Using Diachronic Static Word Embeddings to Detect and Characterize Term Toxification on Reddit

TEAM CYBERLANG

Terranova Oh, Akhilesh Puranam

Dr. Erik Nesse

Thesis submitted in partial fulfillment of the requirements of the Gemstone Honors Program, University of Maryland, Year 2026

Committee: Dr. Erik Nesse, Dr. Anton Rytting, Dr. David Loshin, Christabel Acquaye, Liancheng Gong

Abstract

Toxic discourse on the internet has a powerful influence on its users, and recognizing increasingly negative or derogatory usage of specific terms–such as ‘Karen’ or ‘woke’–can help signal potential “toxification” of an online space. Our research demonstrates that a novel approach building on prior work on diachronic static word embedding analysis can assist in tracking and understanding term toxification on Reddit, a popular online social media platform, by observing how the usages of words with multiple meanings (including a pejorative/negative one) change over time. This study can be expanded by using contextual word embeddings, expanding comparisons to other social media platforms, and examining word toxification in different languages.

Acknowledgements

We would like to express a sincere thank you to the Gemstone Honors Program for providing the opportunity for a four year research project. Specifically, thank you to Dr. Lovell, Dr. Lansverk, Nevenka Zdravkovska, and all Gemstone Honors Program staff and faculty. We would also like to thank our mentor Dr. Erik Nesse for guiding us through this research as our mentor and role model. Thank you to all of our discussants for their invaluable feedback. We would like to thank our friends and family for supporting us as we conducted our research. This research is specially dedicated to Yash Thakur. Without the support of those mentioned, this research would not have been possible.

Table of Contents

Chapter 1: Introduction

Background

The advent of the internet has fundamentally changed how humans communicate with each other. In the past, communication across the world took days or weeks, but now it happens within seconds. Historically, language and culture have developed in relatively isolated bubbles defined by geographic locations, resulting in regions with their own languages, dialects, accents, and slang. American English and British English once shared much more in common than in recent times, where they differ both in accent and (in many cases) vocabulary.. Now, the advent of the internet has begun to shape the linguistic divergence and convergence previously controlled mostly by geography: English speakers in the UK and America often make use of similar internet slang, for instance, and new words and cultural trends are no longer stuck in one place. Instead, they spread rapidly around the world, and linguistic trends that might otherwise develop over decades can now play out in months.

Although widespread digital access provides the advantage of instant communication, it simultaneously introduces significant challenges. Much of the online discussion accounting for a great deal of communication today happens anonymously. This anonymity often gives users a sense of freedom from consequences, empowering them to use more offensive or hostile language that they might otherwise use. As a result of that anonymity (in combination with many other factors–social dynamics, technological advances, internet access on the go via smartphones, greater utilization, etc.), the internet is a complex social environment, and one in which negativity is not only commonplace, but also sometimes used to gain attention and status (Suler, 2004; Schöne et al., 2021).

In any given online community, one way that hostility, negativity, or general "toxicity" (the term we will use to describe pejorative, hostile, derogatory, and other similarly negative uses of language) plays out is through the transformation of otherwise "normal" or relatively neutral words into variants with darker or more negative meanings. Toxification of language in that sense is not new, but unlike traditional language change, which happens slowly, toxification may occur much more quickly through the internet. Harmful terms that start in small online communities can quickly spread to the general public, users may be exposed to them quickly and repeatedly by algorithms shaping the materials they see, and the ability of the internet to rapidly transform and spread negative language (and its effects on users) creates a volatile environment that requires careful study.

Research Problem

As the majority of communication moves to the internet, the mass production and dissemination of toxic language online has the potential for widespread and varied impacts on the internet's users. For the purposes of this research, we define "toxic language"/"toxic speech" as written language used to disparage, marginalize, or diminish an individual or group, whether overtly or delivered in a "joking" or sarcastic manner. Toxic language carries risks for the social and mental well-being of individuals (Dadvar et al., 2012), for example, and its impact is not contained online. Online hostility may be normalized and transferred into offline discourse (Vidgen & Yasseri, 2020), just as internet memes and trends have been integrated into daily vernacular, and constant exposure to digital negativity may desensitize users to toxicity in general. This in turn may propagate harmful behaviors in the offline world where the

consequences for hate speech or aggression can be severe. Given the centrality of the internet to our discourse today, the potential for this normalized negativity to cause widespread social damage is substantial. Therefore, it is imperative to study toxic language and the processes influencing its development and dissemination online to fully understand its origins and impacts.

The benefits of better understanding the evolution and dissemination of toxic language are multifaceted and relate both to social safety and academic inquiry. First, understanding when and where discourse appears to become increasingly toxic allows for proactive harm reduction. By identifying emerging negative terms before their use becomes widespread and potentially normalized, platforms can implement strategies to prevent their spread and some of the harm entailed by their use. This preventive approach is essential for fostering a more positive and inclusive digital ecosystem.

Practically, understanding toxic language is vital for the advancement of content moderation tools. Cyberbullying and online harassment are persistent issues with severe real-world ramifications (Wulczyn et al., 2017; Dinakar et al., 2021). Current moderation systems are often reactive, flagging terms only after they are widely recognized as offensive, but understanding the semantic shifts (for instance) that may precede overt toxicity could aid in the development of tools that identify harmful language in its early stages, potentially mitigating cyberbullying before it escalates.

Beyond immediate safety applications, this research contributes to the academic field of computational linguistics by building upon prior methods to test a novel method for focused semantic change detection in a high-impact domain. Modeling the development of toxic language may reveal interesting patterns related to the development of internet language in

general. Our approach may provide a valuable tool for linguists studying the rapid semantic shifts the internet continues to cause. Finally, we believe our methods may enhance future studies of the relationship between language and community dynamics. Specifically within the context of Reddit, observing how language becomes negative and permeates a subreddit could offer insight into how online communities radicalize or become toxic over time.

Prior Work

Many current toxicity detection systems struggle to identify emergent toxicity because they treat language as static. These systems are designed to flag insults, hate speech, or harassment by checking text against a pre-existing lexicon of known toxic words or by applying machine learning models trained on examples labeled as toxic by human annotators. These systems typically use one of two methods: some systems check words and phrases in a post or article against an existing lexicon, while others use purpose-built machine learning (ML) models to flag toxicity based on more complex linguistic/contextual patterns. Systems that check against an existing lexicon use a pre-established dataset of known toxic words to determine matches. Machine learning models are generally trained to identify *patterns* indicative of toxic language, usually based on examples labeled as toxic or not toxic by human annotators.

While those approaches are effective at identifying content already known to be toxic, they struggle to identify emergent toxicity in such a dynamic environment as the internet. Once given terms, phrases, or training examples identified as toxic, they function (at search or inference time) as if a toxic word were toxic in all times and places (unless guided by sophisticated rules regarding surrounding context), and they can only "see" toxicity post hoc–once it has been deemed toxic by inclusion in a lexicon or training set. However, perhaps especially in online communities, words change quickly—and terms with benign senses are sometimes assigned new, more malicious meanings, all of which may be used alongside each other in the same online space. This process of semantic shift and since Toxicity detectors's reliance on pre-existing datasets for their definitions and training and (especially in the case of

lexically-driven systems) sometimes context-unaware operation is at odds with the realities of semantic shift.

Additionally, although large language models (LLMs) are potentially better "judges" of toxicity than lexical systems or traditional ML algorithms, they are also computationally (and therefore monetarily) expensive to run constantly or repeatedly on the vast amount of data in an online forum such as Reddit.. Moreover, even if LLMs might serve as effective *judges* of toxicity, their judgments may not necessarily provide (without extensive research and testing) additional information about *how* toxification may occur within a community (such as a subreddit) and infect the broader platform. Our research aims to contribute to filling gaps in existing toxicity-detection systems and potentially to understanding toxification at a middle ground between simplicity and expense.

Hypothesis

- We hypothesize that we can effectively measure and quantify the toxification of online discourse on Reddit using static word embeddings. We believe that by using computational tools, specifically word embedding analysis, nearest neighbor comparison, and cosine similarity, we can track the spread of negative words and provide concrete data on how the language on the platform has evolved.
- We hypothesize that our methodology can be used to detect and characterize how words change over time, even if the change is not toxic.

Objectives and Aims

The primary objective of this research is to validate the feasibility and efficacy of using static (non-contextual) word embeddings to detect and characterize toxification of online discourse. We aim to determine in particular whether such word embeddings (generated by packages like Word2Vec), when computed for terms at different times and in different contexts and made comparable by appropriate methods, are sufficiently sensitive tools reasonably suited to those tasks. We intend to accomplish this through experiments designed to observe of a change in a specific word's usage from generally non-negative to toxic over time by analyzing the (Word2Vec) vector representations of the word at different times , the contexts in which the words tend to occur (via qualitative nearest-neighbor investigations), and the sentiment usually surrounding them (via off-the-shelf sentiment analysis tools). We seek to confirm whether our methods can reliably detect not just that a change occurred, but potentially also reveal

characteristics of the change itself that signal the direction, speed, or other aspects of toxification that might predict or explain it in general..

Beyond validating that methodology, this study aims to quantify the evolution of toxicity on Reddit within the limited scope defined by our data and choice of terms to analyze. Currently, discussions about online toxicity are sometimes based on anecdotal evidence or simple prevalence statistics of negative words. We hope to improve upon this by examining alternative metrics for measuring toxification using what we learn about its relationship to static word embeddings. By calculating the cosine similarity between neutral terms and known toxic clusters across different time slices, for instance, we may be able to assign a quantifiable value to the toxicity or toxification of specific communities. This allows us to move from vague observations of hostility to data-driven assessments of platform health.

Scope of the Study

The scope of this research is defined by the dataset, timeframe, and methods selected for analysis. This study utilizes textual data harvested exclusively from Reddit, covering the period from **2005** to **2024**. By focusing our analysis on texts from a single platform, we can observe the evolution of specific terms within a relatively consistent community structure, although we acknowledge the limitations to generalizability of any findings resulting from our observations. The internet is a vast network of interconnected platforms and this study does not address data from other major social media sites. As a result, we cannot examine potentially rich sources of information about toxification, such as the cross-platform migration of terms (e.g., slang originating on Twitter or 4chan before appearing on Reddit. Instead, our focus is entirely contained within the Reddit ecosystem to track how words spread and evolve once they have entered this specific domain.

Methodologically, this research is also bounded by the use of non-contextual word embeddings, specifically the Word2Vec algorithm. We explicitly exclude the use of contextualized embedding models such as BERT. This exclusion is intentional, as a central goal of our research is to investigate whether static embeddings (which are relatively inexpensive computationally and, in some respects, better studied) can be leveraged to detect and characterize toxicity changes over time. For example, we aim to determine if non-contextual embeddings, when aligned across distinct time periods, can effectively capture meaningful information about toxification through changes in the set of words most likely to appear in the same contexts as a word whose usage becomes more negative (i.e., its “nearest neighbors”). Contextual word

embeddings from transformer-based models such as BERT introduce significant complexities beyond the scope of this work.

Chapter 2: Literature Review

Overview

This chapter explains previous work related to toxicity detection, semantic change detection/evaluation, word embedding alignment through orthogonal procrustes, and vector-based analysis such as cosine similarity and nearest neighbors. These methods and foundational concepts are the basis of our research, and this literature review will define related concepts, discuss its presence in current academic research, and explain its significance and application in this study.

Toxicity Detection

Similarly to the definition we use in this paper, “toxic language” refers in the literature to language that is insulting, hateful, harassing, threatening, or harmful in interpersonal contexts (Wulczyn et al., 2017). Detection of toxic language is an area of interest in recent years because online platforms that include user generated posts need to moderate millions (even billions) of posts, comments, and general discourse. Toxicity detection research looks to algorithmically identify harmful language to create a safer environment for their users, protect them from harassment, and encourage positive interaction..

A review article by Fortuna and Nunes (2018) examined different approaches to automated toxicity detection. Overall, the review indicates that toxic language detection tends to focus on general hate speech, racism, sexism, and toxic language related to religion (Fortuna and Nunes, 2018) and identifies three categories of detection approaches:

1. Supervised machine learning approaches
2. Lexicon/dictionary based approaches
3. Context/transformer based approaches

Supervised Machine Learning Approaches

Supervised machine learning approaches are among the most prominent tools in toxicity detection. A foundational study used crowdsourcing and machine learning to analyze personal attacks in online platforms, specifically using Wikipedia as a case study (Wulczyn et al., 2017). In general, their idea was to crowdsource the identification of personal attacks within a sample of discussion comments, aggregate the votes into labels that the supervised machine learning models would train on, evaluate the performance, and then apply the final model to Wikipedia to quantify toxicity. This study looked to answer key questions, including:

1. What is the impact of anonymity?
2. How do attacks vary with the quantity of a user's contributions?
3. Are attacks concentrated among a few highly toxic users?

Through their research, they developed the Wikipedia Detox dataset and found that toxicity is predictable at scale via supervised machine learning. They also found that:

1. Anonymous contributions were much more likely to be an attack, although overall they contribute to less than half of attacks.
2. Users with both low and high levels of contribution were responsible for attacks.
3. The majority of the attacks are widespread among Wikipedia users.

Their findings introduced a large, high-quality dataset for toxicity research that is widely cited and used by other machine learning models for toxicity detection. It also established a transparent and reproducible pipeline for labeling toxic content.

Similarly, Davidson et al. (2017) introduced a hate-speech dataset, with their study highlighting linguistic nuances in hate speech (actual hate speech, offensive but not hateful, sarcasm, ambiguous insults) and their challenges for automated classification systems for toxicity detection.

Both of these foundational studies rely heavily on labeling language or texts considered toxic. However, new derogatory terms emerge rapidly in online communities (Vidgen & Yasseri, 2020), supervised machine learning models fail to detect them until training data is updated, and it would be resource intensive to do so constantly. It is also noted that "Language evolves quickly, in particular among young populations that communicate frequently in social networks" (Fortuna and Nunes, 2018) which suggests a need for alternate approaches to toxicity detection over time on online platforms such as Reddit in particular, where both toxic language itself and trends in its development may not match other platforms like Wikipedia.

Lexicon/Dictionary Based Approaches

Lexicon/Dictionary based approaches to toxicity detection use preestablished lists of harmful expressions and language to classify and detect toxic language by searching for those words in a text. In general, these systems compute toxicity scores based on term occurrence counts, prevalence, or other similar metrics. This method has been used on negative words such

as insults and expletives (Liu and Forss, 2015), to discover the frequency of profane words in texts (Dadvar et al., 2012), and to label/find specific language frequently used in verbal abuse (Dinakar et al., 2011).

Research papers using lexicon/dictionary based approaches to toxicity detection often highlight their low computational cost and robustness. However, some also indicate that (just as with training data for machine learning approaches) lexicons used in this way are static and cannot keep up with the constant linguistic change that happens on the internet and online platforms. Social media platforms like X, Facebook, Instagram, and Reddit have communities that can create and repurpose words at a fast rate. Additionally, counts and static lexicons of toxic terms cannot alone indicate any potential semantic drift of the words they contain. Notably, Fortuna and Nunes (2018) reviewed research using lexical/dictionary approaches and mention using vector-based approaches such as word embeddings in future work to understand semantic drift of toxic words.

Context/Transformer Based Approaches

In recent years, transformer-based language models have introduced a form of contextualized or "context-aware" embedding that encodes information both about a particular piece of text and its surrounding context. Using these embeddings has been shown to improve toxicity detection accuracy. Zampieri et al. (2020), Mozafari et al. (2019), and Caselli et al. (2020) explored the use of contextual word embeddings from models such as BERT, BERT

model variants (BERT-base and BERT-large (uncased), RoBERTa-base and RoBERTa-large, XLMRoBERTa), and ALBERT models.

These studies demonstrate that contextual embeddings outperform earlier approaches using non-contextual word embeddings, especially when it comes to detecting subtle or indirect toxicity, where the aforementioned approaches struggled. However, there are drawbacks with context/transformer based approaches:

1. Using large, performant transformer-based models can be computationally expensive and training them requires significant computational resources.
2. Comparison of contextual embeddings over time is not straightforward, as the embeddings change (by design) depending on the totality of the surrounding context; the nature and meaning of any given change in the embedding is not necessarily intelligible.

Overall, toxicity detection methods are advanced in classification accuracy, but offer limited tools and methods for examining how toxic words and their usages change over time. Studies such as Vidgen and Yasseri (2020) and Fortuna and Nunes (2018) acknowledge that negative language terms evolve rapidly on the internet, but few studies propose methods for addressing the evolution and change. This current study aims to contribute to this by combining known toxicity detection and semantic change techniques to examine negative language and language toxification on Reddit.

Semantic Change

Semantic change (also known as semantic shift) is a "special case of lexical change where an existing form (a lexeme) acquires or loses a particular meaning, i.e., increasing or decreasing polysemy" (Traugott & Dasher 2001, Fortson 2003, Newman 2016, Traugott 2017). Semantic change may occur due to social, cultural, political, or linguistic factors.

The theoretical foundation for much modern computational semantic change research is often summarized by a maxim: "A word is characterized by the company it keeps" (Firth, 1957). Firth's study of words and their contexts of use provided a foundation for modern methods to produce numerical representations (vectors/embeddings, see below) meant to capture semantic information about the words they encode (Hamilton et al., 2016). Firth's principle suggests that word "meaning" can be understood through patterns of occurrence with other words. Computationally speaking, this means that words appearing in similar contexts should have similar embeddings, and in general, changes in a word's typical contexts signal changes in its meaning. Modern research often uses word embeddings to study how words change online.

Word Embeddings

Word embeddings are vector representations of words in a document or text corpus. From the vector representation, one can calculate different statistics such as cosine similarity or solve word analogies through vector arithmetic.

Upon the basis of word embeddings, Hamilton et al. (2016) proposed statistical laws of semantic change, which state that 1) frequently-used words tend to change more slowly than

rarely-used words and 2) words with more polysemous or complex semantics change more rapidly. Additionally, this study demonstrated a method for comparing static word embeddings generated from one corpus of texts with embeddings generated from a separate corpus using Orthogonal Procrustes alignment. Normally, word embeddings resulting from algorithms such as Word2Vec run on separate data are not comparable, but the study's alignment procedure makes it possible to do so.

Their approach involves training separate word embedding models on corpora from different time periods, then aligning these embedding spaces so that vectors can be directly compared. By measuring how far a word's vector has moved between time periods using metrics like cosine similarity or euclidean distance, researchers can quantify semantic change over time.

Kutuziv et al., 2018 extends and refines that work and provides a comprehensive survey of comparison techniques, alignment strategies, and evaluation methods for diachronic word embeddings (Kutuziv et al., 2018). Their work helped establish best practices for semantic change detection when using word embeddings.

Dubossarsky et al., 2017 demonstrates a common pitfall in quantifying semantic change using word embeddings–some apparent shifts arise from random noise and embedding instability rather than actual linguistic change. They showed that frequency controls and robust alignment procedures are essential for distinguishing real semantic change from false positives.

Application to Online Language

While non-computational semantic change studies often examine long-term evolution in formal written language such as newspapers, books, or historical documents, more recent research has turned to online platforms where meanings of words can shift rapidly. Shoemark et al. (2019) validated word embedding alignment techniques for social media language, demonstrating that semantic shifts can be detected reliably in the informal, rapidly evolving language of online platforms, extending methods previously tested only on historical corpora spanning decades. Additionally, this study demonstrated that these word embedding techniques that were initially tested on historical corpora spanning decades can be successfully applied to social media data spanning much shorter time periods.

Another study observed semantic variation across different subreddit communities and found that communities develop their own localized meanings for common terms (Del Tredici and Fernández, 2018). Their work shows that semantic change occurs not only over time but also across different subcommunities within the same platform, which is an integral part of our research to study local (subreddit-specific) and global (platform-wide) spaces.

These findings highlight two principles important for our research design. First, online semantic change can occur on the scale of months rather than decades, making it possible to study relatively short time periods. Second, local communities in online spaces often develop meanings that diverge from global platform usage, and understanding that divergence could contribute to the understanding of online language and how it changes.

Orthogonal Procrustes

When word embeddings are trained on separate corpora, they are not directly comparable because the embedding spaces are rotationally invariant and do not lie in the same vector space. Orthogonal Procrustes resolves this issue by calculating a rotational matrix to transform the embeddings into the same space. After this is done, vector analysis methods such as cosine similarity and nearest neighbors can be used directly.

The Orthogonal Procrustes approach finds the orthogonal matrix R minimizing:

$$\|XR - Y\|$$

where X and Y are matrices (Schönemann, 1966). Orthogonal Procrustes ensures that distances and angles are preserved between the vector representations and thus the comparisons will reflect semantic differences (if any) rather than arbitrary coordinate rotations. The key of Orthogonal Procrustes in combination with word embeddings lies in anchor words, which are terms assumed to have stable meaning across the time periods being compared. Typically, these are high-frequency words unlikely to change quickly or significantly in meaning, such as (but not limited to) “the”, “and”, and “a”.

Multiple studies have used Orthogonal Procrustes to observe semantic change using word embeddings. For example, one study expanded upon Schönemann’s original Orthogonal Procrustes solution and allowed the solution to be computed and applied with user-friendly code (Hamilton et al., 2016). Another study applied the Orthogonal Procrustes solution to align word embeddings and calculate the statistical significance of semantic change in news corpora (Kulkarni et al., 2015). The use of Orthogonal Procrustes has also been applied to social media

corpora, where another study validated Orthogonal Procrustes for short time spans and noisy social media data, demonstrating that semantic shift can be reliably detected from the rapidly evolving online corpora (Shoemark et al., 2019).

For our study of negative language on Reddit, Orthogonal Procrustes allows us to compare embeddings from different subreddits and different time periods while making sure that any observed differences in embeddings of the same word are likely to be due to semantic change rather than inherent to the embedding generation process. Through Orthogonal Procrustes, we are able to track how the meanings of specific negative words have evolved in time and across communities.

Vector Similarity Measures

Vector similarity measures refer to the way vectors can be compared. There are many different methods that can be used, however cosine similarity and nearest neighbors are widely used methods for vector comparison within computational linguistics (Manning et al., 2008; Hamilton et al., 2016).

Cosine Similarity

Cosine similarity measures the cosine angle between two vectors, which provides a metric of their directional similarity that is independent of their magnitude. In the context of

word embeddings, this signifies their semantic similarity where words with similar meanings have vectors pointing in similar directions, which result in higher cosine similarity scores.

The cosine similarity score typically ranges from -1 to 1, where 1 indicates identical direction, 0 indicates orthogonality, and -1 indicates opposite direction. In practice, when working with word embeddings, the embeddings themselves are typically normalized, which means the score will range from 0 to 1 instead, where a higher score still indicates greater semantic similarity. This technique has been widely used in multiple domains of research.

In semantic change research, it quantifies how much a word's position in semantic space has been shifted over time, which implies how much the word has changed in meaning (Hamilton et al., 2016). When comparing time-aligned word embeddings, by extension, lower cosine similarity across time periods indicates a greater semantic drift.

For this study, cosine similarity is the primary quantitative measure of semantic change. By calculating the cosine similarity between a word's embeddings at different time periods and across different communities, we can quantify the degree of semantic shift and potentially determine whether negative meanings are spreading from community to the broader platform or vice versa.

Nearest Neighbors

While cosine similarity provides a single numeric score that captures semantic change, nearest neighbors vector analysis offers a qualitative insight into how the meaning has shifted. This method identifies which words lie closest to a target word in the word embedding space,

which indicates that the "neighbors" appear in the most similar contexts to the target. Changes in a word's nearest neighbors across time may provide valuable insight into semantic change when it occurs. Nearest neighbors comparison has been used to assess the degree and nature of semantic similarity between words in diachronic word embedding research (Hamilton et al., 2016; Kutuzov et al., 2018).

Nearest neighbors analysis relies on the same assumption Firth formalized as a "distributional hypothesis" of word meaning (see also above; Firth, 1957). Although not the first method of deriving word embeddings based on that assumption, the development of Word2Vec, a machine learning architecture to create word embeddings (Mikolov et al., 2013), rendered the process (and thus nearest neighbors analyses as well) accessible to many more researchers. Additionally, nearest neighbors was later tested as a semantic change metric, showing that shifts in neighbor sets can reveal information about the nature of semantic evolutions (Hamilton et al., 2016).

It is crucial to answer the question of how a word changed in meaning after discovering a change. Cosine similarity discovers the change, while nearest neighbors analysis helps validate and explain the change. With this study, it may not be apparent how the target word changed from the cosine similarity score alone. For example, if a word's nearest neighbors shift from being more positive to negative, this may indicate that the word is acquiring negative connotations—even if the overall magnitude of change (as measured by cosine similarity) is modest. This qualitative dimension complements our quantitative measures and provides insight into the specific nature of semantic shifts.

Related Work

While substantial research exists on toxicity detection and on computational semantic change separately, relatively few studies have attempted to combine both to analyze how word meanings shift toward negativity in online communities. This study aims to contribute to that gap.

Several studies approach aspects of our research question but differ in important ways. Shoemark et al. (2019) analyzed semantic change on X (formerly known as Twitter) using similar word embedding alignment techniques, however their focus was not specifically on negative language or toxicity detection. Rather, they examined semantic change regardless of whether target terms were negative or positive. Rather, they looked to find semantic change in general, regardless of whether their target terms were negative or positive. Another study examined how different Reddit communities develop distinct meanings for words, but they did not study temporal evolution nor focus on negative language or toxicity detection (Del Tredici and Fernández 2018). Research on extremist communities has investigated how coded language evolves to evade detection, but these studies typically examine highly specialized communities and do not compare local usage against broader platform trends (Waseem et al., 2017; Vidgen & Yasseri, 2020).

Our searches did not discover prior work combining the following in a single study:

1. Comparison of sub-community (subreddit-specific) versus global (platform-wide) Reddit semantics
2. Analysis across distinct time periods to track temporal evolution

3. Specific focus on polysemous terms that have existing negative usages alongside neutral ones, making them candidates for toxification over time
4. Use of non-contextual embeddings with Orthogonal Procrustes alignment
5. Integration of both cosine similarity (quantitative) and nearest-neighbor analysis (qualitative)

With this combination of techniques and methodological structure, this study aims to track how negative meanings emerge in specific communities and potentially spread or shrink to broader platform usage. We hope that this study will contribute to the tools for earlier detection of harmful language evolution and deeper understanding of how online communities shape language change.

Chapter 3: Methodology

Overview

This study tracks the potential toxification of ten target words on Reddit between 2010 and 2025 by applying diachronic word embedding analysis across two axes: time and community. The methodology is split into six stages. First, data is collected from archived Reddit corpora and partitioned into defined time slices and community-level datasets. Second, the collected text is preprocessed to remove noise that would misrepresent the corpora as a whole. Third, separate static Word2Vec models are trained on each dataset, which are then aligned using Orthogonal Procrustes so that vectors from different word embeddings can be directly compared. Fifth, cosine similarity is calculated between target word vectors across corpus pairs to quantify the degree of semantic change. Sixth, nearest neighbor analysis is conducted to characterize the nature of any shift. These results are then cross-validated with a sentiment-based basis check using VADER to confirm the direction of detected changes.

The target words used in this study were chosen based on a reasonable likelihood that one or more of their usages had become more toxic between approximately 2011 and 2024. Our research did not find any unambiguous, objective method for assessing whether a word usage has in fact become more toxic over time, and so the judgment of likelihood was subjective by necessity. We selected a limited set of words that trends in popular discourse on online platforms, in popular culture, etc., strongly suggest have become more toxic (negative, pejorative, dismissive, etc., as discussed above), but we note that the selection criteria for these terms necessarily influences our results and the generalizability of our findings.

Dual-Axis Methodological Approach

A central feature of this study's design is that semantic shift is examined along two independent axes rather than one. The first axis is temporal; embeddings trained on corpora from an earlier period (ex: t1, covering 2010 to 2013) are compared against embeddings from a later period (ex: t2, covering 2022 to 2025). This comparison captures how word usage has shifted across more than a decade of Reddit activity.

The second axis is community-level. A subreddit-specific corpus ("local") is analyzed alongside a platform-wide corpus ("global"). The local corpus consists of comments drawn from r/teenagers, which is a large, high-activity subreddit whose users are typically younger, and thus the language tends to adopt internet slang earlier than the broader platform. The global corpus consists of comments drawn from Reddit at large, which represents mainstream platform-wide usage.

Combining these two axes produces four distinct embedding spaces for each target word: t1 local, t1 global, t2 local, and t2 global. The interactions among these four spaces allow the study to address the following:

1. Comparing t1 and t2 within the same community type reveals how the usage has evolved over time
2. Comparing local and global at the same timeslice reveals how much the subreddit's usage diverges from or aligns with the broader platform

3. Comparing how t2 global relates to both t2 local and t1 global reveals whether a word's current mainstream usage resembles its earlier mainstream usage or has converged toward the subreddit's more recent usage.

The latter point would signify a term whose meaning has spread outward from a subcommunity to the broader platform. In addition to this main four-way comparison, year-over-year timeslice analyses were conducted using individual yearly corpora from 2011 through 2024 to track the trajectory of change.

Research Design

Reddit was selected as the data source for several reasons. First, it is one of the largest English-language text platforms in the world, structured around topic specific communities called subreddits, which makes it possible to study both subcommunity and platform-wide language. Existing archival datasets cover Reddit's full history with sufficient density for training word embedding models, and the frequent usage of informal language suit the goals of this research to understand toxification of words on online spaces.

Word2Vec was chosen as the embedding architecture–as a static embedding model, it produces a single, stable vector representation for each word in a corpus, making those representations directly comparable across time periods once alignment is applied. Static embeddings trained on large corpora have been demonstrated to reliably detect semantic shifts across time periods, as highlighted in foundation work on diachronic word embeddings (Hamilton et al., 2016; Kutuzov et al., 2018). They are also computationally tractable at the scale

of tens to hundreds of millions of tokens, and although the usage of contextual word embeddings is not out of the question, it would take significantly greater resources to train contextual embeddings. The skip-gram variant of Word2Vec was used, as it has been shown to learn higher-quality representations for rare words.

The implementation used Gensim's Word2Vec library in Python, with the following parameters: a vector dimensionality of 200, a context window of 7, a minimum word frequency count of 2, and 15 training epochs. The vector dimensionality of 200 was selected as a balance between representational richness and corpus size. Three hundred dimensions was used by Hamilton et al. (2016) on a corpus of billions of tokens, and so using 300 dimensions risked producing sparse representations on our datasets. A window of 7 was used to capture broader topical and semantic associations rather than narrower syntactic meanings, which is necessary for informal social media language where meaningful co-occurance can span several words in comparison to standard language. Minimum count was kept low at 2 to retain early, rare occurrences of slang terms in the t1 period. Raising this threshold would drop the infrequent early usages that our study is interested in tracking. Fifteen epochs provided additional training passes to compensate for the relative compactness of the subreddit-level corpora compared to the size of the global dataset.

Ten target words were selected for analysis: **incel**, **woke**, **based**, **karen**, **sus**, **cope**, **clown**, **bot**, **ratio**, and **radical**. The selection criteria required that each word have both a non-pejorative or neutral usage and a negative, pejorative, or slang usage, with some presentation of the earlier neutral sense in the t1 corpus and evidence for an emerging or established negative connotation by t2. Words were chosen to span a range of "toxification" types: some represent the emergence

of entirely new identity-based terminology (incel, karen), others represent slang repurposing of existing words (cope, clown, ratio), and others were included as partial controls or methodologically informative cases (based, bot) to test the sensitivity of the method to shifts of different magnitudes and directions.

Data Collection

This study uses the Pushshift Reddit dataset, an archival collection of Reddit comments spanning from June 2005 to December 2024, accessed via Academic Torrents. Pushshift systematically preserved Reddit's public comment history through continuous API ingestion, and its longitudinal coverage makes it well suited for the diachronic analysis this study requires. We chose this dataset over direct, more targeted collection of data via the Reddit API for two reasons. First, Reddit's 2023 API policy changes substantially restricted access to past comment data, making it no longer feasible to retrieve years of comment history through the API. Second, the Pushshift dataset already contains verified, complete data spanning the full study period, which direct collection could not have provided.

We developed custom Python scripts to extract relevant comments from the dataset and produce structured output for word embedding training. We constructed two distinct corpora: “global” and “local”. The global corpus consists of a 0.1% random sample from over 10 billion comments drawn from Reddit at large, without restriction by subreddit. This sampling rate was chosen to yield a corpus large enough for reliable Word2Vec training while being computationally feasible to process in a reasonable timeframe. The local corpus consists of all

comments from r/teenagers without sampling. r/teenagers describes itself as a community run by and targeted at users between the ages of 13 and 19 (r/teenagers, n.d.), making its user base younger than the platform average on Reddit as a whole. Because younger users tend to adopt internet slang earlier than broader online populations (Fortuna & Nunes, 2018), this community serves as a plausible leading indicator of emerging changes and toxification, especially in informal language, allowing us to examine whether new or pejorative word meanings appear there before spreading to the broader platform.

The study period spans 2011 through 2024, 2011 representing the first year in which r/teenagers produced sufficient comment volume for reliable Word2Vec training, and 2024 representing the most recent year for which complete archival data was available. We note that the Pushshift dataset contains Reddit data going back to 2005, and studying a different subreddit with earlier activity could extend the study period further back in time, but doing so was out of the scope of this study. Dividing each corpus into slices by calendar year then produced 28 subcorpora in total: one local and one global subcorpus for each calendar year from 2011 through 2024. A separate Word2Vec model was trained on each corpus, yielding 28 independently trained embedding spaces.

Following that training, we conducted two types of comparisons for each of our target words. The first compares word embeddings between local and global subcorpora from the same year (2011_local vs 2011_global), measuring how closely r/teenagers' usage of a word aligns with platform-wide usage at any given point in time. The second compares words' embeddings between subcorpora from the same community (local or global) over time by examining adjacent years (2011_local vs 2012_local and 2011_global vs 2012_global), tracking how word usage

evolves incrementally within the two contexts. Together these comparisons allow the study to track both how individual words evolve over time and how subcommunity usage diverges from or converges toward the broader platform across the study period.

To protect user privacy, author usernames and post identifiers were anonymized during data extraction. Each Reddit post ID was replaced with a 16-character hash derived from the original ID using SHA-256, and each author username was mapped to a consistent pseudonym of the form user_ followed by 10 hexadecimal characters, also derived via SHA-256. This anonymization is deterministic, meaning the same original author maps to the same pseudonym throughout the dataset, but irreversible, ensuring that no real usernames are retained in any of the processed files used for analysis.

Text Preprocessing

After data collection, we preprocessed the extracted comments to remove non-semantic characters that would otherwise inflate the model vocabulary without contributing meaningful distributional information, and to normalize token forms across corpora. Preprocessing involved three operations applied in sequence: lowercasing all text, removing characters outside a defined set of alphanumeric and Latin-extended Unicode characters, and tokenizing the resulting strings by splitting on whitespace. Lowercasing ensured that capitalization differences did not result in duplicative vocabulary entries. For instance, Karen and karen are treated as the same token. The character filtering step removed punctuation, mathematical symbols, and formatting characters that appear frequently in Reddit text but carry no semantic content relevant to word meaning for our purposes. Latin-extended characters were retained to accommodate non-English content

present in the global corpus. Tokenization produced the list-of-tokens format required by Gensim's Word2Vec implementation.

An important methodological distinction between our preprocessing approach and similar approaches in the literature is that we did not collapse inflected word forms through lemmatization or stemming. Lemmatization, which maps words to a common base form such as reducing “bullying” and “bullied” to “bully”, is common in NLP tasks where recognizing morphological variants as equivalent improves generalization. We chose not to apply lemmatization for two reasons. First, in internet language specifically, inflected and variant forms of a word frequently carry meaningfully different connotations (Fortuna & Nunes, 2018) . The internet slang usage of a term often manifests in specific forms, such as a nominalization, a deliberate respelling, or a construction unique to a particular meme, that would be collapsed into a base form shared with the word's conventional usage, which may risk obscuring the type of semantic shift/toxification this study focuses on. Second, retaining original word forms ensures that high-frequency function words remain present in every model's learned vocabulary. These words serve as the anchor set for the Orthogonal Procrustes alignment described in the following section, and their absence from the vocabulary would make alignment impossible.

Word Embedding Creation

A separate Word2Vec model was trained independently on each subcorpus. Training models independently on each subcorpus rather than continuously or jointly is consistent with the approach recommended by Kutuzov et al. (2018) for diachronic comparison–this makes sure

that each model captures the distributional semantics of its specific corpus without contamination or influence from other time periods or communities. The trained models were not pre-initalized from existing general purpose embeddings because training from scratch on the target corpora ensures that the resulting vectors reflect the specific vocabulary and usage patterns of Reddit rather than those of whatever corpus a pre-trained model was derived from.

Each model was trained using Gensim's Word2Vec implementation with the skip-gram architecture, which predicts surrounding context words from a target word. Skip-gram is advantageous over the continuous bag-of-words for this project because it produces higher quality representations for infrequent words; notably, several target words (incel, karen) rarely appear in the earlier corpus in t1. The context window of 7 means the model considers up to seven tokens on either side of each target word during training, capturing board topical associations appropriate for the informal language of Reddit discourse.

Word Embedding Alignment

Word embeddings trained on separate corpora cannot be directly compared without alignment. Because the training process involves initialization using random values, independent models will place the same word at entirely different positions in their respective vector spaces as a result, even if the word's usage contexts are identical across both corpora. The vector spaces are rotationally arbitrary; any rotation of the entire space produces an equally valid solution to the training objective, and independent training runs will choose different rotations. Because

direct comparison of unaligned vectors would produce meaningless results, there is a need for an alignment to allow for comparison.

Orthogonal Procrustes analysis addresses this issue by computing the optimal rotation matrix that maps one embedding space to another. Given two matrices X and Y containing vector representations of a shared vocabulary, Orthogonal Procrustes finds the orthogonal matrix R that minimizes $\|XR - Y\|$. Note that an orthogonal rotation preserves all angles and distances within the embedding space being rotated, which ensures that the internal structure of the source embedding (including the relative positions of all the word vectors) is unchanged by the alignment. Only the orientation of the space changes, not the relationships within it.

This alignment process relies on anchor words. These are a set of terms assumed to have stable meanings across the corpora being compared, which will serve as a reference point for commuting the rotational matrix. High frequency function words such as “the”, “and”, “of”, and “a” are the standard choices for anchor words, since their meaning is unlikely to shift significantly over relatively short time periods. The rotation matrix computed from the anchor words is then applied to all words in the source embedding, including the targeted words of interest. This means that the quality of alignment depends on the anchor word set being both stable and sufficiently large to calculate an accurate rotation. For the corpora used in this study, the intersection of vocabularies across corpus pairs ranged from several thousand to tens of thousands of shared tokens, providing a robust anchor basis.

The Orthogonal Procrustes computation was implemented using SciPy's orthogonal_procrustes function. After computing the rotational matrix from the anchor word matrices, the rotation was applied to all vectors in the source embedding space, producing a

rotated version of the earlier or subreddit-specific model that occupies the same coordinate system as the reference model. Then, the cosine similarity and nearest neighbor analyses were conducted on these aligned representations.

Cosine Similarity

Cosine similarity measures the cosine of the angle between two vectors. Typically, this yields a value between -1 and 1 where values closer to 1 indicate that the vectors point in the same direction. In the context of word embeddings, greater directional similarity between two word vectors is taken as a proxy for greater similarity in the distributional-semantic contexts in which the word appears, which implies similarity in meaning. Because the word vectors in our research are L2-normalized within the embedding training process, all cosine similarity values fall in the range [0, 1].

For each target word in each corpus pair, cosine similarity was computed between the word's aligned vector from the earlier/source corpus and its vector in the later or reference corpus. A value close to 1 indicates that the word occupies nearly the same position in both embedding spaces after alignment, reflecting stable usage across the two corpora. A value substantially below 1 indicates that the word's "neighborhood" has shifted, meaning that it appears in different contexts. For the purposes of this study, we consider values above 0.8 as stable with no meaningful shift, values between 0.6 and 0.8 to indicate minor semantic drift, and values between 0.4 and 0.6 to indicate a moderate and meaningful shift. We treat values below

0.4 as indicative of a major semantic shift where the word's meaning (again in distributional-semantic terms) has changed substantially between the compared corpora.

Cosine similarity was computed across all four primary pairings:

1. t1 local vs t2 local (semantic change over time within r/teenagers)
2. t1 global vs t2 global (semantic change over time across Reddit generally)
3. t1 local vs t1 global (community divergence at the earlier time point)
4. t1 local vs t2 global (dual-axes comparison)

The 4th comparison is of interest–if t2 global is more similar to t1 local than to t1 global, it suggests that global Reddit's current usage of a term has converged toward the subreddit's earlier usage, consistent with the meaning having spread outward from the subcommunity to the broader platform. Cosine similarity calculations were also conducted for each consecutive year-pair in the timeslice analysis to track the gradual trajectory of change per year.

Nearest Neighbors

While cosine similarity provides a quantitative measure of how much a word's distributional position has changed, it does not reveal what kind of change occurred. A word whose cosine similarity drops from 0.85 to 0.35 between t1 and t2 has clearly shifted, but the score alone cannot indicate whether it acquired a negative connotation, a neutral slang meaning, a technical meaning, or something else entirely. Nearest neighbor analysis helps to address this by identifying which words lie closest in the embedding space to a target word, which helps to characterize the "semantic neighborhood" in which it lies.

For each target word in each aligned embedding space, the ten nearest neighbors were identified by computing cosine similarity between the target word's vector and the vectors of all other words in the vocabulary, then ranking by similarity and selecting the top ten. The neighbor lists from different corpus pairs were compared qualitatively to determine whether the change in cosine similarity corresponded to the acquisition of a negative or pejorative meaning. If the neighbors in t2 were seen to be generally more negative, insulting, or slang-adjacent than those in t1, the change was deemed likely to be in a negative (toxifying) direction. Conversely, if a word tended to be used in the same context as overall more positive words or neutral words, the shift was considered likely to be positive, or neutral/unrelated to toxification if neutral and/or simply different.

This qualitative examination of neighbor lists is one established method of diachronic word comparisons using embeddings, such as in Hamilton et al. (2016), which demonstrated that changes in nearest neighbor sets reliably reflect genuine semantic change. In our study, neighbor lists serve both as a check on the cosine similarity results and as the primary source of interpretive evidence about the direction of any detected shift. For target words showing major cosine similarity drops, the neighbor lists provide an indication of what the word has come to mean in one or more usages–for example, by confirming that a word like "Karen" previously mostly used as a neutral personal name has "changed neighborhoods". If most uses of "Karen" were as a relatively neutral personal name, we would expect the nearest neighbors of "Karen" to also be personal names and/or other terms not necessarily used in contexts that include negative language. If the usage of "Karen" becomes more toxic, we would expect the nearest neighbors of "Karen" to include fewer personal names (if it is used more and more frequently as an epithet

instead) and potentially more negative terms or terms used in contexts that suggest they are more toxic than neutral as well..

Sentiment Validation

To supplement the findings from the word embeddings and nearest neighbors analysis, and to provide an independent basis for validating the direction of detected shifts, a sentiment-based analysis was conducted using VADER (Valence Aware Dictionary and sEntiment Reasoner) (Hutto & Gilbert, 2014). For each target word and each corpus, all sentences containing the word were extracted and scored using VADER's compound sentiment metric, which assigns values between -1 (maximally negative) and +1 (maximally positive). The average compound score for all sentences containing the word in each corpus was then compared across corpus pairs, and the change in average sentiment between the two corpora was noted.

VADER was selected for this role for two reasons: first, it was developed and validated for informal, social media style text, making it well suited for Reddit data, and second, it operates "off the shelf", or without any training on the target corpus, which makes its output fully independent of the Word2Vec models. A shift labeled as NEGATIVE by the VADER analysis indicates that sentences containing the target word became, on average, more negatively charged in the later of two corpora compared. Use of a term in a sentence with overall negative sentiment/valence does not *guarantee* that one or more usages of the term itself have become more toxic over time, but in aggregate, we consider changes sentence sentiment a plausible point of supporting evidence of whether the word is used more frequently in hostile, derogatory, or

high conflict contexts. VADER scores entire sentences instead of the target word in isolation, which means that (for example) a sentence containing “woke” could be classified as NEGATIVE because of other negative words, rather than “woke” itself. This is a known limitation, but sentiment scores are nevertheless informative–if the sentences in which a word appears have become more negatively scored on average, that supports the word’s context of use being shifted in a negative direction. We selected a shift threshold of 0.10 on the compound scale to represent a meaningful directional shift in sentiment. The value is arbitrary and intended to reduce the potential influence of variations in sentiment due to the sometimes hyperbolic and informal discourse of Reddit (unrelated to toxicity), and future work would be required to validate it empirically. In sum, convergence between major cosine similarity drop and a VADER detected negative shift provides stronger evidence of toxification, and an overall positive shift in context sentences’ sentiment suggests the opposite may be true.

To complement VADER's lexicon-based approach to sentiment classification, we additionally scored all extracted sentences using a DistilBERT sentiment classifier, specifically distilbert-base-uncased-finetuned-sst-2-english, a transformer model fine-tuned for binary sentiment classification. To prepare the data for this analysis, the preprocessed corpus CSV files were read in chunks and each comment's text was passed through spaCy's sentence segmentation model, which splits raw comment text into individual sentences. Each sentence was then checked for the presence of the target keyword, and only sentences containing the keyword were retained for scoring. This filtering step ensured that the sentiment scores reflect the specific contexts in which each target word appears rather than the overall tone of the comment as a whole.

The matched sentences were then passed to the DistilBERT classifier for GPU inference, which generates a probability of positive/negative sentiment for each sentence based on its full content as a whole as compared to DistilBERT's training data. Sentences were processed in large batches to take advantage of GPU throughput across the volume of text in all corpora. Each sentence received a severity classification based on its negativity probability: high (at or above 0.80), medium (at or above 0.50), low (at or above 0.20), and minimal (below 0.20). The average negativity probability for each target word in each corpus was then compared across corpus pairs in the same manner as the VADER scores.

To analyze trends across the full study period, the sentence-level scores from both tools were aggregated by year across all local and global corpora. For each year, the average VADER negativity score and average transformer negativity probability were computed across all matched sentences, weighted by sentence count so that years with more data contribute proportionally more to the trend. These yearly averages were saved for both the local and global corpora and combined into a single dataset for comparison. To visualize the trajectories, grouped bar charts were produced for each metric, displaying local and global yearly averages side by side with a weighted linear regression line overlaid on each group. This allowed us to examine not only whether sentences containing a target word became more negative over time, but whether that trend differed between the subcommunity and the broader platform, and whether the two converged or diverged across the study period.

Chapter 4: Results

Overview

Across the ten target words, the analyses reveal three broad patterns connected to subsets of the targets. The first group consists of *incel*, *karen*, *sus*, *bot*, and *clown* which show major or moderate cosine similarity drop between the earlier and later corpus periods, supporting a substantial shift in meaning. For these words, the nearest neighbor list confirms that the nature of the change is meaningful–in general, the words *incel* and *karen* appear to move from proper-name neighborhoods into clusters of negative, identity-based vocabulary. *Sus* appears to have shifted as well from Spanish usage into being used more frequently in gaming slang contexts centered on Among Us. *Clown* shifted from literal carnival usage into general-purpose insults. *Bot* moved from subreddit specific automation references to generic automated vocabulary.

The second pattern relates to *woke*, *based*, *cope*, *ratio*, and *radical*, and shows moderate to minor cosine similarity drops, with the degree of the shift varying between r/teenagers and Reddit as a whole. For these words, the nearest neighbors suggest that new meanings have developed and become more common, but perhaps not dominant; they appear to be used in parallel with others rather than completely replacing the earlier usage(s).

The third pattern, visible across the timeslice analyses, appears to show a period of rapid change in the mid 2010s followed by stabilization. In the year-to-year comparison, change in most words' meanings appears to decrease significantly after 2018 to 2020, suggesting that some form of semantic stability may have been achieved around that time.

Table 1 below shows cosine similarity values for all four primary corpus comparisons alongside the VADER based sentiment shift direction for each word. Cosine similarity values are interpreted as follows: values at or above 0.80 indicate stable usage with no meaningful distributional shift; values between 0.60 and 0.80 indicate minor drift; values between 0.40 and 0.60 indicate a moderate and substantively meaningful shift; and values below 0.40 indicate a major shift in which the word's distributional neighborhood has changed substantially. Table 2 presents the top five nearest neighbors for each word in each corpus.

Word	t1L → t2L	t1G → t2G	t1L → t1G	t1L → t2G	VADER Local Shift	VADER Global Shift
incel	N/A*	0.369	N/A*	N/A*	Unknown	Negative
woke	0.651	0.618	0.830	0.576	Negative	Minimal
based	0.747	0.884	0.850	0.806	Negative	Minimal
karen	0.363	0.391	0.373	0.371	Negative	Negative
sus	0.233	0.431	0.325	0.391	Positive	Minimal
cope	0.613	0.701	0.745	0.635	Negative	Minimal
clown	0.495	0.674	0.533	0.434	Negative	Minimal
bot	0.299	0.358	0.482	0.225	Minimal	Positive
ratio	0.440	0.745	0.590	0.550	Negative	Minimal
radical	0.499	0.794	0.490	0.463	Negative	Minimal

Table 1. Cosine similarity scores across all four primary corpus comparisons and VADER sentiment shift direction. t1L = r/teenagers 2010–2013; t2L = r/teenagers 2022–2025; t1G = Reddit global 2010–2013; t2G = Reddit global 2022–2025. N/A = word absent from t1 local corpus.*

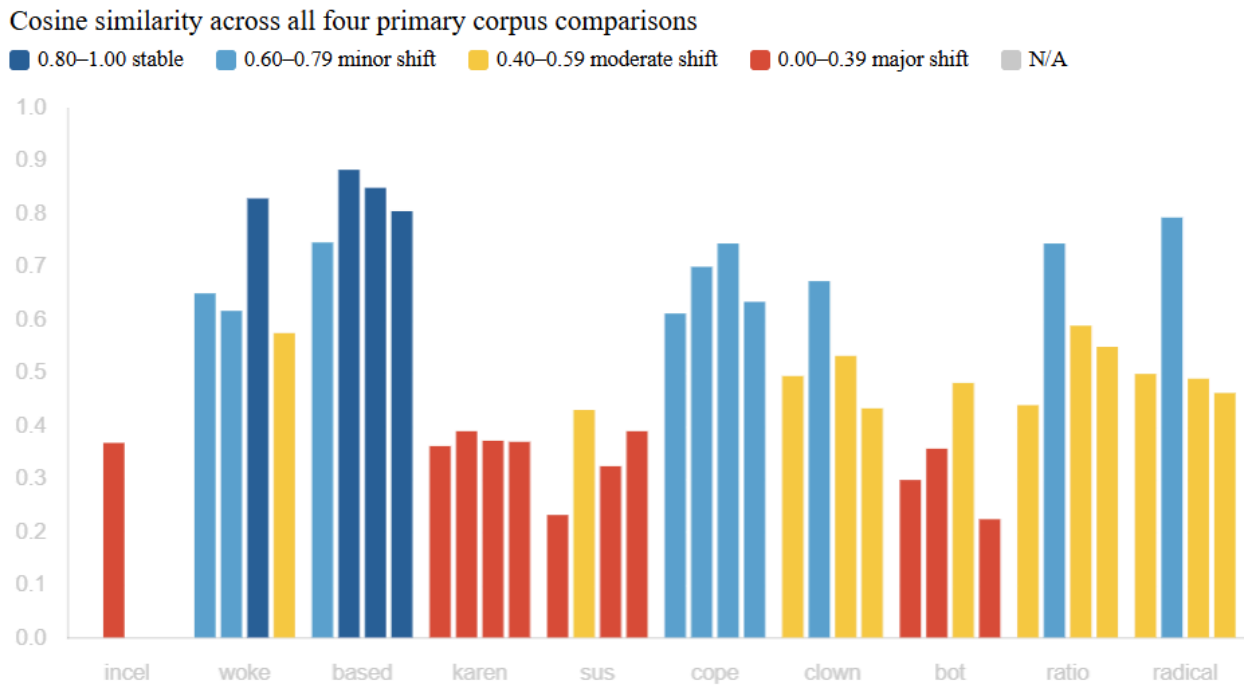

Figure 1: Cosine similarity across all four primary corpus comparisons

Cosine similarity scores for all ten target words across the four primary corpus pairings. t1L = r/teenagers 2010–2013; t2L = r/teenagers 2022–2025; t1G = Reddit global 2010–2013; t2G = Reddit global 2022–2025. Color bands correspond to the four interpretive categories defined in Chapter 3.

VADER shift labels reflect the direction of the average compound sentiment delta between the two corpora being compared (local: t1L vs. t2L; global: t1G vs. t2G). Negative indicates the word appeared in more negatively valenced sentences in the later corpus; Positive indicates the reverse; Minimal indicates a delta below the 0.10 threshold.

Cosine Similarity Results and Comparison

Comparing t1 local to t2 local and t1 global to t2 global reveal that the degree of semantic change varies widely across words and differs between the subcommunity and the platform at large. The local corpus shows greater volatility for most words–seven out of the 10 targets have lower cosine similarity in the t1L-t2L comparison than in the corresponding t1G-t2G comparison. This pattern is consistent with r/teenagers adopting or developing new slang usages earlier or more intensely than broader Reddit, although the magnitude of the difference varies.

The words with the most pronounced shifts in the global corpus are *incel* (0.369), *bot* (0.358), and *karen* (0.391), all of which fall in the major shift range. These three words share the characteristic that their dominant usage in 2010 to 2013 was different from their 2022 to 2025 usage. We speculate that *incel* may have been initially a marginal term with no stable semantic neighborhood and that *bot* likely referred to specific subreddit-level automation scripts. *Karen* appears clearly to have been used predominantly as a personal name. By 2022–2025, each term had a new and distinct primary usage. *Radical* (0.794), *based* (0.884), and *woke* (0.618) show the smallest shifts in the global corpus, though for potentially different reasons. Based on the nearest neighbors, *radical* and *woke* both appear to retain substantial portions of their earlier usage, while *based* is discussed separately below as a case where multiple meanings appear to suppress the measured shift.

In the local corpus, *sus* (0.233) and *karen* (0.363) show the largest shifts, followed by *bot* (0.299). The *sus* cosine similarity in the local comparison is the lowest of any word in the dataset. Based on the nearest neighbor analysis (see below), the shift is likely the result of the coincidence between the popularization of *sus* in Among Us gaming slang as shorthand for

"suspicious" and the spelling of *sus*, a possessive adjective ("your", as in "your friends") in Spanish.

The cross-community comparison at t1 (comparing r/teenagers against global Reddit in the 2010 to 2013 period) provides a baseline measure of how different the subcommunity's vocabulary was from the platform at the earlier time point. For most words, the t1L to t1G cosine similarity is moderate, suggesting that the two communities used these words in relatively similar ways in 2010 to 2013. Notable exceptions are *radical* (0.490) and *sus* (0.325), both of which showed substantially different distributional patterns between local and global even in t1. In the case of *radical*, the local corpus shows a different neighborhood that diverges from the global corpus's more conventional political framing. In the case of *sus*, the difference reflects the Spanish language skew of both local and global usages, which were not aligned with each other.

The comparison between t1L and t2G captures the difference between how r/teenagers used a word in 2010–2013 and how most of Reddit used it a decade later. If a word's meaning spread from the subcommunity outward to the broader platform, we would expect t2G to show greater similarity to t1L than t1G does to t2G; that is, the later platform-wide usage should be closer to the earlier subcommunity usage than it is to the earlier platform-wide usage. For *karen* and *clown*, the t1L–t2G cosine similarity (0.371 and 0.434, respectively) falls in a comparable range to the t1G–t2G value, suggesting the shift in the global corpus is largely independent of the local corpus's early usage. For incel, t1L–t2G cannot be computed because the word is absent from t1 local entirely. The clearest case where t2G usage has moved in a direction consistent with subreddit-originating vocabulary is *cope* and *radical*, where the t2 global neighbor lists contain terminology that is more prominent in the local corpus's vocabulary than in the earlier

global usage. These findings suggest that the global shift appears to have occurred alongside or parallel to the local shift rather than as a direct consequence of it.

Nearest Neighbor Results and Comparison

Word	T1G Neighbors	T2G Neighbors	T1L Neighbors	T2L Neighbors
incel	sodini, americanly, welltravelled, scienceatheism, saulpaul, galphanore, mangia, motherfuker, niemoller, actressactor	incels, misogynist, misogynistic, redpill, sjw, femcel, femaledatingstrategy, manosphere, terf, neckbeard	N/A	N/A
woke	wake, waking, woken, morning, wakes, partyingdrinking, ambian, awoke, hyperventilated, wokeup	wake, waking, woken, umaleficentsclone, pser, jathas, 4305am, hemorrhoidectomy, lockem, orbed	woken, wake, 500am, couchbut, morning, 1140pm, 4pm, waking, 45am, 530am	woken, wake, waking, freethem, eyepads, mxster, yagirl, morningin, 1140am, saladed
based	basing, solely, predicated, depend, relies, relying, dependent, gtbased, rely, depending	basing, dependent, depending, depend, depends, purely, relying, funvibes, predicated, idealogies	basing, solely, afflicting, dichotomies, gtinitial, depending, metrics, lenders, dramatize, holistically	basing, solely, genotypephenotype, purely, sigmal, premised, unfathomaly, isgreen, dadpilled, relying
karen	gillan, decrow, exnun, gadbois, thelma, mccaffrey, katheryn, schwimmer, firths, mmmmmmhm	sarah, ashley, karens, kristen, straughan, lauren, oberman, indistinctly, sheila, douchebag	gillan, aniston, brie, lerman, winstead, alexa, cumberdale, kathy, isabelle, johanson	filippelli, filipelli, fukuhara, mcbitchface, gillan, becky, antilullaby, karens, renjith,

				emily
sus	nadie, alguien, tienen, tener, gente, cosas, aquí, política, todos, algunos	los, propias, importan, suswhereany, secundarias, brinda, hijas, conciudadanos, compinches, oberto	encuentro, ido, estado, cual, dieresis, natzi, trato, espana, inglés, decir	sussy, suspicious, susss, amogussus, wherewhere, sususususus, suss, wherewherewhere, suswhereany, susred
cope	empathize, overcome, autismaspergers, coping, dissonnance, deal, communicate, dyspraxia, handle, dealing	coping, empathize, suicidial, struggle, sympatize, bealive, seethe, pajet, overcome, informationurl	coping, coped, depression, philiophobia, anxiety, angerstress, selfhate, dealt, agression, grieve	coping, seethe, copey, jimgtgym, seithe, copes, cooe, slill, americancant, situaitons
clown	ronalds, clowns, woopity, shticks, costume, kinko, facepaint, gtfans, richkid, smileand	moron, clowns, loser, lmao, lmfao, jackass, idiot, dumbass, douchebag, bafoon	globs, pagliacci, bodysuit, likejust, sissy, pikaaaaachuuuu, hotboxing, greaser, teethe, roaches	shaggy2dope, clowns, possie, posse, siah, juggalette, circus, boppo, gtfuckin, clownin
bot	bot-faqurl, videoinfo, permabaned, experimenturl, scootacheerheres, amp91did, creatorurl, faqurl, problemurl, devianturl	action, performed, am, goodluck12, goodluck51, jerksessions, goodluck80, automatically, removed, linkurl	transcriptions, bgimgurl1, bgimgurl2, bgimgurl3, bgimgurl4, bgimgurl5, plaintext, quickmeme, instantreddart, bgimgurl6	action, performed, automatically, kirukato, removed, rbotsleuthbot, sploogem, am, posthttpurl, httpurl
ratio	ratios, debttoincome, stacktopot, trollnorma, cokerum, 702010, strengthweight, pricebook, waisttohip,	ratios, capequity, capincome, domint, 058g, stopspread, proteinfatcarb, powertoweight, waisttohip, amylopectin	squaring, 253, avocados, 13135, molar, centimetre, 2535, 9998, measurement, femalemale	ratios, famuly, eatio, teenagerpedo, 69420f, slill, 715x, amongsus, killdeath, upvotetocomment

	loantovalue			
radical	extremist, feministic, conservative, hegelians, islamism, gtfundamental, rootstrikers, anti1, antimilk, southernstrategyurl	liberal, progressive, extremist, conservative, ideology, burkean, progressives, conservatism, statist, neoliberal	marisa, extremist, authoritarianlibertarian, antistate, anticapitalist, centreleft, nonhateful, aligns, centrist, nonpartisan	extremist, leftist, gtleftist, extremists, antiright, leftists, liberal, conservative, farright, progressives

Table 2. Top ten nearest neighbors for each target word in each corpus. Neighbors are derived from the aligned embedding spaces and reflect the words most frequently appearing in similar contexts to the target word.

The nearest neighbor lists provide qualitative evidence about the direction and nature of the shifts quantified by cosine similarity. As with cosine similarity measures, several distinct patterns emerge across the ten words with respect to their "neighborhoods".

The clearest evidence of toxification appears with *karen* and *incel*. In both t1 corpora, *karen* functions as a personal name, with its nearest neighbors consisting of other female personal names and actresses (gillian, aniston, brie, lerman, decrow). By t2, the global corpus's neighbors include *karens* , *douchebag*, *straughan*, and *pearlclutcher*, which suggest that the term is much more closely associated with the pejorative social stereotype (i.e., *pearlclutcher*) than before. The t2 local corpus reflects a more intensified version of the shift, with neighbors including *mcbitchface* and *filipelli*, where terms such as *mcbitchface* suggest the pejorative usage is more entrenched in the subcommunity. For "incel", the t1 global corpus shows no clear theme in its semantic neighborhood–the neighbors are usernames and account names (*sodini*,

americanly, *scienceatheism*), which indicate the negative sense of the word was uncommon at best. By t2, its global neighbor set is focused on misogynistic vocabulary: *incels*, *misogynist*, *misogynistic*, *redpill*, *sjw*, *femcel*, *manosphere*, and *neckbeard*.

The changes in *clown*'s neighborhoods illustrate a different type of shift. In t1, both the local and global corpora show a semantic neighborhood appearing to relate mostly to entertainment or clown-like characters, with *ronalds* (Ronald McDonald), *costume*, *facepaint*, *kinko*, and circus-adjacent terms. In t2 global, this neighborhood is replaced with general-purpose insult vocabulary such as *moron*, *loser*, *lmao*, *jackass*, *idiot*, *dumbass*, *douchebag*, with only *clowns* (plural) surviving from the previous neighborhood. The t2 local corpus shows a different trajectory, with the neighbors clustered around Insane Clown Posse fan culture (*shaggy2dope*, *juggalette*, *possie*) alongside circus terms, suggesting that the pejorative usage is more prevalent in the global corpus than the local one for this word. This is one of the few cases in which the shift toward a negative usage appears stronger in the platform-wide corpus than in r/teenagers.

For *cope* and *radical*, the neighbor lists may show semantic expansion rather than clear replacement–the earlier usage is not displaced, but rather supplemented by new senses. The t1 local corpus places *cope* in a mental health context–*coping*, *depression*, *anxiety*, *philiophobia*, *selfhate*. By t2 local, the mental health vocabulary persists (*coping* remains the closest neighbor) but is joined by *seethe*, *cope*, *seithe*, and *copes. This may reflect* a shift in usage toward the sense of dismissal, as when someone is told to "cope and seethe" in response to a perceived grievance, but does not evidently suggest toxification. The t2 global corpus shows a similar expansion, where *coping* and *empathize* remain, but *seethe* enters the neighborhood. This pattern of

coexistence between the older and newer meanings implies that the cosine similarity score (0.613 local, 0.701 global) likely could underestimate the extent of adoption of new usage(s), where the newer word embedding may have been "pulled" in more than one direction simultaneously. For *radical*, the t1 local neighborhood is based in an anarchist and anti-state vocabulary (*antistate*, *anticapitalist*, *authoritarianlibertarian*), while t2 local and both global corpora seem to have converged toward conventional partisan-political vocabulary (*liberal*, *progressive*, *conservative*, *ideology*, *extremist*), suggesting possible normalization of the word's political usage across both communities over time.

The results of *woke* and *based* show one methodological limitation of static word embeddings and nearest neighbor analysis when applied to terms with many possible meanings. For *woke*, the nearest neighbors in both t1 and t2 across all four corpora are dominated by sleep-related vocabulary (*wake*, *waking*, *woken*, *morning*). The political sense of *woke*, which became much more prominent by 2022 to 2025, does not appear to dominate the neighborhood in either the local or global corpus, which may simply be because the sleep-related usage of "woke" is continuously relevant and very common. The Word2Vec model produces a single vector that averages all usages, and the politically charged sense may be in the minority of occurrences, meaning that the averaged vector might shift less than might be expected intuitively when judging by how prominent newer, more toxic usages "feel" based on trends in popular discourse. The VADER analysis nevertheless provides corroborating evidence that something has changed (and become more negative) with respect to *woke*, despite the moderate cosine similarity scores–sentences containing *woke* in t2 local corpus are more negatively valenced on average compared to those in t1 local (delta = -0.129), which would be expected if the word increasingly appears alongside political criticism, mockery, or hostile rhetoric even if the sleep

usage remains frequent. The case of *based* is similar: the neighbor lists across all comparisons appear most related to "standard" English usage of *based* as a predicate adjective (*basing*, *solely*, *depend*, *predicated*) with only occasional traces of the internet slang sense in t2 local. The globally stable cosine similarity (0.884) reflects the dominance of the conventional usage, even if the slang sense could be well established in specific subcommunities.

The results for *sus* are distinct from all other words in the dataset in that the shift does not appear to involve the acquisition of negative meaning. In the t1 period across both corpora, *sus* appears overwhelmingly in Spanish contexts, with the nearest neighbors consisting of Spanish pronouns, determiners, and common nouns (*nadie*, *alguien*, *tienen*, *tener*, *gente*, *cosas*). This reflects the incidence of Spanish speaking users on Reddit whose comments contain *sus* as a Spanish possessive adjective ("your"). In the t2 local corpus, this usage has been almost entirely overshadowed by its use in contexts related to Among Us gaming slang, the neighbors being *sussy*, *suspicious*, *amogussus*, *susss*, and *wherewhere*, likely reflecting the word's viral adoption following the game's peak popularity in 2020. The t2 global corpus does not show the same degree of difference, as Spanish neighbors persist, which accounts for the higher cosine similarity in the global comparison (0.431) relative to the local one (0.233). The VADER shift is labelled positive in the local comparison, reflecting the fact that the Spanish contexts in the t1 local were negatively weighted, while the Among Us gaming contexts in t2 local are more neutral. This word functions as a methodological reference case demonstrating that word embeddings can encode rapid community specific semantic change even when the shift involves no toxification.

The results for *bot* and *ratio* both show major cosine similarity drops apparently driven by changes in conventions rather than cultural or social meaning shifts. For *bot*, the t1 corpora show neighborhoods that are dominated by subreddit specific automation bot references (long URL strings pointing to individual bot FAQ pages, transcription bot references, and mirror bot documentation), which reflect the specific infrastructure of Reddit in the early 2010s. By t2, the neighborhoods in both corpora converges on generic automation vocabulary (*action*, *performed*, *automatically*, *removed*, *am*). The VADER analysis labels this shift as positive in the global corpus, which, although difficult to interpret, does not evidently clash with a change in context from platform-bot interactions to neutral automated-moderation vocabulary. For *ratio*, the t1 local corpus neighborhood skews toward a mathematical and demographic usage, where the nearest neighbors are *squaring*, *centimetre*, *femalemale*, *percentage*, while the t2 local corpus shows the internet slang usage of *ratio* as the measure of social media response disparities, the neighbors including *upvotetocomment*, *killdeath*, and mostly memetic tokens. In comparison, the t1 global corpus shows a financial and scientific mathematical usage that persists largely unchanged into t2 global (*debttoincome*, *pricebook*, *capequity*, *amylopectin*), producing the notably higher cosine similarity in the global comparison (0.745 versus 0.440 for local).

Year-over-year Timeslice Results and Patterns

The year-over-year timeslice analysis, conducted for each consecutive annual pair from 2011–2012 through 2023–2024, reveals the trajectory of change for each word. A consistently high year-over-year cosine similarity indicates that a word's meaning in the vector space has a consistent pattern. A sequence of low year-over-year scores followed by rising scores is

indicative of a word that underwent a period of active semantic change before settling into a new usage pattern.

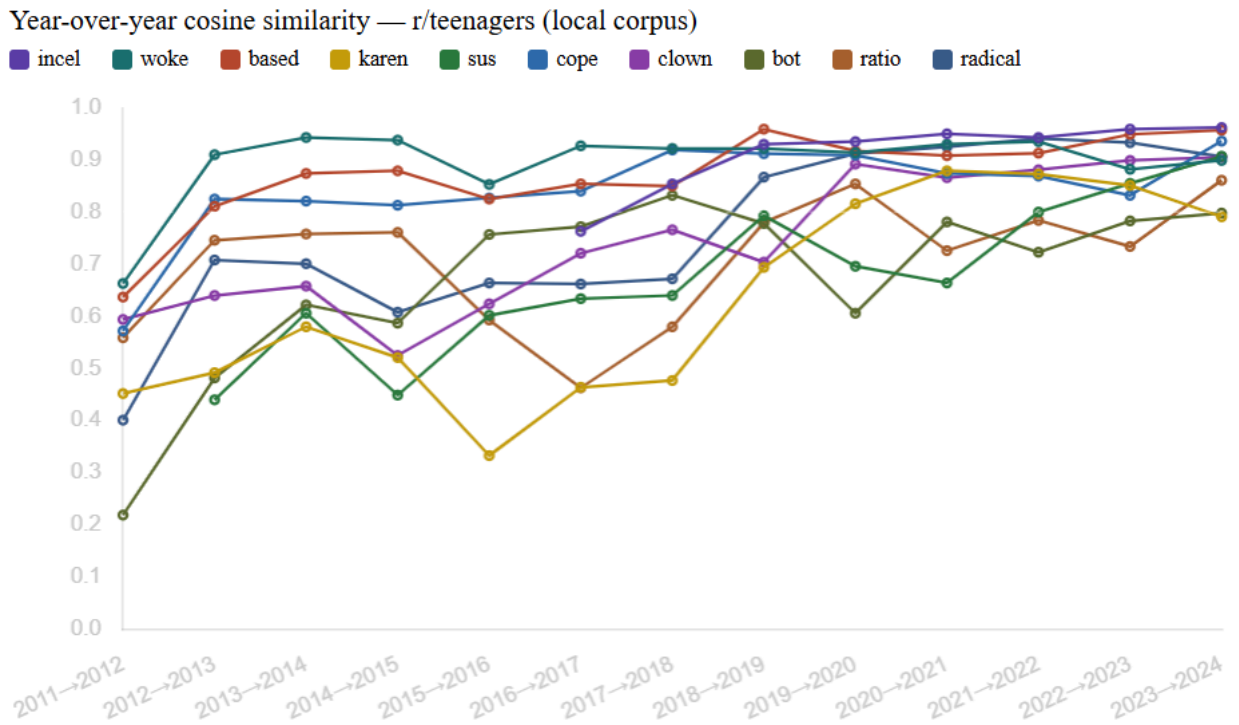

Figure 2: Year-over-year cosine similarity – r/teenagers (local corpus)

Year-over-year cosine similarity for each target word in the r/teenagers corpus, 2011–2024. A consistently high value indicates stable usage from one year to the next. Words absent from a given year are omitted from that point. The dashed line marks the 0.80 stability threshold.

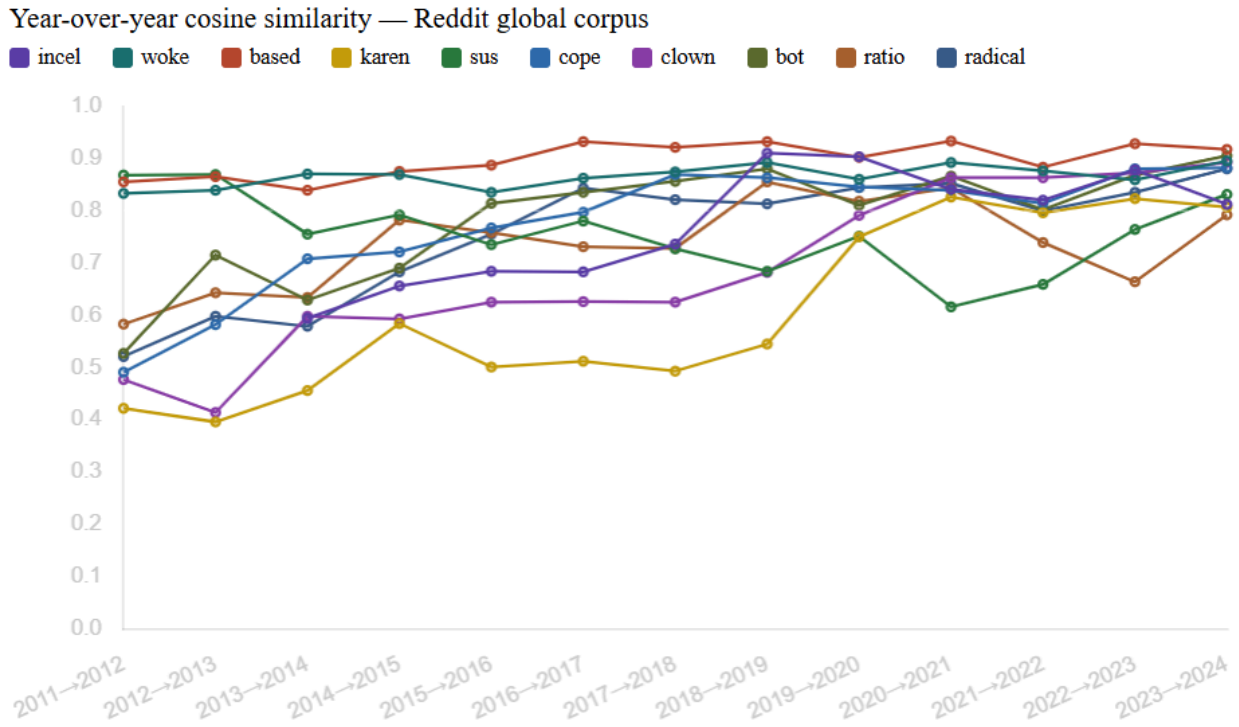

Figure 3: Year-over-year cosine similarity – Reddit global corpus

Year-over-year cosine similarity for each target word in the Reddit global corpus, 2011–2024. Patterns of initial instability followed by stabilization indicate periods of active semantic change before a new usage consolidated.

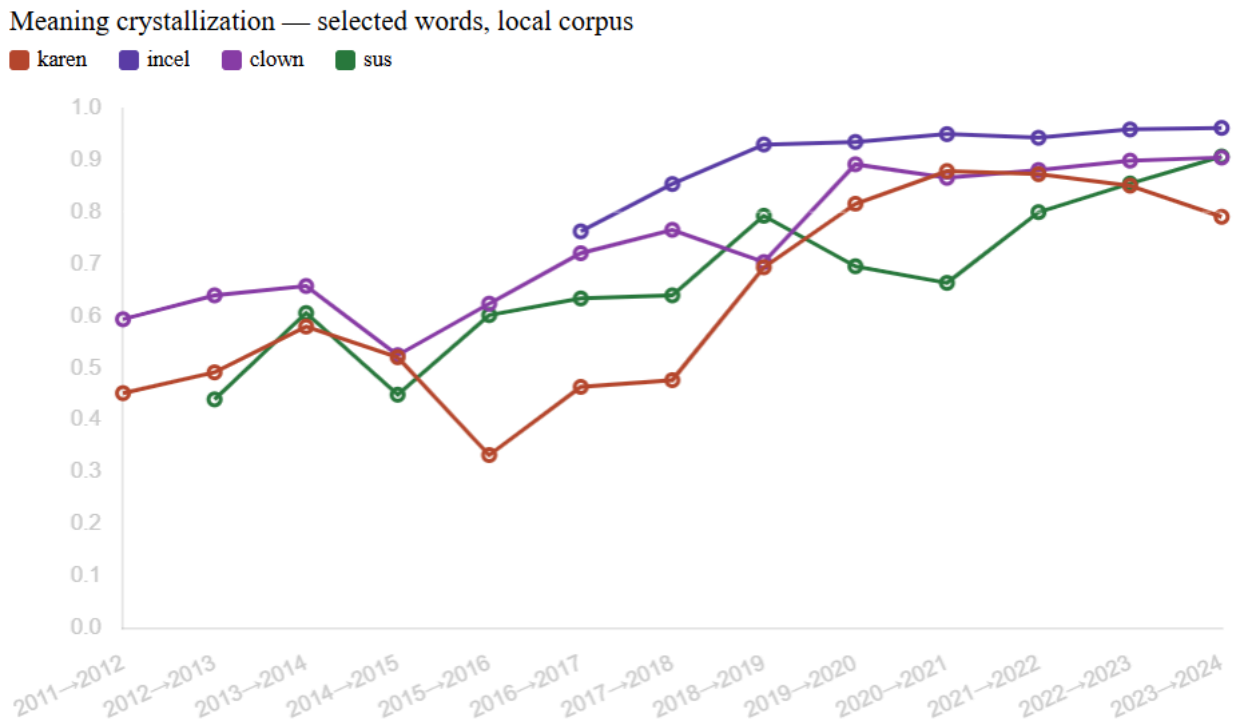

Figure 4: Meaning crystallization – selected words, local corpus

Year-over-year cosine similarity in the r/teenagers corpus for four words demonstrating the crystallization pattern: an extended period of below-threshold instability followed by sustained stabilization. The dashed line marks the 0.80 stability threshold.

Karen shows the most pronounced version of this pattern. Year-over-year cosine similarity in the local corpus averages 0.508 across the first three timeslices (2011–2014), reflecting unstable and inconsistent usage across years, before rising to an average of 0.838 in the final three timeslices (2022–2024). In the local corpus, consecutive-year similarity first crosses and sustains above 0.80 in the 2019–2020 transition, and in the global corpus this stabilization occurs somewhat later, in the 2022–2023 transition. Speculatively, this timing seems

to align with the "cultural moment" in which the term *karen* entered popular discourse as a recognizable social archetype. The delayed stabilization in the global corpus relative to the local corpus is consistent with the word being adopted and consolidated in subcommunity usage before spreading to the broader platform.

Incel first appears in the r/teenagers timeslice data at the 2016 to 2017 transition, where the year-over-year cosine similarity is 0.763, which shows that it entered the subreddit's vocabulary with a relatively stable meaning. This score in the local corpus rises to 0.954 across the final three timeslices, which suggests that the term's meaning did not change. In the global corpus, *incel* first appears in the 2013 to 2014 timeslice with a cosine score of 0.594, showing that the meaning in the global corpus is less stable initially, which then changes to a consistent high-stability above 0.80 from 2018–2019 onward.

Clown and *sus* both show low year-over-year stability in early timeslices, which is then followed by increased stability for seemingly different reasons. For *clown* in the local corpus, the average cosine similarity across the first three timeslices is 0.631, and this rises to 0.895 in the final three, with stability first sustained above 0.80 in the 2019 to 2020 transition–the period in which using *clown* as a general-purpose insult perhaps became a standard internet convention. For *sus*, the meaning stabilizes in the local corpus and sustains above 0.80 only in the 2021 to 2022 transition, possibly reflecting the post-2020 stabilization of the Among Us gaming sense. The global corpus shows a different pattern for *sus* where the stability was already high in the earlier timeslices (0.868 for 2011 to 2012), potentially because the Spanish possessive adjective usage was consistent across years, and there is no clear step-change in the global series associated with the gaming sense.

Among the words showing the least change over the full period, *woke* and *based* both had minimal change and maintained high stability throughout the years. *Woke*'s local year-over-year similarity averages 0.839 in the first three timeslices and 0.905 in the last three; based averages 0.774 and 0.939 respectively. These patterns are consistent with the interpretation that the dominant usage of these words (sleep-related for woke, predicate-adjective for based) remained stable throughout the study period at the level of embeddings, even as culturally different meanings developed.

Local versus Global Community Comparisons

The year-by-year local-versus-global comparisons, which measure how closely r/teenagers' usage of each word aligns with platform-wide Reddit usage in the same year, reveal a general pattern of convergence for several key target words over the study period, with notable differences in the rate and extent of convergence across words.

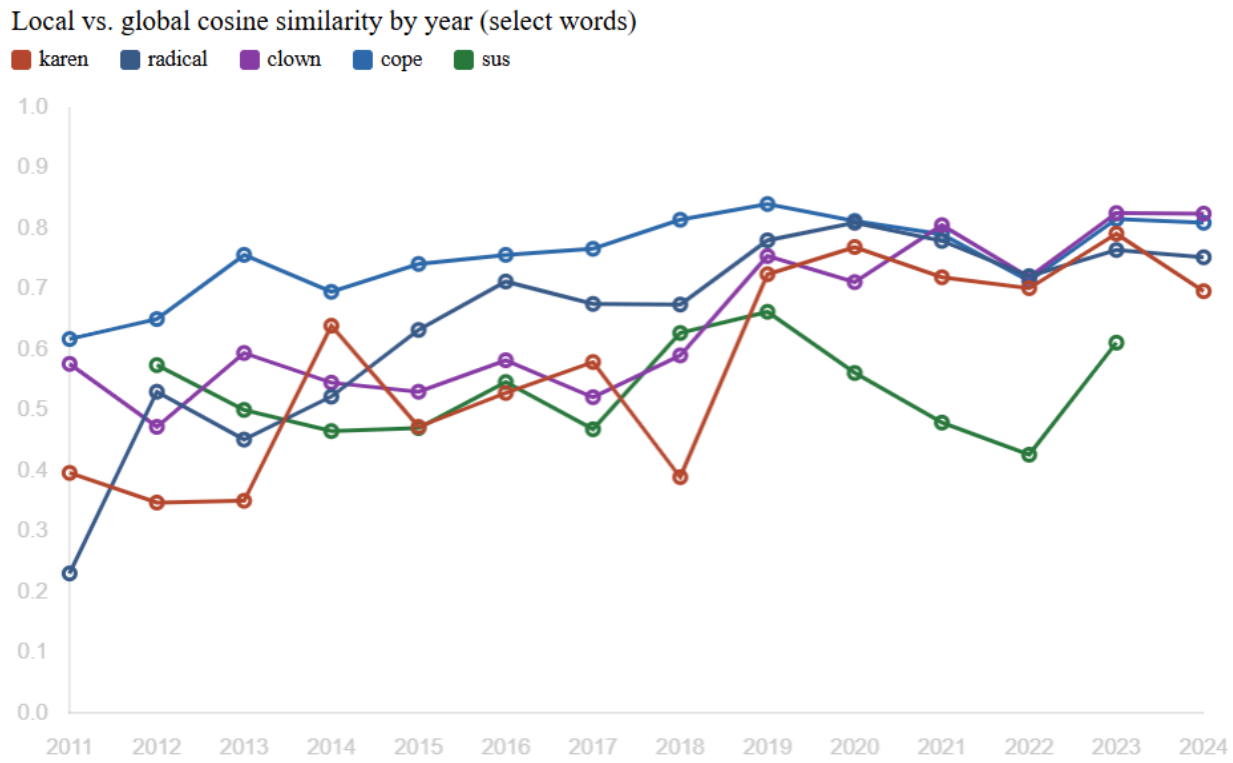

Figure 5: Local vs. global cosine similarity by year (select words)

Annual cosine similarity between the r/teenagers (local) and Reddit global corpora for the five words showing the most notable convergence or divergence trends. Rising values indicate local and global usage becoming more similar over time.

Karen shows the most substantial convergence, with local versus global cosine similarity averaging 0.365 in the first three years of the study (2011–2013) and 0.730 in the final three (2022–2024), a difference of +0.365. In the early periods, the two communities use *karen* primarily as a personal name but with different associations: r/teenagers' neighborhood reflects actress and celebrity culture, while the global corpus reflects different sets of public figures. By the later time period, both communities have converged toward the pejorative sense, which

caused a higher similarity. *Radical* shows a comparable convergence trajectory (+0.342), moving from a state in 2011 where local and global usages were quite different (0.231, with local anchored in anarchist vocabulary and global in mainstream political extremism) to broad alignment by the early 2020s as both converged on conventional partisan framing.

Clown also shows meaningful convergence (+0.242), likely reflecting the adoption of the insulting usage in both communities. However, the global corpus's neighbor lists suggest the pejorative usage may have been adopted in global Reddit at least as early as in r/teenagers–the t2 global neighbors (*moron*, *loser*, *jackass*, *dumbass*) are more insult-adjacent than the t2 local neighbors (*shaggy2dope*, *clowns*, *posse*), suggesting the shift in r/teenagers has been shaped in part by ICP fan culture specific to that subcommunity.

Woke, *based*, and *cope* show relatively stable local-versus-global similarity, with no strong convergence or divergence trend. For *woke* and *based*, this stability likely reflects the persistence of dominant non-slang usages in the embeddings for both communities. For *cope* in particular, the local-versus-global similarity ranges from 0.618 to 0.841, with a small upward trend (+0.105), consistent with being somewhat more commonly used in the dismissive sense in both communities at similar rates.

Bot and *sus* show consistently low local-versus-global similarity, averaging 0.481 and 0.525 respectively in their early years and showing no substantial convergence over time. For *bot*, this reflects the consistently different vocabulary in the two communities. For *sus*, local and global communities seem to diverge: the global corpus has not adopted the Among Us gaming sense to the same degree as r/teenagers, keeping the local-versus-global similarity moderate at best.

Sentiment Validation

The VADER-based sentiment analysis provides an independent signal for several of the shifts identified through cosine similarity and nearest neighbor analysis. For the two words showing the strongest evidence of toxification (*incel* and *karen*), the global corpus VADER results both show a negative shift direction, meaning the sentences in which these words appear became more negative between t1 and t2. The delta for "incel" is -0.277 (from +0.261 in t1 global to -0.015 in t2 global), and for *karen* it is -0.277 (from +0.292 to +0.016). Both represent relatively large shifts from positive or neutral sentence contexts to near neutral average sentiment, consistent with the transition from mostly personal name usage in neutral/varied conversational contexts to more frequent usage in negative contexts more likely to be associated with arguments, complaints, and online conflict.

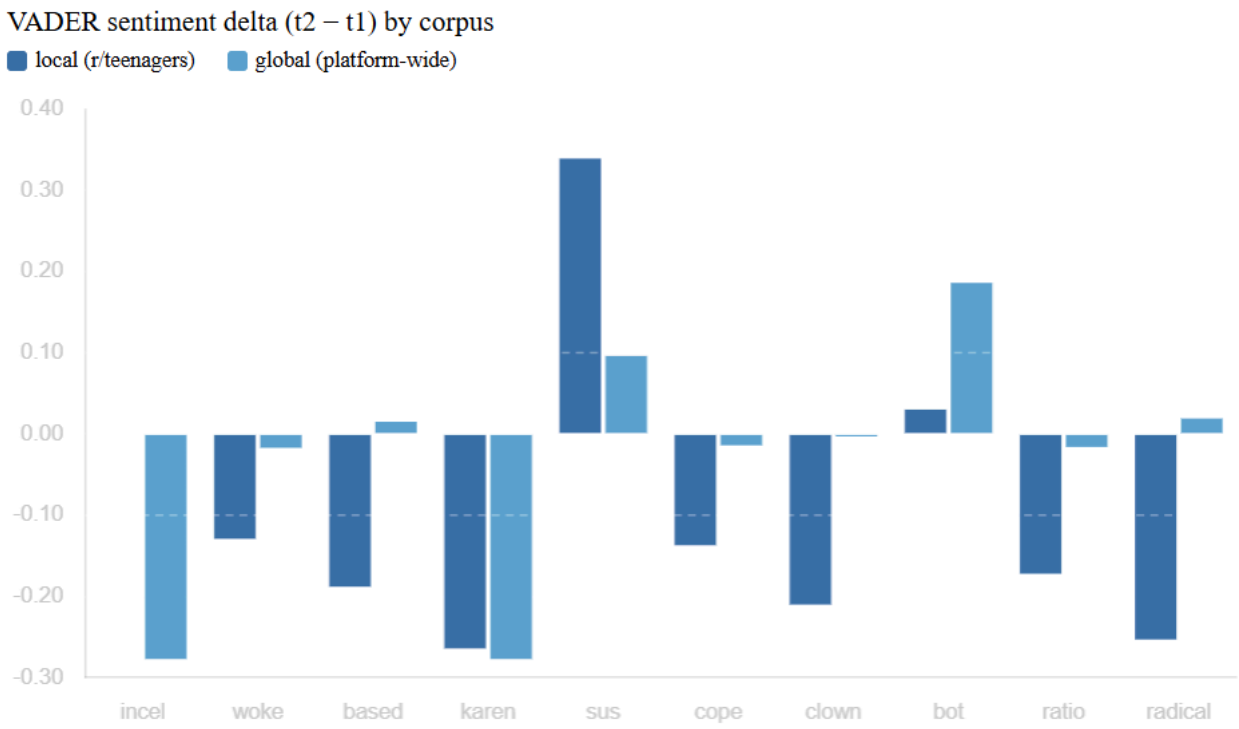

Figure 6: VADER sentiment delta (t2 - t1) by corpus

VADER compound sentiment delta (later corpus minus earlier corpus) for each target word in both the local and global corpora. Negative values indicate sentences containing the target word became more negatively valenced in the later period. The dashed line marks the ±0.10 significance threshold.

The local corpus VADER results show stronger and more widespread negative shifts than the global corpus results, with seven out of the nine words for which t1 local data is available showing negative or positive directional shifts. Words showing negative shifts in the local corpus include *woke* (-0.129), *based* (-0.188), *karen* (-0.264), *cope* (-0.137), *clown* (-0.210), *ratio* (-0.172), and *radical* (-0.253). The global corpus shows negative shifts for *incel* and *karen* only,

with most other words remaining in the minimal range. This divergence between local and global results is consistent with the interpretation that semantic change in the direction of negativity has been more pronounced in r/teenagers than across Reddit as a whole, but it may also reflect the VADER analysis's sensitivity to sentence context, which could amplify shifts in subcommunity language where the informal register makes tonal shifts more legible to the VADER lexicon.

Two words show positive VADER shifts. *Sus* shifts positively in the local corpus (+0.339), which could be interpreted as consistent with a shift from Spanish-language contexts of varying sentiment (since *sus* is a term likely to occur in many contexts) toward the Among Us gaming usage, where average sentiment may be less varied and more neutral or positive due to the game-focused contexts. *Bot* shifts positively in the global corpus (+0.186), consistent with the normalization of bot vocabulary from adversarial or disruptive early automation contexts to routine automated-moderation language. In the early years of Reddit, automated accounts were frequently associated with spam, vote manipulation, and other unwanted platform behavior, meaning sentences containing the word bot were more likely to appear in negative or combative contexts. By the later period, *bot* had shifted towards routine moderation language, producing a more neutral surrounding sentiment. Interestingly, both of these cases underscore the need for careful consideration of the meaning of any shift, since the pipeline detects semantic change in any direction, and also demonstrate that not all words in the study underwent the type of toxification to which we hypothesized they might have been subjected.

Taken together, the VADER results provide meaningful supporting evidence of toxification for words whose embedding changes (and nearest neighbors) also show the clearest relevant shifts (*incel*, *karen*, *clown*, and the local corpus results for *woke*, *cope*, and *radical*). For

words where embedding similarity changes are moderate and polysemy appears to be influencing the shift strongly (*woke*, *based*), the VADER analysis indicates a non-negative directional shift, which is consistent with the observed neighbor list changes and suggests that “more negative” results from sentiment analysis may in fact be reflective of toxification.

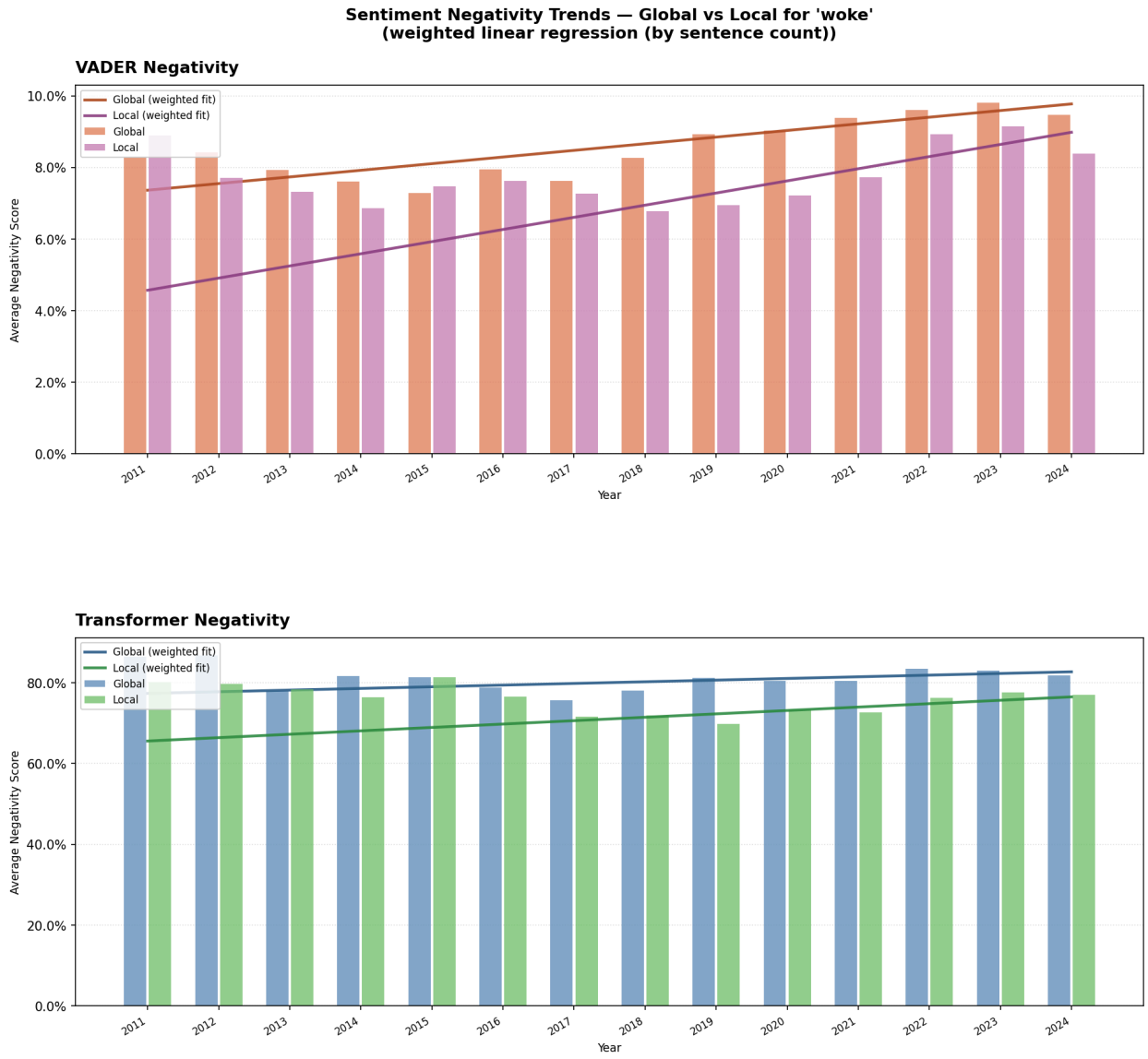

Figure 7: Sentiment Negativty Trends – Global vs Local for 'woke'

(weighted linear regression (by sentence count))

Bars show the change in average sentence sentiment for each word between the earlier and later corpus. The line shows the weighted average negativity trend over time, where years with more data contribute more to the trend.

Our additional sentiment analysis using DistilBERT alongside VADER served as an independent check on the direction of the sentiment shifts detected by VADER. For *woke*, the two tools show broadly consistent findings, with both detecting negative trends in the global and local corpora across the study period, which strengthens confidence in the VADER results and supports the general idea that sentiment analysis coupled with word embeddings may produce usable signals with respect to toxification. The sentiment trends for *woke*, shown in Figure 7, reveal two distinct patterns across the two tools. The VADER negativity scores show a clear upward trend in both corpora across the study period. Sentences containing *woke* in the global corpus rise from approximately 8% average negativity in 2011 to approximately 10% by 2024, while the local corpus follows a similar but steeper trajectory, rising from approximately 4% in 2011 to around 8% by 2024.Weighted linear regression lines, fitted to the yearly averages with each year weighted by its sentence count so that years with more data contribute proportionally more to the trend, confirm a positive trend in both corpora, indicating that sentences containing woke became incrementally more negative over time on both the platform as a whole and in r/teenagers. This is consistent with subjective observations of the word increasingly appearing in politically charged, critical, or hostile contexts over the study period, even though its neighborhood in the word embedding analysis remained anchored in sleep-related vocabulary throughout.

The DistilBERT negativity scores for *woke* tell a different story. Both corpora maintain high baseline negativity throughout the study period, with the global corpus averaging around 80% negativity score throughout and local corpus averaging around 65% in 2011 and increasing to 80% in 2024. Unlike the VADER scores, both DistilBERT regression lines are relatively flat, indicating that contextual negativity was already elevated at the start of the study period and

remained stable rather than increasing over time. In both measures, the global corpus scores higher than the local corpus across nearly every year in the study period. The difference between sentiment analysis results using different tools suggests that, while the use of sentiment analysis in supporting/analyzing toxification appears sound conceptually, model choice, training data, and other considerations should be carefully studied.

Chapter 5: Discussion

Interpretation of Results

This study determined whether static word embeddings, aligned using Orthogonal Procrustes, can detect and characterize semantic change (primarily toxification) in terms associated with negative or pejorative usage on Reddit, and whether any of those changes show evidence of diffusing from a subcommunity to the broader platform. Our findings partially support our hypothesis that static word embeddings can be used for this purpose (in combination with other tools) while also revealing meaningful limits in what static embedding methods can observe.

The most direct evidence that our methodology can detect toxification comes from *karen* and *incel*. Both words underwent major shifts (cosine similarity scores of 0.391 and 0.369 in the global comparison, respectively) and the nearest neighbor lists provide strong qualitative evidence of the direction of change. *Karen* transitioned from a personal name whose neighborhood consisted of other female names and celebrity references into a neighborhood organized around a social archetype and its associated negative vocabulary. *Incel*, which was absent from r/teenagers in 2010 to 2013 and uncommon in the global corpus, entered a neighborhood of misogynist vocabulary by 2022 to 2025. The VADER analysis corroborates both cases, showing the negative sentence sentiment changes of -0.277 in the global comparison for each word, meaning that the sentences in which these terms appeared shifted substantially toward negativity. The convergence of word embedding evidence (indicating change occurring), neighborhood analysis (suggesting toxification), and sentiment evidence (supporting the finding

of toxification) for these two words represents the clearest demonstration that the methodology is capable of detecting what we hypothesized it might be used to detect.

A second group of words–*clown*, *ratio*, and *radical*–show moderate to major semantic change and significant toxification , and support the idea that our methodology could be used to understand *patterns* in toxification/the spread of toxic uses as well. For *clown*, the shift from a literal semantic neighborhood (*ronalds*, *costume*, *facepaint*) to an insult-adjacent one (*moron*, *loser*, *jackass*, *dumbass*) is relatively consistent across the local and global corpora, with the changes in the global corpus appearing "complete" than in the local corpus by t2. This suggests that the pejorative usage of *clown* as a general purpose dismissal may have spread broadly across Reddit rather than outward from a single "source" subreddit like r/teenagers. For *ratio*, the potential shift in toxicity specific to r/teenagers: movement from the mathematical usage to slang usage (*upvotetocomment*, *killdeath*, *amongsus*) within that local community suggests a possibly different pattern of toxification, while the global corpus's cosine similarity (0.745) indicates that mathematical usage remains dominant platform-wide. This divergence between local and global could indicate that the more pejorative sense of ratio became more prevalent only in communities of younger or more internet native users.

The results for *woke*, *based*, and *cope* show a specific limitation of static word embeddings when applied to words with multiple meanings. Static word embeddings represent individual words using one vector, and therefore imply limitations with respect to polysemous words, but our findings suggest that this limitation could be particularly acute for potentially "toxifying" words that acquire a new sense without it overshadowing earlier or other meanings. For all three words, cosine similarity scores are moderate rather than major, with *woke* at 0.618

and 0.651, *based* at 0.884 and 0.747, and *cope* at 0.701 and 0.613 for global and local, respectively. The nearest neighbors of each word suggest that toxification may be occurring to an extent, but earlier/other usage(s) remain(s) prevalent enough to pull the averaged word vector towards its original position, which then reduces the potential signal of toxification provided by a static word embedding vector. For *woke*, sleep related neighbors (*wake*, *waking*, *woken*) dominate the neighbor lists in both t1 and t2 despite our expectation that the political sense of the word would have become (subjectively, judging by popular discourse) significantly more prominent. For *based*, neutral, expected neighbors (*basing*, *solely*, *depend*) dominate throughout. For *cope*, mental health vocabulary (*coping*, *depression*, *anxiety*) persistently appears among nearest neighbors, though seethe enters the t2 local neighborhood, probably due to *cope*'s increasing usage in a dismissive sense. The VADER analysis shows negative shifts in the local corpus for all three terms (*woke*: -0.129; *based*: -0.188; *cope*: -0.137), indicating that even where the averaged vector has not moved dramatically, the sentences containing these words have become more negatively charged. This pattern is consistent with the idea that new, more negative usages may represent a growing proportion of the terms usages, even if they do not yet dominate the overall distribution, and illustrate that our method can produce suggestive results that are nevertheless difficult to interpret.

The sentiment trend analysis for *woke* offers two additional observations that may help resolve some ambiguities in patterns of semantic change, neighbor change, and sentiment change, although further analysis would be necessary to draw strong conclusions. First, the high baseline negativity scores from 2011 onward may reflect something about the word's non-toxic, "conventional" usages rather than any emerging toxification. Waking up and being woken up are experiences that reddit users may describe negatively in general, influencing the scores, meaning

sentences containing "woke" in its sleep-related sense would naturally score as negative under both tools from the very beginning of the study period. This suggests that elevated baselines of that type could be used to temper interpretations of other evidence of political toxification, firstly. Secondly, attention to phenomena such as the consistently lower sentiment scores in the local corpus compared to the global corpus across both tools might signal the presence of other potentially confounding factors that could be studied and controlled for, perhaps including demographic differences between the two communities. r/teenagers skews toward a younger user base that engages less with political discourse than the broader Reddit platform, meaning that notionally charged or hostile politically-oriented usages of "woke" that might drive higher negativity scores in the global corpus could be less likely prevalent in r/teenagers.

Sus and *bot* are the two words from our experiment that produced results most clearly departing from our hypothesis. Both show major cosine similarity drops with *sus* at 0.233 in the local comparison, *bot* at 0.299 and 0.358, but neither change is in the (expected) negative direction. For *sus*, the shift is from Spanish possessive adjective usage to Among Us gaming slang, and the VADER results are positive in the local corpus. For *bot*, the changes in usage and sentiment appear to reflect a kind of normalization of automated-moderation vocabulary, with VADER showing a positive global shift. These two cases demonstrate that our methodology must be refined (such as by filtering source documents by language) to account for unforeseen factors in the data, that it is sensitive to meaningful distributional change regardless of the direction, and that the set of ten target words selected ultimately included examples of multiple types of semantic change, only some of which probably involve toxification.

Relationship to the Study's Hypotheses

The study's first hypothesis–that terms' toxification on Reddit can be detected and characterized based on a foundation of static word embedding analysis coupled with additional techniques–recieves partial support from our results. Taken together, careful consideration of results stemming from methods such as ours (embedding similarity, nearest neighbors, and sentiment scores) show promise as a means of identifying and understanding toxification in online discourse. For five out of the ten words selected for experimentation (*incel*, *karen*, *clown*, *cope locally*, and *radical* locally), there is evidence from both cosine similarity drops and VADER sentiment deltas of an expected shift toward toxification. For two words (*woke* and *based*), word embedding differences and nearest-neighbor differences indicate change, and sentiment scores provide evidence of negative sentiment drift, but our results are at least partially ambiguous and suggest that our embedding-based approach could underestimate semantic change/toxification especially in cases of strong polysemy. For two words (*sus* and *bot*), the changes are major but non-negative, and for one word (*ratio*), the shift is local specific with no corresponding global change. The pattern of changes we observe through our method is thus not uniform across the target words; this is simultaneously a powerful indicator that our methods could (with refinement) be used to uncover and explore meaningful differences in the process of toxification, where there is strong evidence of it, and that they likely do not account for enough of the many possible sources of variation in the factors affecting toxification for those methods to be used uncritically.

One aspect of our hypothesis–that this methodology can effectively measure and quantify semantic *shifts* in words' prevalent usages over time on Reddit–recieves stronger support. For all ten words, the combination of cosine similarity, nearest neighbor analysis, and VADER basis check provide a coherent and reasonably interpretable method to assess whether meaningful changes in a term's typical usages occur. Our timeslice analysis demonstrates that the methodology can detect not only that change occurred but approximately when it occurred, with the stabilization of year-over-year cosine similarity providing a flag of when (a) new usage(s) seem(s) to become dominant. The cases we interpret to be strongly influenced by polysemy (*woke*, *based*, *cope*) reveal a meaningful limitation of the approach for words that acquire new senses without losing old ones, but they also show that sentiment analysis in conjunction with our other methods can serve as a complementary signal of meaningful change in usage. The methodology, taken as a whole, is capable of detecting and characterizing the kind of semantic change it was designed to track, with the caveats described below.

Comparison to Existing Literature

The findings of this study are broadly consistent with several lines of prior work in diachronic word embedding research. Shoemark et al. (2019) validated word embedding alignment techniques for social media text on shorter time spans than those used in historical corpora research, demonstrating that the methodology can detect semantic change reliably in the noisier, more informal language of online platforms. Our results support that finding on the relatively short time scale of our Reddit data and suggest that the embedding alignment approach produces coherent and interpretable results in this context. The identification of meaningful

neighbor-list transitions–such as the progression from personal names to social-archetype vocabulary for *karen*, or from circus imagery to insult vocabulary for *clown*–demonstrates that the approach succeeds at the qualitative characterization task Shoemark et al. described as well.

Del Tredici and Fernández (2018) showed that different subreddit communities develop distinct word meanings, and that these local meanings can diverge substantially from platform-wide usage. Our study finds evidence of this phenomenon in the t1 local versus t1 global comparisons: for *radical*, the local corpus in 2010–2013 was anchored in anarchist and anti-state vocabulary that differed from the global corpus's mainstream political framing. For *ratio*, by t2 the local corpus had adopted an internet-slang usage that appears to remain marginal in the global corpus. These results are consistent with Del Tredici and Fernández's finding that subreddit-level communities can function as semi-autonomous linguistic environments. Our research adds a temporal dimension to that finding: tracking these divergences and convergences over time, and examining whether subcommunity-specific usages propagate outward, adds a layer of analysis that prior work on within-platform semantic variation did not address.

Research on the evolution of coded or hateful language in extremist communities, including work on how specific terms are repurposed to evade content moderation (Waseem et al., 2017; Vidgen & Yasseri, 2020), has consistently found that semantic change in toxic vocabulary is rapid and community-driven. Our results are consistent with this broader literature: the words that showed the clearest negative shifts (incel, karen, clown) all moved from relatively neutral or contextually unremarkable usages in the early 2010s to heavily charged vocabulary in the early 2020s (a relatively rapid change), in a pattern consistent with subcultural language becoming mainstream internet discourse. However, the present study makes a methodological

rather than conclusive contribution; it demonstrates that this change likely can be tracked quantitatively rather than only described qualitatively or detected retrospectively after terms have already become widely recognized as problematic.

Implications of Results

The findings of this study have implications in several areas. For the field of computational linguistics, the results provide evidence that static word embedding alignment, even with its limitations, can serve as a useful tool for tracking semantic change and toxification in informal online text. The combination of cosine similarity and nearest neighbor analysis strengthens this finding: our results suggest that cosine similarity scores do provide a meaningful, quantitative signal for the degree of change in a term over time, while nearest neighbors provide qualitative evidence to support that finding as well as to characterize the direction and nature of the change. The addition of sentiment analysis as an independent variable to corroborate qualitative findings of toxification (or the absence thereof) addresses a specific gap in embedding comparisons/cosine similarity metrics: namely, their inability to indicate the meaning of any changes they reveal.The year-over-year timeslice findings have potential implications (with refinements, as discussed above) with respect to online content moderation. First, our method may contribute to understanding patterns in toxification that could help predict toxification and/or differing need for content moderation at different “stages” of a particular toxification trend’s evolution. The year-over-year cosine similarity data shows that for several words, a period of low and unstable consecutive-year similarity preceded the emergence of a stable new meaning. This pattern could serve as an early-warning signal for terms that are in the

process of acquiring new, potentially harmful meanings, before those meanings have become widespread enough to be flagged by other toxicity detection methods, such as those relying on lexicons of already-recognized toxic language. The case of *karen* supports this; year-over-year stability was low in the 2011 to 2019 period and first crossed the 0.80 threshold in the local corpus at the 2019 to 2020 transition. It is plausible (though not tested by our study) that this kind of pattern of change could signal toxification before a term becomes commonly recognized as an insult or problematic social archetype. Whether a monitoring system built on this approach would be practically feasible at scale is a question for future work, but the results suggest the underlying signal is present in the data and discoverable through our methods.

Our observation of divergence between local and global results for several words also has implications for how platforms might monitor language evolution. The finding that certain usages appear in subcommunity corpora before or more intensely than in platform-wide corpora suggests that the subreddit-level monitoring may provide an earlier indication for emerging vocabulary shifts–our study shows this pattern with *ratio*, *radical*, and *cope*. However, given our results for *clown*, where the insult usage appears more prevalent in the global corpus than in r/teenagers, we note that it is probably not the case that all toxification originates in subcommunities before entering the mainstream. Some shifts appear to develop in parallel across the platform, or may originate in different communities than those under study, and the exact mechanisms governing toxification are an intriguing area for future study, potentially supported by the type of analyses we have performed.

Limitations of Study

Several limitations of the study's design and execution are important to acknowledge. The most significant methodological limitation is the use of static, non-contextual word embeddings. Word2Vec produces a single vector representation for each word in its vocabulary, collapsing all usages of a word and any distinctions in context between those usages into a single average position in the embedding space. For words that have acquired a new sense without losing earlier ones (part of the polysemy problem for *woke*, *based*, and *cope*) this averaging means that the measured cosine similarity shift could be small relative to a (notional) expectation of significant change accompanying a new usage, because the old sense (and its contexts) may strongly influence the word's embedding. Sentiment analysis might partially compensate for this limitation, but a more complete approach–perhaps involving contextual embeddings from transformer-based models such as BERT, which generate different representations for the same word depending on its surrounding sentence–might add valuable information. However, the computational cost of training or tuning models (if necessary) on corpora of the size used here is significant and beyond the scope of this study, and interpretation of contextual embeddings presents challenges beyond the scope of this study as well. Notably, contextual embeddings for one word differ based on complex interactions with every other word in a text, and the sense of the differences are not transparently interpretable. Nevertheless, an updated approach using contextual embeddings might prove powerful if studied carefully.

A second limitation concerns the scope of the data. This study draws exclusively from Reddit, which means that cross-platform dynamics between platforms such as X, 4chan, or TikTok into Reddit, or the reverse are not observable. The study can establish that a term's usage

changed within Reddit and can compare subcommunity usage against platform-wide usage, and we suggest that our approach is capable of providing valuable information about term toxification in principle, but we cannot be certain it will produce useful results on all platforms. Similarly, since we examined only one platform, we cannot be sure that any patterns of change we detected (such as shifts in usage seeming to indicate spread of terms from subreddits into the Reddit mainstream) apply to other platforms, or whether (for instance) the factors influencing changes and spread on Reddit are associated with Reddit in particular or potentially linked to activities on other platforms as well. For words like *incel*, which are associated with specific online communities present across multiple platforms, restricting the analysis to Reddit means that the earliest stages of the term's semantic development may predate its Reddit visibility and therefore fall outside the study's observational window.

A third limitation concerns the target word selection procedure. The ten words selected for analysis were chosen based on our knowledge of internet language and cultural context along with the VADER results, and therefore represent a set of terms whose semantic shifts were already known or suspected prior to the analysis. This means the study demonstrates that the methodology can detect and characterize known shifts, which is a necessary and meaningful result, but it does not demonstrate that the methodology can identify previously unknown shifts from a large vocabulary without prior knowledge of which words to examine. A prospective application of this methodology would require additional validation work to address issues of multiple comparisons and false positive rates.

The Orthogonal Procrustes alignment procedure also introduces a limitation. The quality of the alignment, and therefore the interpretability of all cosine similarity comparisons, depends

on the assumption that the anchor words used to compute the rotation matrix have stable meanings across the corpora being compared. High-frequency function words (the, and, of) are the standard choice for this role, and their stability is well-supported in the literature. However, if a major proportion of high-frequency words shift in meaning the alignment may be imperfect in ways that are difficult to detect without ground-truth validation. Dubossarsky et al. (2017) demonstrated that some apparent semantic shifts in word embedding research arise from noise and embedding instability rather than genuine linguistic change; the minimum word frequency filter (min_count = 2) used in this study reduces but does not eliminate this risk, particularly for words that appear very rarely in one of the two corpora being compared.

Finally, the VADER sentiment validation operates at the sentence level rather than at the level of the target word's specific usage within a sentence. A sentence containing “woke” that receives a negative VADER score may be negative because of other words in the sentence rather than because of the specific sense in which “woke” is being used. This limitation is partially addressed by the fact that VADER is used as a corroborating basis check rather than a primary measure, and by the fact that convergence between cosine similarity drops and VADER sentiment shifts is treated as stronger evidence than either signal alone. Regardless, interpretation of the VADER results measures the sentiment of the context in which a word appears rather than the sentiment of the word's usage.

Future Work

Several extensions of this study would strengthen or broaden its findings. The most consequential would be the incorporation of contextual word embeddings. As discussed above, the polysemy limitation of static embeddings affects words like “woke” and “based”, where new senses are developing alongside persistent old ones. A hybrid approach–using the Word2Vec pipeline developed here as a computationally efficient first-pass filter to identify candidate words showing distributional change, then applying a contextual model to those candidates specifically –would allow the granular sense-level analysis that static embeddings cannot provide, while keeping computational costs manageable. This approach would be valuable for distinguishing cases of semantic expansion, in which a new sense develops alongside an existing one, from cases of semantic replacement, in which the old sense is displaced.

A second extension concerns the anchor word assumptions underlying Orthogonal Procrustes alignment. While high-frequency function words are the established standard for this role, relatively little empirical work has examined how sensitive the resulting alignment is to variation in the anchor word set or to the size of the shared vocabulary used to compute the rotation matrix. Varying these parameters and measuring the effect on cosine similarity outcomes for known-stable and known-shifting words would provide valuable information about the robustness of the alignment procedure and help establish more principled guidelines for its application in future diachronic studies.

Cross-platform validation represents another important direction. Applying the same methodology to corpora from platforms including X, 4chan, and Discord would allow examination of whether the semantic shifts observed on Reddit originated on those platforms and

eventually migrated to Reddit, or developed independently across multiple communities simultaneously. For terms like “incel”, where the community associated with the word has a documented presence across multiple platforms, cross-platform analysis could illuminate the mechanics of how platform-specific subcultural vocabulary enters mainstream internet discourse.

The timeslice analysis conducted here also opens the door to prospective monitoring applications. The finding that several words showed a pattern of unstable year-over-year cosine similarity before crystallizing in a new sense suggests that monitoring year-over-year stability for a broad vocabulary could serve as a warning detection for vocabulary shifts. Implementing and evaluating such a monitoring system, including assessment of its precision and recall against a set of known semantic shifts, would be a natural and practically significant extension of the present work.

Finally, extending the framework to languages other than English would both broaden the study's applicability and allow examination of whether the patterns observed here are specific to English-language Reddit or represent more general properties of how meaning changes in large online communities. Reddit hosts substantial non-English-language communities, and the Pushshift dataset includes this content, making the data infrastructure for such an extension already available.

Chapter 6: Conclusion

Summary

The primary objective of this research was to observe, quantify, and characterize the toxification of language on Reddit. Specifically, we sought to track specific target words over time to determine whether static word embeddings and analyses of nearest-neighbor terms could reveal both useful information about the process of toxification and a reliable signal allowing toxification to be detected.. By analyzing the semantic trajectory of these words, our goal was to measure the degree to which their meanings shifted toward negativity. This study aimed to move beyond anecdotal evidence of online toxicity by providing a data-driven framework for observing how language evolves within digital communities.

To achieve these goals, we developed a computational pipeline utilizing non-contextual word embeddings. We processed a comprehensive archival dataset of Reddit comments and submissions, generating static Word2Vec models for distinct time slices. To ensure comparable vector spaces, we aligned these models using Orthogonal Procrustes analysis. Our analytical framework employed a dual-metric approach. We utilized cosine similarity to calculate the quantitative magnitude of semantic change and nearest neighbor analysis to qualitatively assess the nature of that change. Furthermore, by comparing models trained on specific local subreddits against models trained on the global platform, we established a method to track the spread of semantic shifts from niche communities to the broader digital ecosystem.

The results demonstrate that non-contextual word embeddings, when applied with Orthogonal Procrustes alignment and supplemented by sentiment validation, constitute a viable and computationally tractable framework for tracking term toxification in large online corpora. The methodology successfully detected meaningful semantic change across all ten target words, with the clearest evidence of toxification appearing in words like karen and incel, and the sentiment analysis providing a signal in cases where polysemy suppressed the embedding-based measure. The study has limitations worth acknowledging, namely that static embeddings average across all senses of a word, the analysis is bounded to Reddit, and the target words were selected based on prior knowledge rather than discovered prospectively. Nonetheless, the pipeline developed here provides a foundation for future work incorporating contextual embeddings, cross-platform analysis, and prospective monitoring applications that could identify harmful semantic shifts before they become entrenched in the broader lexicon.

References

Bartunov, S., Kondrashkin, D., Osokin, A., & Vetrov, D. (2015). *Breaking Sticks and Ambiguities with Adaptive Skip-gram* (arXiv:1502.07257). arXiv. https://doi.org/10.48550/arXiv.1502.07257

Beck, C. (2020). DiaSense at SemEval-2020 Task 1: Modeling Sense Change via Pre-trained BERT Embeddings. In A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, & E. Shutova (Eds.), *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 50–58). International Committee for Computational Linguistics. https://doi.org/10.18653/v1/2020.semeval-1.4

Bochkarev, V. V., & Shevlyakova, A. V. (2024). How Fast Do Distribution and Semantics of Polysemic Words Change? *Journal of Physics: Conference Series*, *2701*(1), 012099. https://doi.org/10.1088/1742-6596/2701/1/012099

Bootstrapping without the boot. (n.d.). https://doi.org/10.3115/1220575.1220625

Camacho-Collados, J., & Pilehvar, M. T. (2018). On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis. In T. Linzen, G. Chrupała, & A. Alishahi (Eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 40–46). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5406

Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2021). HateBERT: Retraining BERT for Abusive Language Detection in English. *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, 17–25. https://doi.org/10.18653/v1/2021.woah-1.3

Dadvar, M., Jong, F. de, Ordelman, R., & Trieschnigg, D. (2012, February 23). *Improved Cyberbullying Detection Using Gender Information*. https://pure.knaw.nl/ws/files/458160/2012.DIR.dadvar.pdf

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 512–515. https://doi.org/10.1609/icwsm.v11i1.14955

Dinakar, K., Reichart, R., & Lieberman, H. (2021). Modeling the Detection of Textual Cyberbullying. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(3), 11–17. https://doi.org/10.1609/icwsm.v5i3.14209

Dubossarsky, H., Tsvetkov, Y., Dyer, C., & Grossman, E. (2015). *A bottom up approach to category mapping and meaning change*. http://www.openu.ac.il/ISCOL2015/downloads/ISCOL2015_submission20_c_11.pdf

Dubossarsky, H., Weinshall, D., & Grossman, E. (2017). Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1136–1145. https://doi.org/10.18653/v1/D17-1118

Erk, K., McCarthy, D., & Gaylord, N. (2009). Investigations on word senses and word usages. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - ACL-IJCNLP '09*, *1*, 10. https://doi.org/10.3115/1687878.1687882

Erk, K., McCarthy, D., & Gaylord, N. (2013). Measuring Word Meaning in Context. *Computational Linguistics*, *39*(3), 511–554. https://doi.org/10.1162/COLI_a_00142

Erk, K., & Padó, S. (2008). A structured vector space model for word meaning in context. *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08*, 897. https://doi.org/10.3115/1613715.1613831

Erk, K., & Pado, S. (2010). *Exemplar-Based Models for Word Meaning in Context*.

Ethayarajh, K. (2019). *How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings* (arXiv:1909.00512). arXiv. https://doi.org/10.48550/arXiv.1909.00512

Firth, J. (1957). *A Synopsis of Linguistic Theory 1930–1955. In Studies in Linguistic Analysis* (1st ed.).

Fortuna, P., & Nunes, S. (2019). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, *51*(4), 1–30. https://doi.org/10.1145/3232676

Giulianelli, M., Del Tredici, M., & Fernández, R. (2020). Analysing Lexical Semantic Change with Contextualised Word Representations. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics (pp. 3960–3973). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.365

Giulianelli, M., Luden, I., Fernandez, R., & Kutuzov, A. (2023). *Interpretable Word Sense Representations via Definition Generation: The Case of Semantic Change Analysis* (arXiv:2305.11993). arXiv. https://doi.org/10.48550/arXiv.2305.11993

Hagen, S., & De Zeeuw, D. (2023). Based and confused: Tracing the political connotations of a memetic phrase across the web. *Big Data & Society*, *10*(1), 205395172311631. https://doi.org/10.1177/20539517231163175

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016a). Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2116–2121. https://doi.org/10.18653/v1/D16-1229

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016b). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1489–1501. https://doi.org/10.18653/v1/P16-1141

Hu, R., Li, S., & Liang, S. (2019). Diachronic Sense Modeling with Deep Contextualized Word Embeddings: An Ecological View. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3899–3908). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1379

Hutto, C., & 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

Ilić, S., Marrese-Taylor, E., Balazs, J. A., & Matsuo, Y. (2018). *Deep contextualized word representations for detecting sarcasm and irony* (arXiv:1809.09795). arXiv. https://doi.org/10.48550/arXiv.1809.09795

Joseph, B. D., & Janda, R. D. (2008). *The Handbook of Historical Linguistics*. Wiley.

Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/6.pdf

Kilgarriff, A. (1997). *I don't believe in word senses* (arXiv:cmp-Lg/9712006). arXiv. http://arxiv.org/abs/cmp-lg/9712006

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., & Petrov, S. (2014). Temporal Analysis of Language through Neural Language Models. *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 61–65. https://doi.org/10.3115/v1/W14-2517

Kulkarni, V., Al-Rfou, R., Perozzi, B., & Skiena, S. (2015). Statistically Significant Detection of Linguistic Change. *Proceedings of the 24th International Conference on World Wide Web*, 625–635. https://doi.org/10.1145/2736277.2741627

Kutuzov, A., & Giulianelli, M. (2020). *UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection* (arXiv:2005.00050). arXiv. https://doi.org/10.48550/arXiv.2005.00050

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1384–1397). Association for Computational Linguistics. https://aclanthology.org/C18-1117

Laicher, S., Kurtyigit, S., Schlechtweg, D., Kuhn, J., & Schulte im Walde, S. (2021). Explaining and Improving BERT Performance on Lexical Semantic Change Detection. In I.-T. Sorodoc, M. Sushil, E. Takmaz, & E. Agirre (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop* (pp. 192–202). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-srw.25

Laing, R. E. & Illinois State University. (with Baldwin, J. R. & Illinois State University). (2021). *Who Said It First?: Linguistic Appropriation of Slang Terms Within the Popular Lexicon*. Illinois State University ; Ann Arbor.

Lau, J. H., Cook, P., McCarthy, D., Gella, S., & Baldwin, T. (2014). Learning Word Sense Distributions, Detecting Unattested Senses and Identifying Novel Senses Using Topic Models. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 259–270. https://doi.org/10.3115/v1/P14-1025

Lau, J. H., Cook, P., McCarthy, D., Newman, D., & Baldwin, T. (2012). *Word Sense Induction for Novel Sense Detection*.

Lenci, A. (with Sahlgren, M.). (2023). *Distributional semantics* (1st ed.). University Press.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151. IEEE Transactions on Information Theory. https://doi.org/10.1109/18.61115

Liu, Q., McCarthy, D., & Korhonen, A. (2022). *Measuring Context-Word Biases in Lexical Semantic Datasets* (arXiv:2112.06733). arXiv. https://doi.org/10.48550/arXiv.2112.06733

Liu, Q., Ponti, E. M., McCarthy, D., Vulić, I., & Korhonen, A. (2021, April 17). *AM2iCo: Evaluating Word Meaning in Context across Low-Resource Languages with Adversarial Examples*. arXiv.Org. https://arxiv.org/abs/2104.08639v2

Liu, S., & Forss, T. (2015). *New classification models for detecting hate and violence web content*. https://www.scitepress.org/papers/2015/56367/56367.pdf

Liu, Y., Medlar, A., & Glowacka, D. (2021). Statistically Significant Detection of Semantic Shifts using Contextual Word Embeddings. In Y. Gao, S. Eger, W. Zhao, P. Lertvittayakumjorn, & M. Fomicheva (Eds.), *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems* (pp. 104–113). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eval4nlp-1.11

Loureiro, D., Rezaee, K., Pilehvar, M. T., & Camacho-Collados, J. (2021). Analysis and Evaluation of Language Models for Word Sense Disambiguation. *Computational Linguistics*, *47*(2), 387–443. https://doi.org/10.1162/coli_a_00405

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge university press.

Martinc, M., Montariol, S., Zosa, E., & Pivovarova, L. (2020). Capturing Evolution in Word Usage: Just Add More Clusters? *Companion Proceedings of the Web Conference 2020, WWW '20*, 343–349. https://doi.org/10.1145/3366424.3382186

Maslennikova, Y., & Bochkarev, V. (2024). Evaluation of word embedding models used for diachronic semantic change analysis. *Journal of Physics: Conference Series*, *2701*(1), 012082. https://doi.org/10.1088/1742-6596/2701/1/012082

McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). *Learned in Translation: Contextualized Word Vectors*.

Miaschi, A., & Dell'Orletta, F. (2020). Contextual and Non-Contextual Word Embeddings: An in-depth Linguistic Investigation. In S. Gella, J. Welbl, M. Rei, F. Petroni, P. Lewis, E. Strubell, M. Seo, & H. Hajishirzi (Eds.), *Proceedings of the 5th Workshop on Representation Learning for NLP* (pp. 110–119). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.repl4nlp-1.15

Mihalcea, R. F., & Moldovan, D. I. (2001). A highly accurate bootstrapping algorithm for word sense disambiguation. *International Journal on Artificial Intelligence Tools*, *10*(01n02), 5–21. https://doi.org/10.1142/S0218213001000398

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (arXiv:1301.3781). arXiv. https://doi.org/10.48550/arXiv.1301.3781

Mitra, S., Mitra, R., Maity, S. K., Riedl, M., Biemann, C., Goyal, P., & Mukherjee, A. (2015). An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, *21*(5), 773–798. https://doi.org/10.1017/S135132491500011X

Mitra, S., Mitra, R., Riedl, M., Biemann, C., Mukherjee, A., & Goyal, P. (2014). *That's sick dude!: Automatic identification of word sense change across different timescales* (arXiv:1405.4392). arXiv. https://doi.org/10.48550/arXiv.1405.4392

Moore, R. L. (2004). We're Cool, Mom and Dad Are Swell: Basic Slang and Generational Shifts in Values. *American Speech*, *79*(1), 59–86.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019). *A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media* (arXiv:1910.12574). arXiv. https://doi.org/10.48550/arXiv.1910.12574

Panchenko, A., Ruppert, E., Faralli, S., Ponzetto, S. P., & Biemann, C. (2017). Unsupervised Does Not Mean Uninterpretable: The Case for Word Sense Induction and Disambiguation. In M. Lapata, P. Blunsom, & A. Koller (Eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 86–98). Association for Computational Linguistics. https://aclanthology.org/E17-1009

Pantel, P., & Lin, D. (2002). Discovering word senses from text. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '02*, 613–619. https://doi.org/10.1145/775047.775138

Paradis, C. (2011). Metonymization: A key mechanism in semantic change. In R. Benczes, A. Barcelona, & F. J. Ruiz De Mendoza Ibáñez (Eds.), *Human Cognitive Processing* (Vol. 28, pp. 61–88). John Benjamins Publishing Company. https://doi.org/10.1075/hcp.28.04par

Peter Ludlow. (2014). *Living Words: Meaning Underdetermination and the Dynamic Lexicon: First edition*. OUP Oxford. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=e025xna&AN=1200927&site=ehost-live

Peters, M. E., Neumann, M., Zettlemoyer, L., & Yih, W. (2018, August 27). *Dissecting Contextual Word Embeddings: Architecture and Representation*. arXiv.Org. https://arxiv.org/abs/1808.08949v2

Pilehvar, M. T., & Camacho-Collados, J. (2019). *WiC: The Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations* (arXiv:1808.09121). arXiv. https://doi.org/10.48550/arXiv.1808.09121

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, *55*(2), 477–523. https://doi.org/10.1007/s10579-020-09502-8

Popa, D.-N. (2019). *From lexical towards contextualized meaning representation*.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*.

Rahimi, Z., & Homayounpour, M. M. (2023). The impact of preprocessing on word embedding quality: A comparative study. *Language Resources and Evaluation*, *57*(1), 257–291. https://doi.org/10.1007/s10579-022-09620-5

Reddit Data API Wiki. (2023, June 2). Reddit Help. https://support.reddithelp.com/hc/en-us/articles/16160319875092-Reddit-Data-API-Wiki

Riemer, N. (Ed.). (2016). *The Routledge handbook of semantics*. Routledge. https://doi.org/10.4324/9781315685533

Rosenfeld, A., & Erk, K. (2018). Deep Neural Models of Semantic Shift. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 474–484). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1044

Rudolph, M., & Blei, D. (2018). Dynamic Embeddings for Language Evolution. *Proceedings of the 2018 World Wide Web Conference, WWW '18*, 1003–1011. https://doi.org/10.1145/3178876.3185999

Sá, J. M. C. de, Silveira, M. D., & Pruski, C. (2024). *Survey in Characterization of Semantic Change* (arXiv:2402.19088). arXiv. https://doi.org/10.48550/arXiv.2402.19088

Schlechtweg, D., Walde, S. S. im, & Eckmann, S. (2018). *Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change* (arXiv:1804.06517). arXiv. https://doi.org/10.48550/arXiv.1804.06517

Schönemann, P. (1966, March). *A generalized solution of the orthogonal Procrustes problem*. https://web.stanford.edu/class/cs273/refs/procrustes.pdf

Schutze, H. (1998). *Automatic Word Sense Discrimination*. *24*(1).

Shoemark, P., Liza, F. F., Nguyen, D., Hale, S., & McGillivray, B. (2019). Room to Glo: A Systematic Comparison of Semantic Change Detection Approaches with Word Embeddings. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 66–76. https://doi.org/10.18653/v1/D19-1007

Stewart, I., Arendt, D., Bell, E., & Volkova, S. (2017). Measuring, Predicting and Visualizing Short-Term Change in Word Representation and Usage in VKontakte Social Network. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 672–675. https://doi.org/10.1609/icwsm.v11i1.14938

Tahmasebi, N., Borin, L., & Jatowt, A. (2019). *Survey of Computational Approaches to Lexical Semantic Change* (arXiv:1811.06278). arXiv. https://doi.org/10.48550/arXiv.1811.06278

Tahmasebi, N., Borin, L., Jatowt, A., Xu, Y., & Hengchen, S. (2021). *Computational approaches to semantic change*. Zenodo. https://doi.org/10.5281/ZENODO.5040241

Tang, X. (2018). A state-of-the-art of semantic change computation. *Natural Language Engineering*, *24*(5), 649–676. https://doi.org/10.1017/S1351324918000220

Tang, X., Qu, W., & Chen, X. (2016). Semantic change computation: A successive approach. *World Wide Web*, *19*(3), 375–415. https://doi.org/10.1007/s11280-014-0316-y

Tang, X., Zhou, Y., & Bollegala, D. (2023). Learning Dynamic Contextualised Word Embeddings via Template-based Temporal Adaptation. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 9352–9369). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.520

Traugott, E. C. (2017). *Semantic change*. https://traugott.people.stanford.edu/sites/g/files/sbiybj28616/files/media/file/traugott2017a.pdf

Traugott, E. C., & Dasher, R. B. (2002). *Regularity in semantic change*. Cambridge university press.

Tredici, M. D., & Fernández, R. (2018). *Semantic Variation in Online Communities of Practice* (arXiv:1806.05847). arXiv. https://doi.org/10.48550/arXiv.1806.05847

Vidgen, B., & Yasseri, T. (2020). Detecting weak and strong Islamophobic hate speech on social media. *Journal of Information Technology & Politics*, *17*(1), 66–78. https://doi.org/10.1080/19331681.2019.1702607

Wanchoo, K., Abrams, M., Merchant, R. M., Ungar, L., & Guntuku, S. C. (2023). Reddit language indicates changes associated with diet, physical activity, substance use, and smoking during COVID-19. *PLOS ONE*, *18*(2), e0280337. https://doi.org/10.1371/journal.pone.0280337

Wiedemann, G., Remus, S., Chawla, A., & Biemann, C. (2019). *Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings* (arXiv:1909.10430). arXiv. http://arxiv.org/abs/1909.10430

Wijaya, D. T., & Yeniterzi, R. (2011). Understanding semantic change of words over centuries. *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web, DETECT '11*, 35–40. https://doi.org/10.1145/2064448.2064475

Wulczyn, E., Thain, N., & Dixon, L. (2017). *Ex Machina: Personal Attacks Seen at Scale* (arXiv:1610.08914). arXiv. https://doi.org/10.48550/arXiv.1610.08914

Xu, Y., & Kemp, C. (2015). *A Computational Evaluation of Two Laws of Semantic Change*.

YAROWSKY David. (1995). Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of ACL, 1995*.

Zahrotun, L. (2016). Comparison Jaccard similarity, Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method. *Computer Engineering and Applications Journal*, *5*(1), 11–18. https://doi.org/10.18495/comengapp.v5i1.160

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., & Çöltekin, Ç. (2020). *SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020)* (arXiv:2006.07235). arXiv. https://doi.org/10.48550/arXiv.2006.07235