft-delegated-act-of-the-digital-services-act

Teena Xu, Emmanuel Guajardo (2022). *Mission-based Academic Tweeting: Analysis of Content Engagement over One Year,* Open Forum Infectious Diseases, Volume 9, Issue Supplement_2. Retrieved from https://doi.org/10.1093/ofid/ofac492.1135

TechRadar. (2025). X could show if you're using a VPN – here's all we know. TechRadar. https://www.techradar.com/vpn/vpn-privacy-security/x-could-soon-show-if-youre-using-a-vpn-heres-all-we-know

Thomas, T. (2019, March 22). What a U.S. operation against Russian trolls predicts about escalation in cyberspace. *War on the Rocks.* Retrieved from

https://warontherocks.com/2019/03/what-a-u-s-operation-against-russian-trolls-predicts-about-escalation-in-cyberspace/

Thornton, R. (2015). *The changing nature of modern warfare: Responding to Russian information warfare*. The RUSI Journal, 160(4), 40-48.

Tian, L., Zhang, X., Kim, M. M.-H., Biggs, J., & Rizoiu, M.-A. (2025). *X-Troll: eXplainable detection of state-sponsored information operations agents* (arXiv:2508.16021). arXiv. https://doi.org/10.48550/arXiv.2508.16021

TikTok Inc. v. Garland, 604 U.S. ___ (2025). Slip opinion. Retrieved from https://www.supremecourt.gov/opinions/24pdf/24-656_ca7d.pdf

Time. (2024, April 5). *China uses AI to sow disinformation and escalate cyber threats, Microsoft says. Time.* https://time.com/6963787/china-influence-operations-artificial-intelligence-cyber-threats-microsoft/

Turner Lee, N. (2018, April 10). *Facebook's Mark Zuckerberg testifies before Congress: What comes next?* Brookings. Retrieved from https://www.brookings.edu/articles/facebooks-mark-zuckerberg-testifies-before-congress-what-comes-next/

Trump, D. J. (2025, September 30). *Executive Order 14352 of September 25, 2025: Saving TikTok while protecting national security. Federal Register, 90*(190), 67219–67224. Retrieved from https://www.federalregister.gov/documents/2025/09/30/2025-19139/saving-tiktok-while-protecting-national-security

Twenge, J. M. (2017). *iGen: Why today's super-connected kids are growing up less rebellious, more tolerant, less happy—and completely unprepared for adulthood—and what that means for the rest of us*. Atria Books.
Twitter. (2018). Academic Research Access. Twitter. Retrieved from https://developer.twitter.com/en/products/twitter-api/academic-research

Uren, T. (2025, May 30). Russia's cybercriminals and spies are officially in cahoots. Lawfare. https://www.lawfaremedia.org/article/russia's-cybercriminals-and-spies-are-officially-in-cahoots

United States Department of Justice. (2024, September 4). *Two RT employees indicted for covertly funding and directing U.S. company that published thousands of videos in furtherance of Russian interests*. U.S. Department of Justice. Retrieved from https://www.justice.gov/opa/pr/two-rt-employees-indicted-covertly-funding-and-directing-us-company-published-thousands-videos

U.S. Attorney's Office, Southern District of New York. (2021, November 18). *U.S. Attorney announces charges against two Iranian nationals for cyber-enabled disinformation and threat campaign designed to interfere with the 2020 U.S. presidential election*. https://www.justice.gov/usao-sdny/pr/us-attorney-announces-charges-against-two-iranian-nationals-cyber-enabled

U.S. Congress. (1948). Smith-Mundt Act of 1948 (U.S. Information and Educational Exchange Act). Public Law 80-402, 62 Stat. 6, 22 U.S.C. §§ 1431-1448.

U.S. Congress. (2012). Smith-Mundt Modernization Act of 2012. Public Law 112-239, 126 Stat. 1632.

U.S. Cyber Command. (2022, October 25). *Cyber 101: Defend forward and persistent engagement.* U.S. Cyber Command. Retrieved from https://www.cybercom.mil/Media/News/Article/3198878/cyber-101-defend-forward-and-persistent-engagement/

U.S. Department of Homeland Security. (2017, October 14). *DHS and partners convene first election infrastructure coordinating council.* U.S. Department of Homeland Security. Retrieved from https://www.dhs.gov/archive/news/2017/10/14/dhs-and-partners-convene-first-election-infrastructure-coordinating-council

U.S. Department of Homeland Security. (2025, January 20). *Memorandum for all members of Department of Homeland Security advisory committees.* (as reported in Reuters, 2025-01-21). https://www.reuters.com/world/us/us-department-homeland-security-firing-all-advisory-committee-members-letter-2025-01-21/

U.S. Department of Justice. (n.d.). *Foreign Agents Registration Act (FARA)*. https://www.justice.gov/nsd-fara

U.S. Department of Justice. (2018a, February 16). *Grand jury indictment of Internet Research Agency, et al., for conspiracy to defraud the United States* [Press release]. U.S. Department of Justice. Retrieved from https://www.justice.gov/opa/pr/justice-department-announces-charges-against-russian-nationals-and-companies-conspiracy

U.S. Department of Justice. (2018b, July 13). *Grand jury indicts 12 Russian intelligence officers for hacking offenses related to the 2016 election* [Press release]. U.S. Department of Justice. Retrieved from https://www.justice.gov/opa/pr/grand-jury-indicts-12-russian-intelligence-officers-hacking-offenses-related-2016-election

U.S. Department of Justice. (2018c). *United States v. Internet Research Agency LLC, et al.* Criminal No. 1:18-cr-00032 (D.D.C.). Retrieved from https://www.justice.gov/file/1035477/download

U.S. Department of Justice. (2022, March 2). *Attorney General Merrick B. Garland announces launch of Task Force KleptoCapture* [Press release]. https://www.justice.gov/archives/opa/pr/attorney-general-merrick-b-garland-announces-launch-task-force-kleptocapture

U.S. Department of Justice. (2020). *Iranian hackers indicted for disinformation campaign targeting U.S. voters* [Press release].

https://web.archive.org/web/20210301000000/https://www.justice.gov/opa/pr/iranian-hackers-indicted-disinformation

U.S. Department of Justice. (2021, November 18). *Two Iranian nationals charged for cyber-enabled disinformation and threat campaign designed to influence the 2020 U.S. presidential election* [Press release]. https://www.justice.gov/opa/pr/two-iranian-nationals-charged-cyber-enabled-disinformation-and-threat-campaign-designed

U.S. Department of Justice. (2024, July 9). *Two RT employees indicted for covertly funding and directing U.S. company that published thousands of social media posts on behalf of Russia Today.* Office of Public Affairs. https://www.justice.gov/archives/opa/pr/two-rt-employees-indicted-covertly-funding-and-directing-us-company-published-thousands

U.S. Department of Justice. (2025, February 5). *Memorandum on criminal enforcement policies* (Bondi memo; excerpted in Inside Political Law). https://www.justice.gov/ag/media/1388541/dl

U.S. Department of State. (n.d.). *About us – Global Engagement Center* [Archived 2021–2025]. 2021–2025.state.gov. https://2021-2025.state.gov/about-us-global-engagement-center-2/

U.S. Department of State. (2023, September 28). *How the People's Republic of China seeks to reshape the global information environment* (Global Engagement Center special report). https://2021-2025.state.gov/wp-content/uploads/2023/10/HOW-THE-PEOPLES-REPUBLIC-OF-CHINA-SEEKS-TO-RESHAPE-THE-GLOBAL-INFORMATION-ENVIRONMENT_508.pdf

U.S. Department of the Treasury. (2024, November 1). Treasury issues final rule expanding CFIUS coverage of certain real estate transactions near military installations [Press release]. https://home.treasury.gov/news/press-releases/jy2708

U.S. Election Assistance Commission. (2017, May 11). *A new home base for critical infrastructure information.* U.S. Election Assistance Commission. Retrieved from https://www.eac.gov/blogs/new-home-base-critical-infrastructure-information

U.S. Telecom Association v. Federal Communications Commission, 825 F.3d 674 (D.C. Cir. 2016).

Virginia State Board of Pharmacy v. Virginia Citizens Consumer Council, Inc., 425 U.S. 748 (1976).

Wang, J., Zhang, A., & Karger, D. (2021**).** Designing for engaging with news using moral framing towards bridging ideological divides. *arXiv preprint arXiv:2101.11231*. https://arxiv.org/abs/2101.11231

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe. Retrieved from

https://firstdraftnews.org/glossary-items/pdf-wardle-c-derakshan-h-2017-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making-council-of-europe/

Warner, M. (2023, March 7). *Senators introduce bipartisan bill to tackle national security threats from foreign tech* [Press release]. Office of Senator Mark R. Warner. https://www.warner.senate.gov/public/index.cfm/2023/3/senators-introduce-bipartisan-bill-to-tackle-national-security-threats-from-foreign-tech

Weizenbaum Institute. (2025, September 26). *DSA40 Data Access Days.* Weizenbaum Institute. https://www.weizenbaum-institut.de/en/events-1/dsa40-data-access-days/

Wies, S., Bleier, A., & Edeling, A. (2023). *Finding Goldilocks influencers: How follower count drives social media engagement*. *Journal of the Academy of Marketing Science, 51*(5), 1049–1068. https://www.researchgate.net/publication/362956735

Williams, S. (2025, January 22). DHS disbands cyber safety review board, ending one of CISA's few bright spots. *CSO Online*. https://www.csoonline.com/article/3807871/trump-administration-disbands-dhs-board-investigating-salt-typhoon-hacks.html

Westerman, D., Spence, P. R., & Van Der Heide, B. (2012). *A social network as information: The effect of system generated reports of connectedness on credibility on Twitter*. *Computers in Human Behavior, 28*(1), 199–206. https://doi.org/10.1016/j.chb.2011.09.001

Wolters, P., & Zuiderveen Borgesius, F. (2025). *The EU Digital Services Act: What does it mean for online advertising and adtech?* (Working paper). arXiv. https://arxiv.org/pdf/2503.05764.pdf

Xu, M., Cropp, F., & Cameron, G. T. (2023). Dissecting moral judgements: Using moral foundation theory to advance the contingency continuum. Public Relations Review, 49(4), 102370. https://doi.org/10.1016/j.pubrev.2023.102370

Viktoratos, I., & Tsadiras, A. (2021). *Personalized Advertising Computational Techniques: A Systematic Literature Review*. MDPI. https://doi.org/10.3390/info12110480

Zhang, Y., Lukito, J., Su, M.-H., Suk, J., Xia, Y., Kim, S. J., Doroshenko, L., & Wells, C. (2021). *Assembling the networks and audiences of disinformation: How successful Russian IRA Twitter accounts built their followings, 2015–2017*. *Journal of Communication, 71*(1), 60–84. https://doi.org/10.1093/joc/jqaa042

Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., & Blackburn, J. (2019). Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the Web. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion)* (pp. 218–226). ACM. https://doi.org/10.1145/3308560.3316495

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Public Affairs. https://www.hbs.edu/faculty/Pages/item.aspx?num=56791