ABSTRACT

Title of Document: USING SOCIAL MEDIA TO EVALUATE PUBLIC ACCEPTANCE OF INFRASTRUCTURE PROJECTS

Qinyi Ding, Doctor of Philosophy, 2018

Directed By: Dr. Qingbin Cui,
Department of Civil & Environmental Engineering

The deficit of infrastructure quality of the United States demands groundbreaking of more infrastructure projects. Despite the potential economic and social benefits brought by these projects, they could also negatively impact the community and the environment, which could in turn affect the implementation and operation of the projects. Therefore, measuring and monitoring public acceptance is critical to the success of infrastructure projects. However, current practices such as public hearings and opinion polls are slow and costly, hence are insufficient to provide satisfactory monitoring mechanism.

Meanwhile, the development of state-of-the-art technologies such as social media and big data have provided people with unprecedented ways to express themselves. These platforms generate huge volumes of user-generated content, and have naturally become alternative sources of public opinion. This research proposes a framework and an analysis methodology to use big data from social media (e.g. the microblogging site Twitter) for project evaluation. The framework collects social media postings, analyzes public opinion towards infrastructure projects and builds multi-dimensional models around the big data. The interface and conceptual implementation of each component of the framework are discussed. This framework could be used as a supplement to traditional polls to provide a fast and cost-effective way for public opinion and project risk assessment.

This research is followed by a case study applying the framework to a real-world infrastructure project to demonstrate the feasibility and comprehensiveness of the framework. The California High Speed Rail project is selected to be the object of study. It is an iconic and controversial large-scale infrastructure project that faced a lot of criticism, complaints and suggestions. Sentiment analysis, the most important type of analysis on the framework, is discussed concerning its application and implementation in the context of infrastructure projects. A public acceptance model for social media sentiment analysis is proposed and examined, and the best measurement of public acceptance is recommended.

Moreover, the case study explores the driving force of the change in public acceptance: the social media events. Events are defined, evaluated, and an event influence quadrant is proposed to categorize and prioritize social media events. Furthermore, the individuals influencing the perceptions of these events, opinion leaders, are also modeled and identified. Three opinion leadership types are defined with top users in each type listed and discussed. A predictive model for opinion leader is also developed to identify opinion leaders using an a priori indicator. Finally, a user profiling model is established to describe social demographic characteristics of users, and each demographic feature is discussed in detail.

USING SOCIAL MEDIA TO EVALUATE PUBLIC ACCEPTANCE OF INFRASTRUCTURE PROJECTS

by

Qinyi Ding

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Qingbin Cui, Chair
Professor Gregory Baecher
Professor John Cable
Professor Michelle (Shelby) Bensi
Professor Samir Khuller
Professor Yunfeng Zhang

ACKNOWLEDGEMENTS

This dissertation is completed with the support and encouragement of a lot of people. I am grateful to Dr. Qingbin Cui, a wise and inspiring academic mentor, for his continuous support and advice on my research study. We had a lot of enlightening discussions regarding the topic, and I would not be able to complete the dissertation without his guidance.

I would like to thank my committee, for spending time reviewing my work, providing feedback and attending my defense. It is a great honor to present and share my research work in front of these prestigious scholars.

I would also like to express my gratitude to my family. My parents, my wife and my two kids all exhibited tremendous support and care during my study. As a part time PhD candidate, it is difficult to balance family and study. Nevertheless, my family, especially my wife, Xinyuan Zhu, supported me unconditionally throughout the course of my study. Thank you very much.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Infrastructure Projects

Infrastructure is the foundation of the United States' economy. It drives business, communities and people to thrive by providing efficient transportation systems, low-cost, reliable energy sources, and robust water systems (Herrmann, 2013). However, the infrastructure of this nation is falling behind the pace of development, and is resultantly in need of more investment in both finance and labor force. The ASCE report card released in 2013 shows that the GPA of America's infrastructure is D+ and 3.6 trillion dollars of investment are needed by 2020 (Figure 1-1). Infrastructure across multiple sectors is performing below expectations, affecting the efficiency of people's daily lives.

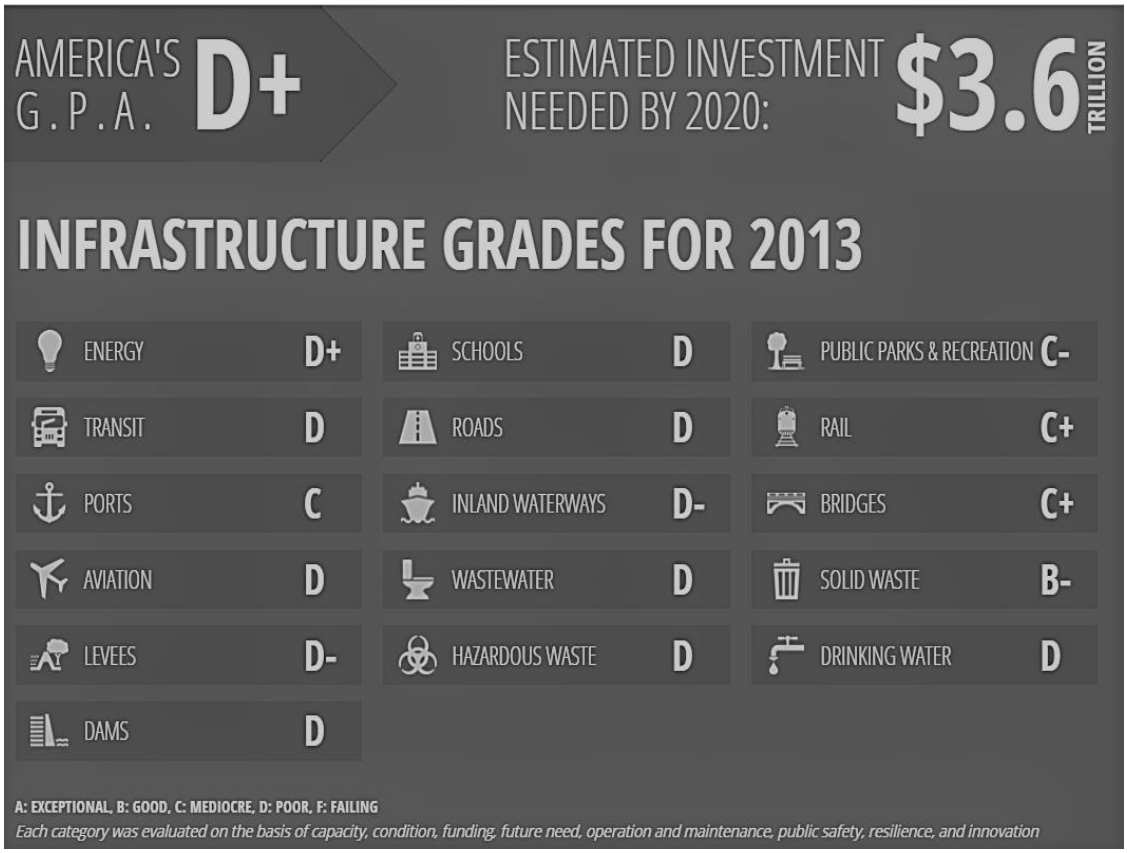

Figure 1-1 America's GPA of infrastructure projects (Herrmann, 2013)

In order to improve the quality of the infrastructure, a lot of new facilities will need to be constructed, and old facilities will need to be repaired or replaced. President Donald Trump has proposed to invest $1 trillion over 10 years on America's infrastructure including highways, bridges, airports, schools etc. (Carnevale & Smith, 2017; Update, 2017) Such a big investment in fixing and upgrading infrastructure is critical to the economy and can foster inspiring results on the economy and improve living conditions. For example, transportation projects bring user benefits such as ease of access and travel time and improved product cost, quality and supplying efficiency (Weisbrod & Weisbrod, 1997). They also contribute to job creation and the boost of local economy (The White House, 2014). Renewable energy projects are crucial to sustain the energy consumption increase caused by the projected population growth, and mitigate environmental issues such as acid precipitation, stratospheric ozone depletion, and greenhouse effect (Dincer, 2000).

While the outlook is promising, infrastructure projects can have negative impacts. For example, some infrastructure projects have to fill wetlands, disrupt wildlife corridors or negatively affect wildlife refuges and recreation areas (Hayes, 2014). Dam projects might have environmental, social, political and economic impacts such as altering global water cycles, migration and resettlement of affected people, changes in rural economy and employment structure, effects on infrastructure and housing, impacts on non-material or cultural life and community health issues, etc. (Bartolomé, De Wet, Mander, & Nagraj, 2000; Cernea, 1988; Gleick, 2012; Tilt, Braun, & He, 2009).

Due to the variety of social impacts of infrastructure projects, it is critical to assess public acceptance when the project is being planned and implemented. Although public acceptance does not determine project success as directly as financial and scheduling factors, it plays an important role in project implementation and maintenance. A lot of researchers list public support as one of the critical success factors of infrastructure projects. (X. Zhang, 2005) identified "supportive and understanding community" and "the project is in public interest" as two success sub factors under "favorable investment environment". (Chua, Kog, & Loh, 1999) also identified "impact on public"

as one of the six success-related factors of the "project characteristics" category, and found that project success is not exclusively dependent on project management, monitoring, and control efforts, but project characteristics and contractual arrangements as well. (Devine-Wright, 2007) recognize public acceptance as an important issue in implementing advanced renewable energy technologies. (Jobert, Laborgne, & Mimler, 2007) list creating trust with local populations as a main challenge. Once created, trust serves as a key to project success. It is, therefore, crucial to pay attention to the public acceptance of infrastructure projects in order to gain public support and avoid barriers, delays and lawsuits associated with public opposition.

1.2 The Importance of Public Opinion on Infrastructure Projects

Martin Richards, Chairman of England's MVA Consultancy, mentioned that "the failure to recognize the importance of public response and to make full allowance for that response is bound to lead to failure of any project" (Federal Highway Administration, 1992). Consulting and collaborating with the public "can lead to reduced financial risk (from delays, legal disputes, and negative publicity), direct cost savings, increased market share (through good public image), and enhanced social benefits to local communities" (IFC, 1998). Being one of the contributors to the success of infrastructure projects, public acceptance is dangerous to ignore. Social opposition could potentially slow down the project (Cohen, Reichl, & Schmidthaler, 2014), or even lead to project cancellation and lawsuits.

The Presidio Parkway project in California is a good example of how public opinion can affect the implementation of the project. The project was set to replace the historic south access road to the Golden Gate Bridge in San Francisco because the road was structurally and seismically deficient (Presidio Parkway, 2016). It was one of the first public private partnership (PPP) projects awarded in California. However, it experienced an unexpected delay for almost a year during phase II, which was partly due to the lawsuit brought forward by the group Public Engineers in California Government (PECG), who sought to stop phase II of the project by arguing that the project was not

authorized by the Streets and Highways Code section 143 (Roberts, 2011). PECG viewed the PPP project as anti-union and anti-public engineer (Maddex, 2012). Even though the District Court of Appeal rejected all PECG's arguments, the litigation delay cost the project almost a year. The lawsuits and construction delays can be avoided by effective planning and clear contract clauses (AECOM, 2007).

As another example, the I-77 Express Lanes project, a $650 million, 26-mile project converting and constructing toll lanes between Charlotte and Cornelius in North Carolina (Dutzik, Bradford, Weissman, & others, 2017), faced a lot of inquiries, ranging from whether the project could effectively control congestion, to concerns about a private / foreign company operating and tolling commuters, to worries regarding real estate values and local business. Inaction by elected leaders generated public concern about the rising cost and concept of P3 / HOT and an organization, Widen I-77, was formed in October 2012 for this reason (Widen I-77, 2016). The North Carolina House overwhelmingly passed a bill to cancel the contract with a private developer (Morrill, 2016).

The Baltimore Red Line Rail project, a 14 mile light rail traveling through downtown and West Baltimore, was also another innovative project that was viewed as a critical step to revitalize local communities by providing access to major employment centers (CPHA, 2015). However, the project was cancelled in June 2016 due to concerns regarding its financial effectiveness, resulting in major disappointment and opposition. Multiple panels and groups demanded an alternative plan or a continuation of the project. The National Association for the Advancement of Colored People (NAACP, n.d.) filed a federal civil rights complaint against Maryland, alleging that the state discriminated against African American residents in Baltimore when the project was killed and the state money was diverted to road and bridge projects elsewhere (Wiggins & Turque, 2015).

These highlighted examples demonstrate the significance of public acceptance and just a few of the ways public opinion can bring fundamental impacts to a project (Dimitropoulos & Kontoleon, 2009; Evans, Parks, & Theobald, 2011; Wüstenhagen, Wolsink, & Bürer, 2007). A project cannot be

successful without a supporting public. The monitoring and tracking of public acceptance is therefore critical to ensure the health of a project.

1.3 Current Methods of Public Acceptance Assessment

Traditionally public acceptance assessments were part of citizen participation and public consultations, a process that gives citizens the opportunity to influence the decision making of public affairs. In the context of projects, this process can be beneficial in reducing financial risk, reducing direct costs, and increasing market share and social benefits (IFC, 1998). There are many techniques and methodologies available that can raise the awareness of the environmental and social impacts of projects, including brochures, advertising, exhibitions, polls, focus group interviews, public hearings, etc. (IFC, 1998). Below we will discuss two of the widely adopted methods, public hearings and public opinion polls, which have proven to be effective despite drawbacks.

1.3.1 Public Hearings

Public hearings are a widely used method to collect public opinion and engage interest groups. It is one of the most traditional methods to allow people to be involved in government activities and projects. Its usage is still increasing and (Checkoway & Van Til, 1978) estimated the number of public hearings each year to be in the tens of thousands. They serve, according to (Heberlein, 1976), four distinct functions, i.e. the informational function to inform citizens about the project, the cooptation function to give people an opportunity to complain about the project, the ritualistic function when the hearing is demanded by law but not by the public, and the interactive function, the ideal function when the agency actually seeks public opinion and responds accordingly.

Even though public hearing is a method extensively relied upon in United States (Cole & Caputo, 1984), its problems cannot be overlooked. (Checkoway, 1981) listed several shortcomings of public hearings, such as that they are not always held at a convenient location, that the terms used are not

easily understood by everyone, that attendees are not representative of the actual population, and that the influence of public hearings on decision making is limited. (Kemp, 1985) argued that instead of serving the purpose of pluralistic decision making, the outcomes of public hearings are likely not rational and objective, and are manipulated for the benefit of dominant groups. (Cole & Caputo, 1984) found that it does not impact citizen behavior or policy choices enough. The potential bias in the results of public hearings is critical to address the effectiveness on predicting public acceptance. For infrastructure projects, interest groups have various channels to have their voice heard. Public acceptance assessment should reach out to the general public who are affected by the project to mitigate any discontent and address issues of concern. From that perspective, public hearings do not provide a setting for everyone to express their opinions, hence this method alone is not sufficient for project acceptance evaluation.

1.3.2 Public Opinion Polls

Another traditional method to assess public opinion is a public opinion poll. Evolved from the straw poll, which is an informal and unofficial vote to assess public opinion (Erikson & Tedin, 2015), modern scientific polls are widely used in politics of the United States. Polls conducted by newspapers, television networks, or other professional organizations attempt to reveal how people view controversial topics such as presidential approval or elections. These polls, especially polls conducted by institutions like Gallup, follow the fundamental principle called equal probability of selection, which states that if every member of a population has an equal probability of being selected in a sample, then that sample will be representative of the population. Under this principle, modern polls typically need about 1,000 adults to serve as a sample of the opinion of the whole nation (Newport, Saad, & Moore, 1997). Even with such a small number of adults, the result of a fine-tuned sample can be highly representative.

In addition to political matters, public opinion polls are also applied to infrastructure projects and policies. The WSDOT conducted several polls to assess public opinion on congestion pricing

including seven 90-minute focus group sessions, executive interviews with opinion leaders, a telephone survey and a group survey targeting a wider audience in the next phase (Ulberg, 1995). The Boeing Company also conducted a 200-person random survey on people in the Los Angeles International Airport to assess the opinion regarding unmanned aerial vehicles for cargo, commercial, and passenger transportation (MacSween-George, 2003). The sampling mechanism of these surveys cannot be as accurate as Gallup poll. After all, it is difficult to obtain the demographic distribution of the targeted audience in the first place. Nonetheless, this does not stop polling from becoming one of the most popular public opinion assessment tools.

Although traditional polling is a popular way of gathering public opinion, it still has its own drawbacks that can negatively affect its performance, especially in regard to polling for infrastructure projects. There are three major defects of using traditional polls to assess public opinion.

- Firstly, it is expensive to collect data. A scientific poll such as those created by Gallup uses methodologies like “random digit dialing”, which creates a list of all possible household phone numbers in America and then selects a subset of numbers from that list for Gallup to call. “Within household selection” is another methodology which selects a random adult from the list of all adults living in the household (Newport et al., 1997). This is an extensive and costly procedure that is not feasible for infrastructure projects. A standard telephone poll of one thousand respondents can easily cost tens of thousands of dollars to run (O’Connor, Balasubramanyan, Routledge, & Smith, 2010). In the world of real projects, according to (Heberlein, 1976), contracted surveys can range up to 50 to 60 dollars per interview; although, using telephones and WATS lines can reduce costs to 10 dollars per interview. Besides the cost of conducting the survey, there are also costs for data input, data analysis, and other managerial expenses.

- Secondly, it takes time to analyze the data and generate results. Due to the complexity of the procedure, the time it takes to reach out to all respondents and collect necessary information is

significant. The time to conduct a survey can range from a month and a half to multiple years (Heberlein, 1976). Such a meticulously designed and executed poll could provide a representative analysis, but should also be too time-consuming to keep up with the pace of project development.

- More importantly, it is difficult to obtain true opinions from interviewees. This difficulty comes from multiple sources. Firstly, the wording of the interview questions is important, as poorly designed questions can result in ambiguous interpretations and misleading outputs. Secondly, the interviewees might not have enough knowledge about the project to have an established opinion and/or the opinion they have established may be easily swayable. Thirdly, the interviewees might not be fully aware of all the consequences of a project and may be making their choices based off the limited knowledge they have when answering questions (Heberlein, 1976). Unlike a political poll where people's votes have a direct impact on the final result of an election no matter how well he/she understands the policies, infrastructure project polls can be more irrelevant and subjective to the respondent. Thus, the fact that their answers to the questionnaire can be highly constrained by their knowledge about the project can critically undermine the accuracy of the polling results.

In summary, traditional polling is not an ideal method for infrastructure projects because of its cost, duration, and the potential to generate inaccurate results. However, it might still be better or more viable than other methods available as long as drawbacks are considered.

The current problems in public opinion analysis inspire this research study, which tries to explore the possibility of using social media as an alternate data source of public opinion. The advancement in technology empowers social media sites to become influential sources of information offering huge advantage in their abundance of data. In this dissertation, we would like to prove the feasibility of evaluating public opinion based on social media and mitigate the aforementioned problems.

1.4 The Advantage of Evaluating Public Opinion Using Social Media

"Social Media is a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content" (Kaplan & Haenlein, 2010). Social media sites allow people to tell their stories (Facebook), share their pictures (Flickr), and publish their videos (Youtube), interact with other people they may or may not know (Twitter), collaborate on knowledge sharing (Wikipedia) and ask and answer questions (Quora). People spend an increasing amount of time on social media sites where they are free to express themselves. In 2015, 65% of American adults use social networking sites with a wide variety of age, gender, and race (Perrin, 2015). Figure 1-2 shows the percentage growth of Internet users and all adults.

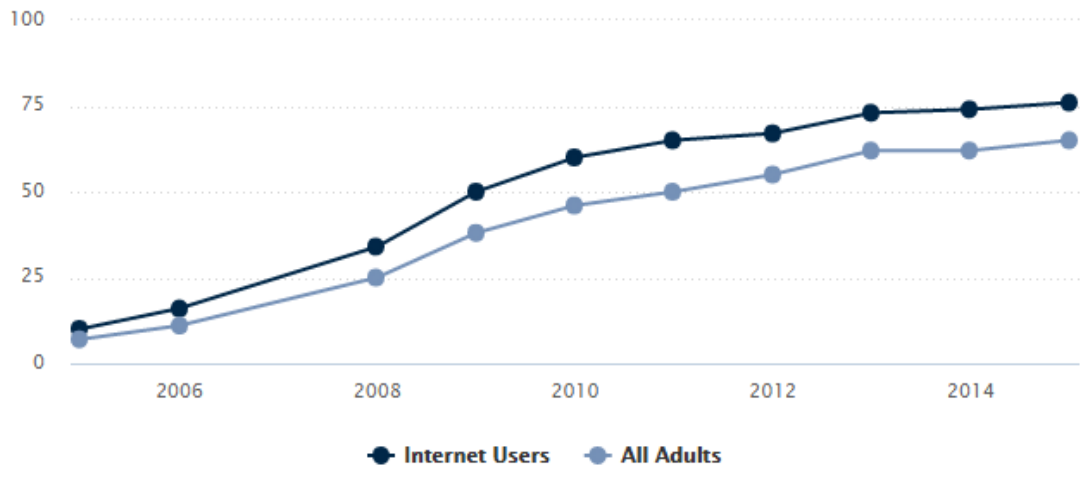

Figure 1-2 Percentage of all American adults and internet-using adults who use at least one social networking site (Perrin, 2015)

The market leader, Facebook, is reporting 1.59 billion monthly active users as of April 2016 (Statistica, 2016). Twitter, the microblog site which we will fetch the data from, also enjoys 320 million monthly active users. Such users are generating a tremendous volume of rich data covering all aspects of people's daily life. Properly filtered, mined, and analyzed, the big data can be used to uncover trending topics, public opinions, widespread attitudes, etc. It is reasonable to argue that the

big data can be applied towards infrastructure projects as well, providing insights from people's genuine expression, which was difficult to achieve previously.

There are several advantages of using Twitter as a supplementary data source rather than traditional methods such as public hearings and polls. First of all, it is almost free to get streaming of data from Twitter when designed properly. Twitter provides various APIs for users to fetch tweets about certain topics, from certain users, or meeting certain search terms. It is not very difficult to write a software program to crawl tweets without having to reach out to people in person and conduct interviews. The number of tweets related to infrastructure projects is relatively low, therefore it is very likely that a Twitter crawler is capable of capturing all the tweets within the free tier of Twitter's rate limit. Historical tweets, however, need to be purchased. Twitter is partnering with Gnip (Gnip, 2017) to provide multiple historical tweet search APIs covering the full history of Twitter since March 2006 as a paid service. The real cost of using Twitter for project evaluation is the developers' time and effort to develop robust software or scripts to fetch, analyze and generate insights from big data, rather than the data itself.

Secondly, tweets can be fetched in a nearly real-time fashion. Different Twitter APIs provide different levels of response time. With Twitter's Streaming API, an application will be pushed with the stream of tweets in real time. With the REST API, the crawler program has to pull tweets from Twitter in a predetermined interval, hence is not truly in "real time". However, the response time and refresh frequency can be at least daily, a huge gain already compared to the monthly cycle of public hearings and polls. The speed of data fetching and refreshing is critical for project managers to closely monitor public opinion and its changes and quickly understand the reason behind these changes in order to make timely data-driven decisions.

Thirdly, given the popularity of Twitter usage, tweets can truly represent the general public rather than certain interest groups, experts, or professionals. Anyone impacted by the project can speak freely on Twitter regardless of their knowledge or education. People actively post on social media

to express their true selves, whereas they are passively asked to speak in public or answer a stranger's questions on the telephone. Therefore, tweets are genuine user expressions of their opinion and sentiment.

Lastly, research shows that tweets are able to generate similar results to traditional polls. (O'Connor et al., 2010) has conducted research to show that a relatively simple sentiment ratio based on related Twitter data can capture the trends of polls. For both consumer confidence and presidential job approval polls, Twitter data analysis has proven itself to be a good supplement for traditional polls. (Kryvasheyeu et al., 2016) demonstrated that in the event of a disastrous occurrence like Hurricane Sandy, the per-capita number of Twitter messages corresponds to disaster-inflicted monetary damage. With a simple metric of counting tweets, broken down into multiple dimensions such as proximity and time, social media provides a good indicator of the impact of the disaster. These research studies bring confidence to use Twitter to describe the status quo and predict the future.

It is worth mentioning that the advantages of social media based analysis address all the intrinsic problems of traditional methodologies. The time to deliver insights and analysis is shortened, the update speed is much higher, the cost of conducting analysis is greatly reduced, and the result could potentially be equally representative, if not more. The benefits of new technology ensure that tweets can be used in infrastructure projects to provide insights of public opinion and help evaluate project risks. Using social media is an innovative procedure that can re-define current project evaluation practices by providing a more data driven decision making process.

1.5 Research Needs

As discussed earlier, traditional methods of public acceptance assessment are costly and time-consuming. On the contrary, social media has shown great potential in providing similar trending to its traditional counterparts, with the benefit of being more cost-effective and providing faster

response. Social media also allows the voice of the public to be expressed and heard, and the big data it generates can be used to evaluate public acceptance.

Although previous research has already studied subjects such as presidential approval ratings, elections, natural disasters, etc., little research has targeted infrastructure projects, which are in sorely needed to improve the project evaluation process and understand public support and opposition. In order to bridge the gap and prove the feasibility, there are some fundamental questions we need to answer in this research, from both technical and modeling perspective, as shown in Figure 1-3.

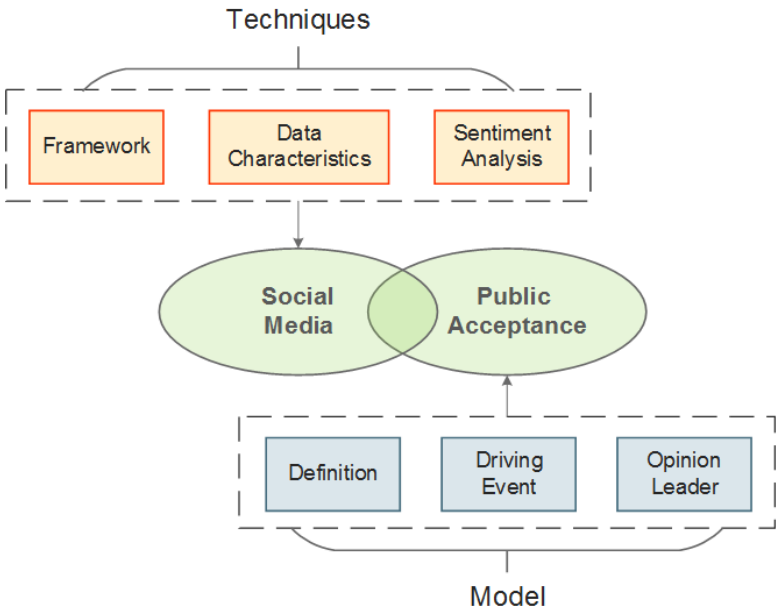

Figure 1-3 Research Structure

Firstly, despite several research studies on social media analysis, there is a lack of standard framework for the predefined workflow, data structure and analytical metrics for social media data analysis. This research aims to define a process which can be reused and customized so that various infrastructure projects can be accommodated and benefitted. Thus, a conceptual framework defining components, interfaces and database structures for general purpose analysis needs to be established, as well as sample analytical metrics to be used to describe public acceptance and its derived measurements.

Secondly, it is questionable whether the methodologies used in previous research studies are also applicable to infrastructure projects. Infrastructure projects are more geographically limited and less impactful compared with large-scale events such as presidential elections and natural disasters. The data fetching paradigm and the data volume can be very different compared with those events. It is worth studying the data characteristics of infrastructure projects to investigate if they are comparable with previous research studies, and propose or recommend new techniques if they are not. It is also necessary to examine the efficiency of data fetching mechanisms for infrastructure projects to improve their efficiency and accuracy.

Thirdly, the basis of public acceptance is sentiment analysis. There are a lot of methodologies available for sentiment analysis, including machine learning based approaches and lexicon based approaches. Considering the uniqueness of infrastructure projects, the performance of different analysis techniques can be tested to find out the most suitable methods for infrastructure projects.

With the help of the building blocks for public acceptance analysis, we are equipped to assess public acceptance of infrastructure projects using social media. The rest of the research addresses this problem by answering the following three questions.

- What is public acceptance in social media?

The dataset collected from social media is a list of postings by social media users. A public acceptance model needs to be developed to translate hidden attributes of these postings into numeric measurements of public acceptance. Different public acceptance models can be developed following various political principles and different user characteristics. These models are then validated to find the one most suitable for infrastructure projects.

- What is driving the change in public acceptance?

Understanding and measuring public acceptance initializes project evaluation. It is equally important to understand the reasons behind the change of public acceptance over time. Taking

advantage of the big data generated from social media, this research explores the driving force behind the change of public acceptance to have it defined and modeled. Understanding the real time fluctuation of public acceptance helps to alarm the occurrence of emergent occasions when public acceptance takes a dive.

- Who is driving the change in public acceptance?

In addition to knowing what is swinging the public acceptance, it is of the same importance to find out who is leading these changes. Social media data reveals not just postings, but also the users who composed them and their characteristics. Therefore, we can discover the leaders influencing changes in public acceptance, their distinctions with other users, and predict potential leaders.

In summary, this research is dedicated to address the problem of costly and time-consuming public acceptance assessment by introducing social media as the data source. It proposes processes and techniques to meet the need for infrastructure projects, and furthermore, how to conduct public acceptance analysis. Three aspects, the public acceptance itself, the driving factor of public acceptance changes and the driving individuals, are modeled and assessed in a real-world case study. This research brings forward a valuable alternative to the current project evaluation process.

1.6 Dissertation Structure and Research Plan

The problems discussed above justify the research need on this topic. Addressing these issues from the research need, this research starts with the project evaluation framework, providing a template social media analysis procedure. It then compares and proposes optimal tools and techniques for data fetching and sentiment analysis for infrastructure based datasets. Then, it is followed by a comprehensive analysis of public acceptance evaluation modeling three objects: the public acceptance, social media events driving the change of public acceptance and social media users that lead such changes. The structure of the dissertation is outlined in Figure 1-4.

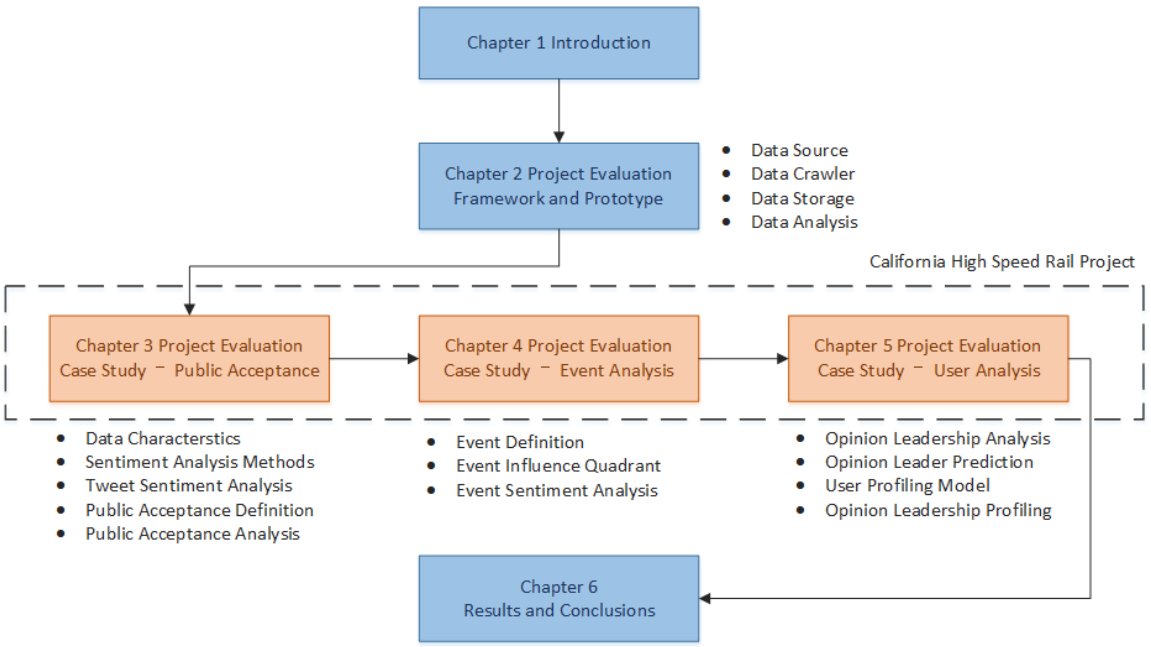

Figure 1-4 Dissertation Structure

Chapter 2 starts with proposing the conceptual framework of project evaluation using social media and big data. Four core modules of the framework are introduced and sample analysis is proposed. Chapter 3 first introduces the case study project, the California High-Speed Rail project, and discusses its characteristics and the reason for selection. Then there is the discussion of data characteristics and the performance of different search terms used in data fetching. It is followed by a discussion of sentiment analysis techniques using machine learning and lexicon based approaches and the performance comparison of several methods. Later in chapter 3, three possible definitions of public acceptance are proposed and applied to the case study. The validity of each model is tested and the most suitable one is recommended for infrastructure projects. Chapter 4 defines the driving force, the social media event, and models its impact using the event influence quadrant. Chapter 5 focuses on user analysis and build models to describe different opinion leadership types and predict opinion leaders. It also abstracts user profile into a multi-dimensional model to better depict individual users.

Chapter 2. Evaluation of Project Acceptance Using Big Data – Framework and Prototype

2.1 Introduction

Given the huge deficit in maintaining and improving current infrastructures in this country (Engel, Fischer, & Galetovic, 2011; Herrmann, 2013), a large number of infrastructure projects can be expected to launch in the near future to bring changes to the status quo. While tracking cost and schedule remains the core requirements of a successful project, the monitoring of public acceptance which leads to public relation risks, is equally important due to the impactful nature of infrastructure projects. After all, an unsupportive community contributes negatively to the project and might significantly impair the implementation and operation of the project.

The current practices to assess public acceptance include methods such as public hearings and public opinion polls. These traditional methods, although being widely used, are too slow and costly to estimate public acceptance. In order to overcome the drawbacks of these methods, and furthermore, enrich the dimensions of public acceptance analysis, this research studies the use of social media and the big data it generates as an alternate data source of public acceptance assessment. Taking advantage of the amount of active users of social media and the fast response, this methodology is expected to provide the capability to produce insights on public acceptance in a real time and cost-effective manner, and bring more in-depth analysis to the field. In this chapter, we would like to propose a conceptual framework to standardize the process to fetch, store, and analyze social media data for infrastructure projects. Some sophisticated analyses utilizing the framework are proposed to evaluate infrastructure projects. Subsequent chapters apply the framework on a real-world case study to demonstrate its feasibility and versatility.

2.1.1 Framework Architecture

The overall architecture of the project evaluation framework is first introduced. The data flow among all four components, data source, data crawler, data storage and data analysis, is then discussed in general. It is followed by detailed descriptions of each component, including various available web services as data sources, corresponding processes of data crawlers and a standard data schema for data storage. A user knowledge library is conceptualized to accommodate customizations and domain specific enhancements of the techniques to be used, which can be developed in a crowd sourcing manner.

2.1.2 Sample Data Analysis

This chapter is followed by sample analyses the framework is able to perform as the end result and deliverable. This section demonstrates the versatility of the framework, which is able to conduct not only public acceptance analysis, but also event, user and project risk evaluation etc. This is a significant enrichment to the current practice which is struggling in the former task alone. To summarize, this chapter proposes a framework to standardize the process and data structure for project evaluation in order to provide guidance for future research and implementation.

2.2 Literature Review

2.2.1 Infrastructure Projects

Infrastructure projects are key components in civil engineering and are vital to the development of economy and technology. The national population growth requires more infrastructure with better reliability, durability, and efficiency. Constructing new infrastructures and repairing or replacing existing ones are both important actions to take to ensure infrastructures in service can support individual and business activities. These infrastructure projects can also create new jobs and boost the local economy (The White House, 2014).

However, the status quo of infrastructure in U.S. is not optimistic after accumulated deficit in repair and maintenance (Engel et al., 2011). ASCE rated the overall GPA of America's infrastructure as D+ and estimated an investment of 3.6 trillion dollars being required by 2020 (Herrmann, 2013). Bridges in the United States have an average age of 42 years with one in nine being structurally deficient. 42% of major urban highways are congested, causing an estimated $101 billion of wasted time and fuel. These old or flawed infrastructures in this country are in dire need of being renovated to prevent tragedies such as the I-35W Mississippi River bridge collapse (Hao, 2009). Inspections and evaluations excluded the under-designed gusset plates which are already fractured over years, causing the tragedy on August 1, 2007, with 13 people dead and more than 100 injured (Astaneh-Asl, 2008). Aging infrastructure also increases the vulnerability to threats posted by common environmental conditions, extreme natural hazards, and terrorism (Homeland Security, 2010).

Numerous infrastructure projects can be expected to launch in this country to improve its infrastructure quality. It is crucial to be equipped with the knowledge and methodologies to ensure the success of these projects, and public acceptance evaluation is an important piece of the puzzle.

2.2.2 Assessing Public Acceptance

Due to the scale of infrastructure projects and the potential impacts on the local community both environmentally and economically, public acceptance is one of the key factors in determining the success of a project. (Diekmann & Girard, 1995) include public interference, occurred when a large group of people are minimally impacted or a small group of people are greatly impacted, as a predictor of project dispute. (Chua et al., 1999) include "impact on public" as one of the success related factors under project characteristics. (Hardcastle, Edwards, Akintoye, & Li, 2005) identifies political support (Qiao, Wang, Tiong, & Chan, 2001; W. R. Zhang, WANG, TIONG, Ting, & Ashley, 1998) and social support (Frilet, 1997) as critical success factors for project success. (X. Zhang, 2005) identified "Supportive and understanding community" and "the project is in public interest" as two of the 47 success sub-factors for public private partnership infrastructure projects.

It is therefore important to address the public discontent and improve public acceptance of infrastructure projects. Although infrastructure projects bring more job opportunities and better facilities, they will inevitably affect certain groups of people who need to be properly compensated or educated in order for them to get onboard with the projects. For some controversial projects, people also question their financial feasibility, potential corruptions, and the real benefit they will bring. The Presidio Parkway project in California (Roberts, 2011) was delayed for almost a year in its phase II, partly due to the lawsuit brought forward by the Public Engineers in California Government which viewed the project as anti-union and anti-public engineer and sought to stop the phase II of the project. The I-77 High Occupancy Toll (HOT) lane project in North Carolina was questioned regarding a private/foreign company operating and tolling commuters, and the concerns on real estate values and local businesses (WidenI77, n.d.), contributing to the bill to cancel the contract with a private developer (Morrill, 2016).

The academic research studies and industrial examples reinstated the importance of public acceptance assessment. Given the change of people's opinion and the emergence of media events, an assessment methodology is needed to not only complete the assessment, but also complete it fast and inexpensively.

However, the current practice to assess public acceptance is not satisfactory. One of the most widely used methods in infrastructure projects is the public opinion poll (MacSween-George, 2003; Ulberg, 1995). While it has evolved over a long history to have a scientific methodology yielding good results, it still suffers serious drawbacks such as the expensiveness to conduct it (O'Connor et al., 2010) and the amount of time it takes to return the result (Heberlein, 1976). In some cases, poll results are not guaranteed to be accurate because people may hide their true feelings when answering poll questions. The 2016 presidential election is a good example of how the majority of polls predicted the opposite result.

The limitations of the public opinion poll make it unable to provide continuous assessment of public acceptance both economically and timely. Such a dynamic feedback loop can help project managers make data driven decisions and measure public acceptance changes according to changes of policies and public awareness. This research turns to social media to propose a light-weight, fast and inexpensive version of public opinion assessment.

2.2.3 Social Media

With the advance in Internet technologies and increased engagement on the web, social media has become a major portal to exchange ideas and share updates. Social media is a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and allow for the creation and exchange of user generated content (Kaplan & Haenlein, 2010). As of 2015, 90% of young adults and 35% of those 65 or older, 68% of women and 62% of men are using social media. Overall, 65% of adults, almost 10 times more than 10 years ago, are social media users (Perrin, 2015).

Social media not only provides an abundance of data from millions of users, but also offers interfaces to stream data with a low cost (depending on the volume) and a high speed. It close the gap of the aforementioned drawbacks of traditional public opinion assessment methods. Moreover, there are already research studies trying to extract information from social media and compare it with other traditional methods. (Asur & Huberman, 2010) uses a linear regression model to predict movies' box revenue based on almost 3 million tweets, with a result outperforming the predictions of the Hollywood Stock Exchange. (Ritterman, Osborne, & Klein, 2009) use Twitter to model the public belief that the swine flu will become a pandemic. Promising results were given to suggest that noisy social media is able to reflect public opinions. (Signorini, Segre, & Polgreen, 2011) demonstrated that Twitter can be used to qualitatively monitor public interest in influenza and quantitatively estimate disease activity in real time. (O'Connor et al., 2010) used Twitter data to capture the trend of polls for both consumer confidence and presidential job and Twitter data

analysis has proved to be a good supplement of polls. (Kryvasheyeu et al., 2016) used Twitter to estimate damage distribution of natural disasters, and demonstrated advantages of using social media such as speed, low cost and simplicity. All in all, the successful use cases of social media activities in describing and predicting the public bring confidence that it can also be used in infrastructure projects to assess public acceptance.

2.3 Project Evaluation Framework

To use social media for public acceptance evaluation, or furthermore, for project evaluation in general, we would like to develop a framework to define the components and functionalities, the workflow connecting them, and their interfaces. The framework aims to accommodate different infrastructure projects for their timely and cost-effectively evaluation. The purpose of the framework is to provide data driven project evaluation, including public acceptance analysis, public event analysis, user analysis, project legal risk analysis, etc.

2.3.1 Architecture

The architecture of the project evaluation framework is shown in *Figure* 2-1. It starts with Twitter, the primary data source of the framework. There are two major types of data to be retrieved from Twitter: the tweets, including the date and time when they are posted, the user who posted them, and the text of the tweet; and the user profiles, including users' location, follower count, website, etc. In addition, the raw data collected from Twitter can be further enriched to serve the needs of various analyses. For example, the web pages referenced in some tweets are wrapped in distinct tiny URLs. They can be restored to the original full URL, from which the title and the text of the web page can be retrieved. As another example, the user profile contains location information entered arbitrarily by user. They can be geocoded and normalized using geocoding services.

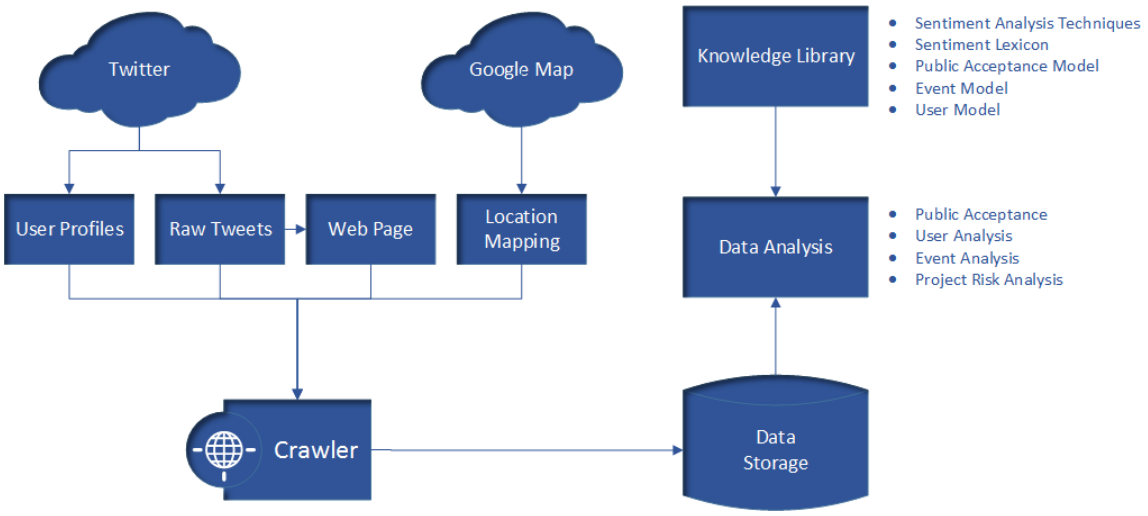

Figure 2-1 Architecture of the Project Evaluation Framework

A few web crawler applications need to be developed to fetch real-time data from Twitter and other proprietary data sources. Crawler is a software program which pulls data automatically from web pages to create a local repository (Cho & Garcia-Molina, 1999). For the Twitter crawler, the user could use one or multiple search terms to query Twitter API for relevant tweets. The search terms are provided by the user and typically contain keywords, user accounts and/or hashtags. The user account search (the @ sign) is used to search for tweets related to a certain user; the hashtag search (the # sign) is used to search for tweets of a certain topic, and the keyword search is a general purpose search. In our case study, the California High Speed Rail project, we will use a combination of keyword: "california high speed rail", user account: "@CaHSRA", the official organization account, and hashtag: "#CaHSRA". On top of the Twitter crawler, additional crawlers for enriched data dimensions are also required depending on the analysis to be performed. In our case study, we crawl the original website URL and its title to enable the grouping of web pages. We also utilized Google Maps to geocode location information within user profiles. Additional information can be crawled to support advanced analysis, including but not limited to, census data, political affiliation data and Wikipedia data, etc.

The data storage module is essentially a database management system which provides a repository for data input and output. It is responsible for storing the raw data from crawlers, running extract, load and transform (ETL) processes to prepare and join data, and assisting metric calculate for various analyses. It is also responsible for serving various requests from the data analysis module by providing data in the desired format.

The data analysis module is the core component of the framework. There are three key parts in this module: the data to be analyzed, the analysis methodology, and the knowledge library. The data is provided by the data storage module with proper aggregation, filtering and sorting. The analysis methodology is dependent on the specific evaluation. For example, semantic analysis uses various national language processing (NLP) techniques and machine learning techniques such as neural network (Collobert & Weston, 2008). Sentiment analysis uses lexicon based or machine learning based techniques (Pang & Lee, 2008). Other analyses include topic analysis, word frequency analysis, user segmentation and even manual screening etc. The framework user is responsible for choosing the methodology best suited for the analysis. In our case study, we examine multiple techniques and compare their advantages and disadvantages for infrastructure projects.

The knowledge library is a user defined model to describe features and characteristics of the analysis target and result interpretation. For example, our case study initializes the effort of establishing an infrastructure specific dictionary for sentiment words to be used for tweet sentiment analysis. We also define public acceptance in the context of social media and infrastructure project, and the calculation based on tweet sentiment analysis. Furthermore, we model social media event, opinion leader, opinion follower, original contributor and user demographic model, all of which contribute to the content of the knowledge library.

2.3.2 Data Source

The data sources used in this research include Twitter, our primary data provider, and other proprietary data sources such as Google Maps and the Internet. Twitter provides various application programming interfaces (APIs) for programs to access Twitter data, including tweets, users, entities, and places (Twitter. Inc, 2016). By authenticating a Twitter application and providing a list of input parameters, the APIs return a list of objects or attributes in JSON format.

There are currently two types of APIs supported by Twitter, the REST API and the Streaming API. The REST API provides access to read and write Twitter data, post or retweet tweets, and read or modify user profiles. A REST API call pulls data a single time, and is subject to the API rate limit, which is based on the number of requests per a 15-minute window. In this research, we use one of the REST APIs, the Search API, which allows users to query against a sampling of recent Tweets published in the past. Three pricing models are available for the Search API with the search window of 7 days, 30 days and full archive (back to 2006) respectively. It provides similar functionality like the search feature in Twitter mobile or web client. The Streaming API, on the other hand, allows users to monitor the stream of tweets and likes in real time by pushing tweets and other messages to the client. Two pricing models are available for the Streaming API with different limitations on filters (Twitter. Inc, 2016).

Streaming API is more suitable for short-term events such as the World Cup, a presidential election, or major product releases. It is expected to generate a burst of tweets in a short period of time where streaming is better than constant searching. Infrastructure projects, however, can be better supported by the Search API. These projects last longer, and the volume of tweets is expected to be less than popular public events, hence intermittent searching from Twitter should be able to cover all target tweets.

The Search API is built around query parameters that include keywords, hashtags, logical operators, and time frames, etc. A sample result of a tweet is shown in Figure 2-2.

```
{ ⊟
   "statuses":[ ⊟
      { ⊟
         "coordinates":null,
         "favorited":false,
         "truncated":false,
         "created_at":"Mon Sep 24 03:35:21 +0000 2012",
         "id_str":"250075927172759552",
         "entities":{ ⊞ },
         "in_reply_to_user_id_str":null,
         "contributors":null,
         "text":"Aggressive Ponytail #freebandnames",
         "metadata":{ ⊞ },
         "retweet_count":0,
         "in_reply_to_status_id_str":null,
         "id":250075927172759552,
         "geo":null,
         "retweeted":false,
         "in_reply_to_user_id":null,
         "place":null,
         "user":{ ⊞ },
         "in_reply_to_screen_name":null,
         "source":"Twitter for Mac",
         "in_reply_to_status_id":null
      },
      { ⊞ },
      { ⊞ },
      { ⊞ }
   ],
   "search_metadata":{ ⊞ }
}
```


Figure 2-2 Sample Tweet Result

A sample request of the Search API is shown below.

GET https://api.twitter.com/1.1/search/tweets.json?q=%23freebandnames&since_id=24012619984051000&max_id=250126199840518114 5&result_type=mixed&count=4

As shown in Figure 2-2, a tweet has multiple attributes such as create time, creator, text, and/or geographical tagging, etc. In this research, we primarily use create time, creator, and text for sentiment analysis, event analysis and user analysis. The rich set of dimensions, e.g. the geotagging

of the tweet, is able to provide extra information for further analyses which is out of the scope of this study.

A sample of a user returned by the Search API is shown in Figure 2-3. Similarly, a user has multiple attributes. In this research, we primarily use user name, user URL, user location, and counting metrics like followers_count, friends_count, and favourites_count for segmentation purposes. Other information such as the user network can help build the relationship among users which can be useful in social network analysis.


```
{ ⊟
   "id":2244994945,
   "id_str":"2244994945",
   "name":"TwitterDev",
   "screen_name":"TwitterDev",
   "location":"Internet",
   "profile_location":null,
   "description":"Developer and Platform Relations @Twitter. We are developer advocates. We can't answer all
   "url":"https://t.co/66w26cua1O",
   "entities":{ ⊞ },
   "protected":false,
   "followers_count":429831,
   "friends_count":1535,
   "listed_count":999,
   "created_at":"Sat Dec 14 04:35:55 +0000 2013",
   "favourites_count":1713,
   "utc_offset":-25200,
   "time_zone":"Pacific Time (US & Canada)",
   "geo_enabled":true,
   "verified":true,
   "statuses_count":2588,
   "lang":"en",
   "status":{ ⊞ },
```


Figure 2-3 Sample User Result

Another important data source referenced in this research is Google Maps. It is a web service providing APIs for directions, elevations, and places, etc. In this research, we use the geocoding API to formalize the location entered by users and categorize them into administrative levels such as country, state, and county. A sample geocoding request is shown as follows (Google, n.d.):

https://maps.googleapis.com/maps/api/geocode/json?address=1600+Amphitheatre+Parkway,+Mountain+View,+CA&key=YOUR_API_KEY

A sample return result from Google Maps is shown in Figure 2-4. A hierarchical location structure is constructed using Google Maps search, which is very helpful in unifying and grouping geographical information.


```
{
   "long_name":"Santa Clara County",
   "short_name":"Santa Clara County",
   "types":[
      "administrative_area_level_2",
      "political"
   ]
},
{
   "long_name":"California",
   "short_name":"CA",
   "types":[
      "administrative_area_level_1",
      "political"
   ]
},
{
   "long_name":"United States",
   "short_name":"US",
   "types":[
      "country",
      "political"
   ]
},
```


Figure 2-4 Sample Google Map Result

As mentioned above, Twitter and Google Maps are only two of the many data sources available from the Internet. Numerous data sources can be included in the framework such as public surveys, census, and state level statistics etc.

2.3.3 *Data Crawler*

Four data crawler applications are developed to fetch data from the data sources mentioned above. The purpose of these software applications is to periodically update the repository of new tweets, new user profiles, new websites referred to in new tweets, and new locations from new users.

Every 7 days, the Twitter crawler program is triggered to query the Search API using a pre-defined search term. A list of new tweets collected during the time frame are fetched and stored into a staging area in the data storage module. Data cleansing ETL processes are necessary to remove bad data and duplicate entries and merge all tweets into one primary dataset where each tweet is uniquely identifiable. New users are discovered from new tweets and their profiles are fetched by the user crawler using user API. New tiny URLs from the increment tweets are fed into the URL crawler to restore full URLs. Furthermore, new locations from new users are processed by the location crawler. This process is illustrated in Figure 2-5 and Figure 2-6.

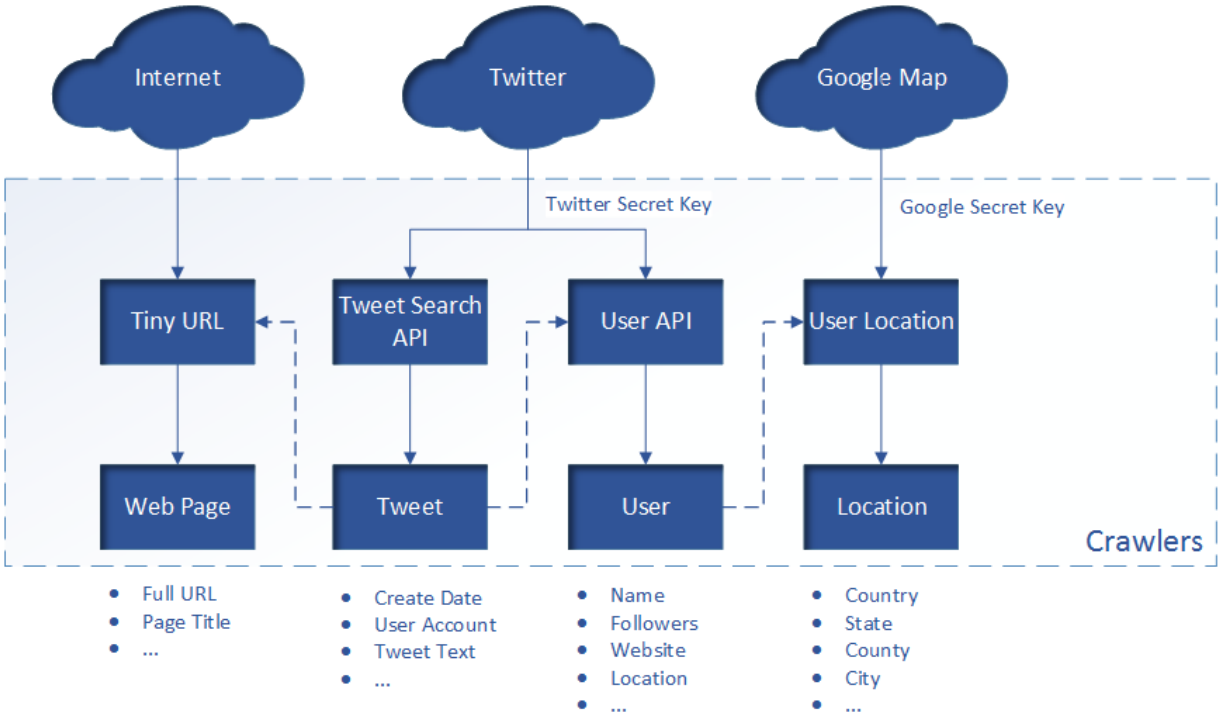

Figure 2-5 Crawler Workflow

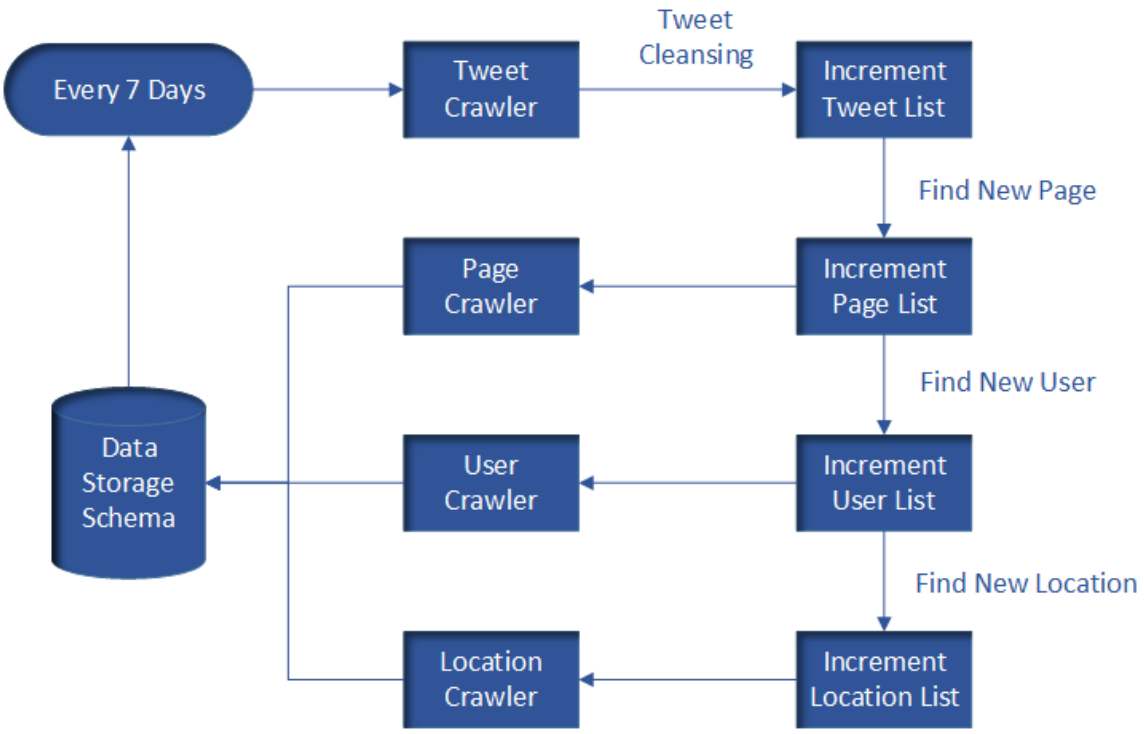

Figure 2-6 ETL Workflow

2.3.4 Data Storage

The data storage module is a database management system serving as an interactive data repository of the crawler, the data analysis module, and other subsequent reporting/presentation layers. The data storage module essentially defines the entities and relationships of the objects fetched from social media and proprietary data sources. It is also responsible to execute scheduled ETL jobs for various tasks. The entity-relationship model of the social media is shown in Figure 2-7.

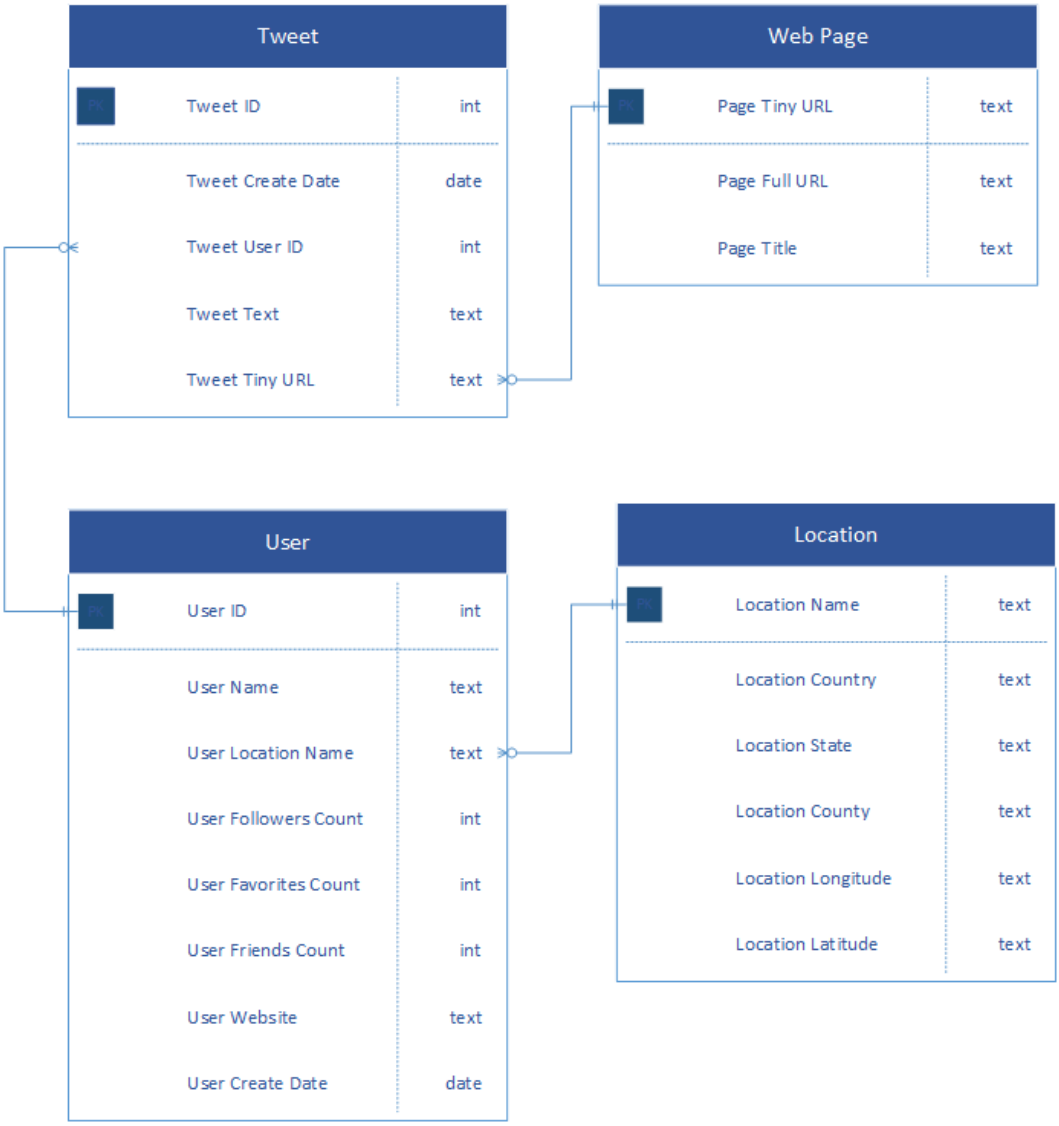

Figure 2-7 E-R Model of Social Media Objects

The entities identified in this framework are tweet, user, web page and location, where tweet is a child of user as well as a child of web page. User is a child of location. As mentioned in the data crawler section, the data storage module is responsible to clean up tweets by trimming and removing unnecessary characters like white space and carriage returns, merging the staging table with the final result table, and removing duplicate entries. It is responsible for identifying new users and new locations from the stream of new tweets. In addition, the data storage module supports all the query needs of the data analysis module and other presentation layers.

2.4 Data Analysis

The data analysis module is the core module of the framework because all the discoveries and predictions are based off the results of this module. Depending on the goals and requirements of the project evaluation task, the data analysis module determines the types of analysis to perform. With the help of the user knowledge library which provides the modeling and technical support customized for infrastructure projects, as well as the data feed from the data storage, the data analysis calculate various measurements and produce analytical results.

The foundation of social media analysis is text analysis. There are different types of text analyses available, such as word frequency analysis, sentiment analysis, text clustering, and entity recognition etc. (Zoss, 2017) listed some popular approaches used in processing of text. There are also a lot of tools available for text analysis, including R, an environment integrating software for data manipulation and calculation (Venables, Smith, Team, & others, 2004), WEKA, an open source system built for machine learning and data mining (Hall et al., 2009) and RapidMiner, an open source data science platform for data analysis, text mining, web mining, and sentiment analysis (RapidMiner, 2014). Besides these specialized text analysis platforms, most of the popular computer languages have text mining and natural language processing libraries and toolkits available, such as the natural language toolkit (Bird, Steven, Loper, & Klein, 2009) and the topic modelling library (Rehurek & Sojka, 2010) for Python.

Many of the text analysis platforms and toolkits sue general purpose techniques and training sets, which might not perform well with tweets, or infrastructure project related tweets. To improve the performance of the analysis, the knowledge library needs to be developed to determine the most effective techniques and metrics to measure and monitor. Given the lack of research on infrastructure project assessment using social media, it is critical to start the endeavor to build a domain specific knowledge library for this subject area. In chapters 3, 4 and 5 we contribute to the establishment of the knowledge library of infrastructure project evaluation by optimizing the

techniques in terms of data fetching and tweet sentiment analysis. Customized sentiment lexicon is developed to improve the accuracy of the sentiment analysis algorithm. We also develop a series of models around public acceptance, including the public acceptance model, social media event model, opinion leadership model, and the user profile model.

Discussed below are some analyses which can be supported by the framework. Sentiment analysis, event analysis, user analysis and project legal risk analysis are highlighted, though the framework is a general purpose one and more sophisticated analyses can be supported.

2.4.1 Sentiment Analysis

Sentiment analysis, also referred to as opinion mining, is a field of study which applies machine learning, natural language processing, and text analysis to identify "what other people think". Sentiment analysis can be used in applications to review related websites, as a sub-component technology, in business and government intelligence, and across other domains (Pang & Lee, 2008). It fits naturally with public acceptance evaluation which is based on the positivity and negativity of the public opinions.

Sentiment analysis is typically conducted at three different levels: document level, which identifies the sentiment expressed in a whole document; sentence level, which determines the polarity of each sentence and/or the subjectivity of the sentence; and entity and aspect level, also called feature level, which identifies not only the sentiment of the expression, but also the specific target of the sentiment (Liu, 2012).

Generally there are two types of sentiment analysis techniques: unsupervised and supervised. (Turney, 2002) created an unsupervised learning technique and (Pang, Lee, & Vaithyanathan, 2002) reviewed three supervised machine learning techniques including Naive Bayes, maximum entropy classification, and support vector machines. While these research studies are focused on document

level sentiment analysis, (Hu & Liu, 2004) proposed an opinion summarization technique for sentence level analysis.

This research focuses on sentence level sentiment analysis because Twitter once had the limitation of 140 characters per tweet. Although there have been changes made to relax this limitation, it is still valid to the corpus obtained in this research. Both techniques will be implemented in order to baseline the performance. The machine learning based approach takes advantage of the Aylien text analysis module in RapidMiner, which uses supervised machine learning techniques for the sentiment classifier (Barnaghi, Ghaffari, & Breslin, 2016). The training set behind the module is generally trained for all tweets, not specifically for infrastructure project related ones. Given the early stage of this research, the number of tweets available for training is not sufficient to support the training for infrastructure projects. The work has been started, but this research still relies on general machine learning toolkits.

This research also develops a lexicon based sentiment analysis process similar to the one proposed by (Hu & Liu, 2004). This application treats tweets as bag of words and does not require training datasets to work effectively. There are a few sentiment dictionaries publicly available for research use. In this research, we uses the dictionary maintained by (Hu & Liu, 2004), and tries to adapt the dictionary to infrastructure projects to yield better results. In future work, more tuning will be performed to build a domain specific sentiment dictionary for all infrastructure projects. The details of this application and the tuning of the dictionary will be discussed in chapter 3.

With the calculation of sentiment polarities for all tweets, a public acceptance evaluation model is also developed to assess public acceptance in a time series. Public acceptance can be measured by a one-vote-per-tweet model or by a one-vote-per-user model. Details of the public acceptance model are demonstrated in chapter 3 where the case study and the application of different public acceptance models are discussed.

2.4.2 Event Analysis

It is observed that public acceptance fluctuations are often triggered by massive retweet of certain web pages such as news articles and announcements. In addition to public acceptance analysis, it is interesting to analyze the events behind the scene which drives the burst of tweets and how they influence public acceptance.

Taking advantage of the web page crawler which restores tiny URLs to original URLs, the tweets can be grouped by web pages to analyze the group behavior. This research develops an event model which defines, detects, and categorizes events. A two-dimensional model of event influence measurement is developed to measure both the impact and duration of an event. Based on the measurement, strategies to mitigate the impact of negative events and enhance that of positive ones is also discussed. A detailed demonstration of the event analysis is provided in chapter 4 along with the case study.

2.4.3 User Analysis

Twitter is not only a collection of tweets, but a dynamic network of tweets and their posters. The user is the agent who spreads the events and causes public acceptance changes. Users in social media are not merely strangers or virtual accounts, they have their own behaviors and characteristics which can be revealed by social media itself. Taking advantage of the abundance of data generated by social media, different groups of users and their behaviors and influence can be analyzed.

Two possible analyses is proposed by this research to cluster users in different groups. The opinion leadership model, which is based on the number of retweets of a certain user, is designed to discover opinion leaders, opinion followers and original contributors. The user profiling model, which is based on multiple demographic attributes including sentiment, popularity, institution and location, is developed to describe user attributes. User analysis is an important study to find out the leading

people in the world of social media for targeted campaigns and lobbies. The detailed discussion of user analysis can be found in chapter 5.

2.4.4 Project Legal Risk Analysis

The analyses proposed above attack project evaluation from different perspective. Integrating them together, it is possible to derive a project legal risk analysis based on social media. Traditionally, project risk evaluation relies heavily on expert opinion, which is subjective and the result could vary among different experts. With the help of social media, it is possible to evaluate project risk using a data driven approach.

For example, a project's legal risk can be determined by multiple factors. Firstly, the overall public acceptance sets the tone of legal risk to be likely or unlikely. Secondly, institution accounts' sentiment can be investigated separately to reveal whether there are organized oppositions. Thirdly, certain threatening words can be detected from institutional accounts' tweets to alarm possible legal actions that are underway. The development of the legal risk model is out of the scope of this research, however, it is a valuable application in future research.

2.5 Conclusion

In this chapter, we propose a new project evaluation framework based on big data generated by social media. This framework is aiming to solve the problem of the current expensive and time-consuming process of retrieving public opinion and evaluating the project. The architecture of the framework is introduced, and the components of the framework including data source, data crawler, data storage, and data analysis are discussed in detail concerning their responsibilities, workflows, and data structures.

Following the discussion of the framework, several sample analyses are proposed to facilitate and enrich the evaluation of a project. Sentiment analysis, event analysis, user analysis and project legal

risk analysis are discussed to demonstrate the capability of the framework. More discussions of these models can be found in subsequent chapters.

Meanwhile, we would like to initialize the effort of building the knowledge library for infrastructure projects. Our contributions to the library include the development of the domain specific sentiment dictionary, the public acceptance model, the event model, and the user model, all serving to provide multi-dimensional evaluation of infrastructure projects. Although the models proposed are derived from the case study, the framework is versatile to conduct analyses beyond the listed ones. By providing a standard process to obtain, store, and analyze data from social media, we expect the future project evaluation to be nimbler and data driven, hence more accurate and useful to infrastructure project development.

Chapter 3. Evaluation of Public Acceptance Using Big Data – A Case Study on Public Acceptance

3.1 Introduction

In order to solve the problem of a real-time and cost-effective public acceptance assessment of infrastructure projects, a project evaluation framework was proposed in the last chapter to collect and analyze public opinion using social media and big data. After the description of conceptual modules and analyses, we would like to apply this framework on a real-world project to examine its usability and limitations. In this chapter, we would like to address the concerns listed below from the case study.

3.1.1 Feasibility to Retrieve Quality Data from Twitter

(O'Connor et al., 2010) and (Kryvasheyeu et al., 2016) have done research regarding presidential elections, presidential approval, and natural disasters. (O'Connor et al., 2010) used 1 billion tweets from 2008 to 2009 (100,000 to 7 million messages per day) collected by querying the Twitter API and the "Gardenhose" real-time stream. (Kryvasheyeu et al., 2016) obtained the raw data for Hurricane Sandy from a tweet archiving company with the hashtag "#sandy" and with a set of specific keywords. They retrieved 52.55 million messages from 13.75 million unique users posted in 2012 between the 15th of October and 12th of November.

However, infrastructure projects are different from these events in previous research studies. First of all, infrastructure projects are rarely national or international, hence the people who are interested in and/or affected by these projects form a relatively small population. A typical infrastructure project generates much fewer tweets when compared to national and international events. Secondly, infrastructure projects last much longer than those short-term events. Infrastructure projects usually take a few years to complete, which makes the data collection process a long-term and continuous

effort. The requirement of the system is different than that of systems designed for a one-time outburst of tweets. Thirdly, on the technical side, tweets related to infrastructure projects tend to be noisier since they often refer to keywords such as road or venue names. Tweets containing these keywords could be blended with other unrelated topics such as traffic and accident reports, which might impair the quality of the data collected.

Similar to (O'Connor et al., 2010) and (Kryvasheyeu et al., 2016), we also query Twitter API to get the majority of the data feed. In addition, we also use the Twitter account and hashtag topics to search relevant tweets. The contribution of these search terms to the corpus could vary a lot, hence it is necessary to examine the data volume and data quality of the search terms to provide guidance for future research and applications. This chapter uses the case study to investigate data retrieval characteristics and study how to obtain quality data from Twitter.

3.1.2 Sentiment Analysis Methodology

In this research, sentiment analysis is the most important analysis to help determine public acceptance and project risk. Performing sentiment analysis on tweets is difficult due to the lack of context, the limitation on the length of the tweet and the use of Internet slangs. As mentioned in the last chapter, sentiment analysis techniques include the machine learning based approach and the lexicon based approach. Many research studies have been conducted to create numerous methods in modeling and implementation. This research applies 3 different algorithms on the case study data set retrieved and compares their performance. A lexicon based sentiment analysis algorithm is developed and, with the help of an infrastructure project specific sentiment dictionary, is customized from a general purpose sentiment lexicon.

3.1.3 Public Acceptance Model

Public acceptance of infrastructure projects uses the result of sentiment analysis to depict the degree of support of the public. Different political perspectives, i.e. the pluralistic model and the elite

model, provide different interpretations on how individual tweet sentiment reflects an individual's sentiment inclination, and how an individual's sentiment should be aggregated to calculate the overall acceptance. This research discusses different mapping strategies from an the sentiment polarity of individual tweets to public acceptance by using the number of tweets, the number of users and the weighted number of users by user popularity, and compares the performance of these strategies.

3.2 Literature Review

3.2.1 Using Social Media for Prediction

Multiple players on the Internet, especially the social media industry, contributed to the explosion of information. As of 2014, in every minute, Google receives over 4 million search queries, Youtube users upload 72 hours of new video, and Twitter users tweet 277,000 times (James, 2014a).

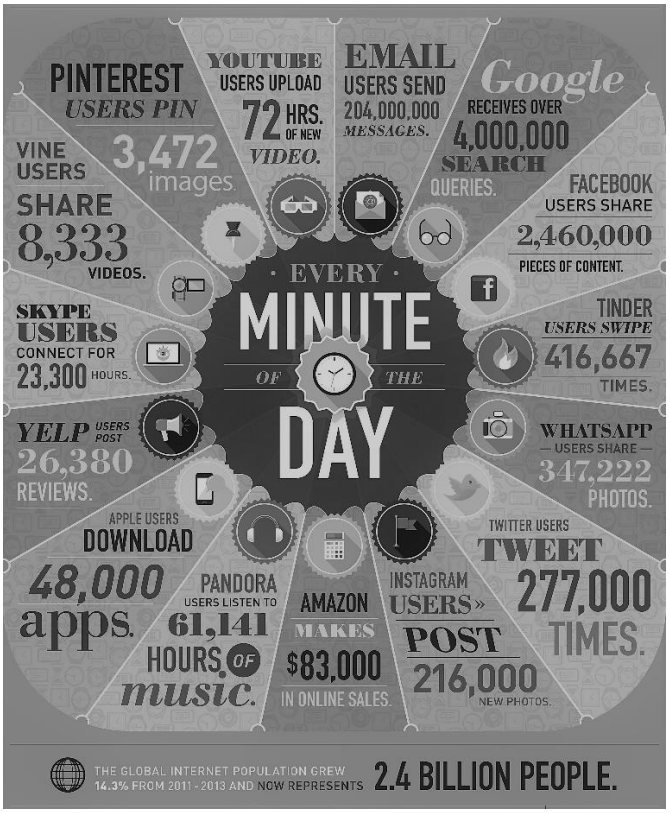

Figure 3-1 Data Generated Every Minute (James, 2014b)

Among all the popular social media sites, Twitter has been extensively used as a research platform for opinion predictions. It has a privacy policy favorable to research, a relatively large user base, and a mature set of APIs. (Asur & Huberman, 2010) used a keyword search from movie titles to extract 2.89 million tweets for 24 different movies. On top of the data they built a linear regression model to predict box-office revenues of movies. (Wang, Gerber, & Brown, 2012) searched and collected tweets posted by traditional news stations and newspapers. They used a semantic role labeling method to conduct a criminal incident prediction. (Ritterman et al., 2009) crawled about 1 million tweets per day and used the prediction markets as the aggregation mechanism and the support vector machine as the classification system to predict a swine flu pandemic. (O'Connor et al., 2010) collected 1 billion tweets from Twitter API for 2008 and 2009 to analyze consumer confidence and presidential approval polls. They found that using the ratio of positive versus negative messages on the topic, the analysis resulting from the tweets can replicate the results of traditional polls. (Kryvasheyeu et al., 2016) used hashtag "#sandy" and a keyword search in about a month and collected 52.55 million messages from 13.75 million unique users. They found that "the per-capita number of Twitter messages corresponds directly to disaster-inflicted monetary damage."

Therefore, Twitter acts not only as a social media, but also a research platform to enable crowdsourcing analysis on various subject areas. Most of the available research studies yield positive results, i.e. despite the noise and inaccuracy in the raw Twitter data, tweets are able to reflect and predict public acceptance, stock, revenue or other predictive measurements. Based on these previous studies, this research would also use Twitter as the source of data to construct a real-time project evaluation system to facilitate the decision making of project stakeholders, and study the similarities and differences between infrastructure data sets and others.

3.2.2 Twitter Sentiment Analysis

Much attention in academia has been given to tweet sentiment. Generally speaking, there are two types of methods, lexicon based methods and machine learning based methods. The lexicon based method uses a pre-defined dictionary, which contains a list of sentiment words and sentiment polarities, and applies algorithms to negate or intensify the sentiment to determine the sentiment of a sentence. The machine learning based approach uses the algorithms for text categorization and applies them to sentiment classification (Tang et al., 2014). (Pang et al., 2002) favors the machine learning based approach because of the subtle nature of sentiment expression. Sentences with words containing no obvious sentiment, such as "How could anyone sit through this movie?", can naturally have a strong sentiment. Machine learning algorithms can better "understand" the meaning of the sentence than the lexicon based approaches. However, (Thelwall, Buckley, & Paltoglou, 2012) argue that machine learning algorithms depend too much on the training datasets, which is usually human coded whose accuracy is skeptical. Machine learning classifiers are also optimized to a specific domain, e.g. Iraq, Iran, Palestine and Israel, which could signal negative indicators to a trained political classifier, though these words are by themselves neutral in other contexts. Another representative lexicon based approach is from (Turney, 2002) who assessed Pointwise Mutual Information based on words "excellent" and "poor", and used the difference to determine their sentiment orientation. This approach reached an accuracy of 74.39%.

3.2.3 Infrastructure Projects and Public Acceptance

Research on public acceptance assessment of infrastructure projects is limited. Most of the existing studies rely on traditional opinion gathering methods such as public hearings and public opinion polls (Cole & Caputo, 1984; Heberlein, 1976). These methods are primarily developed to offset the

difficulties in data collection in the past. The methodologies focus on problems such as how to reach out to certain groups of people, how to ask fitting questions to collect their feedback, and how to analyze the received responses. These methods are still valid and effective, however, the times have changed dramatically from data deficiency to information explosion. The past difficulties in data collection will be mitigated by using the abundance of user-generated social media data, and the real challenge now is how to effectively filter out the data in need and process them in a timely fashion.

Recently, some Chinese scholars have pioneered an endeavor that assesses a large hydropower project using social media. (Jiang, Lin, & Qiang, 2015) proposed a project sentiment analysis (PSA) system to assess public opinion of the Three Gorges Project. The system collects, processes, and classifies data from a Chinese social media site using a lexicon-based approach and provides intelligence such as frequently used words and word cloud analysis.

(Jiang, Qiang, & Lin, 2016) extends the system to not only give the sentiment value of each text, but also to analyze the spatial and temporal sentiment post intensity and sentiment polarity. A list of topics exhibited in the negative and positive corpus is also constructed to illustrate the implications of the hydropower project on the public.

Previous research brings confidence in applying social media on infrastructure project evaluation. This research studies data retrieval strategies, sentiment analysis methodologies and public acceptance models using a real-world case study, the California High-Speed Rail project. The case study proves the feasibility of the project evaluation framework and observations are made in comparing different strategies and models.

3.3 Case Study of the California High-Speed Rail

3.3.1 Overview of the California High Speed Rail Project

The California High-Speed Rail (CAHSR) is the first high-speed rail system in the nation (California High-Speed Rail Authority, 2016a). It will connect northern California (San Francisco and Sacramento) to southern California (Los Angeles and San Diego) and major cities in the state. The system will have a total length of 800 miles with up to 24 stations. It will operate with a speed up to 220 miles per hour (350 km/h) and provide service from San Francisco to Los Angeles in under 3 hours.

Figure 3-2 Map of the California High-Speed Rail (California High-Speed Rail Authority, 2016b)

The Californian pursuit for a high-speed rail can be dated back to 1981 (California High-Speed Rail Authority, 2017). In 1996, the California High-Speed Rail Authority was created by the California Legislature to plan, design, and operate a high-speed rail system to connect California. In 2008, Proposition 1A, the bill which authorized a $9.95 bond measure to support the initial

construction of the California High-Speed Rail Project, was approved by the Californian voters. In 2009, $8 billion in national funding was established according to the American Recovery and Reinvestment Act (ARRA) and California secured $3.3 billion. In 2012, the Legislature approved almost $8 billion for construction in the Central Valley. Now that the construction is under way, the initial operating segment (IOS) is expected to complete in 2022. The phase 1 blended system is expected to complete in 2029, which will connect San Francisco with Los Angeles.

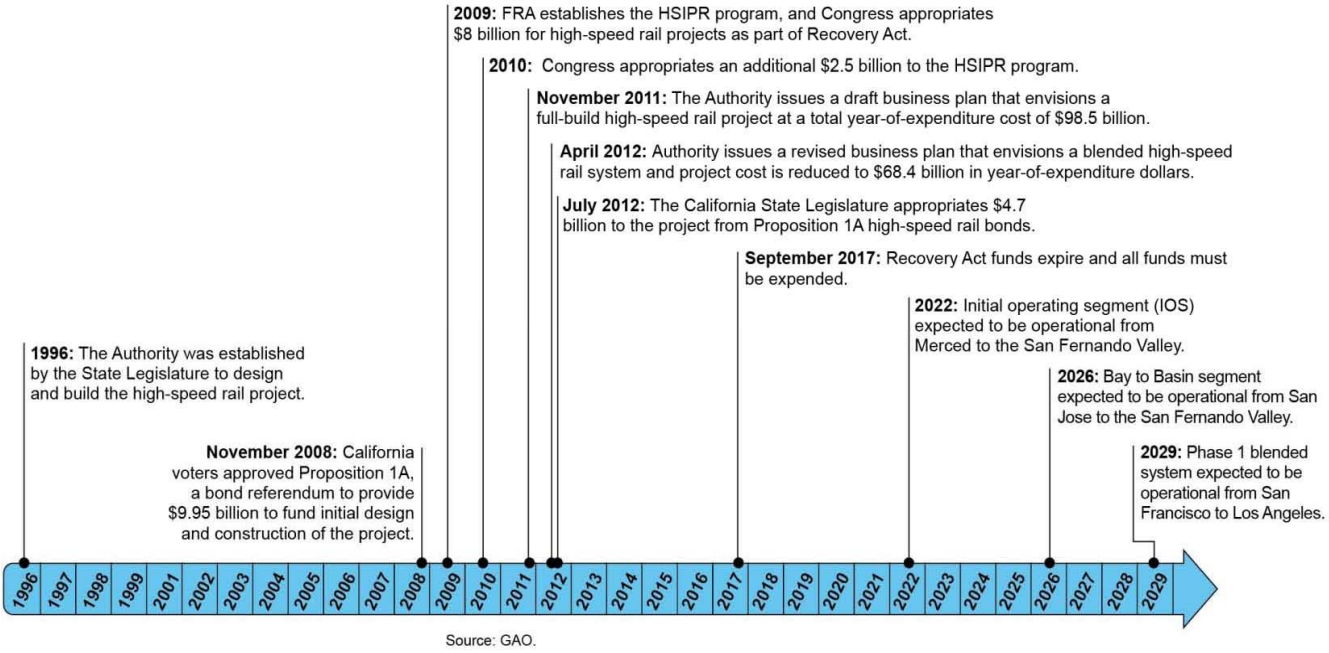

Figure 3-3 California High-Speed Rail Timeline (United States Government Accountability Office, 2012)

The CAHSR project can be roughly divided as five segments: Silicon Valley to Central Valley, Bakersfield to Burbank, Burbank to Anaheim (altogether the Phase 1 of CAHSR), Los Angeles to San Diego, and Sacramento to Merced (altogether the Phase 2 of CAHSR). The total cost estimate of the Phase 1 is $64.2 billion (California High-Speed Rail Authority, 2016b). Currently under construction is the Silicon Valley to Central Valley line, with an estimated cost of $20.679 billion.

Table 3-1 Capital Cost Estimates: San Jose – North of Bakersfield (Silicon Valley to Central Valley Line) (in Millions) (California High-Speed Rail Authority, 2016b)

FRA STANDARD COST CATEGORIES	2015$	YOE$
10 – Track structures and track	$7,038	$7,851
Civil (10.04–10.06, 10.08, 10.18)	*$1,061*	*$1,153*
Structures (10.01–10.03, 10.07)	*$5,147*	*$5,769*
Track (10.09, 10.10, 10.14)	*$830*	*$929*
20 – Stations, terminals, intermodal	$279	$308
30 – Support facilities: yards, shops, administrative buildings	$193	$219
40 – Sitework, right-of-way, land, existing improvements	$4,910	$5,309
Purchase or lease of real estate (40.07)	*$1,302*	*$1,345*
50 – Communications and signaling	$468	$528
60 – Electric traction	$1,108	$1,258
70 – Vehicles	$774	$865
80 – Professional services (applies to categories 10–60)	$2,994	$3,249
90 – Unallocated contingency	$985	$1,091
100 – Finance charges	-	-
TOTAL	**$18,749**	**$20,679**

Currently the funding sources of the Phase 1 of CAHSR include federal grants of $3.48 billion, Proposition 1A bond proceeds of $9.95 billion, and Cap and Trade proceeds of about $500 million per fiscal year. Taking into account the appropriations for environmental related activities, the funding available to construct the Silicon Valley to Central Valley line is listed in Table 3-2.

Table 3-2 Funding Available for Planning and Construction for San Jose – North of Bakersfield (Silicon Valley to Central Valley Line) (California High-Speed Rail Authority, 2016b)

FUNDING SOURCE	AMOUNT (IN MILLIONS)
APPROPRIATED FUNDS	
State Bonds (Proposition 1A)	$2,609
Federal Grants (ARRA/FY10)	$3,165 *
Planning Funds	$338 **
COMMITTED FUNDS	
State Bonds (Proposition 1A)	$4,166
Cap and Trade (Through 2024)	$5,341
FINANCING PROCEEDS	
Long-term Cap and Trade (2025-2050)	$5,237
Total Sources of Funds	**$20,856**
Construction Cost (see Section 5)	$20,680
Reserve	**$176**

3.3.2 Oppositions and Legal Challenges

The CAHSR project is a large-scale infrastructure project which impacts a lot of residents and businesses and could potentially incur strong oppositions and lawsuits. Citizens for California High Speed Rail Accountability (CCHSRA), an organization of people affected by the CAHSR project, is working to hold CAHSRA accountable for the economy, environment, and other impacts brought on by the project (CCHSRA, n.d.). They led the lawsuit of *John Tos, Aaron Fukuda, County of Kings v. California High Speed Rail Authority, et al* which argued that CAHSR was supposed to be a dedicated, not blended, track system, that the current plan cannot support the proposed speed limit, and that the Proposition 1A voted for is not what is actually being executed (CCHSRA, 2016). This lawsuit cost the project about $63 million and 17 months of delay (The Fresno Bee, 2016). There were a few other lawsuits involving the environmental certification, the use of cap-and-trade money; and the preemption of enforcing the California Environmental Quality Act (The Fresno Bee, 2015). CAHSRA is winning those lawsuits, though with heavy setbacks in the schedule and extra costs. Moreover, as the project proceeds, new lawsuits are likely to emerge.

CAHSR is one of the most highlighted infrastructure projects in the nation, which is expected to be controversial and newsworthy on both traditional media and social media. This is one of the reasons why it was picked as the case study project of this research. Meanwhile, it can be beneficial for CAHSR to take advantage of the social media sensation it generates to discover the public acceptance of the project.

3.3.3 Selection of CAHSR as the Case Study Project

This research selects the CAHSR project as the case study project with consideration of the aspects listed below.

- CAHSR is an ongoing project sparking continuous discussion on Twitter. Unlike previous research studies, where a huge dump or even the full archive of data is downloaded for researching,

this research focuses more on the real-world application of the framework by following the progress of CAHSR. Such an ongoing project provides the opportunity to test the real time implementation, and is more cost-effective compared to purchasing a large amount of historical data.

- CAHSR is a large-scale project that has attracted a lot of attention. Many people are impacted by this project, which could trigger a large volume of Twitter activity. Data volume is critical to the accuracy of data analysis, hence a large-scale project is more suitable for research purpose than smaller ones. As a pioneer research study, we would like to select the CAHSR project that guarantees the adequacy of data.

- CAHSR is controversial. The benefits brought by the project are evident as it could strategically improve the transportation and economy of the state of California. On the other hand, the adverse impacts of the project are real and tangible. Questions about damages, financial feasibility, and scandals will impair the public image of the project. Therefore, discussions, debates, and lawsuits are always part of the project, providing data from both supporters and opponents. Such a sentimental data feed matches perfectly with our research needs.

Based on the selection criteria of an ongoing, large-scale and controversial infrastructure project, CAHSR is the best candidate among all the projects evaluated. It stands out with the volume and quality of the data. In future research, we will include other projects for evaluation as well.

3.4 Data Characteristics

3.4.1 Data Volume

Typically, Twitter-based research targets trending events that will spur a massive number of tweets. The aforementioned events such as presidential elections, sporting events, and natural disasters can easily generate millions of tweets within a few months and sometimes lasting for only a few days. Infrastructure projects, however, last much longer but get much less attention. Projects as large as CAHSR are still local and can only generate local tweets. The length of the construction lifecycle

also contributes to the loss of momentum. After all, infrastructure projects are not as exciting as entertainment or political events. For our case study, we were able to collect 24,855 tweets between *2016-06-10* and *2017-10-22* with an average of 49.8 tweets per day. Without retweets, there are 10,403 original tweets with an average of 20.9 tweets per day. We pulled data for a few candidate projects and CAHSR generated the highest number of tweets. The daily data volume is shown in Figure 3-4.

The data volume acquired for CAHSR is much lower than in previous research. However, this is still a big improvement to the status quo. Compared with polls and public hearings which take months to gather information, Twitter is able to provide a stream of data from 14,546 people over 17 months. What makes it more valuable is the continuous feed of data, which depicts the dynamic change of public acceptance over time, is something traditional methodologies cannot provide. The status and the effect of actions become more measurable using the social media approach.

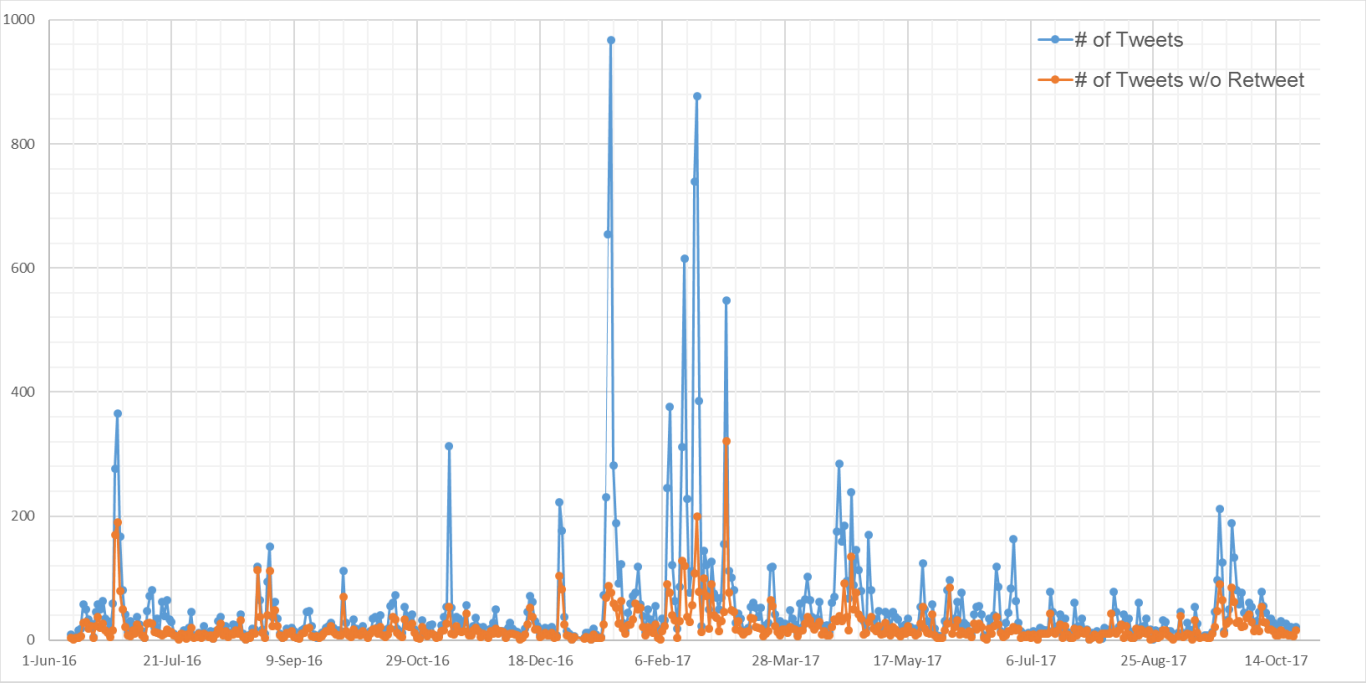

Figure 3-4 Daily Twitter Activities

3.4.2 Data Retrieval Difficulty

It is not easy to retrieve desired tweets for infrastructure projects. Most Twitter based research studies use hashtags and keywords to search for tweets, but this approach might increase the noise level for infrastructure projects due to potential name conflicts. Take the I77 HOT lane project as an example, when searching for keyword or hashtag "I77", most tweets returned are traffic congestions or accidents on I77 rather than sentimental expressions to the HOT project. CAHSR is selected partly because we can still use the traditional approach to get data, however, in future work a more sophisticated data filtering / noise cancellation mechanism need to be developed to accurately locate tweets for any infrastructure projects.

3.4.3 Search Terms

Three types of search terms are used in the case study:

- Search for a specific account (the @ sign). @CaHSRA is the official account for California High-Speed Rail Project.
- Search for a specific topic (the # sign). #CaHSRA is a hashtag topic people refer to when posting about CAHSR.
- Search based on keywords. In the case study, keyword string "california high speed rail" is used.

Different search terms can achieve different performance in data retrieval. The returned volume of each search term is shown in Table 3-3. Using the sentiment analysis result in the later part of this chapter, the volume of each sentiment polarity per search term is shown in

Table 3-4. Please note that the search terms are not mutually exclusive, and one tweet could satisfy multiple search terms.

Table 3-3 Twitter Activity Volume Comparison among Different Search Terms

Search Term	Twitter Activities	Twitter Activities w/o Retweets
@CaHSRA	5725 (23.0%)	2154 (23.5%)
#CaHSRA	1538 (6.2%)	198 (2.2%)
California High Speed Rail	19743 (79.4%)	7210 (78.5%)
Total	**24855**	**9184**

Table 3-4 Twitter Activity Sentiment Comparison among Search Terms

Search Term	Positive	Neutral	Negative
@CaHSRA	1602 (34.2%)	3444 (26.4%)	679 (9.5%)
#CaHSRA	705 (15.0%)	768 (5.9%)	65 (0.9%)
California High Speed Rail	3494 (74.5%)	9890 (75.9%)	6359 (89.1%)
Total	**4690**	**13026**	**7139**

As can be seen from Table 3-3 and

Table 3-4, keyword search contributed almost 80% of the total tweets in the corpus. Hashtag search, on the other hand, contributes the least (6.2%). As observed from the case study, the most critical search term for infrastructure project is the keyword string. One should not expect a lot of tweets mentioning official project accounts or hashtags.

3.5 Sentiment Analysis Methods

In the last section, it is proven that Twitter is able to provide project related feedback with a much larger volume than traditional methods to be used for further analysis and intelligence. Once the source data is retrieved from Twitter, it is crucial to apply effective data analysis techniques to interpret the data. Out of the many data analysis techniques, sentiment analysis is the most important one since the sentiment polarity directly lead to the result of public acceptance of an infrastructure project.

Sentiment analysis builds a classifier which categorizes text strings into different sentiment polarities i.e. positive, neutral, and negative. As discussed in the literature review, there are two types of sentiment analysis approaches, including the machine learning based approach, a supervised method needing pre-labeled data, and the lexicon based approach, an unsupervised method where a sentiment lexicon is used which pre-marks the sentiment polarity of certain words. The latter approach is more generic and relatively easy to implement, but in the case of sarcasm and sentimental expression without signaling words, the performance might be worse than the former. Undoubtedly, the quality of the sentiment lexicon is critical to the success of the lexicon based approach.

F score is a common measure of the performance of the classifier instead of accuracy alone. It is a combined measurement of precision and recall that is especially useful when the classified classes are highly skewed. Precision and recall are defined as (Olson & Delen, 2008):

$$\text{Precision} = \frac{tp}{tp + fp} \tag{3.1}$$

$$\text{Recall} = \frac{tp}{tp + fn} \tag{3.2}$$

Accuracy and the F score are defined as:

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \tag{3.3}$$

$$\text{F} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.4}$$

In this research, we test 3 sentiment analysis tools and algorithms: the Aylien text analysis API, the SentiStrength text analysis application, and our customized sentiment analysis algorithm. A sentiment analysis baseline dataset is constructed (Appendix E) with a subset of tweets labeled manually of their sentiment. These candidate tools and algorithms are applied to the baseline and their performance scores are measured to determine the most suitable method. The best algorithm

discovered in baselining is then applied to the corpus to obtain the sentiment of each individual tweet. Due to the lack of a training set and a sentiment lexicon specifically for infrastructure projects, ready-to-use software packages and dictionaries are primarily used in this study.

3.5.1 Sentiment Analysis Baseline

A baseline dataset was constructed to measure the performance of different sentiment analysis methods. 400 unique tweets were randomly selected, and their sentiment have been manually marked. During the process, unusable data were removed from the dataset, resulting in the final dataset containing 347 unique tweets, yielding 5.2% margin of error, and distributed with 226 (65.1%) negative, 40 (11.5%) neutral, and 81 (23.3%) positive tweets. The overall polarity is skewed towards negativity, which represents the distribution of the whole corpus. The entire list of tweets with manual sentiment tagging can be found in Appendix D.

Marking the polarity of the tweets is difficult. The subtlety and ambiguity of tweets as well as the use of slangs and/or abbreviations makes it difficult to flag a tweet as positive, neutral, or negative. Some tweets have such vague sentiments that different participants could have contrary opinions on their polarity. These opinions were gathered and considered when determining the final polarity of each tweet.

Moreover, the sentiment polarity of the tweet can be different depending on how people view the project. For example, the statement “I don’t like the fact that the government stops funding CAHSR” is a negative statement but positive towards the project. Some tweets speak positively about other projects to criticize CAHSR, while some blame the government for supporting the project. Within the baseline dataset, 47 (13.5%) out of 347 tweets have a sentiment towards the project be different from the sentiment of the tweet. Because of the relatively low percentage, the sentiment of the tweet rather than the sentiment towards the project is the main focus of the baselining exercise. Differentiating both types of sentiment is a valuable research topic for future research.

3.5.2 Aylien Text Analysis API

This research tests three sentiment analysis methods to compare their performance. These methods include two third-party toolkits, the Aylien text analysis API and the SentiStrength text analysis application, and a self-developed algorithm using the lexicon based approach. For each method, the principle of the algorithm is introduced, the application on the baseline dataset is conducted, and the performance is evaluated and discussed.

Aylien is a company based in Dublin, Ireland. It provides a comprehensive set of text analysis APIs featuring sentiment analysis, classification, extraction, summarization, etc. (Aylien Ltd, n.d.). The text APIs are integrated with platforms such as RapidMiner and Google Sheets, providing a neat way to access from GUI.

The sentiment analysis model used by Aylien is shown in Figure 3-5:

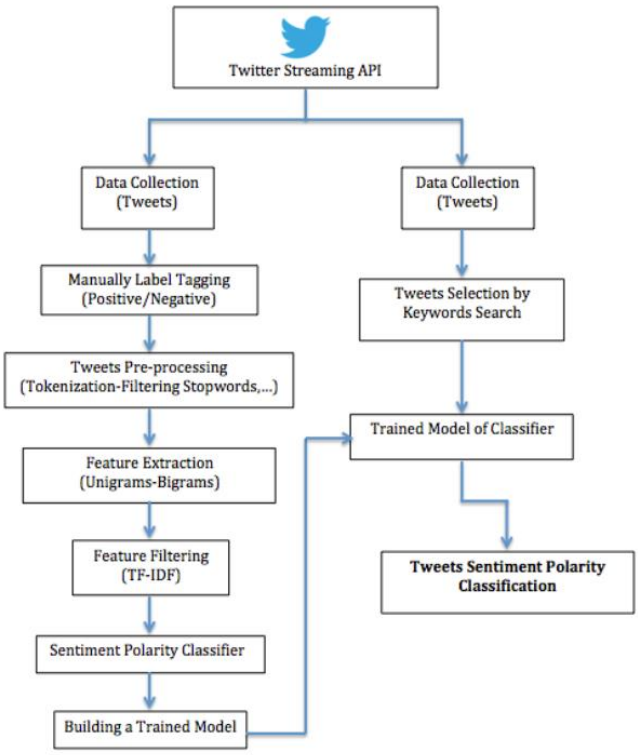

Figure 3-5 Aylien Sentiment Analysis Model (Barnaghi et al., 2016)

With the sentiment operators dedicated to tweet sentiment analysis, it is straightforward to build a RapidMiner process for sentiment analysis, which is shown in Figure 3-6.

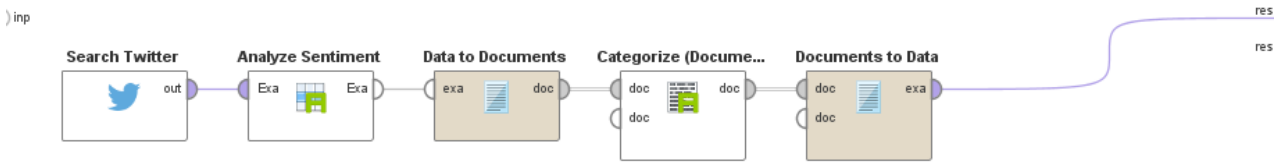

Figure 3-6 Sample RapidMiner Process for Tweet Analysis (Waldron, 2015)

Applying the Aylien text API on the sentiment baseline, the sentiment predictions and the F scores are shown in Table 3-5 and

Table 3-6.

Table 3-5 Sentiment Analysis Result Using Aylien Text API

	Predicted		
Actual	Positive	Neutral	Negative
Positive	20	56	5
Neutral	1	37	2
Negative	11	161	54

Table 3-6 F Score Analysis of Aylien Text API Result

Measures	**Accuracy**	**Precision**	**Recall**	**F1**
Positive	19.6%	62.5%	24.7%	35.4%
Neutral		14.6%	92.5%	25.2%
Negative		88.5%	23.9%	37.6%

The Aylien text API results in very low accuracy, low precision for neutral classes, and low recall for positive and negative classes. Therefore, the F1 scores for all polarities are lower than 40%. The recall of the neutral class is very high, indicating that the API tends to classify most tweets as neutral. Overall, this machine learning based attempt is not satisfactory in the context of infrastructure projects.

3.5.3 SentiStrength Text Analysis Application

SentiStrength is a lexicon-based application designed to detect sentiment polarity and strength in short informal social web text (Thelwall et al., 2012). It uses a lexicon which codes sentiment words on a scale of -5 to +5 for their prior polarity. Besides the lexicon basis, SentiStrength also uses non-lexical features such as spelling correction, idiom list, and emotion list. For each text (tweet), SentiStrength returns two values with range of 1 to 5 for both positive and negative sentiments. We choose SentiStrength as a lexicon based sentiment analysis tool. The difference of the positive value and negative value is used to determine the polarity of the text. When it is 0, the polarity is neutral, otherwise the polarity is the same as the sign of the difference.

Applying the SentiStrength text analysis application on the sentiment baseline, the sentiment predictions and the F scores are in Table 3-7 and

Table 3-8.

Table 3-7 Sentiment Analysis Result Using SentiStrengh Text Analysis Application

	Predicted		
Actual	Positive	Neutral	Negative
Positive	40	31	10
Neutral	11	19	10
Negative	34	76	116

Table 3-8 F Score Analysis of SentiStrengh Text Analysis Application Result

Measures	**Accuracy**	**Precision**	**Recall**	**F1**
Positive	50.4%	47.1%	49.4%	48.2%
Neutral		15.1%	47.5%	22.9%
Negative		85.3%	51.3%	64.1%

The lexicon-based approach improves the result compared with the Aylien text analysis API. The F1 scores of positive and negative sentiments have significantly increased to 48.2% and 64.1%, and the recall of the positive and negative classes are both increased. The accuracy reaches 50.4%, which is 31% more than the Aylien API.

3.5.4 Customized Lexicon Based Approach

Besides using third-party tools and packages, we have also developed our own sentiment analysis application using the lexicon based approach. The workflow of the algorithm is as follows. For each tweet, the word list is extracted by splitting the sentence by spaces. This is followed by a pruning process which handles the case of the letters, numbers, carriage returns, and special characters. Each tweet then contains a list of standardized words, which will be used to match the lexicon to calculate the sentiment score. The sentiment score of a tweet is defined as the difference between the number of positive words and the number of negative words. This workflow is depicted in Figure 3-7.

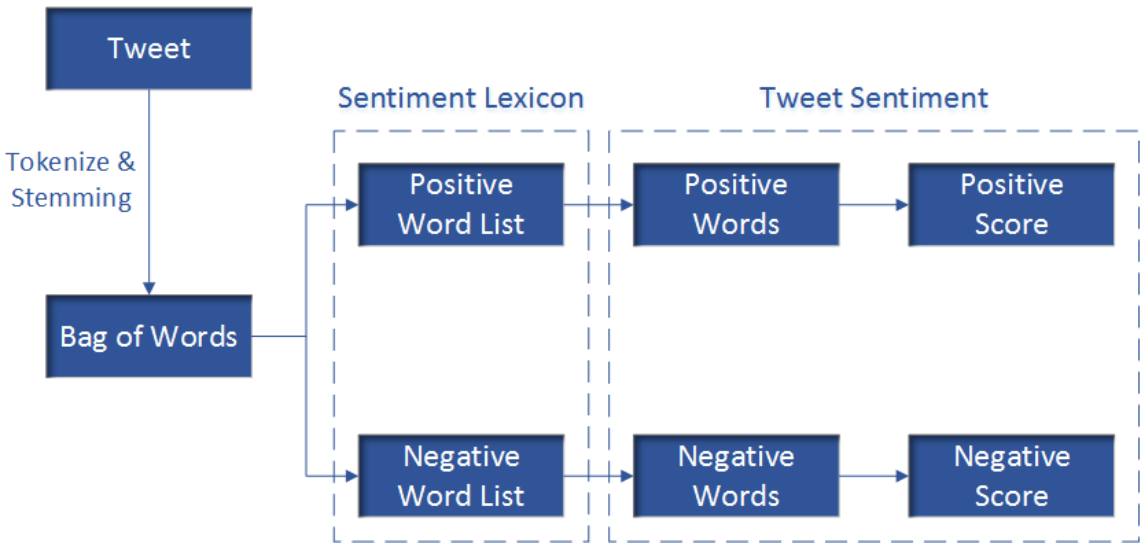

Figure 3-7 Lexicon Based Sentiment Analysis Workflow

The sentiment word list compiled by (Hu & Liu, 2004) is used, which contains about 6800 sentiment words, and has accounted for past tense verbs and common misspellings of social media. There are many other lexicons, such as SentiWordNet or Harvard General Inquirer, all of which should serve the need of this algorithm.

Even though lexicon based approaches are generally not domain specific, a tailored lexicon will function better when focusing specifically on infrastructure projects. We have examined the lexicon and made the following changes to improve the performance of the algorithm.

- Removed *like, trump, work* from the positive word list

In discussions about infrastructure projects, *like* is more often used to compare with something instead of expressing fondness. *Trump* is removed because it coincides with the name of the current president. *Work* is mostly used to refer to some real work, rather than if something "works". In the context of infrastructure projects, these common words tend to have neutral meanings and are therefore removed from the positive list.

- Removed *critical* from the negative word list:

Critical often refers to something important rather than criticizing in infrastructure projects. This word should have neutral sentiment instead of negative.

- Added *derail* to the negative word list

Derail is a commonly used pun when highway/railroad projects are adversely impacted by certain events. It is negative by itself but it is not a commonly used word, and therefore is not originally included in the list. Including this word is meaningful and can improve the performance of the algorithm.

Working with the updated word list with the changes in the dictionary, the sentiment tagging results and the F scores are shown in Table 3-9 and Table 3-10.

Table 3-9 Sentiment Analysis Result Using Customized Lexicon Based Algorithm

	Predicted		
Actual	Positive	Actual	Positive
Positive	72	Positive	72
Neutral	13	Neutral	13
Negative	50	Negative	50

Table 3-10 F Score Analysis of the Customized Lexicon Based Algorithm

Measures	**Accuracy**	**Precision**	**Recall**	**F1**
Positive	68.3%	53.3%	88.9%	66.7%
Neutral		34.6%	45.0%	39.1%
Negative		91.9%	65.0%	76.2%

This customized algorithm shows a significant improvement compared with the SentiStrength text analysis application. The overall accuracy was increased by 17.9% and the F1 score has reached around 70% for positive and negative sentiments and almost 40% for neutral. This method is therefore the most favorable approach among the three and is used to conduct sentiment analysis on the entire dataset.

3.5.5 Sentiment Analysis Discussion

The key metrics of different sentiment analysis methods are listed in Table 3-11.

Table 3-11 F Score Comparison of Sentiment Analysis Methods

Method	**Accuracy**	**F1 Score**		
		Positive	Neutral	Negative
Aylien Text API	19.6%	35.4%	25.2%	37.6%
SentiStrength	50.4%	48.2%	22.9%	64.1%
Customized Dictionary	68.3%	66.7%	39.1%	76.2%

The results of third-party tools, especially the Aylien Text API, do not meet their expected performances. One of the major contributors to this is the introduction of the neutral sentimental class. Neutral class is less discussed in previous research than the positive class and the negative class. Some of the lexicon based approaches take neutrality into consideration (Ding, Liu, & Yu,

2008), and some conduct sentiment analysis after neutral class was determined (Wilson, Wiebe, & Hoffmann, 2005). A lot of research, however, tends to filter them out to focus only on positive and negative sentiments and get better performance (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011) and (Pang et al., 2002).

However, (Koppel & Schler, 2006) suggested that the neutral class cannot be ignored and all three sentiments need to be identified when performing sentiment analysis. Neutral class plays a critical role in our infrastructure dataset. Unlike movie review and product review datasets, we do not have other rating indicators such as number of stars to help determine the overall polarity of the review text. Forcing neutral tweets to be labeled as either of the poles will introduce skewing into the result.

With the inclusion of the neutral class, (Vryniotis, 2013) showed that the majority of the classifiers have performance degradation with the 3-class classification compared with binary classification. This is one of the reasons why both third party applications underperform in the case study.

There are two reasons why our text analysis algorithm outperformed the others. Firstly, we observed that most of the tweets regarding infrastructure projects are simple and straightforward. Sarcasm and puns are not very common in this corpus compared with other review datasets, thereby reducing the amount of negation and cancellation of sentiments. In other words, the sentiment of the tweet heavily depends on the sentiment words rather than the structure of the sentence and the parts of speech of the words. The machine learning algorithms trying to understand these tweets sometimes misunderstand it, while the straightforward bag-of-words approach performs better. Using the sentiment word list by (Hu & Liu, 2004) directly, the accuracy slightly drops to 64.6%, which is still an acceptable ratio.

Secondly, a customized sentiment dictionary is able to boost the performance of the algorithm. Adding and removing 5 words from the dictionary contributed to a 3.7% increase in accuracy. Once the user habits in tweeting about infrastructure projects are studied in-depth, a more comprehensive word list could be built, which is expected to further improve the performance. Similar

methodology applies to the machine learning based approaches, which are known to be domain specific. In future research, after collecting enough tweets related to infrastructure projects, the effort of building and training a domain specific classifier can be started as an alternative to the lexicon based approach.

3.6 Tweet Sentiment Analysis

The customized sentiment analysis algorithm outperforms the third party tools, and is therefore selected to conduct a full spectrum sentiment analysis on the entire dataset. By applying this algorithm upon the 24,855 tweets collected between 2016-06-10 and 2017-10-22, the sentiment of each tweet is labeled as positive (labeled 1), neutral (labeled 0), or negative opinions (labeled -1). The sentiment value for any given tweet is determined by positive words PW and negative words NW in the tweet.

$$V_t^{i,j} = sgn(PW - NW) \tag{3.5}$$

where $V_t^{i,j}$ is the sentiment value of tweet *i* of user *j* on time *t*. If PW is larger than NW or the tweet shows a positive attitude toward the project, $V_t^{i,j}$ is assigned 1. If PW is smaller than NW or the tweet shows a negative attitude toward the project, $V_t^{i,j}$ is assigned -1. $V_t^{i,j}$ equals to zero when there is no difference between PW and NW.

Figure 3-8 plots the trending of the sentiment, aggregated by day, for the entire time range of the case study.

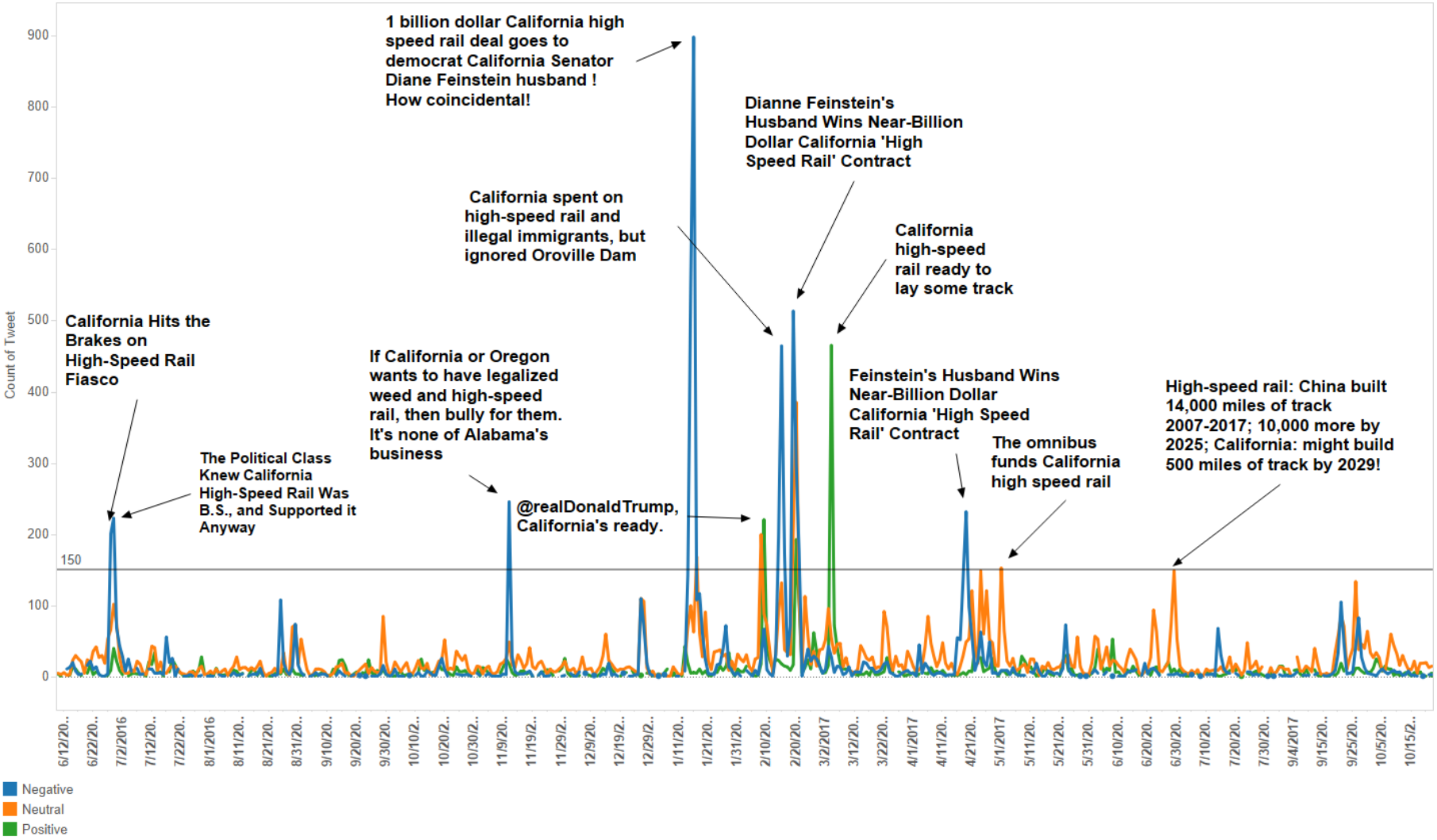

Figure 3-8 Tweet Sentiment Analysis Result

3.6.1 Tweet Sentiment Trending

In Figure 3-8 the tweet sentiment was aggregated by polarity and date. Polarities are color coded such that negative is blue, neutral is orange, and positive is green. As can be observed from the figure, tweet sentiment trending is not smooth or steady over time. They go through shocks which are mostly associated with one or more major events. Events or spikes with more than 150 tweets are annotated in the chart. Some key events are discussed below.

The first negative spike with over 150 tweets is a combination of two events which both occured on or around Jun 28, 2016. The first event is a report from Bloomberg by Virginia Postrel (Postrel, 2016) titled "California Hits the Brakes on High-Speed Rail Fiasco". The second event is a follow up article from reason.com, "The Political Class Knew California High-Speed Rail Was B.S., and Supported it Anyway" (Welch, 2016). Both articles were immediately retweeted in massive numbers, generating 201 and 223 retweets, respectively.

The most impactful event happened on Jan 16, 2017. The original tweet was posted by account "Whitehouse Plumber (@rharrisonfries) reading "1 billion dollar California high speed rail deal goes to democrat California Senator Diane Feinstein husband! How coincidental! #MAGA". A similar event happened again on Feb 19, 2017 by Thomas Lifson on conservative50.com (Lifson, 2016). This news article started trending right away and many retweeted "*Dianne Feinstein's Husband Wins Near-Billion Dollar California High Speed Rail Contract*". A third spike happened once again on Apr 19, 2017. These events attracted 899, 514, and 232 retweets, respectively.

The biggest positive event took place on Mar 04, 2017. The original news is from Associated Press, "California high-speed rail ready to lay some track" (Thompson, 2017). It was reprinted by other media and was retweeted by at least 466 people, marking a critical milestone of the project.

People use Twitter to voice their opinions, and one of the easiest ways is retweet an existing tweet (or like, as on Facebook). By supporting someone else's statement, their preferences are also

expressed. Observed from the major events, media (news agencies and news websites) plays a critical role in covering trending topics and swinging public opinion. This shows that the public still relies on media to collect information and form their opinions.

3.6.2 Tweet Sentiment Polarity Distribution

The neutral class dominates the corpus. There are 13,026 neutral tweets out of a 24,855 total tweets, which is 52.4% of the total. The neutral class is made up of tweets with no sentiment at all, as well as tweets with both positive and negative sentiments that canceled each other out. Among the 13,026 neutral tweets, 11,777 (90.4%) of them do not have any sentiment words. However, a lack of sentiment words does not necessarily mean the tweet does not contain a sentiment. Using the last event as an example, the tweet reads "High-speed rail: China built 14,000 miles of track 2007-2017; 10,000 more by 2025; California: might build 500 miles of track by 2029!" Even with no sentiment word, this tweet is still slightly negative. In future research, we will aim to understand these tweets better to map their sentiment more accurately.

Making up the second largest class in the corpus is the negative class. It contains 7,139 (28.7%) tweets. In the case of events, however, negative events generate much more traffic than the other two. The biggest event during the case study was a negative one, with 899 negative tweets on Jan 16, 2017. Out of 11 major events marked in Figure 3-8, 7 of them are negative. Therefore, negative events are the major event type and are the most impactful. Moreover, negative tweets in our setting bring more information than neutral ones, such as questions, problems, and complaints. It is critical to understand negative events in order to understand public opinion regarding the project.

Last but not least, the positive class is the smallest class of all three. It contains 4,690 (18.9%) tweets. Positive tweets show people's support, praise, and affirmation towards the project. The biggest positive event is the exciting milestone "California high-speed rail ready to lay some track". The lack of positive events means that the overall opinion towards the project is negative.

3.7 Public Acceptance Analysis

The work of (Calais Guerra, Veloso, Meira Jr, & Almeida, 2011) on the opinion holder bias prediction is based on the assumption that users express their opinions through endorsements. While their assumption is specific to one user agreeing with the other when retweeting, we would like to extend the assumption and assume that user endorses a certain message at a given time. This assumption, along with the previous observation that the majority of tweets are consistent between the sentiment of the tweet and the sentiment towards the project, enables the aggregation of all tweet sentiments to derive the public acceptance.

The public acceptance measures whether the general public supports or opposes a certain infrastructure project. Provided that social media is able to feed real time data flow, the public acceptance in this framework is designed to be a time series metric which depicts the level of support and opposition over time.

3.7.1 Public Acceptance Definition

The public acceptance is defined as a ratio of positive counts over the summation of both positive and negative counts. The breakeven point for this formula is 50%, where the number of positive votes is the same as negative ones. Neutral class is not included in the formula for simplicity, hence the public acceptance can be interpreted as the supporting ratio, and the difference between 1 and the public acceptance is the opposing ratio. It is still important for the algorithm to be able to identify and exclude neutral tweets.

Intuitively, daily sentiment should be used to calculate daily public acceptance. However, counting tweets on a specific day returns volatile public acceptance. The data volume, number of tweeters, and sentiment could all change dramatically, resulting in drastic fluctuations. By following the same methodology used in public opinion polling as (O'Connor et al., 2010), daily sentiment is

replaced by the weekly moving average to smooth the Public Acceptance ratio (PA_t), as shown in formula 3.5.

$$PA_t = \frac{\sum_{i=0}^{6} PS_{t-i}}{\sum_{i=0}^{6} PS_{t-i} + \sum_{i=0}^{6} NS_{t-i}} \tag{3.6}$$

where PA_t is the public acceptance ratio on time point t. PS_t is the positive score and NS_t is the negative score on time *t*.

Different definitions of PS_t and NS_t can be derived from different perspectives. It is common to use tweet volume i.e. the number of messages as an indicator of public acceptance, for example, (Jiang et al., 2015). This is different from the electoral equality principle of "one person one vote" used in public polls. Although there is a clear difference in the determination of public acceptance, both methods can be valid and reflect two classic perspectives on the role of media and polls on public policy, i.e. elite model and pluralist model. The elite model assumes that elite groups dominate politics and society and therefore, public opinion is subservient to political elites. Whereas, the pluralist model assumes that power is dispersed throughout society so that no one group dominates. As such, public opinion should be independent from political influence (Robinson, 2008). It remains unknown how variant public opinion can be viewed through these methods. This research considers both methods and defines PS_t and NS_t by tweet, user, and user influence.

3.7.2 Project Acceptance by Tweet

Public opinion can be evaluated through all relevant tweets regardless of the people who post them. To the extreme, there are cases where one user posts hundreds of tweets, and cases when many users post one tweet each. This method considers these tweets with equal weight. Each tweet has a sentiment value $V_t^{i,j}$ determined in the tweet sentiment analysis process. The daily positive count

value PS_t is then determined by counting the number of all positive tweets from every user of that day. Similarly, NS_t is determined by all negative tweets.

$$PS_t = \sum_j \sum_i V_t^{i,j} \ \ for\ all\ V_t^{i,j} = 1 \tag{3.7}$$

$$NS_t = \sum_j \sum_i \left| V_t^{i,j} \right| \ \ for\ all\ V_t^{i,j} = -1 \tag{3.8}$$

Although it is common and intuitive to use the number of tweets as the basis of calculating public acceptance, this approach could potentially be built on a biased sample. Firstly, this approach treats all tweets equally, leading to the result that people who post more have higher weight than others. It is expected that majority of people or accounts with high tweet volumes are interest groups advocating or opposing projects. Placing them in a more important bucket could potentially cause bias for "loud" voters and against the general public, who tend to be quiet most of the time. Secondly, this approach does not remember people's endorsement. A lot of users tweet only a few times during the case study time frame, and their inclination is only considered on the day their tweets are posted. As time goes by, their voice is diluted and their importance decreases. However, everyone should have an equal vote no matter how vocal this person is. Therefore, the main disadvantage of this model is the overlook of the human factor.

3.7.3 Project Acceptance by User

An alternative model is proposed to address the potential issues in the by tweet approach. Instead of using tweets as the basis, users are used following the electoral equality principle. In this method, one person can only vote once per day, no matter how many tweets are posted. Frequent tweeters are treated equally as common people. Taking advantage of big data technologies and avoiding more sampling bias, all users collected should be factored into the calculation.

Moreover, a user's stance remains if no changes are made later on. By default, all users start with neutral position. Once a sentimental tweet is posted on a certain day, the user is treated as positive or negative accordingly for that day, and for all the future days. This position holds until a new tweet is posted by the same user with different sentiment polarities, and the user's position changes and stays going forward.

The calculation of the public acceptance consists of two steps. Daily user sentiment V_t^j of person *j* is calculated first, by summing up sentiment values of all tweets $V_t^{i,j}$ on day *t*. If person *j* posts no tweets on day *t*, then $V_t^j = V_{t-1}^j$, as shown in formula 3.9.

$$V_t^j = \begin{cases} sgn\left(\sum_i V_t^{i,j}\right) \; if \; \exists V_t^{i,j} \\ V_{t-1}^j \; if \; \nexists V_t^{i,j} \end{cases} \tag{3.9}$$

After V_t^j is determined, PS_t and NS_t are calculated as the number of positive and negative people on day *t*. People with neutral sentiment are treated as abstain from voting.

$$PS_t = \sum_j V_t^j \;\; for \; all \; V_t^j = 1 \tag{3.10}$$

$$NS_t = \sum_j |V_t^j| \;\; for \; all \; V_t^j = -1 \tag{3.11}$$

Following the electoral equality principle, the by user approach addresses the major issues of the by tweet approach. However, the electoral equality principle might not fit perfectly in the cyber space. In terms of the size of the broadcast audience and the ability to influence people, some people such as celebrities and public figures have a stronger influence than normal people. By assigning everyone equal weight on their votes, this approach does not consider the difference in influence among people.

3.7.4 Project Acceptance by Influence

To take into consideration the level of influence of Twitter users, the third approach assigns different coefficients according to a user's followers count, compared with equal distribution in the last approach. (Cha, Haddadi, Benevenuto, & Gummadi, 2010) found that a user's degree of influence in social media follows a power-law scale, hence a logarithm scale is used to measure a user's influence based on the number of followers. The user influence is defined as:

$$I^j = 1 + \log(1 + F^j) \tag{3.12}$$

where F^j is the number of followers for user *j*. Some special handlings are made for people with 0 followers. From the modeling perspective, the number of followers at time *t* (F_t^j) should be used to describe the expansion of a user's influence radius. However, due to the throttling rate of Twitter API it is very difficult to track user's followers' count over time. For this case study, it is assumed that user follower is stable enough that a snapshot in time could be used to represent user influence during the case study time frame.

Similar to the by user approach, the user's vote is either the aggregated sentiment of the day or previous day's sentiment in the case of no tweets. The public acceptance by influence approach then normalizes the PS_t and NS_t by user influence, as shown in formula 3.13 and 3.14.

$$PS_t = \frac{\sum_j (I^j * V_t^j)}{\sum_j I^j} \; for\ all\ V_t^j = 1 \tag{3.13}$$

$$NS_t = \frac{\sum_j (I^j * |V_t^j|)}{\sum_j I^j} \; for\ all\ V_t^j = -1 \tag{3.14}$$

All three approaches represent different voting mechanisms under different assumptions. There is not one approach which is superior to the others. We apply all these models against the case study dataset to investigate the validity and difference of these models in the following section.

3.7.5 *Project Acceptance Analysis Result*

The result is generated by applying all three models on the whole case study dataset. There is no random sampling process involved since the tweeters are already a sample of the whole population, and the big data technology allows us to quickly process the data of such volume. This analysis results in three sets of daily metrics PS_t and NS_t, which are used to calculate 7-day moving total respectively, which is then used to calculate the daily public acceptance using equation 3.6. The fluctuation of public acceptance under all three methods is plotted in Figure 3-9.

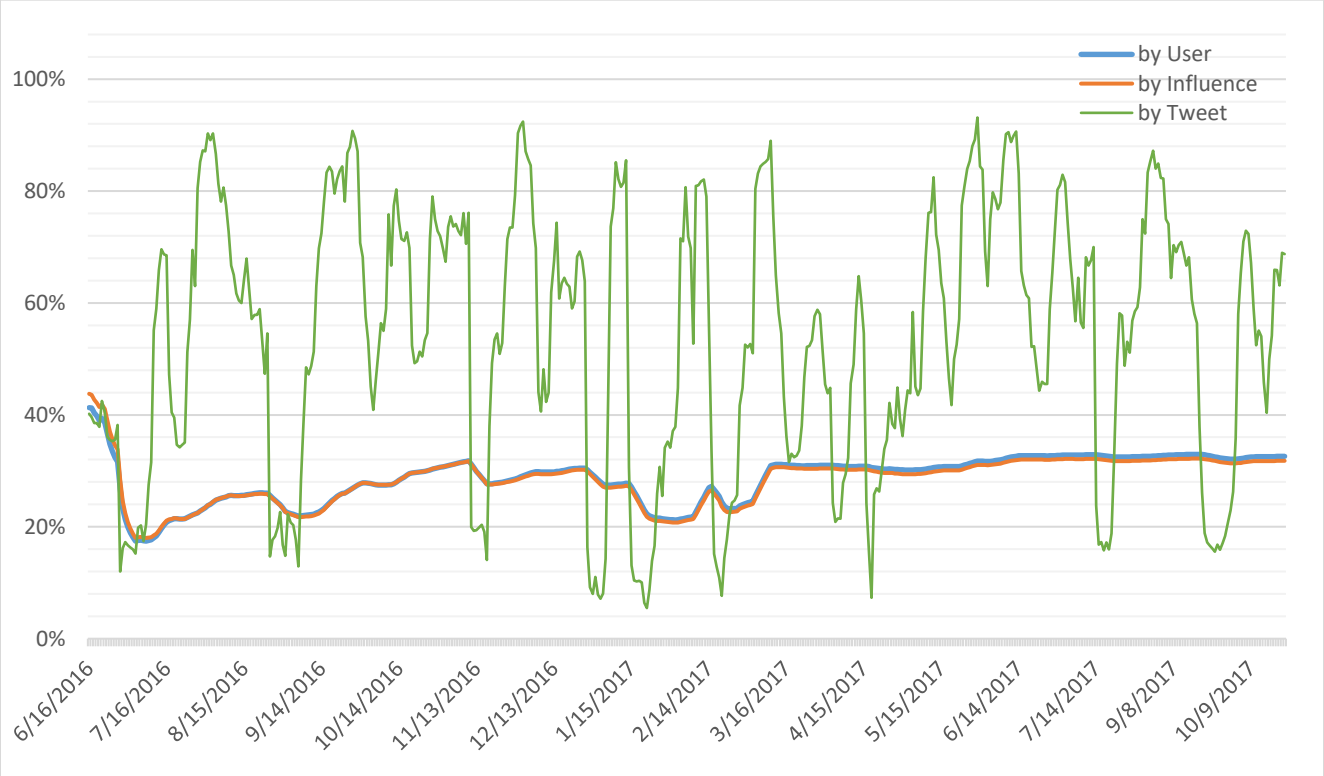

Figure 3-9 Public Acceptance Analysis by Tweet, User and Influence

This daily tracking of public acceptance provides a real-time impact monitor of relevant events. For example, there are several peaks and valleys where public opinion moves toward an opposite direction. Relevant tweets and events may explain why public opinion has shifted on a specific time point.

- 2016-07-12, Diana Gomez from @CaHSRA explains the progress and challenges of building high-speed rail in California! #IWillRide
- 2016-07-19, @CaHSRA From construction to outreach, take a look at everything #CAHSRA has accomplished in the last 6 months. #Iwillride
- 2016-08-30, High-speed rail critics question the first route segment, which will end in an almond orchard.
- 2016-08-31, #Californias #CapandTrade Program is sick and will take #HighSpeedRail down with it via @Forbes #Env #Transit #OpEd
- 2017-06-22, @CaHSRA Congratulations on writing a great California Government Tweet: (Ranked 43rd for Jun 20.)
- 2017-07-10, @CaHSRA: Another reason connecting the Silicon Valley to the Central Valley is so important... better access to more affordable housing.

Several interesting observations can be made according to Figure 3-9. As clearly shown, public acceptance by tweet is more volatile than the other two. The measurement can jump from 35% to 85% in 7 days, or from 85% to 13% in 3 days. Even though a moving total is used to smooth the curve, the flip of the acceptance polarity is still very frequent. This is attributed to the lack of memory of this approach, i.e. the public acceptance only considers the tweets of a given day, when the number of tweeters and tweets are mostly random. On the contrary, the other two approaches keep a user's vote until it changes, therefore old tweets could still have impact on future days, resulting in much lower volatility and steadier curve.

There is still some consistency among all these methods. Some choppy uptrend of the by tweet approach is represented by a stable increase in the other two. They rise and dip due to the same set of events, but the magnitude of the by user and by influence approaches are significantly smaller than the by tweet approach.

ANOVA analysis is conducted to examine the differences among these three measures. Table 3-12 summarizes the ANOVA analysis results, which shows a significant difference among acceptance by tweet, user, and influence.

Table 3-12 ANOVA Analysis of Public Acceptance Measurements

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
by User	464	133.1443	0.286949	0.001838
by Influence	464	131.6245	0.283674	0.001809
by Tweet	464	247.8665	0.534195	0.055117

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	19.16357	2	9.581787	489.1619	1.5E-161	3.002203
Within Groups	27.20797	1389	0.019588			
Total	46.37155	1391				

On average, public acceptance towards CAHSR is around 28% using the by user and by influence approach, but the by tweet approach disagrees and is reporting a 53% acceptance. Again the memoryless feature of this approach is driving the difference. Under this approach, a major event generating hundreds of retweets has only a few days of impact radius, and will be overridden by lesser events in the future.

As mentioned before, 50% is the breakeven point between positive and negative acceptance. In this case study these methods report two different results, positive from the by tweet approach and negative for the by user and by influence approaches. Judging by the amount of negative events and negative users, which will be further discussed in next chapter, the overall public acceptance is expected to be negative. The by user and by influence approaches are hence believed to be more accurate in this case.

A separate ANOVA test regarding the difference between the by user and by influence model is also conducted, yielding statistically insignificant results, as shown in Table 3-13. The public acceptance by user is slightly higher than the by influence measurement, indicating that higher influential people tend to be more pessimistic than the average.

Table 3-13 ANOVA Analysis of Public Acceptance by User and by Influence

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
by User	464	133.1443	0.286949	0.001838
by Influence	464	131.6245	0.283674	0.001809

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.002489	1	0.002489	1.364813	0.243006	3.851521
Within Groups	1.688722	926	0.001824			
Total	1.691211	927				

Three models discussed above provide different flavors on public acceptance measurement. Acceptance by tweet tracks sentiment of any given day, acceptance by user remembers any users' last vote, and acceptance by influence takes user's influence (followers) into consideration. Considering both accuracy and simplicity, acceptance by user is the recommended model for public acceptance calculation.

3.8 Conclusion

In this chapter, we have conducted the first phase of the case study using the project evaluation framework. California High-Speed Rail project is selected due to its scale, controversy, and coverage in social media compared with other infrastructure projects. The data characteristics of infrastructure projects are compared with other common topics in social media analysis, which we found to have much lower volume in tweets. The performance of different search terms used to retrieving tweets is discussed. Keyword searching is the most effective search term, taking as much

as 80% of the data retrieval volume. Therefore, carefully selected keywords are critical to the project evaluation task.

Three sentiment analysis techniques, the Aylien text analysis API, the SentiStrength text analysis applications and the customized lexicon based algorithm are tested using the sentiment baseline where each individual tweet is manually tagged of its sentiment polarity. The accuracy and F1 scores of all the techniques are compared, and the customized lexicon based sentiment analysis algorithm yields the most satisfactory results with 68% accuracy and around 70% F1 score. We also initialize the contribution to customize a domain specific sentiment dictionary for infrastructure projects. The customized algorithm is applied to the whole corpus to obtain the sentiment over 16 months. Observations are made regarding the event-based nature of public sentiment fluctuation and the distribution among positive, neutral and negative polarities.

Based on the tweet sentiment analysis, the public acceptance model is developed by defining the measurement using the number of positive and negative tweets within a moving window. Three public acceptance models, by tweet, by user, and by influence, were proposed, applied, and examined using the case study. The by user model and the by influence model generate more smooth curves than the by tweet model. They also result in statistically significant public acceptance readings than the by tweet model, and the former measurement is closer to reality. The by user model is the most favorable model of public acceptance in consideration of accuracy, curve smoothness and simplicity.

Chapter 4. Evaluation of Public Acceptance Using Big Data – A Case Study on Social Media Events

4.1 Introduction

In chapter 3, a public acceptance model is defined in the context of social media and the by user model is the most favorable one to measure public acceptance. While the knowledge about the time series public acceptance is important to project managers, it is also valuable to reveal the cause of its change so that actionable items can be taken to improve public acceptance. To answer the question of WHAT drives the change of public acceptance, the scope of the analysis is extended beyond text analysis to reveal the driving factors of public acceptance fluctuation.

In section 4.3, we start social media event analysis by defining event itself and its influence. A two dimensional event influence measurement is proposed, leading to the development of event influence quadrant to be introduced in section 4.4. Section 4.5 discusses the overall sentiment of individual events, and section 4.6 discusses the strategies to use events to improve public acceptance.

4.1.1 Social Media Events

As shown in the public acceptance analysis, tweet sentiment peaks and valleys as a result of breaking news, articles and announcements. These incidents are shared, referenced and spread in the social media world and contribute to the fluctuation of public acceptance. Although they are not as wide spread as events such as presidential elections, Olympic Games or natural disasters, they form small scale social media events which diffuse through the same channel. Due to the relatively low volume in tweets, these events play a critical role in public acceptance by generating massive number of retweets of the viral articles and spreading the information to a broader audience.

In order to describe these social media events, the tweet model needs to be modified and new dimensions need to be added.

4.1.2 Extending the Project Evaluation Framework

Built around the object tweet, the project evaluation framework needs to be extended to accommodate social media event analysis. The web page is a critical component in event analysis, however, they are masked in tiny URL and cannot be grouped together. A separate crawler is developed to restore the original web page URL from the tiny URL used in tweets, and the full URL is included in the data model as another dimension of tweet. The social media event model is established on top of these web pages. The data structure of the framework is thereby expanded to support the multi-dimensional data structure from Twitter and its periphery.

4.1.3 Event Influence Analysis

Event analysis provides project managers with a target to act on to improve public acceptance. The impact of negative events need to be mitigated whereas the impact of positive events should be amplified. The project evaluation framework equips managers with a tool to monitor social media events as well as the effect of any policy changes. In this chapter, we focus on the evaluation of event influence so that events can be categorized and processed in a prioritized order. It is shown that event influence cannot be determined solely by the number of tweet it generates, therefore an event influence quadrant is proposed to cluster events together in a two dimensional model. We also discuss strategies to promote and demote events across different quadrants.

4.2 Literature Review

An event is an occurrence of something noteworthy. Event has its social attributes as they involve participation of people. Traditional media was viewed as the original distributor of major events. With the emergence of social media, however, new technologies and platforms are becoming more and more influential in spreading event information. As social media plays an increasingly

important role in events, research studies started to pay attention to the behavior of the online community at the occasion of events.

On the subject of emergency events, (Palen, Hiltz, & Liu, 2007) found that local citizens are not only the first responders, but also show continuous support during such events. With the help of information sharing from social media, online groups and forums are able to provide stronger support in disaster recovery. (Palen, 2008) studied two disastrous events, the mass shooting at Virginia Tech and the 2007 southern California wildfires, and found that social media is able to support the distribution of critical event information and identify victims quickly and efficiently. (Yates & Paquette, 2011) concluded that social media is able to support collaborative knowledge sharing and reuse and facilitate decision making.

In terms of political events, (Vaccari, Chadwick, & O'Loughlin, 2015) found that social media creates more exposure of debates to respondents, and that commenting and engaging with Twitter hashtags is correlated with political engagement. (Chadwick, 2011) showed that informal social network activists are part of the growing force of Britain's politics. (Larsson & Moe, 2012) suggest that Twitter provides an outlet for minorities and the general public to express their political opinions, a privilege which used to belong to only elites and politicians. There are studies showing limited impact of Twitter, for example, (Larsson, 2013) found Twitter to have limited impact on changing journalistic norms or practices, and (Larsson & Moe, 2012) pointed out that the Twitter population is much lower in foreign countries.

Despite the concerns about the unregulated rules and the novelty of the technology and community, the results of previous research studies are overall positive that Twitter, as well as other "new media", is able to impact and reform the traditional ways events are diffused and conceived. It is thus necessary to include event in our research scope.

4.3 Event Analysis

As shown in the tweet sentiment analysis result in Figure 3-8, tweet sentiment fluctuates over time. Manual inspection of the tweets on the days of dramatic changes reveals that the majority of the driving tweets are similar retweets of certain web pages or other users' tweets. When catching articles and occasions take place, people tend to retweet the web pages or opinion leaders' tweets referring to the articles, creating a massive traffic of retweets and comments, which then affects the public acceptance. We define such clustered retweets as events. They are islands in the sea of social media which draw a lot of attention and generate a large amount of retweets. They also spread across a wide audience, thus having greater impacts on public acceptance than individual tweets. Some events might turn into public relation crises and eventually jeopardize the project if they are not closely monitored and properly handled, and the project evaluation framework is able to provide real time event monitoring and alarming.

4.3.1 Event Definition

Observations from Figure 3-8 triggers the investigation of the root cause of the fluctuation of public acceptance. Manual inspection of the raw data reveals that the changes can be mainly attributed to the retweet of one or several web pages. However, it is difficult to find the original person who shared the web page. According to the two-step model of information diffusion (Katz & Lazarsfeld, 1966), the web page is tweeted by a set of opinion leaders, whose tweets are further retweeted by their followers, and cascade through the hierarchy of followers. It is entirely possible that the retweets cross reference each other, constructing a graph of retweet network. Hence, it is both difficult to find the retweet of the original web page, and difficult to traverse the whole retweet network to collect all the tweets of an event.

In order to cluster all tweets of a certain event, instead of searching for tweets from specific users, this research focuses on the original article / web page referenced in the tweets. In other words, a Twitter event is defined as a set of tweets referencing the same web page or news article.

$$E_p = \{t : tweets\ referring\ to\ web\ page\ p\} \quad (4.1)$$

Due to the previous limitation on the number of characters allowed in tweets, all tweets use tiny URLs instead of full URLs. Tiny URL is a web service targeting shortening URLs so that social media such as Twitter can use the concise version of URL which still redirect to the original page (Galper, Goyal, & Gilbertson, 2013). Once a URL is included in a tweet, it is automatically converted into a 23-character tiny URL even if the original URL has less than 23 characters. Furthermore, all tiny URLs are different even though they are referring to the same web page. To be able to group tweets by web pages, extra information cleansing is necessary to translate tiny URLs back to their original form.

A web page crawler is developed to restore the full URL from tiny URL.

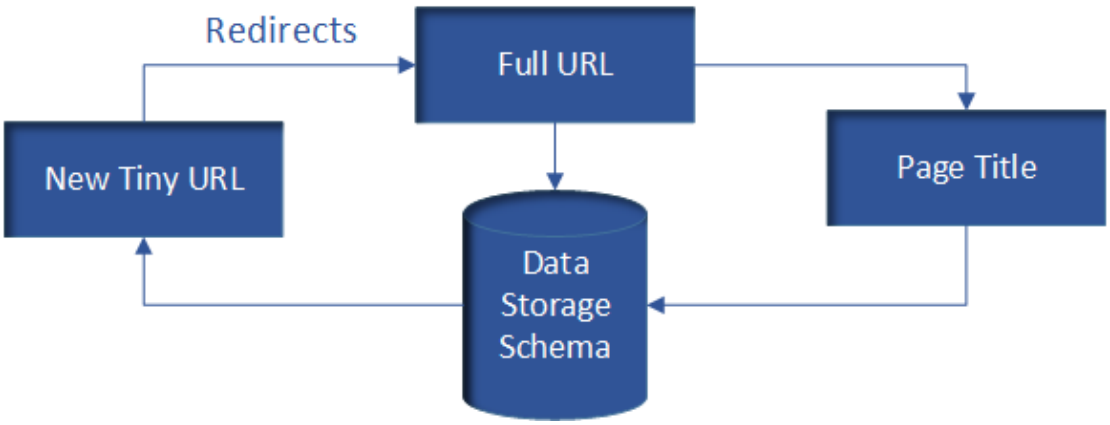

Figure 4-1 Web Page Crawler Workflow

The idea of the crawler is simple, for every new tiny URL included in the system, the crawler visits the tiny URL and goes through all the redirects until it reaches the final web page. It then fetches the full URL of that page and its title, and performs some text cleansing. It then sends the result

back to the data storage as an enrichment of tweets. Figure 4-1 demonstrates the workflow of the web page crawler.

By applying the crawler on the CAHSR case study, all events, web pages which are referred to at least once, are collected and clustered together. As a result, Figure 4-2 plots the histogram for different retweet counts. A total of 3,103 events are derived from the data set, with the total reference by tweet ranging from 1 to 1,279, and the total reference by users ranging from 1 to 1,227. Among all these events, 2840 of them have only 1 to 5 retweets, and are not included in the histogram due to the overwhelming volume.

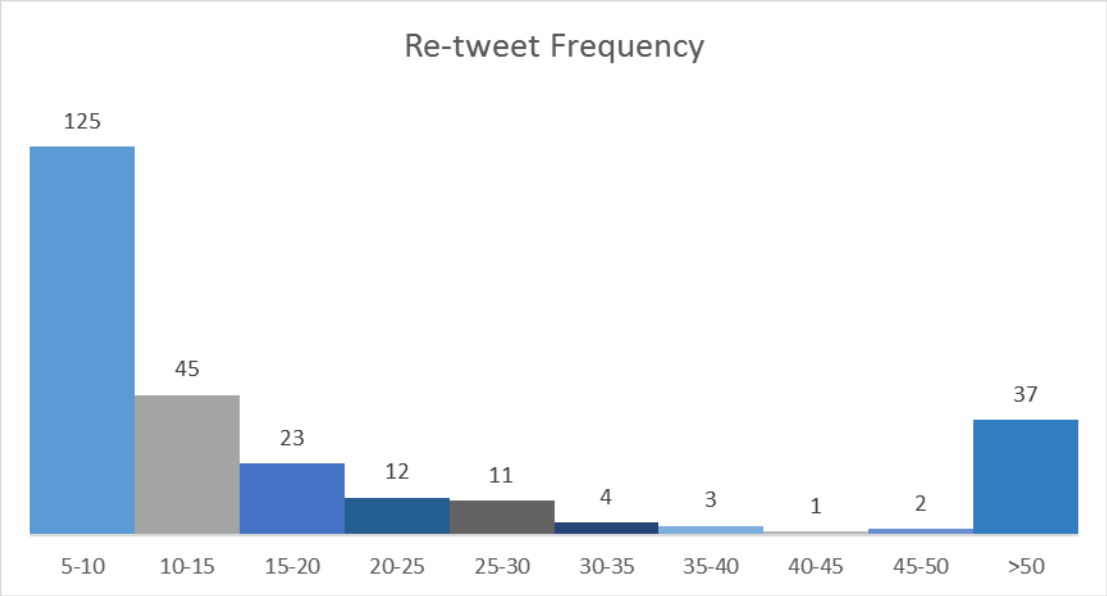

Figure 4-2 Histogram of Event Tweets

Considering all 3103 events, the mean number of tweets referring each event is 4.0, and the standard deviation is 28.6. In order to include more events for analytical purpose, we use the range of 1-sigma, which means events with 33 or more tweets. A total of 45 events are identified and the top 10 of them are shown in Table 4-1.

Table 4-1 Top 10 CAHSR Events by Number of Tweets

Event	Retweet
Dianne Feinsteins Husband Wins Near-Billion Dollar California High Speed Rail Contract	1279
California Hits the Brakes on High-Speed Rail Fiasco - Bloomberg	504
Lame-Duck Obama Admin Rejects CA High-Speed Rail $15 Billion Loan	364
Trump administration halts Californias plans for high-speed rail and infrastructure improvements	337
The Hill on Twitter: "Trump laments lack of high-speed rail in US during meeting with top airline execs [tinyurl]"	303
California High Speed Rail Authority - State of California	239
ABC News – Breaking News, Latest News, Headlines & Videos	148
Oroville Dam flood danger recedes; state criticized for spending on rail, illegals - Washington Times	147
CA High-Speed Rail Contractor Gets 18% Raise After Missing Completion Date - Breitbart	147
The Political Class Knew California High-Speed Rail Was B.S., and Supported it Anyway - Hit & Run : Reason.com	132

4.3.2 Event Influence

The definition of event enables the analysis on the joint impact of all the tweets within an event. Intuitively, the number of retweets or the number of retweeting users are good measurements of event influence. However, this measurement alone might not be sufficient enough to depict how influential an event is. Some events create viral distributions, generating a large number of retweets in a short period of time. Some events, on the other hand, might not have that many retweets, but instead enjoy a longer life span as people constantly refer to it. These events are also influential since they could accumulate enough pressure from the number of people involved and the continuous interest it incurs. In this chapter, we propose a two dimensional measurement of event influence: by magnitude and by duration.

- Event Magnitude

Event magnitude is the amount of attention associated with an event. A typical event has the highest magnitude in the first a few days, and diminishes over time. Similar to the public acceptance analysis, we propose and examine three measurements i.e. by tweet, by user, and by user influence. That is, counting the number of tweets generated by an event, the number of unique users tweeted on the event, and the number of users with weighted on the logarithm of their followers. The definition is formulated in formula 4.2, 4.3 and 4.4.

$$\mathrm{EMag(p)} = \sum_i T(p)_t^i \tag{4.2}$$

$$\mathrm{EMag(p)} = \sum_j U(p)^j \tag{4.3}$$

$$\mathrm{EMag(p)} = \sum_j 1 + \log\left(1 + F(p)^j\right) \tag{4.4}$$

where $T(p)_t^i$ is tweet i on time t referring event p, $U(p)^j$ is user j who has retweeted event p, and $F(p)^j$ is the followers count of user j who has retweeted event p.

- Event Duration

Event duration measures how long an event lasts, from the days of the first tweet after an event is started, and the time of the last tweet belonging to the same event.

$$\mathrm{EDur(p)} = Date(T(p)_n) - \mathrm{Date}(T(p)_1) \tag{4.5}$$

where $T(p)_n$ is the last tweet referring event p and $T(p)_1$ is the first.

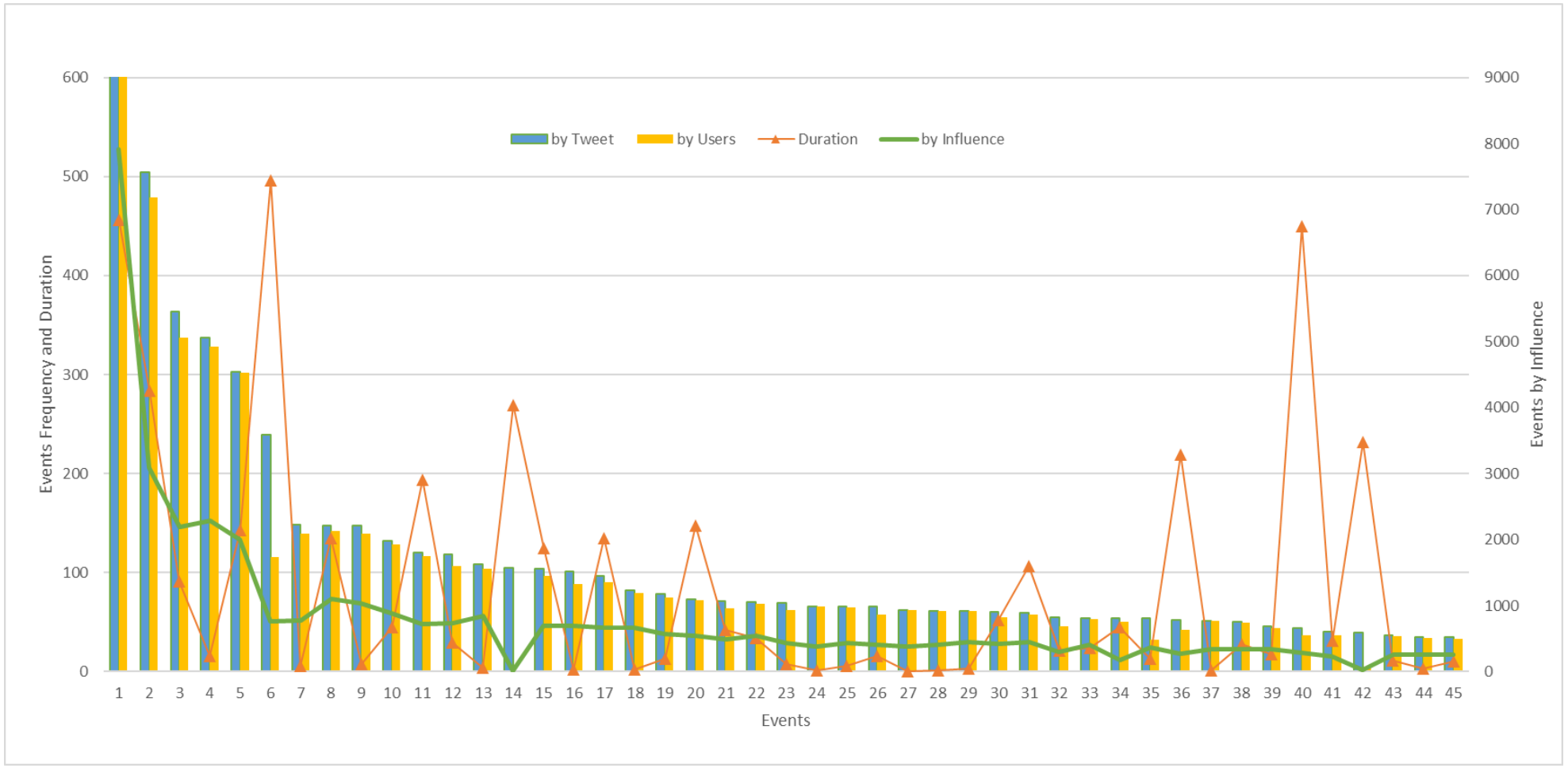

Figure 4-3 Event Influence Measurments

Since the last tweet changes over time, the measurement is a continuous metric and some events might have a sudden gain in duration after some silence.

Considering all 4 measurements (3 for event magnitude and 1 for event duration) of event influence, Figure 4-3 plots them regarding the 45 highly influential events identified in section 4.3.1. The x-axis is the sequence number of an event, and the y-axis is the corresponding measurements.

Three event magnitude measurements yield very similar results. The by user metric ranges from 1 to 1,227 with a mean of 3.7 and standard deviation of 27.1. The by influence metric ranges from 1 to 7,908.2 with a mean of 27.4 and standard deviation of 181.8. Outside of the 1-sigma range the by tweet approach returns 45 events, the by user approach returns 43, and the by influence approach returns 45. The majority of the events overlap with some slight differences, indicating that most people only retweet an event once.

A close look at these top events uncovers a problem with the by tweet approach. Event #14, "*Emerging Challenges and Opportunities of High Speed Rail Development on Business and Society (Advances in Civil and Industrial Engineering)*", scores 105 using the by tweet measure but 1 using the by user measure, and 5.7 using the by influence measure. Only 1 user is actively tweeting the message for as many as 105 times. Similarly, event #42, *"LA Times"*, has 39 retweets with only 4 users and 21.1 influence score. Hence the by user and by influence approaches are better than the by tweet approach in detecting and excluding fake events triggered by a small number of people.

Also it is necessary to use both magnitude and duration to capture and measure events. Using magnitude alone qualifies event #27, "*Caltrains FTA Grant Delay Smacks of Partisan Politics - CityLab*", which received 62 retweets by 62 unique accounts all on the same day, 2017-02-23. However, all the accounts are named as "CHNG[City Name]" (such as CHNGAustin and CHNGBerkeley) and are obviously a group of accounts belonging to the same organization. The duration measurement for that event is 1, which easily distinguishes this event from other normal events. Similarly, event #24, "*Jerry Brown Vetoes Bill to Improve High-Speed Rail Oversight*", and

event #37, "*Visiting California governor looks to China for high-speed rail inspiration - Peoples Daily Online*", are all one-day events with minimum temporal impacts. Therefore, it is crucial to combine event magnitude using the by user approach with event duration to find and prioritize genuine events. However, Figure 4-3 is not very friendly to read, and the relationship between the event magnitude and event duration is not clearly visible. In the next section we introduce a new presentation of event influence to address this issue.

4.4 Event Influence Quadrant

Due to the difficulty to analyze event influence using bar charts since magnitude and duration are two completely different dimensions, we propose to use an event influence quadrant to visualize the impact of an event. The x-axis of the quadrant is the event duration measured as the number of days, and the y-axis is the magnitude measured using the by user approach. The reference line of average duration and average magnitude of the top events divides the space into four quadrants with different characteristics. Figure 4-4 illustrates the distribution of the 43 events identified using the 1-sigma range of the by user approach.

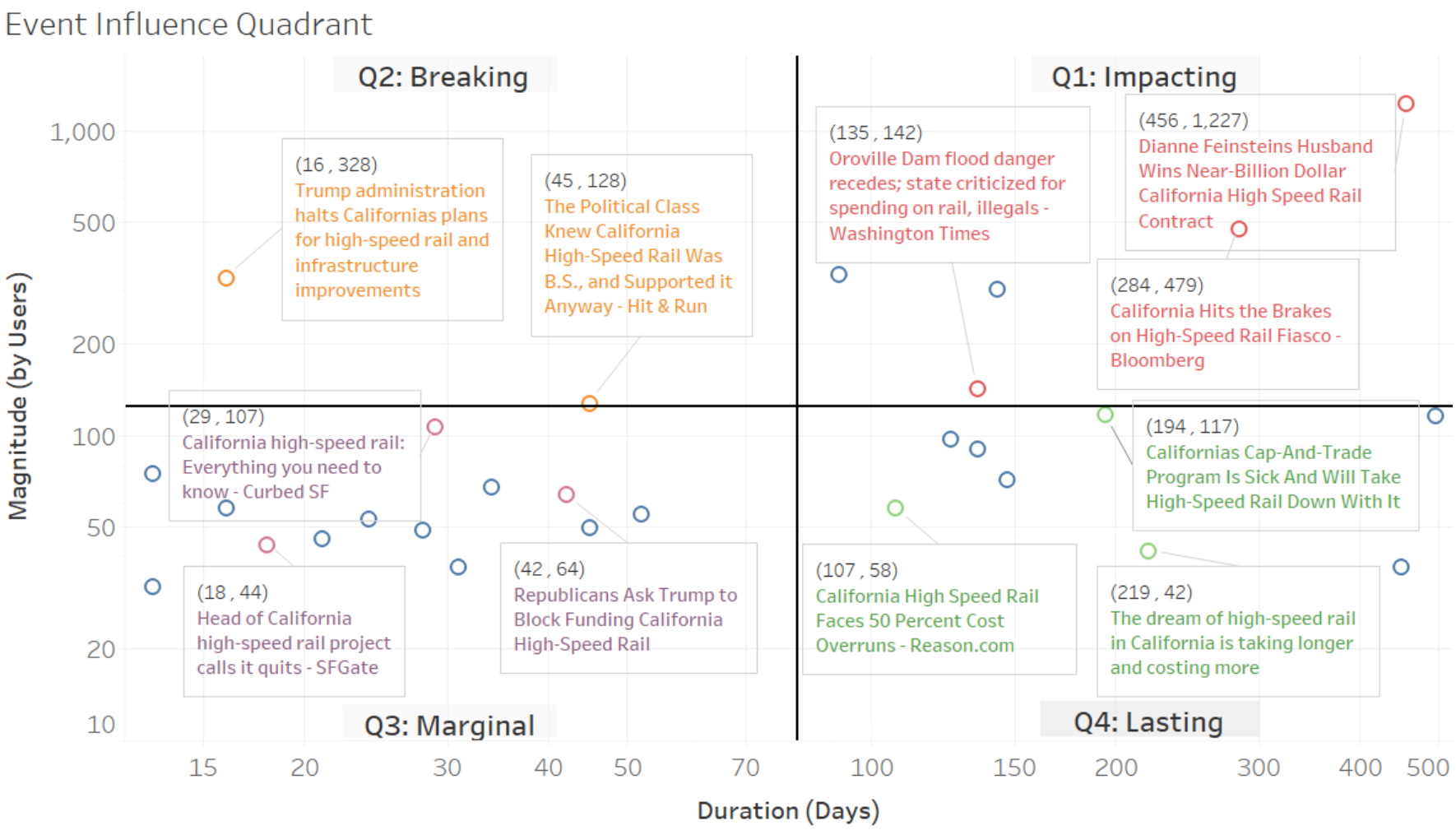

Figure 4-4 Event Influence Quadrant

The event influence quadrant groups events into four categories:

Quadrant 1 (the impacting quadrant): a set of events score more than average in both the number of retweeting users and duration. These events are undoubtedly the most important events due to the scale of the audience and the lasting period. They attract a lot of attention from the public and are deterministic in the public acceptance of the project. Almost all of these events mark serious concerns, issues, and achievements of the infrastructure project which are worth paying close attention to. All five Q1 events in the case study are highlighted below:

- *Dianne Feinstein's Husband Wins Near-Billion Dollar California High Speed Rail Contract.* This is a potential scandal regarding a senator. Combined with another similar quadrant 4 event, the impact of this incident is even higher.

- *California Hits the Brakes on High-Speed Rail Fiasco – Bloomberg.* An article from main stream media with serious questions regarding the financial feasibility of the project.

- *Oroville Dam flood danger recedes; state criticized for spending on rail, illegals - Washington Times.* This article criticizes the state of California of spending money on the CAHSR project instead of the Oroville Dam project which is in need of reinforcement.

- *The Hill on Twitter: "Trump laments lack of high-speed rail in US during meeting with top airline execs"*. This is a news article regarding a president meeting with airline executives.

- *Lame-Duck Obama Admin Rejects CA High-Speed Rail $15 Billion Loan.* This is a news article regarding the failure of the project to get financial support before president Trump takes office.

All Q1 events reveal the focus of the public attention. The questions and concerns raised in these events need to be addressed appropriately to avoid public relation crisis.

Quadrant 2 (the breaking quadrant): a set of events which score above average in the number of retweeting users but lasts a relatively small period of time. Events in this quadrant are similar to breaking news, which draw attention from a lot of people immediately, and the heat dissipates when people turn to other trending topics. Events in this quadrant are important due to the amount of attention they drive. Two Q2 events are identified in this case study:

- *Trump administration halts California's plans for high-speed rail and infrastructure improvements*.
- *The Political Class Knew California High-Speed Rail Was B.S., and Supported it Anyway - Hit & Run*.

Quadrant 4 (the lasting quadrant): a set of events with a relatively small number of tweeting users but an above average duration. Events in this quadrant were kept being mentioned, indicating a continuous interest in a certain topic. Some events in the case study are still being discussed after a year. Investigating the list of Q4 events, we discovered the topic of cost overrun being a constant concern of the project:

- *CA High-Speed Rail: Over Budget, Behind Schedule – Breitbart.*
- *California High Speed Rail Faces 50 Percent Cost Overruns - Reason.com*
- *The dream of high-speed rail in California is taking longer and costing more.*

The following event almost reaches quadrant 1 with 8 less users. It can still be viewed as a highly influential event:

- *California's Cap-And-Trade Program Is Sick And Will Take High-Speed Rail Down With It.*

Quadrant 3 (the marginal quadrant): a set of events which score below average in both magnitude and duration. This is the quadrant of the least impactful events compared with other quadrants. Nevertheless, they are still important enough to be positioned in the quadrant, however,

more resources should be spent on events in other quadrants. Similar to the last event mentioned in quadrant 4, there is an event which almost reaches quadrant 2.

- *California high-speed rail: Everything you need to know - Curbed SF*.

The event influence quadrant provides a clear illustration of the importance of individual events. It is a powerful tool to categorize and prioritize tens of events so that stakeholders can allocate their time and resources wisely. It is worth mentioning that due to the continuous nature of the project evaluation framework, the events and the influence quadrant also evolve over time, and the actions to take on events should be updated accordingly.

4.5 Event Sentiment Analysis

The web pages referred by the events have their own sentiment towards the project. However, in the dataset, these web pages are only tiny URLs, and the sentiment of an event is determined by the text of the tweet, not the web page itself. Therefore the sentiment of an event can be completely different from that of the web page. The trending of event sentiment is studied by accumulating sentiment scores of certain events over time. For CAHSR, we observed a highly consistent sentiment inclination, i.e. tweet sentiment towards a certain follow a steady direction.

Figure 4-5 depicts the sentiment time series of the major negative event, "*Dianne Feinstein's Husband Wins Near-Billion Dollar California High Speed Rail Contract*". With some initial negative tweets, the sentiment score falls dramatically in Feb 2017, triggering by two opinion leaders @xsevenx and @PamelaD66560527 and their retweets of this news. Their tweets are massively distributed, throwing a negative bomb to the event sentiment.

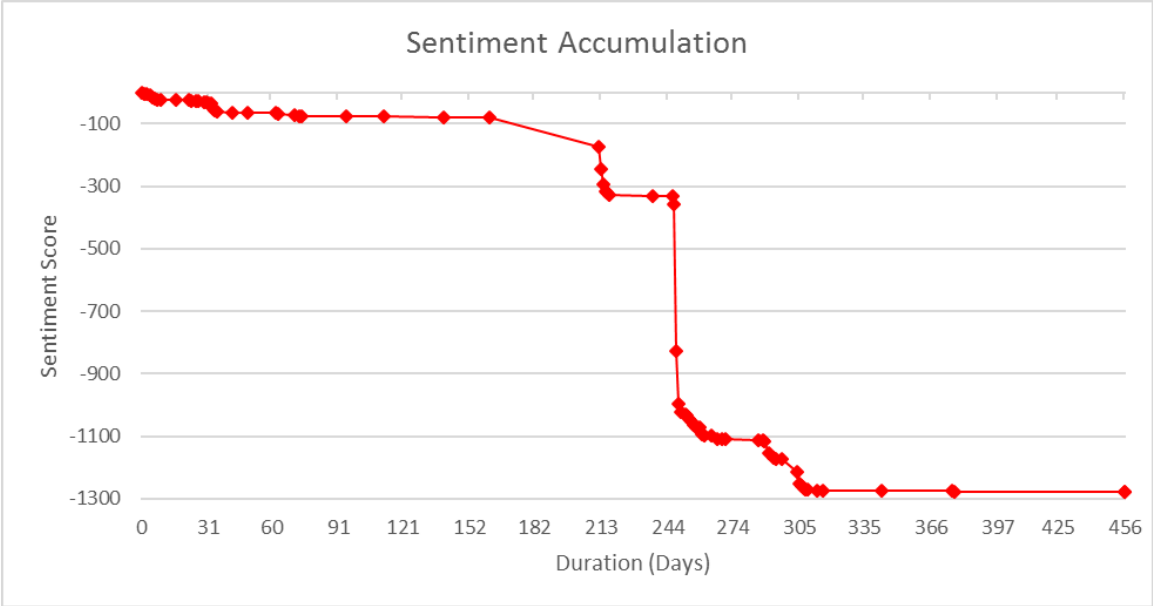

Figure 4-5 Sentiment Accumulation of Event "Dianne Feinstein's Husband Wins Near-Billion Dollar California High Speed Rail Contract"

Figure 4-6 shows the time series sentiment of one of the major positive events "*Trump administration halts California's plans for high-speed rail and infrastructure improvements*". While the sentiment of the article is negative, the tweets referring to the article reads "California is ready". Therefore, the sentiment of the event is flipped and is shown as a positive event.

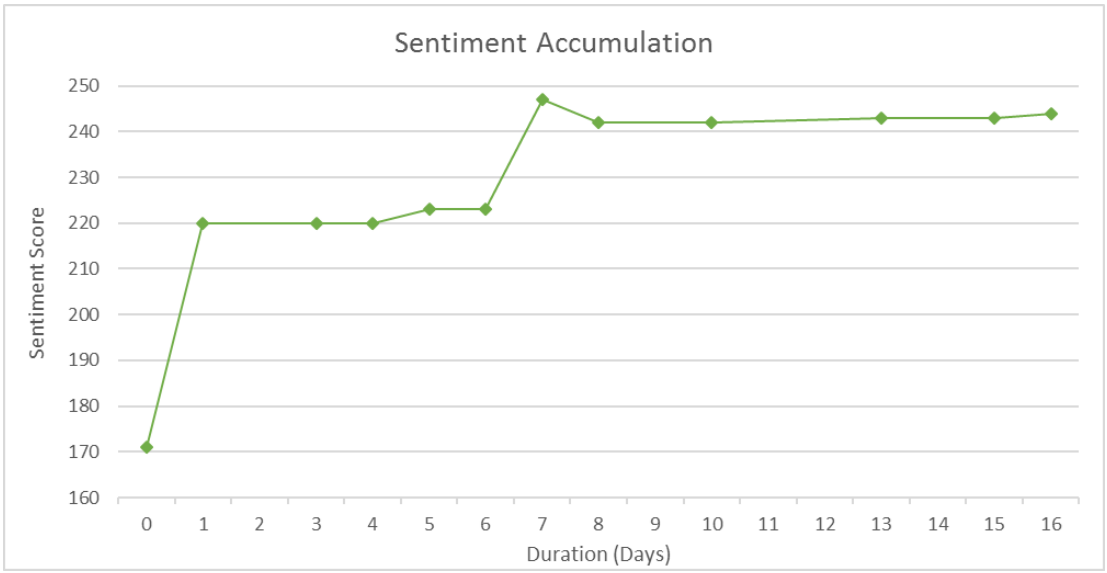

Figure 4-6 Sentiment Accumulation of "Trump administration halts California's plans for high-speed rail and infrastructure improvements"

Most events are studied for their sentiment, and the sentiment growth of two typical events are shown in Figure 4-7 and Figure 4-8. The majority of the tweet activity takes place in the first few days, and is followed by a more flattened growth with much fewer tweets. The constant trend of events demonstrates the importance of early interference. Setting the tone of an event in its early stage is more effective than turning the tide later.

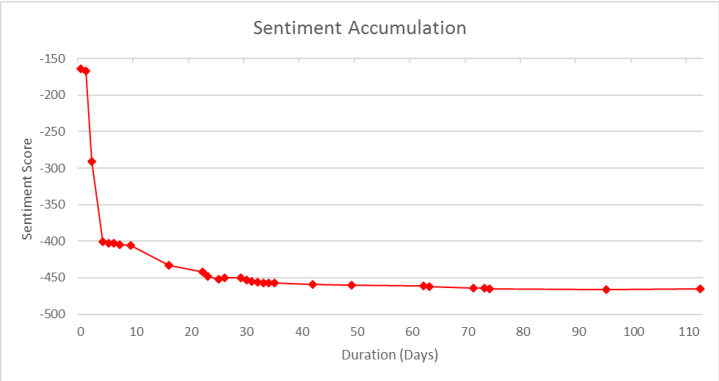

Figure 4-7 Sentiment Accumulation of "California Hits the Brakes on High-Speed Rail Fiasco - Bloomberg"

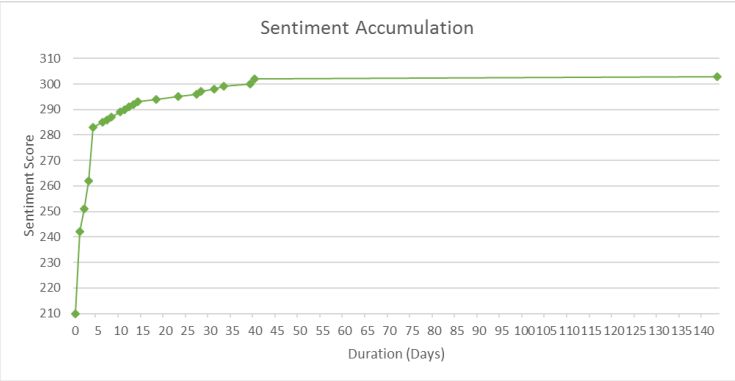

Figure 4-8 Sentiment Accumulation of "Trump laments lack of high-speed rail in US during meeting with top airline execs"

4.6 Event Altering Strategy

While it is important to observe events and their growth in influence, it is more meaningful to take actions to improve the public image of an infrastructure project. By taking advantage of social media and the event influence quadrant, it is fast and easy to identify the targets. The following strategies are proposed to utilize events to maximize public acceptance.

4.6.1 Positive Events

Positive events improve public acceptance and recognize the success stories of the infrastructure project. It is desired for these events to be highly influential and publicly aware. Therefore, the strategy to handle positive events is to expand its influence, in both magnitude and duration, to promote these events into more influential quadrants. Event magnitude can be augmented by marketing campaigns to increase media coverage and public awareness, or by lobbying opinion leaders to advocate their followers. Spreading positive events to a greater audience aims to promote

these events vertically from Q3 to Q2 or Q4 to Q1. To further promote events horizontally, a steady coverage, rather than an intense tweet eruption, is preferred to keep public attention on the events.

4.6.2 Negative Events

Conversely, it is desired to demote negative events out of highly influential quadrants. Negative events express public concerns on potential issues and rumors around the project. Addressing these issues and concerns are certainly the first action to take. Fast and well executed actions help braking the events from growing into more influential quadrants.

Besides focusing on the magnitude and duration, another possible strategy to mitigate the negative influence is to change the sentiment of the event. As mentioned in the event sentiment analysis, event "*Trump administration halts California's plans for high-speed rail and infrastructure improvements*" turns a negative article to a positive event. The side effect of this approach, however, is that if the effort to change sentiment fails, the attention brought to the event might further accrue its magnitude and duration, making it more negatively impactful instead.

4.7 Conclusion

In this chapter, we analyzed the driving factor of public acceptance fluctuation, the social media events. The project evaluation framework is extended to include event analysis, and a new object, the web page referenced by tweet, is accommodated in the data crawler and data storage schema design.

The social media event is defined and a two-dimensional model is proposed to combine event magnitude and event duration to measure event influence. Three different strategies to measure event magnitude, by tweet, by user, and by user influence, are evaluated, and they yield very similar results. The by user approach is slightly better due to its ability to detect fraudulent events.

To better illustrate event influence, an event influence quadrant is proposed to divide events into four quadrants, the impacting quadrant, the breaking quadrant, the lasting quadrant, and the

marginal quadrant. Events in different quadrants have different characteristics. The events of the CAHSR case study are discussed based on their corresponding quadrant. The event quadrant is not a tool to suggest actions to be taken on social media events. It serves as a real-time, cost-effective, and direct tool to monitor event influence changes and facilitate the decision making on public relation affairs. The emergence of events and the change in event influence and sentiment can be quickly caught, and the movement across quadrants alert project managers of potential escalation of events.

Finally, the sentiment trend of social media events are discussed. Typical events follow a single directional sentiment trend, hence it is important to intervene early. Utilizing the event influence quadrant, event altering strategies are discussed regarding how to promote positive events and demote negative events.

Chapter 5. Evaluation of Public Acceptance Using Big Data – A Case Study on Social Media Users

5.1 Introduction

Social media sites such as Facebook and Twitter are platforms which allow individuals and communities to share and discuss user-generated content (Kietzmann, Hermkens, McCarthy, & Silvestre, 2011). Some cleverly composed tweets or videos can have big impact on products and companies (Weber, 2010). User is the most fundamental element of social media and is the very source of creativity. Previous chapters discussed the public acceptance of infrastructure projects and the driving force, event, behind the fluctuation of public acceptance. In this chapter, we would like to discuss the human factor of the equation, and answer the question of WHO are driving the change of public acceptance.

We start the social media user analysis with opinion leadership analysis in section 3. Opinion leadership theory is originally developed by Paul Lazarsfeld and Elihu Katz (Katz, 1957). Three factors, expression of values, professional competence, and nature of their social network contributed to the role of opinion leaders (Katz & Lazarsfeld, 1966). Not only important in traditional media communication, opinion leadership plays a critical role in social media as well. (Turcotte, York, Irving, Scholl, & Pingree, 2015) found through a Facebook experiment that social media recommendations increases the level of trust for particular media, and sharing by opinion leader further amplifies the effect. Opinion leadership in this chapter is not limited to only opinion leader, although it is the most important group of all. Other opinion roles such as opinion follower and original contributor are also defined and discussed. In section 4, an a priori prediction methodology is proposed to filter potential opinion leaders once they emerge in the targeted infrastructure project dataset.

In section 5, a multi-dimensional user characteristic model is introduced to describe user profiles in parallel to opinion leadership. User sentiment, popularity, institutional attribute and location data are collected and analyzed to reveal the distribution among these characteristics. Finally, in section 6, we conclude by discussing the opinion leadership in the context of the user characteristic model.

5.1.1 Twitter Opinion Leader Analysis

Naturally, the number of retweets is much larger than the number of original tweets since the effort involved in both activities is different. Most of the models proposed in this research so far do not distinguish retweet from original tweet, which potentially ignores people who are able to initiate the chain of retweets. On the other hand, those opinion leaders, influential people whose posts get a lot of retweets, play a critical role in improving public relations and building positive public images for infrastructure projects. In this chapter, two methods are proposed to identify opinion leaders. In addition, opinion followers and original contributors are identified as two distinct opinion leadership types. An opinion leader prediction model is also built to detect opinion leaders using a priori indicators.

5.1.2 Twitter User Analysis

Besides opinion leaders, it is interesting to study social media user characteristics. Twitter user API provides rich user profile information to make it much easier than traditional methods to know the respondents. A user profiling model, a model including user sentiment, user popularity, user institution, and user location is developed to describe the demographic features of users. It is then combined with the opinion leadership model to depict these important users.

5.2 Literature Review

5.2.1 Opinion Leadership

Opinion leadership theory was developed by (Lazarsfeld, Berelson, & Gaudet, 1948) whose research objective was to determine how political information are received from the source. A two-step information diffusion process was discovered where the information was, as the first step, received by a minority of opinion leaders, who then pass the information onto opinion followers who are less involved in the topic. On the contrary of the intuition that information is being transmitted directly to the receivers, opinion leaders relay the information to the large population.

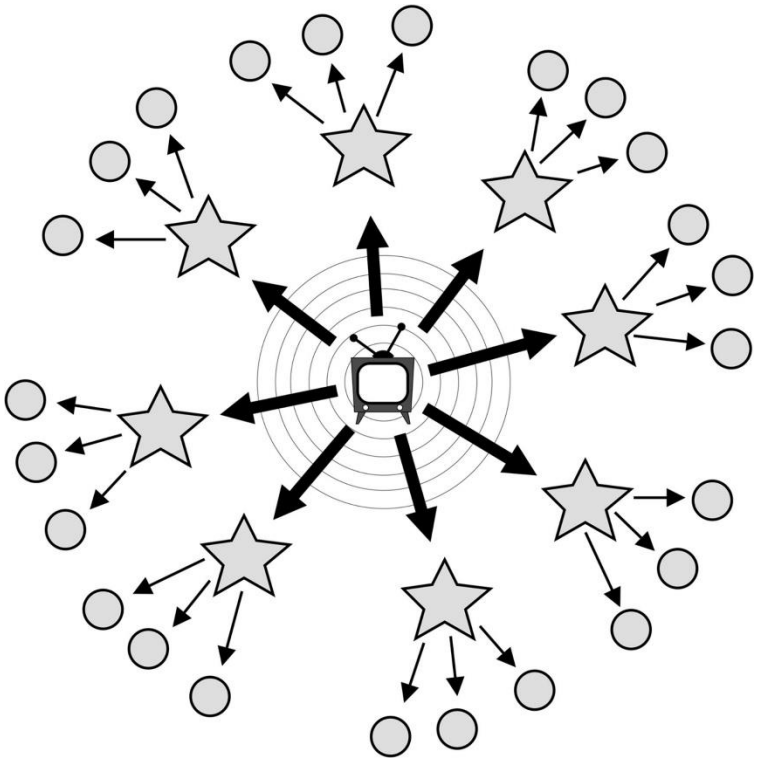

Figure 5-1 Two-Step Flow Model of Influence (Watts & Dodds, 2007)

Opinion leaders are originally suggested as engaged, knowledgeable and to be trusted (Lazarsfeld et al., 1948). (Katz, 1957) found that opinion leaders often belong to the same social groups as their followers. They are influential in their interest area, and the role of influencer and influencee could be exchanged in different situations. In addition to passing information, opinion leaders can also

give advice and serve as role models (Weimann, 1994). (Nisbet & Kotcher, 2009) summarized three categories of opinion leaders as issue specific opinion leaders (Childers, 1986), influence as personality strength (Weimann, Tustin, Van Vuuren, & Joubert, 2007), and Roper ASW's influential (Keller & Berry, 2003).

(Domingos & Richardson, 2001) however, argue that people are often strongly influenced by their peers, friends, and acquaintances rather than opinion leaders. Leveraging the social network value of each customer is more cost-effective than marketing through the influentials. (Watts & Dodds, 2007) found, through mathematical simulations, that the conditions under which influentials trigger large-scale information diffusion are exceptional rather than usual. They argue that influentials are modestly more important than ordinary people, but are not as deterministic as suggested in the conventional theory. Sometimes, influentials are accidental opinion leaders and the trend rely more on the society than any specific person who started it.

5.2.2 Opinion Leadership on Social Media

Traditionally, opinion leadership and information diffusion are discussed based on media such as TV, newspaper and magazines. The emergence of social media brings fundamental changes to traditional media, and correspondingly how information spread through the network. (Bennett & Manheim, 2006) altered the conventional two-step flow model to a one-step message passing paradigm in social media. They suggest that opinion leaders are less likely to lead due to the capability of content generators to deliver messages directly to individuals through more narrow and efficient channels.

On the other hand, researchers suggest that social media has changed the role of opinion leaders, from a first-hand information relay to a filter or a personalized information transmitter of the plenty of information on the Internet. The modern technology and social networks make it easier for people to exercise opinion leadership within their contact circle (Mutz & Young, 2011). (Forbes,

2013) found that people make their buying decisions depending more on recommendations within their connected friends than traditional opinion leaders. (Turcotte et al., 2015) emphasized the importance of opinion leadership in social media as they serve an informing and educating role to the public. (Cha et al., 2010) studied the influence measurements of Twitter users, and the spatial-temporal analysis on how influentials interact with different topics and their followers.

5.3 Opinion Leadership Analysis

Previous research focus mostly on opinion leaders and major events such as the Iranian presidential election, the outbreak of the H1N1 influenza, and the death of Michael Jackson (Cha et al., 2010). For infrastructure projects, topics with relatively low social media activities, the validity of these models needs to be examined and customized. In this research, we employ two different indicators and discuss their effectiveness and advantages. Moreover, it is interesting and meaningful to understand ordinary users and other opinion leadership types, and compare them with opinion leaders. After all, under the pluralist model, they have equal votes on the topic as the opinion leaders. Therefore in this research we extend the scope to three different opinion leadership types, namely opinion leaders, opinion followers and original contributors. Multiple indicators are used to define these opinion leadership groups, apply on the dataset to get the list of users and analyze their group characteristics. For opinion leaders specifically, a predictive model is proposed to identify potential opinion leaders using a priori indicator.

5.3.1 Opinion Leader

An opinion leader in social media is an engaged and trustworthy individual or organization who can influence the general public by his/her opinions. Opinion leaders can be politicians, news media, experts or knowledgeable and respected people. Arguably, they are conceived to play a critical role in information diffusion as pivots of the network.

By definition, opinion leaders' opinion are followed by a large amount of people. (Cha et al., 2010) used three indicators to describe user's influence. Indegree influence is determined by the number of followers of a Twitter user, retweet influence is determined by the number of retweets generated under the user's name, and mention influence is determined by the number of mentions of a user among comments with other users. The study found that indegree influence alone has very little relevance about a user's influence. Tweets related to infrastructure projects generate very limited number of conversations, hence the number of mentions is not a good influence candidate. Therefore, tweet influence is the best indicator among all three.

(Cha et al., 2010) also discussed the indicator of normalized number of retweets by total tweets, which they found to rank local opinion leaders higher than users with highest number of retweets. However, the normalized indicator could work for infrastructure projects due to the low volume of tweets. There might not be a clear distinction between local opinion leaders and global opinion leaders in infrastructure projects. Therefore, both the absolute number of retweets and the normalized measure are tested in opinion leader identification.

The opinion leader score based on the sheer number of retweets can be defined as

$$\mathrm{OL}_k = \sum_t \sum_j R_t^{k,j} \tag{5.1}$$

where $R_t^{k,j}$ is a retweet of user *j* mentioning (@) user *k* on time *t*.

Similarly, the score based on the normalized number of retweets can be defined as

$$\mathrm{OL}_k = \frac{\sum_t \sum_j R_t^{k,j}}{\sum_t T_t^k} \tag{5.2}$$

where T_t^k is a tweet of user *k* on time *t*.

Applying formula 5.1 and 5.2 on the CAHSR data set, we obtained 1,121 users out of 13,396 users, a subset of total users excluding those failed to be crawled by the user API. These 1,121 users have

their tweets retweeted at least once during the data collection time frame. The opinion leader score using the absolute number of retweets range from 1 to 1,802 with the mean of 12.0 and the standard deviation of 69.6. Using the 3-sigma rule, we identify opinion leaders under this score to have at least 221 retweets, resulting in the following 9 opinion leaders.

Table 5-1 Opinion Leaders Identified by Number of Retweets

User	Description	Retweet	Normalized Retweet
CaHSRA	Official Twitter for California's High-Speed Rail Project. [tinyurl]	1802	7.12
rharrisonfries	retired television broadcast mgmt. & USN	1049	262.25
PamelaD66560527	Freedom Fighter	461	230.5
Bud_Doggin	Conservative IT professional. Go TRUMP! Followed by @JessieJaneDuff @FiveRights @TEN_GOP @GenFlynn #TRUMP #MAGA Prouly Blocked by @williamlegate	374	374
JerryBrownGov	On Facebook at: [tinyurl]	340	340
iowahawkblog	Karma's janitor	279	279
activist360	Singer-songwriter, musician, activist, poet, yogi, fierce paladin for social justice and the environment. NEW RELIGION is available at [tinyurl]	266	266
DaytonPubPolicy	Kevin Dayton is the President & CEO of Labor Issues Solutions, LLC and the Dayton Public Policy Institute in California.	263	1.23
2020fight	Teacher & Advocate. Fighting for 2020...	259	129.5

Similarly, the opinion leader score using the normalized retweets range from 0.01 to 374 with the mean of 5.71 and the standard deviation of 25.2. Using the 3-sigma rule, we identify opinion leaders under this score to have at least 82 retweets per tweet, resulting in the following 16 opinion leaders.

Table 5-2 Opinion Leaders Identified by Normalized Number of Retweets

User	Description	Retweet	Normalized Retweet
Bud_Doggin	Conservative IT professional. Go TRUMP! Followed by @JessieJaneDuff @FiveRights @TEN_GOP @GenFlynn #TRUMP #MAGA Prouly Blocked by @williamlegate	374	374
JerryBrownGov	On Facebook at: [tinyurl]	340	340
iowahawkblog	Karma's janitor	279	279
activist360	Singer-songwriter, musician, activist, poet, yogi, fierce paladin for social justice and the environment. NEW RELIGION is available at [tinyurl]	266	266
rharrisonfries	retired television broadcast mgmt. & USN	1049	262.25
PamelaD66560527	Freedom Fighter	461	230.5
primalpoly	Evolutionary psych professor; wrote some books. Mate choice, sexual politics, Effective Altruism, freedom. Most tweets are ironic & don't reflect anyone's views	206	206
ramzpaul	Video maker and speaker. Youtube channel: [tinyurl] Support my Patreon: [tinyurl]	144	144
2020fight	Teacher & Advocate. Fighting for 2020...	259	129.5
peddoc63	TexasPatriot,Nurse,Jesus,Family +Guns,Vets, Blue lives, Israel PASSIONATE boutMyCountry! Honored2BfollowedBy @AlvedaCKing @RealJamesWoods @peddoc63 @A_M_Perez	119	119
RedNationRising	Welcome to the Official Red Nation Rising Twitter page! Grassroots organization for Education, Constitution and Civics. #RedNationRising	214	107
jimEastridge1	Christian~ Conservative ~ TeaParty ~ Patriot~ constitutionalist ~ There would not be a 1st Amendment without the 2nd Amendment~ ????? ???? Trump Supporter !!	212	106
RMConservative	Senior Editor at [tinyurl] Conservative writer, policy analyst, new book Stolen Sovereignty [tinyurl]	106	106
gehrig38	Whatever it Takes 9-11 am M-F [tinyurl] call in live! [Tel] Liberals welcome to argue IN REALITY!	206	103
ALT_USCIS	The account #trump came after. Immigration, stuff they don't want you to see, facts, patriotism. Not the views of DHS-USCIS. #altgov #SAVEDACA	92	92
tomesimpson	"The heart of the wise inclines to the right but the heart of the fool to the left." Ecclesiastes 10: 2. FB: @DineshDSouza @TRobinsonNewERA @NatPoliceAssoc	84	84

Except for the number of people, the list of opinion leaders identified by both indicators have a good overlap. 8 out of 10 people in Table 5-1 are also in Table 5-2. It is noticeable that most of the opinion leaders do not have a lot of tweets in the dataset. Therefore, two frequent tweeters, CaHSRA and DaytonPubPolicy have their ranks dropped after the normalization.

Politicians and traditional media accounts are still playing a critical role as influencers for the public. However, grass root organizations and individuals take a large portion of the list. Identifying these people and organizations is a distinguished contribution of this research, as they can be easily overlooked in traditional public opinion assessments. These accounts might not be as powerful as governors and mainstream media, but their impact on the public acceptance of CAHSR cannot be underestimated. Following their tweets reveals trending topics of the project, and furthermore, lobbying them is an effective way to improve public acceptance of the project.

5.3.2 Opinion Follower

Similar to opinion leaders, opinion followers are the majority of the public, consisting of consumers searching for information for guidance from sources such as the media. It might not be as straightforward as opinion leaders on why it is important to understand opinion followers. However, (Kellerman, 2007) shows that leaders and followers cannot be conceived separately without knowing the other. Followers have their own interests, power and influence, just as the leaders do, even though their authority is relatively lower. It is therefore meaningful to identify top opinion followers to at least better understand opinion leaders.

We define the opinion follower score is the absolute number of retweets posted by a certain user. There is little value to normalize this score since the retweets are a subset of the total tweets of this user. The opinion follower score is formulated as:

$$\mathrm{OF}_j = \sum_t \sum_k R_t^{k,j} \tag{5.3}$$

where $R_t^{k,j}$ is retweet of user *j* mentioning (@) user *k* on time *t*.

By applying formula 5.3 on the CAHSR data set, the opinion follower score is calculated for all users collected. 9,613 users have retweeted at least once, much higher than the number of user being retweeted. The opinion follower score range from 1 to 186 with the mean of 1.4 and the standard deviation of 3.5. The 3-sigma rule gives the threshold of 13, identifying 61 top opinion followers. Table 5-3 lists the top 19 opinion followers with at least 30 retweets.

Table 5-3 Top 19 Opinion Followers

User	**Description**	**Tweet**
CAGovTweets	Using data to highlight great California government communication. From @measuredvoice	186
cahsr_scam	Put the Brakes on California's High-Speed Rail	117
dougqdrozd	@CaHSRA by day. Kings, Dodgers, Raiders follower the rest of the time. My own opinions.	90
dougq_d	N/A	87
ca_trans_agency	California State Transportation Agency develops and coordinates state transportation policies and programs to meet safety, mobility and air quality objectives	75
lmburcar	Wife, Friend, @CAHSRA Press Secretary & All Around Fun Chick	65
CaHSRA	Official Twitter for California's High-Speed Rail Project. [tinyurl]	64
JaCastruccio	Racing enthusiast, novice horsewoman, mother of sailor, chef and pilot. Proud OTTB sponsor and owner of a retired Cal Bred. I put my money where it counts.	63
jvvine	Passionate about family, close friends, traveling and golf.	61
USHSR	The premier organization advocating for high speed rail in the USA, with connecting transport networks & urban smart growth. Join us for #WCRail17!	58
ClaySharps	formerly young Marine Cpl. now grouchy conservative Grandpa. I believe in God, USA, Family, Corps..No, I will never be p.c. or sensitive.	51
ericdchristen	Business owner, husband of Lt. Col Karyn Christen, father and homeschool dad of Damian, Sophia and Gabriel. Lover of Christ and free markets.	44
Minky42659	N/A	39

vergie49398619	N/A	37
CaWater4All	Fixing CA's long-term water problems by increasing storage capacity for all water users #Water4All #moreDAMstorage #HighSpeedFail	34
jetsison	Licensed California Real Estate Broker with over 20 years in the business. If it is land you seek, it is I you should speak... to! :)	32
RailfanGuy	Journalist, writer, overjoyed Cubs fan. Views are my own. Follows and RTs are not endorsements.	32
DaytonPubPolicy	Kevin Dayton is the President & CEO of Labor Issues Solutions, LLC and the Dayton Public Policy Institute in California.	30
meli_fig	Press Secretary. Personal assistant to 7 yr old superhero and nocturnal twins. Wife. @Dodgers fan. @calpoly Mustang Faithful. Views are my own.	30

Interestingly, the lists of top opinion leaders and top opinion followers are two nearly exclusive user groups. As mentioned before, opinion leaders tend to tweet in a very low volume, where opinion followers are the opposite. Two opinion leaders, by sheer number of retweets, CaHSRA and DaytonPubPolicy, are listed as top opinion followers as well. @CaHSRA is an exception since it is the official tweet account of CAHSR. @DaytonPubPolicy appears to be an active information conveyer of CAHSR by processing large amount of information and leading a large amount of people.

Three organizations, two opposing organizations (@cahsr_scam and @CaWater4All) and one advocating organization (@USHSR), are found in this list as well. Unlike the organizations in opinion leaders, these organizations have strong sentiment towards the project. Identifying opinion followers is therefore observed to be an effective way to identify interest groups and organizations.

5.3.3 Original Contributor

The iconic indicators of opinion leaders and opinion followers are based on retweets, either from or referencing a Twitter account. Retweet also plays a critical role in both public acceptance analysis and event analysis where they are the dominant contributor to those metrics. Although

retweet is a key component of social media analysis, focusing only on retweets overlooks the endeavor of original content generators who spent more effort composing a message than simply clicking a button.

Original contributors are users who write original tweets instead of retweeting someone else's. They represent the group of people who are willing to express themselves on social media. They are not necessarily opinion leaders, which are determined mostly by their followers, however, they are certainly part of the interest group of the infrastructure project who can provide original insights.

We define the original contributor score by the number of original tweets posted by a certain user.

$$\mathrm{OC}_j = \sum_t OT_t^j \tag{5.4}$$

where OT_t^j is an original tweet of user *j* on time *t*, i.e. a tweet not retweeting any other tweets.

Applying formula 5.4 on the CAHSR data set, 5,556 users were found to have composed at least one original tweet during the data collection time frame. The original contributor score range from 1 to 430 with the mean of 2.0 and the standard deviation of 8.9. The 3-sigma rule gives the threshold of 29, qualifying 25 top original contributors. Table 5-4 lists the top 17 original contributors with at least 40 original tweets.

Table 5-4 Top 17 Original Contributors

User	Description	Original Tweets
RobertDolezal	High-tech content executive and startup advisor to top teams	430
cahsr_scam	N/A	287
CaHSRA	Official Twitter for California's High-Speed Rail Project. [tinyurl]	189
DaytonPubPolicy	Kevin Dayton is the President & CEO of Labor Issues Solutions, LLC and the Dayton Public Policy Institute in California.	183
DrRajSelladurai	#Jesus-Follower, Husband, Dad, Business Professor, Author, Fulbright Specialist. #ServantLeadership; #HighSpeedRail; #Businesshealthcare Collaboration.	159

CAGovTweets	Using data to highlight great California government communication. From @measuredvoice	136
CALHSR	Grassroots #transit advocates, mostly California High Speed Rail. #FOIA good, data good, good govt good #Persisters who insist we can do better.	103
caledlawgroup	California's Premier Eminent Domain Law Firm. Practicing exclusively eminent domain in California.	64
alevin	[tinyurl]	63
CCHSRA	Citizens for California High Speed Rail Accountability (CCHSRA) is a nonpartisan advocacy group. RT's/Mentions/Follows are not necessarily endorsements.	61
RAILMag	The most extensive coverage of North American passenger rail on Twitter	60
shedmaster48	[tinyurl]and [tinyurl]..sharing international railway news stories..not necessarily endorsements.You decide...	58
USHSR	The premier organization advocating for high speed rail in the USA, with connecting transport networks & urban smart growth. Join us for #WCRail17!	58
narprail	NARP is a 23,000-plus-member nonprofit that seeks a modern, customer-focused, national passenger train network to provide a travel choice Americans want.	54
derekhandova2	Content marketing and writing. Interested in B2B space and technology stories. See website link for white paper and case study examples.	53
suldrew	San Francisco, CA	45

Besides the organizations identified previously (@cahsr_scam and @USHSR), two more organizations emerges in the top original contributor list with one advocating for (@CALHSR) and one opposing (@CCHSRA) the project. Investigating the original contents generated by these top contributors revealed what they are after and helps the project managers understand the focus of these public interest groups. Top opinion followers and top original contributors are both effective tools to identify actively engaged individuals and organizations.

Besides pulling a list of original contributors, we would like to extend the analysis further to the content of these original tweets. Unlike opinion leaders and opinion followers whose tweets are dominated by major events, original contributor is least affected in that matter since retweets are not considered in the score formula.

Word cloud technique is an appealing visualization to provide an overview of high frequency texts (Heimerl, Lohmann, Lange, & Ertl, 2014). It is used in analyzing the content from original contributors. All of the original tweets from the case study dataset are tokenized, stemmed, stop word filtered, and visualized in word cloud, as shown in Figure 5-2.

Figure 5-2 Word Frequency Analysis for Original Content

Some interesting observations can be made regarding Figure 5-2.

- The most valuable people being quoted are, in descending order: @jerrybrowngov, @caltrain, @realdonaldtrump, @potus and @alevin. Except for @alevin, other accounts are all government or organization accounts. President Trump and his policies have big impact on CAHSR, and @alevin is a hidden opinion leader whose opinions are being referred to the most by the community.

- The most popular hashtag topics are, in descending order: #california, #cahsra, #highspeedrail, #iwillride, #bullettrain, #hsr, #ca, #rail and #transit. Besides some obvious hashtags, #iwillride stands out as a popular topic when discussing about CAHSR. These hashtags are candidates to be included in the crawler search terms to enrich the dataset.

- Financial readiness, including both funding and cost, is a highly concerning topic since "funding", "cost", "budget" and "money" are all high frequency words.

- When comparing CAHSR, people always refer to Texas and the hyperloop project for the perspective of state and technology.

Interesting and hidden observations can be made in analyzing the original tweets. While retweets demonstrate people's endorsement on others' opinion, original tweets show independent thinking of the public and provide a different perspective to observe the project.

5.4 Opinion Leader Prediction

Section 5.3 demonstrated the capability of the project evaluation framework to identify opinion leaders, opinion followers and original contributors, and as a side effect, most active public interest groups. However, it is worth noticing that the opinion leaders are identified a posteriori, i.e. they need to accumulate their tweets in the dataset before being identified, which could potentially delay necessary responses. A timely opinion leader identification process is important so that once they enter the data collection, the system can mark them as potential opinion leaders in order to take proper actions.

An a priori indicator is needed in order to predict potential opinion leaders. Taking into consideration the data available from social media, two possible indicators, the absolute number of retweets and the normalized number of retweets of any random 7-day window, are proposed to be the a priori indicators. The indicators are similar to formula 5.2 and 5.3, with two differences. Firstly, the indicator is limited to a 7-day window due to the limitation from Twitter which allows standard API calls to fetch only last 7 days of data. Secondly, it includes all tweets from a given user, no matter if they are related to the infrastructure project or not. Due to the relatively small volume of tweets related to infrastructure projects, limiting the topic would significantly reduce the amount of tweet collected for each user and potentially invalidate the indicators.

As discussed before, absolute and normalized number of retweets yield very similar results in identifying opinion leaders. We will use both indicators to predict opinion leaders and test their effectiveness. Due to Twitter's throttling on data retrieval rate, it is difficult to crawl all of the 13,396 users in the record. 443 users are sampled randomly, and within a 7-day window, their own tweets and the retweeting tweets are crawled. For absolute number of tweets, the 443-user sample has the mean of 122.0 and the standard deviation of 450.7, giving a 3-sigma threshold of 1,475 tweets. For normalized number of tweets, the 443 users has a mean of 0.6 and a standard deviation of 1.1, resulting in a 3-sigma threshold of 3.85.

Accordingly, the top opinion leaders listed in Figure 5-1 and Figure 5-2 are crawled again for all their tweets and all tweets retweeting theirs in a random 7 days window. The resulting tweets and normalized retweets are shown in Table 5-5. The left 3 columns list the top 9 opinion leaders identified by absolute number of retweets, sorting by the number of retweets a priori in descending order. The right 3 columns list top 16 opinion leaders identified by normalized number of retweets, sorting by normalized retweets a priori in descending order.

Table 5-5 Opinion Leader Measurments Comparison

User	**Retweet**	**Norm. Retweet**	**User**	**Retweet**	**Norm. Retweet**
iowahawkblog	8,848	40.77	activist360	3,768	56.24
activist360	3,768	56.24	iowahawkblog	8,848	40.77
2020fight	2,817	5.13	RedNationRising	652	31.05
rharrisonfries	486	0.39	ALT_uscis	2,487	25.38
JerryBrownGov	370	18.5	JerryBrownGov	370	18.5
CaHSRA	242	7.56	primalpoly	1,222	14.9
DaytonPubPolicy	66	0.52	ramzpaul	3,788	13.06
PamelaD66560527	61	0.05	gehrig38	217	10.85
Bud_Doggin	43	2.15	RMConservative	588	8.28
			2020fight	2,817	5.13
			Bud_Doggin	43	2.15

	jimEastridge1	996	1.84
	rharrisonfries	486	0.39
	PamelaD66560527	61	0.05
	peddoc63	118	Not authorized
	tomesimpson	18	Not authorized

Using the absolute number of tweets, only 3 out of 9 opinion leaders are qualified a priori. On the contrary, using the normalized number of retweets, 10 out of 16 opinion leaders are included in the a priori list. Among the 6 unmatched ones, 2 of them do not authorize API calls to their timeline, which are not counter examples of the indicator.

Based on the analysis above, using the normalized number of retweets as the indicator for opinion leader prediction, as well as identification, is favorable compared to the absolute number. Firstly, this indicator gives relatively more a posteriori opinion leaders. Secondly, the type II error using this indicator to predict opinion leaders is significantly lower than using the absolute number. Correspondingly, the workflow to identify opinion leaders in the project evaluation framework is shown in Figure 5-3.

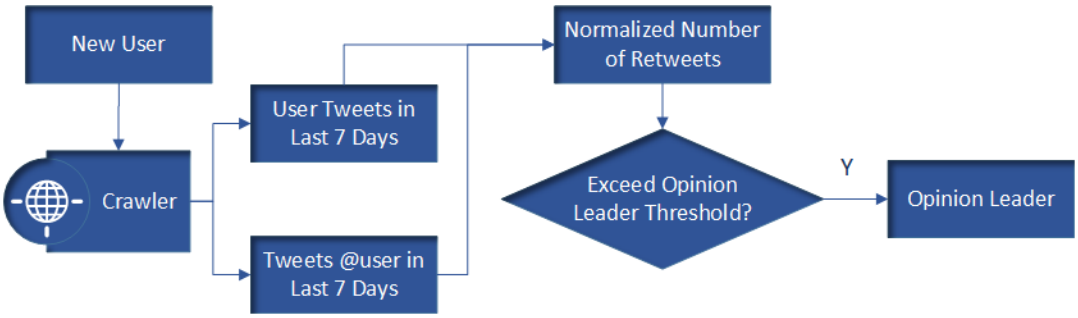

Figure 5-3 Opinion Leader Prediction Workflow

5.5 User Profiling Model

The opinion leadership analysis determines how influential a user is in the topic of infrastructure project. In this section, we will explore four different pieces of information available from our framework, user sentiment, user popularity, user institution, and user location, to further

characterize users. User sentiment determines user's overall attitude towards the project, and also how often a user changes his/her opinions. User popularity shows how large of a social network a user has built. User institution reveals whether a user represents an individual or an organization, and whether the user's opinion is personal or a group one. User location segments users by geographical property, showing regional variance of the acceptance. Dividing users by their demographic and social attributes help project managers understand the distribution of the interest community, which can be used to conduct targeted marketing campaigns or activities to improve the public image of the project.

5.5.1 User Sentiment

Sentiment is previously used to describe the attitudes of tweets, which is then used in calculating public acceptance and event influence. Similarly, sentiment could also be used to describe users' attitude towards infrastructure projects. In this research, we examine the overall aggregated sentiment and sentiment changes of users in order to determine how strong a user's attitude is, and how difficult it is to change it.

- Overall Sentiment

The user's overall sentiment aggregates all sentiment values of tweets posted by a certain user over the entire data collection time frame, resulting in a single sentiment value as a snapshot at a certain point in time. Since the project evaluation framework is a continuous model, as the infrastructure project proceeds, it is possible for users to change their sentimental stance from positive to negative or vice versa. The user's overall sentiment is defined as:

$$S_j = \sum_t \sum_i V_t^{i,j} \tag{5.1}$$

where $V_t^{i,j}$ is the sentiment value of tweet *i* of user *j* on time *t*.

Figure 5-4 shows the histogram of the user sentiment distribution of the case study.

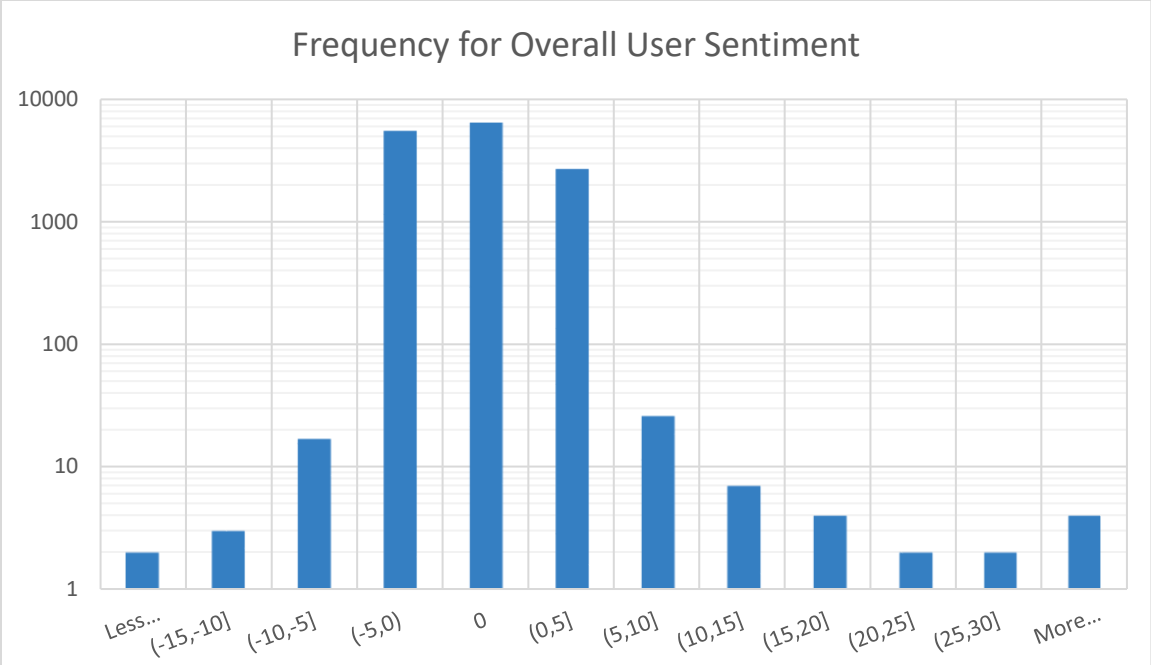

Figure 5-4 User Overall Sentiment Distribution

Among all users 5,477 (37.7%) are negative, 6,344 (43.6%) are neutral, and 2,701 (18.6%) are positive. Coincidentally, if we follow the definition of public acceptance (formula 3.6) and replace tweet count with user count, we can calculate a similar public acceptance by user count.

$$PA(U) = \frac{PU}{PU + NU} \tag{5.2}$$

where PU is the number of positive users and NU is the number of negative users. This formula gives us 33.0% supporting ratio, aligning well with the result of using tweet count.

The overall sentiment of CAHSR dataset users ranges from -82 to 182. The average sentiment score is -0.17, which is slightly negative. The standard deviation of the sentiment is 2.19, hence the 3-sigma range is (-7, 7). Beyond the range, there are 32 most positive users and 9 most negative users. Table 5-6 lists top 9 positive users and negative users in the data set.

Table 5-6 Top Positive and Negative Users

User	Sentiment	User	Sentiment
CAGovTweets	182	CaWater4All	-7
DrRajSelladurai	77	kwilli1046	-7
CaHSRA	67	DaytonPubPolicy	-8
jmorrison9	32	NorCalCrush	-8
ca_trans_agency	29	chuckie_chopper	-10
dougqdrozd	26	stevemongomac	-10
railLAorg	23	ClaySharps	-13
InfoHeaders_met	21	ericdchristen	-16
jvvine	19	RobertDolezal	-82

Table 5-7 Sentiment Score of Top Opinion Leaders, Opinion Followers and Original Contributors

Opinion Leader	Sentiment	Opinion Follower	Sentiment	Original Contributor	Sentiment
activist360	0	CAGovTweets	182	RobertDolezal	-82
iowahawkblog	-1	CaHSR_Scam	-2	CaHSR_Scam	-2
RedNationRising	-2	dougqdrozd	26	CaHSRA	67
ALT_USCIS	-1	DougQ_D	17	DaytonPubPolicy	-8
JerryBrownGov	1	ca_trans_agency	29	DrRajSelladurai	77
primalpoly	0	lmburcar	18	CAGovTweets	182
ramzpaul	1	CaHSRA	67	CALHSR	-2
gehrig38	-1	JaCastruccio	2	caledlawgroup	3
RMConservative	0	jvvine	19	alevin	7
2020fight	1	USHSR	13	CCHSRA	-2
Bud_Doggin	-1	ClaySharps	-13	RAILMag	7
jimEastridge1	-2	ericdchristen	-16	shedmaster48	3
rharrisonfries	-1	Minky42659	-1	USHSR	13
PamelaD66560527	-1	vergie49398619	0	narprail	8
peddoc63	-1	CaWater4All	-7	derekhandova2	0
tomesimpson	0	jetsison	11	suldrew	12

Table 5-7 lists the sentiment score of top opinion leaders, opinion followers and original contributors. Since 16 opinion leaders, 61 opinion followers and 25 original contributors are found in opinion leadership analysis, only top 16 of three categories are listed.

Some interesting observations can be made according to Table 5-6 and Table 5-7. Firstly, none of the opinion leaders are strongly sentimental. In fact, opinion leaders typically have a relatively small number of tweets related to infrastructure projects. The preferred indicator, the normalized number of retweets, outstand small number of tweets and large number of retweets. Thus it is expected for opinion leaders to have mild sentiment.

This is a clear distinction between opinion leaders and the other two categories. Opinion followers and original contributors demonstrate very high correlation with top sentimental users. Just within the top 16 lists, 12 (75%) top opinion followers and 10 (62.5%) top original contributors are also top sentimental users. The characteristics of these categories highlight large volume of tweets, and these users tend to be sentimental at the same time.

Top sentimental users are usually opinion followers. In fact, we observe that if a top sentimental user has a very low opinion follower score, there is a good chance for that user to have bot-like behaviors – automated programs capable of doing human-like activities such as tweeting, retweeting, liking, and following by calling Twitter APIs. These programs are common in Twitter just as other social media platforms (Chu, Gianvecchio, Wang, & Jajodia, 2012).

Looking at these 4 users out of the 16 top opinion followers who are not top sentimental users:

- @DrRajSelladurai, an account with sentiment score of 77 and opinion follower score of 18, mostly retweets two tweets, "*Exciting high speed rail in CA, FL, TX, IL, IN...USA!*" and "*HighSpeedRail: Link to the Future! Travel America by Rail Again!*".

- @jmorrison9, an account with 32 sentiment score and opinion follower score of 0, only retweets "*A free boarder wall, have the railroads pay for it, and install a high speed rail from California to Texas, thanks.*"

- @InfoHeaders_met, an account with sentiment score of 21 and opinion follower score of 0, only tweet about "*The latest On Railways!*"

Therefore, a high sentiment score and a low opinion follower score indicates suspicious accounts with bot-like behaviors, which are at least trivial to be analyzed further.

- Sentiment Change

The overall user sentiment describes the aggregated user attitude at a certain point in time towards an infrastructure project. However, the time dimension is not manifested in the overall sentiment score. One advantage of using social media is the ability to investigate time series changes of user sentiment and observe when and why their sentiment changes. Regarding the CAHSR case study, 8,379 out of 14,546 (57.6%) total users, including the ones cannot be returned by user API, have had sentimental tweets regarding CAHSR. In total, six types of opinion changes are available for analysis, namely neutral to positive, neutral to negative, negative to positive, and their reverse directions. Table 5-8 summarizes the distribution of each type of change and the corresponding user followers.

Table 5-8 User Opinion Changes

User Type	**Number of User**	**Percentage of User**	**Average Followers**
Change opinion	8,379	57.6%	7,052
Neutral to Positive	3,620	24.9%	3,266
Positive to Neutral	983	6.8%	3,231
Neutral to Negative	5,925	40.7%	2,373
Negative to Neutral	779	5.4%	2,418
Negative to Positive	347	2.4%	1,884
Positive to Negative	238	1.6%	2,565

As shown in Table 5-8, about half of the users change their opinion. Among all those changes, the majority goes from neutral to negative and positive. Very few people change from sentimental back to neutral, and even fewer people change their opinion dramatically between negative and positive. Since most people stick to the decision after it was first made, it is critical to inform them at the first place before they form their opinion. Changing people's opinions and attitudes appears difficult after a bad first impression.

To better illustrate user's change of opinion, two legitimate personal users, @dougqdrozd, and @RobertDolezal, are selected from the list of top sentimental users for time series analysis.

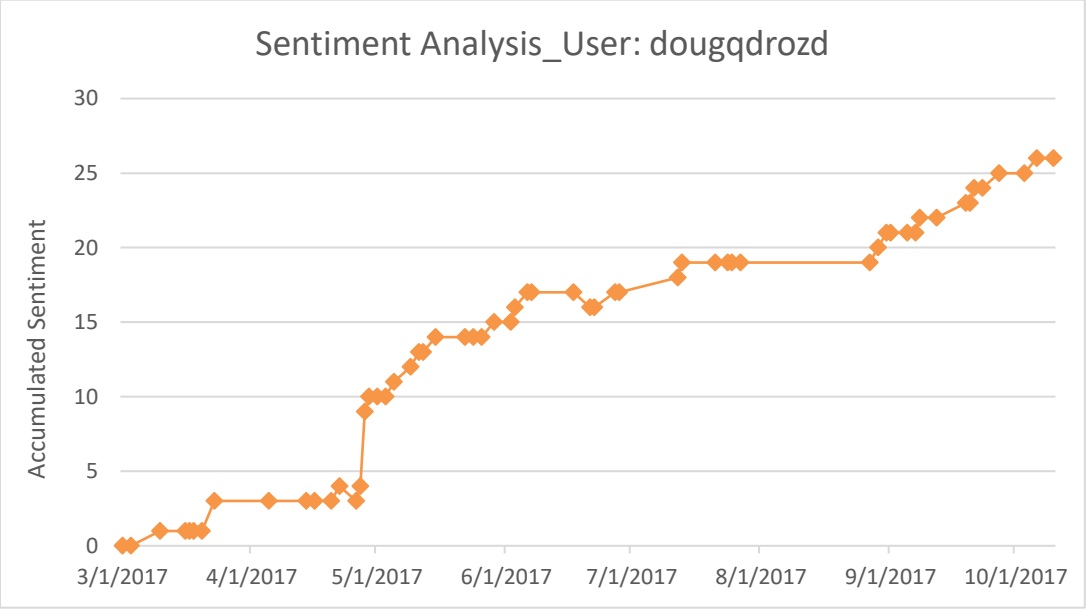

Figure 5-5 Sentiment Analysis of User @dougqdrozd

Figure 5-5 is the sentiment trend of @dougqdrozd, a top positive user, while Figure 5-6 belongs to @RobertDolezal, a top negative user. With some minor fluctuations, the accumulated sentiment of both users are unidirectional, meaning their mind set is predetermined and hence reflected by their tweets. The extensive usage of puns and sarcasm in both users' tweets lead to some sentiment mapping errors, however, the overall trending is solid and representative.

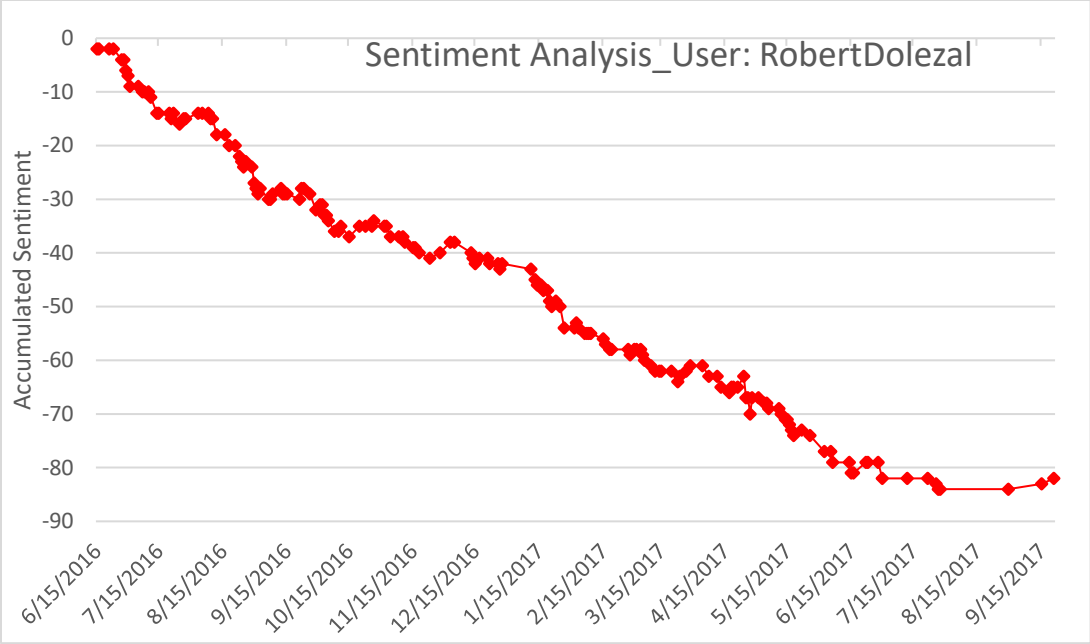

Figure 5-6 Sentiment Analysis of User @RobertDolezal

The difficulty to change people's opinion emphasizes the importance of first impressions and opinion leaders. A well-received first impressions, solid financial plan, considerate public policies or a promising prospect establishes positive images of the project to the public, facilitating high level public acceptance. Conversely, negative project images, once conceived by the public, are also resistant to change and would cost a lot more in the future.

5.5.2 User Popularity

The number of user's followers, also known as indegree influence or user popularity, was briefly discussed in opinion leadership analysis. Although (Cha et al., 2010) concluded that this number alone reveals very little about the user influence, it is still an important measure of user popularity. Among multiple similar metrics available for a Twitter user profile, including followers count, favorite count and friend count, the followers count is the most important one since it is the most difficult measure to be "manipulated". It requires continuous hard work to accumulate followers count.

Following the same power-law characteristic of user influence (Cha et al., 2010), and similar to formula 3.12 and 4.4, users are categorized into 5 follower tiers based on the logarithm scale, as shown in Table 5-9.

Table 5-9 Five Tiers of Followers

Followers Count	**Tier**
[0, 10]	Micro
(10, 100]	Small
(100, 1000]	Medium
(1000, 10000]	Large
(10000, ∞)	Xlarge

The distribution of users and tweets by user follower tiers is shown in Table 5-10:

Table 5-10 User Distribution Based on Followers Tiers

Tier	**Number of Users**	**Number of Tweets**	**Average Sentiment**
Micro	328 (2.4%)	518 (2.3%)	-0.12
Small	1569 (11.7%)	2322 (10.1%)	-0.03
Medium	5431 (40.5%)	9267 (40.4%)	-0.07
Large	5101 (38.1%)	8688 (37.9%)	-0.14
XLarge	967 (7.2%)	2113 (9.2%)	-0.09

The number of users in each tier is relatively proportional to the number of tweets they post. Medium and Large tiers have the majority (around 80%) of users and tweets. XLarge users tweet more frequently than the rest of the tiers. The top 16 opinion leaders have a median follower count of 40,097 and a mean of 136,654, heavily overlapping with XLarge users. Micro and Small tiers contribute the least to the framework, as expected.

5.5.3 User Institution

The Twitter ecosystem consists of both individuals and institutions. It has evolved to be a social media not only for people to express their opinions, but also a platform for traditional media,

companies, and non-profit organizations to take advantage of the population and activeness for marketing, campaigns and advertisements. Personal accounts represent opinions of an individual, whereas institutional accounts represent collective opinions. It is crucial to distinguish both types of accounts to understand the driving force behind the accounts, which could mean different extent of concern and risk. A constantly negative organization could be a warning to the project of potential lawsuits.

It is difficult to classify an account based on its tweeting behavior or followers. A very active user could tweet as frequently, or sometimes more frequently, than an institutional account. A celebrity is likely to have more followers than a company. In order to best classify a user, we turn to their registered website and use the following criteria to distinguish these accounts.

- The account has to have a website.
- The account's website needs to end with ".org", ".gov", ".edu", ".mil", and ".int" (Postel, 1994).

Applying these criteria to the CAHSR, the distribution is listed in Table 5-11.

Table 5-11 User Distribution Based on Institution

Account Type	**Number of Users**	**Number of Tweets**	**Average Sentiment**
Personal	12790 (95.5%)	21072 (92%)	-0.12
Institutional	606 (4.5%)	1836 (8.0%)	0.13

It is clear and expected that personal accounts take up the majority of the population. Unlike the followers count distribution, institutional categorization does not provide a proportional distribution between the number of users and number of tweets. 4.5% of institutional accounts tweet about 8.0% of tweets, indicating that institutional accounts are more active than average personal accounts. Meanwhile, when all popularity tiers exhibit slightly negative sentiment, which aligns with the overall sentiment, institutional accounts goes the opposite and have an average sentiment of 0.13. They are hence more supportive to project, possibly due to the political stance of a governmental or organizational accounts.

5.5.4 User Location

The last characteristic of the profiling model is user location. Location analysis provides the ability to gain insight from the location (geographic) component of user profiles. Given the absence of some basic demographic attributes such as gender and age, user location enriches the user profile model by providing an extra demographic dimension. The geographical information gives the physical location of the user and helps project managers better segment them by latitude and longitude, state, country or urban and rural.

User location is an optional field of a user profile. (Pennacchiotti & Popescu, 2011) found that about 80% of the Twitter population enter some location in their profiles, and (Cheng, Caverlee, & Lee, 2010) estimated that only 26% report a specific location such as city. Conducting a similar search in the case study, we found that 72.8% of users have a non-empty location and they own 74.6% of the total tweets. The high volume of location availability attest the validity and importance to conduct user location analysis. However, there is a caveat that location information in Twitter is entered by the user and Twitter does not check the validity of the location. Users could enter everything, including fake locations and non-location statements, to be one's location profile.

Table 5-12 shows the distribution of users who do and do not have a location entered in their profile.

Table 5-12 User Distribution Based on Availability of Location Information

Location	**Number of Users**	**Number of Tweets**
Yes	9749 (72.8%)	10373 (74.6%)
No	3647 (27.2%)	3538 (25.4%)

Unrelated location information is one of the difficulties to further analyze user location. Even if the locations are legitimate, there is no standard format for location entries. For example, "California" and "CA, US" are two different locations, but refer to the same region and should be treated equally. Google Maps web service is used to geocode location information. User location is collected,

queried against Google Maps API and a structured location is returned and stored in the database. Primarily city, county, state and country are used in our analysis

Google Maps API is a powerful geocoding service, however, it is not able to detect non-locations. For meaningless locations, such as “Everywhere” or “Planet Earth”, Google still tries its best to find the most relevant location instead of marking them as invalid. It is a future research item to eliminate invalid locations to increase the accuracy of the geocoding process. There is also a rate limit on Google Maps APIs of 2,500 free requests and 100,000 paid requests per day (Google, n.d.). However, it is sufficient for infrastructure projects given the number of users involved in social media.

Applying geocoded locations to the CAHSR case study, all users are marked with their geocoded locations. Looking at US users, overall there are 1,498 positive users and 2,766 negative users, noting an acceptance rate of 35.1% using formula 5.2. This is slightly higher than the 28% rate calculated in section 3.7, but still well aligned. In the following sections, we conduct the location analysis at state and county level to uncover more detailed information.

- State Analysis

Figure 5-7 shows the state-wise distribution of users in US. Only users with location are included in the figure, hence the total number of users are less than previous analyses.

Being a California state project, it is natural for CAHSR to get most attention from California people. Overall, Californians contributed 8,795 tweets on this topic within the time frame, more than all other states combined. In terms of sentiment, California people contributed 857 positive tweets and 797 negative tweets, yielding an acceptance rate of 51.8%, much higher than the average.

Table 5-13 compares California users with the rest of the world.

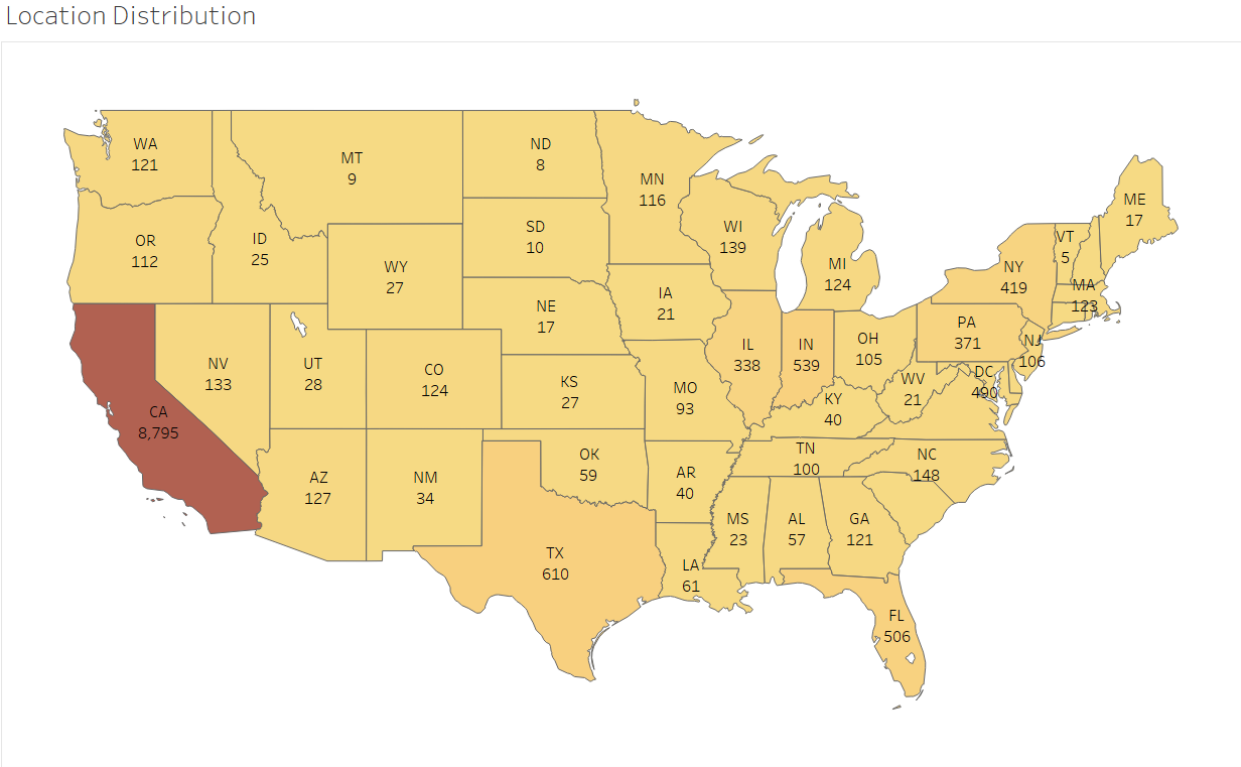

Figure 5-7 National Tweet Distribution

Table 5-13 User Distribution between California and All Other Regions

Location	Positive Users	Negative Users	Public Acceptance
California	857	797	51.8%
Rest of the World	1,180	2,992	28.3%

In fact, California ranks the second in acceptance rate among all US states. Table 5-14 lists the top 3 states of public acceptance.

Table 5-14 Top 3 States of Public Acceptance

Location	Positive Users	Negative Users	Public Acceptance
Wisconsin	21	18	53.8%
California	857	797	51.8%
District of Columbia	44	46	48.9%

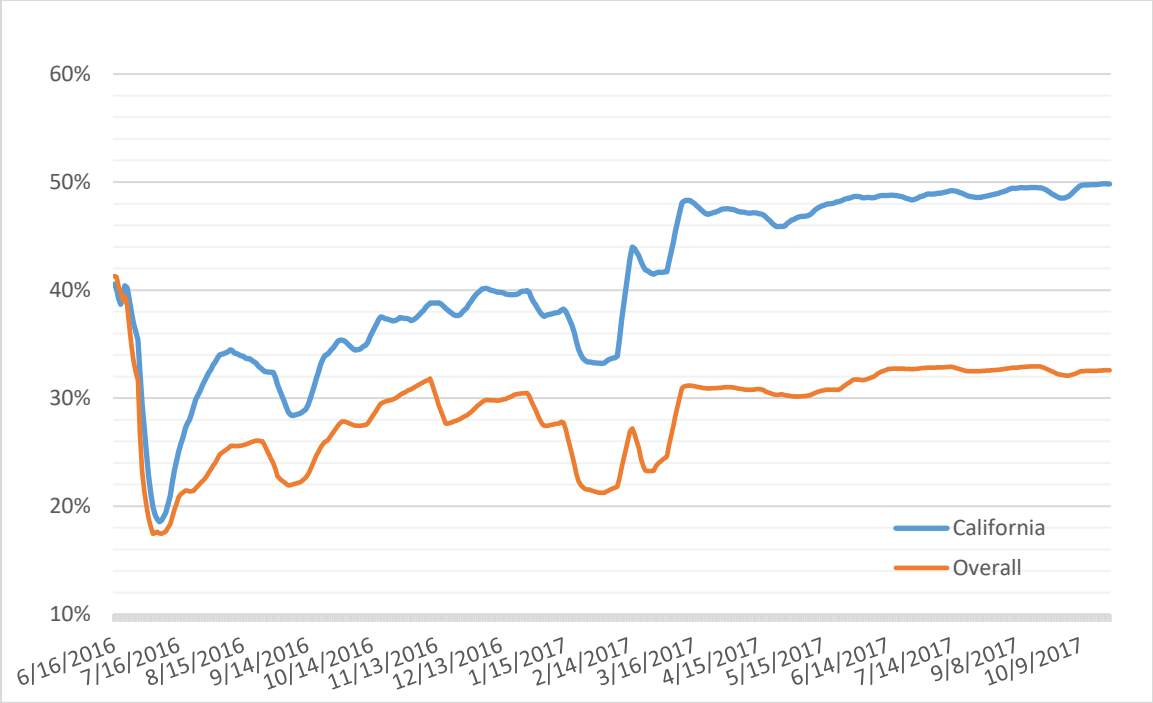

Figure 5-8 Public Acceptance Comparison between California and Overall

Figure 5-8 shows the comparison of public acceptance between Californians and the whole population. It is clear that despite the peaks and dips of public acceptance, Californians demonstrate solid support to the project, much higher than other states. The overall public acceptance is slightly above 50%, leading to a generally supportive stance. This is critical to project managers since the local community is directly impacted by the project, and therefore their support is more important than people from remote states. This should bring reassurance to project managers to certain extent despite the negative news around the project.

Being a highlighted infrastructure project in the nation, other states also pay attention to the project. Texas, Florida, and DC are the top states with interest in this project. Diving deep into the tweets of these states, there is no other driving factors to the volume of the tweet. The top states, however, align with the rank of the population of the states. Table 5-15 shows top 6 states regarding tweet contribution top 6 state with most population (top 5 excluding California, which makes the top 1 in both list coincidentally). District of Columbia is the outlier whose population ranking is 49 and

tweet ranking is 3. This is expected since large-scale infrastructure projects are closely related to politics and policy makers.

Table 5-15 Top States Based on Population

State	# of tweets	State	Population
California	8,795	California	38,332,521
Texas	610	Texas	26,448,193
Florida	493	New York	19,651,127
DC	490	Florida	19,552,860
New York	419	Illinois	12,882,135
Pennsylvania	371	Pennsylvania	12,773,801

- County Analysis

A more detailed county level analysis is also conducted. Due to the significant difference in volume, California is the only state of interest for county-wise research. Using geocoded location, county level user distribution is calculated with top 8 (total users exceeding 100) shown in Table 5-16.

Table 5-16 User Distribution Based on California County

County	Positive Users	Negative Users	Total Users	Public Acceptance
San Francisco County	167	99	521	63%
Los Angeles County	131	143	506	48%
Sacramento County	59	36	181	62%
San Diego County	26	40	131	39%
Fresno County	39	28	126	58%
Alameda County	40	23	124	63%
Santa Clara County	40	21	109	66%
Orange County	27	34	101	44%

The four metropolitans in California, San Francisco, Los Angeles, Sacramento and San Diego undoubtedly take the top 4 rank in terms of the number of users. Users from San Francisco and Los Angeles are significantly more than other counties, showing their interest and engagement with the

project. The distribution is clearly unrelated to population, where San Francisco is only rank 13 in California and Los Angeles has more than 10 times more population. This also does not seem to be related to political affiliations, where these metropolitans are all democratic.

Judging from the acceptance rate, people from Bay Area highly support the project, along with people from Sacramento, Alameda, Santa Clara and Fresno. On the other hand, people from Los Angeles, San Diego and Orange are not as optimistic as other top counties. If CAHSR is to gain more support from Californians, these three counties are the critical ones to fight for.

5.6 Opinion Leadership Characteristics

We conclude this chapter by combining the opinion leadership analysis with the user profiling model and summarize the characteristics of each opinion leadership types. Opinion leadership analysis focuses on the influence score which counts the number of tweets and retweets, but little is known about these users and how they their social behavior is regarding infrastructure projects. Table 5-17 shows the profiling of each opinion leadership types.

Table 5-17 User Profiles of Opinion Leadership Types

Opinion Leadership	Sentiment	Tweet	Median Popularity	Institution
Leader	-0.5	1.5	40,097 (XLarge)	1 / 16 (6.3%)
Follower	11.7	65.3	558 (Medium)	7 / 61 (11.2%)
Original Contributor	16.3	135.7	1,255 (Large)	5 / 25 (20%)
Overall	-0.2	1.8	801 (Medium)	606 (4.5%)

In the CAHSR case study, opinion leaders do not have strong sentiment towards the project. They have a slightly negative sentiment of -0.5. On average, opinion leaders post only 1.5 tweets, even less than the overall average. They are not very active in the world of CAHSR but they are influential enough to be overlooked. Correspondingly, the median number of followers of opinion leaders is 40,097, well qualified as the top tier XLarge for followers. Only 1 out of the 16 identified

opinion leaders are institution accounts. Even though the percentage is still higher than the overall percentage, the vast majority of opinion leaders are not institutional.

Opinion followers are much more positive on this project than opinion leaders. Their average sentiment is 11.7, demonstrating strong confidence in the project. On average, opinion followers post 65.3 tweets, a lot more than opinion leaders and overall average. They are active tweeters of CAHSR, however, their followers on median only falls in medium tier, limiting their influence. 7 out of 61 identified opinion followers are institution accounts, higher than opinion leaders.

Original contributors boost all metrics further up. Their sentiment is 16.3, highest among all opinion leadership types. They post 135.7 tweets per person, doubling the amount of opinion followers. Their median followers fall under large tier, hence they are more active with a larger audience than opinion followers. 5 out of 25 identified opinion followers are institution accounts, and the percentage is the highest of all types again.

In the order of opinion leader, opinion follower and original contributor, the sentiment goes more positive, tweet activeness is higher, and the percentage of organizational users get higher, too. While opinion leaders present a group of highly influential people for CAHSR, opinion followers and original contributors represent a group of active and engaged individuals and organizations who are concerned about the success of the project.

5.7 Conclusion

In this chapter, we discuss the people factor of public acceptance fluctuation, the social media users. After the analysis of what is public acceptance and what is the driving force of public acceptance, we try to answer the question of WHO are driving the changes of public acceptance. Opinion leadership analysis and user profiling analysis are the focuses of this chapter due to the information diffusion theory.

Again, the project evaluation framework is extended to user analysis. As a critical component of social media analysis, the user dimension is accommodated into the data crawler and data storage module. An opinion leadership model is established to define and evaluate opinion leaders under two measurements, by the number of retweets and by the number of normalized retweets based on the total tweets of a user. An opinion leader prediction model is also proposed to identify potential opinion leaders using a priori indicators. It is observed that normalized number of retweets is an effective indicator in predicting opinion leaders, and is therefore the recommended indicator for the opinion leader model. Besides opinion leader, two other leadership roles, the opinion follower and the original contributor, are also defined and discussed. There are a lot of overlapping users between opinion followers and original contributors, but they are almost exclusive to opinion leaders. Opinion followers and original contributors are also an effective way to find interest groups and organizations.

The user analysis also establishes a user profiling model to describe users using their social media demographic information. User sentiment, popularity, institution and location are discussed in detail. Opinion leaders tend to be neutral due to their small number of tweets, whereas opinion followers and original contributors are highly sentimental. It is also observed that people tend to stick with their first impression of the project and it is difficult to change their established opinion. User analysis shows that California people are more supportive to this project than the rest of the country, and within the California bay area shows more support than other counties.

Finally, the opinion leadership model is combined with the user profiling model to characterize different opinion leadership roles. Interestingly, we found that in the order of opinion leader, opinion follower and original contributor, the sentiment is more positive, the number of tweets is higher, and the percentage of institution is higher. User analysis provides a lot of detail to help understand the people involved in the CAHSR project and demonstrates the scalability of the project evaluation framework.

Chapter 6. Conclusion and Discussion

This dissertation proposes a comprehensive framework for project evaluation using social media and big data. The current methodology of public acceptance evaluation is costly and time-consuming, trigging this research to improve the process using advanced technologies. This dissertation starts with a conceptual framework for general project evaluation. The components are discussed and sample crawler workflow, database schema and data analysis are proposed. Chapter 3, 4 and 5 applies the framework on a real-world project, the California High Speed Rail project, to examine its feasibility in different perspectives. Chapter 3 first introduces the necessary techniques to conduct public acceptance analysis, including the data retrieval method and sentiment analysis algorithms. It then defines public acceptance under the context of social media, specifically the microblogging site Twitter, and proposes and compares the performance different measurements. Chapter 4 discusses the driving factor of public acceptance, social media events. Event is defined in social media and two dimensional event influence evaluation is proposed and examined. The original event influence quadrant is introduced to visualize and monitor real time event status change. Chapter 5 furthers the discussion on the driving individuals of public acceptance, the opinion leaders. The opinion leader is again defined under the context of social media along with two other opinion leadership types, opinion follower and original contributor, and a predictive method is developed to identify potential opinion leaders. Finally, a user profiling model is built to describe the demographic attributes of social media users, and is combined with opinion leadership analysis to depict opinion leaders.

This research reforms the existing methodology of public acceptance analysis by taking full advantage of social media activities. It testifies that social media is not only able to provide fast and cost-effective response on the progressive public acceptance of a project, it is also capable of generating in-depth insights and answer the questions of what and who is driving the changes of

public acceptance. The methodology can be used to provide real-time public acceptance monitoring and facilitate data-driven decision makings to improve public acceptance of infrastructure projects.

6.1.1 Summary of the Proposed Methodology and Results

Various methodologies are proposed and applied to the CAHSR case study in this research.

In the definition of public acceptance, three different models are proposed and tested, including public acceptance by tweet, by user and by influence, corresponding to the pluralistic model and the elite model in politics. Public acceptance of the CAHSR project is calculated using these models and an ANOVA test is conducted to compare the statistical significances. The result is in favor of the by user approach, in which one user has only one vote every day and previous day's vote will be used if there is no vote on the current day, due to the accuracy of the result and the ease of implementation.

In the event analysis, a two dimensional model is developed to describe event influence by both event magnitude and event duration. Event magnitude, similar to public acceptance, can be defined using the by tweet, by user and by influence models. Again, the by user model is favorable due to the capability of excluding false events created by bot-like user accounts. An event influence quadrant is then proposed to categorize and visualize social media events into four quadrants based on their influence. The event influence quadrant is a powerful tool for effective real-time event tracking and monitoring.

In user analysis, two models are developed to describe opinion leadership and user demographics. For opinion leadership, a measurement model is established to describe three leadership types, opinion leader, opinion follower and original contributor. Two indicators are used to define opinion leader, namely number of retweets and normalized number of retweets. An opinion leader predictive model is then developed to discern potential opinion leaders using a priori indicators, number of retweets or normalized number of retweets in a random 7-day period. As a result, the

number of normalized retweets is a favorable approach since it gives more consistent result in opinion leader identification and prediction models. The user profiling model describes user demographics of user sentiment, popularity, institution and location, and is used in combination of opinion leadership analysis to reveal the characteristics of different opinion leadership types.

6.1.2 Contribution to the Body of Knowledge and Practical Application

The topic of this research is inspired by the industrial practice of public acceptance analysis and the potential of social media to bring changes to this subject area. Although previous research has been conducted on Twitter analysis on various topics, it is the contribution of this research to bring this technology advancement to project management. Even though the data volume provided by Twitter is much less than other major events, it is still significantly higher than what a typical public opinion poll / survey can offer. This research opens the door of combining social media with project management to improve public awareness of infrastructure projects.

In the context of infrastructure projects, this research contributes to formalize a project evaluation framework using social media and big data. It generalizes the content to be fetched from social media and their relationship, along with the possible analysis to evaluate a project. Technically, this research finds out the most suitable tools and methods for project evaluation. It is observed that 1) infrastructure projects result in fewer tweets than other events; 2) key word search is the most efficient searching method; and 3) lexicon based approach provides better accuracy and F1 score for sentiment analysis. This knowledge could facilitate future research on infrastructure project evaluation combined with social media.

The majority of this dissertation is built around public acceptance from different perspectives. The key contribution of the study is the establishment of a grand model on public acceptance assessment which is able to answer the questions of what is the public acceptance, why does it change and who changes it. In detail, this research contributes to build a public acceptance model with definition

and measurement, a social media event model with definition and influence measurement, and a social media user model containing an opinion leadership model, an opinion leader prediction model, and a user demographic model. All these sub-models created around the core concept of public acceptance analysis are innovations of this research, which organically and logically completes the puzzle of public acceptance analysis using social media.

6.1.3 Limitations and Future Research

There are several directions in which this research can be improved and future research studies can be conducted.

Firstly, the data volume and the variety of projects analyzed is limited in this research. With no special agreements with Twitter, this research performs data retrieval as a normal market researcher, who is constraint to the volume and the time frame within which the data can be fetched. Although this experience is valuable for general purpose project evaluation since not everyone can reach an agreement with Twitter, a complete dataset with a full historical data can be helpful in picturing public acceptance within the project life cycle. A full set of data is also helpful in the predictive model of opinion leader, which in this research is limited to the amount of tweet allowed to be downloaded. Moreover, this research is based on a selected case study. More case studies on different infrastructure projects could refine and improve the project evaluation framework and test its scalability and versatility.

Secondly, the technologies used in this dissertation can be further enhanced and optimized. To be specific, the sentiment analysis methodology applied in this research can be improved for more advanced techniques with better accuracies. Although it is observed that lexicon based sentiment analysis performs better than machine learning based approach, it is entirely possible that properly trained with sufficient data, the machine learning approach could reach, or even outperform the performance of the lexicon one. This is a topic worth studying in future research. As of the lexicon

based approach, it is meaningful to compose a list of words specifically for infrastructure project sentiment analysis. While the work has been initiated in this research, the dictionary needs to be perfected by a thorough investigation of all infrastructure related tweets. With a better performance on the sentiment analysis algorithm, all public acceptance analysis on top of it can benefit from a stronger confidence. It is also mentioned in section 3.5 that there are two types of sentiments, sentiment of the tweet and sentiment towards the project. It is worth studying how to distinguish both sentiments for more accurate sentiment analysis on infrastructure projects.

Lastly, this dissertation treats the events and users individually, without considering the social network nature of these objects. In social media, users and tweets are interconnected, constructing a network of communication. Such a social network can be used to describe events and opinion diffusion mechanisms more clearly. This research does not include the connection into the models. It is, however, a good topic for future research to advance the event and user models.

Appendix A Twitter Crawler Pseudocode

```
Initialize consumer_key
Initialize consumer_secret
Initialize access_token
Initialize access_token_secret
Initialize Twitter client using all the keys

Open file
If file is open
   search_term = ["california high speed rail", "@CaHSRA",
   "#CaHSRA"]

   For each search term
      Call Twitter Search API with the keyword
      For each tweet returned
         Clean spaces and carriage returns
         Write create date, user name and tweet text in file
      End
   End
End

Close file
```

Appendix B Geocoding Using Google Map Pseudocode

```
Initialize hostname
Initialize username
Initialize password
Initialize database
Initialize Google Map key

Establish database connection using hostname, username, password
and database

Get locations to be geocoded from database

For each location to be geocoded
    Call Google Map API using the following url
    url="https://maps.googleapis.com/maps/api/geocode/json?addres
    s=%s&key=%s&language=en" % (address.encode('utf-8'), key)

    Get response JSON string
    Decode JSON string

    Load premise into database
    Load locality into database
    Load street_number into database
    Load route into database
    Load level_2 into database
    Load level_1 into database
    Load zipcode into database
    Load latitude into database
    Load longitude into database

    Commit database changes
End

Close database connection
```

Appendix C Sentiment Analysis Pseudocode

```
Initialize hostname
Initialize username
Initialize password
Initialize database

Establish database connection using hostname, username, password
and database

Get all tweets to be processed

Read negative word file
Construct negative word list
Read positive word file
Construct positive word list

Open file
For each tweet to be processed
    Split tweet into a bag of words

    For each word in the tweet
        If positive word list contains this word
        Put the word in tweet positive words list
        Positive score ++

        If negative word list contains this word
        Put the word in tweet negative words list
        Negative score ++
    End

    Write tweet, positive words list, negative words list,
    positive score, and negative score back to file
End
Close file
```

Appendix D Key SQL Statements

```
-- Tweet Sentiment

SELECT tweet                AS "Tweet",
SIGN(positive - negative)   AS "Sentiment"

FROM tweet t;


-- Event Influence

SELECT url_title                              AS "Page Title",
COUNT(1)                                      AS "by Tweet",
COUNT(DISTINCT t.user_name)                   AS "by Users",
SUM(1 + log(u.followers_count))               AS "by Influence",
DATEDIFF(MAX(create_date),MIN(create_date)) AS "Duration"

FROM tweet t

LEFT OUTER JOIN user u ON t.user_name = u.user_alias

WHERE url_title <> 'N/A'

GROUP BY 1 ORDER BY 2 DESC;


-- Find Opinion Leader

SELECT u.user_alias              AS "User",
MAX(u.description)               AS "Description",
COUNT(*)                         AS "Tweet",
SUM(SIGN(positive-negative))     AS "Sentiment"

FROM tweet t

LEFT OUTER JOIN users u
ON SUBSTRING(SUBSTRING(tweet, LOCATE('@', tweet), LENGTH(tweet)-
LOCATE('@', tweet)), 2, LOCATE(' ', SUBSTRING(tweet, LOCATE('@',
tweet), LENGTH(tweet)- LOCATE('@', tweet))) - 3) = u.user_alias

WHERE LOCATE('RT', tweet) = 1 and u.user_alias IS NOT NULL

GROUP BY 1 ORDER BY 3 DESC;
```

Appendix E Sentiment Analysis Baseline

Sequence	Tweet	Sentiment	Project Sentiment
28	@CaHSRA This is great explanation of high speed rail noise levels. Good reading for @TxAgainstHSR. @TXRailAdvocate https://t.co/axbTVOneuQ	positive	positive
46	"High-speed rail @CAHSRA lawsuit delays cost $63 million, 17 months https://t.co/fW4iLy2MeN via @SFGate"	negative	negative
58	"California's high-speed rail project wins lawsuit, adds $63M to project cost. https://t.co/DhIz7S7Eaj https://t.co/CfyjuFhRcH"	positive	positive
101	"CA's @CapAndTrade auctions, @CAHSRA bullet train funding are on life support, Ã¢â‚¬Å“whole system could failÃ¢â‚¬Â? https://t.co/LlZTIOvRHE @CAWater4All"	negative	negative
111	I'm applying to work for @CaHSRA in Fresno - wish me luck! #Iwillride	positive	positive
156	"@Talkmaster @KasimReed Whoever it was that didn't know, I'll bet they have a lot in common with fans of California high-speed rail."	negative	negative
168	"@realDonaldTrump @CaHSRA #Create jobs! Exciting high speed rail CA, FL, IL, IN, USA! https://t.co/F8D0jnVBBR https://t.co/AyrQ0DtT6l"	positive	positive
263	Thank you @CaHSRA & Lyles College alum Benjamin Camarena for encouraging young engineers to pursue their passions! https://t.co/kSVBslSaKg	positive	positive
335	I liked a @Youtube video https://t.co/ulQA2vKVp5 Train Wreck: California High Speed Rail Path Of Destruction	negative	negative
392	Construction work on high-speed rail in Hanford: HANFORD Ã¢â‚¬â€? Preliminary road work for the California high-speed railÃ¢â‚¬Â¦ https://t.co/RPI90L6mpV	neutral	neutral
464	Even CHP showed up for our HSR all staff meeting! Here's to a great year! #iwillride #buildhsr @CaHSRA https://t.co/H22IDjyywH	positive	positive
465	Free ice cream for lunch. Yes! Ã°Å‚Â?Â¦ #HappyFridayToMe @ California High-Speed Rail Authority https://t.co/CMUMCi4ulq	positive	positive
541	"@CAHSRA ""Stop promising Big Rock Candy Mountain and covering butt when result is a hill of beans"" https://t.co/c23n8mPjWk @CAWater4All"	negative	negative
554	@BorensteinDan @WaltersBee @JerryBrownGov @CaHSRA I'm encouraged that more and more people are seeing this train for what it is. Wasteful	negative	negative
584	"@EastBayOpinion @EastBayTimes @JerryBrownGov @CaHSRA just buy Elon musk's hyperloop. Cheaper, faster, better. Stop doing drugs moonbeam"	negative	negative
627	HawaiÃ¢â‚¬Ëœi. Hello? Knock knock! Hello? https://t.co/ykxMCU6YFM	negative	negative
761	"Yes, but we need good ""within metro"" rail. @vpostrel to CA: Pull plug on high-speed rail fiasco https://t.co/VPLIaZ0myP @BV"	neutral	negative
768	"The Political Class Knew California High-Speed Rail Was B.S., and Supported it Anyway #libertarian https://t.co/OcHKJhFrYv"	negative	negative
806	"Ahhh... California, where bad ideas are performance art paid for by taxpayers. https://t.co/RuzkxtbC9I"	negative	negative

859	This sounds awfully familiar..... https://t.co/WGo5NmOSuW	negative	negative
870	"More evidence that government ""investments"" are often foolish and wasteful. https://t.co/VG1JBQXggo"	negative	negative
895	"Yet again, fate sneaks up on high-speed rail in CA. From China, the ""developed world"" looks so backward. https://t.co/M77mPKVp2Y"	negative	negative
896	"#California ""a classic example of how concentrated benefits and diffused costs shape public policy"" #PublicTransport https://t.co/4xbK4uhvAW"	positive	positive
907	Even California may not be able to lie enough to keep high-speed rail on lifeÃ‚Â support https://t.co/khHRnJRfWP https://t.co/jX2uRJu9Xr	negative	negative
917	California's crazy train! https://t.co/Tunt2vcw8m	negative	negative
1028	"#Alaska The Political Class Knew California High-Speed Rail Was B.S.,... https://t.co/ZzTZDdRm3O https://t.co/SuU89pUDDw"	negative	negative
1129	#highspeed rail looks to me like a final attempt to #Bankrupt #California https://t.co/jxXQVDpiVd	negative	negative
1188	Liberalism: The ideology that lies to people about what is good for them. #tcot https://t.co/6TMSP3kK5W	negative	neutral
1196	See the progress of the seven #CAHSRA construction sites in the June Construction Update at https://t.co/smjXaLef6y https://t.co/ZquH83OhOW	positive	positive
1301	"Building water storage in California is vital. It is more important than funding the ""crazy"" High Speed Rail... https://t.co/PdRl15VqEC"	negative	negative
1303	California Gov Covered Up High Speed Rail Subsidies Warning @fpmag https://t.co/otuvEF1b16	negative	negative
1363	California's high-speed trainwreck...Thanks @GovWalker for saving WI from this disaster. https://t.co/np3mdswXIG https://t.co/m3SnW5tN3a	negative	negative
1388	High speed rail in CA: It was all BS https://t.co/7QcPUren1Y	negative	negative
1455	LibOC: California Needs High Speed Rail: Vice President Joe Biden and Secretary of Transportation Ray LaHood ... https://t.co/EbEDdBo8lg	positive	positive
1458	#HappeningNow! Mechanical engineering professor Dr. The Nguyen talking dynamic modeling & design! #CAHSRA https://t.co/EQsuVxIGjh	positive	neutral
1479	#StatusCheck: We've got an update on the progress of California's #Highspeedrail: https://t.co/ZlDKjspByX	positive	neutral
1485	California is 1 step closer to shutting down the high speed rail scam. Hopefully they kill it. https://t.co/eNUbREqtmS	negative	negative
1500	Paying a fortune for a train to nowhere. https://t.co/gPwCpjdINc	negative	negative
1526	California should pull the plug on high-speed rail fiasco by @vpostrel https://t.co/Q6wRH2vLyL via @BV	negative	negative
1543	@WSJecon part of the problem is mismanagement of existing funds. Look a the waste of billions on high speed rail in California.	negative	negative
1607	I drive California roads every day..... the pot holes and cracks are getting bigger and worse.... but we're getting high speed rail?	negative	positive
1628	@CaHSRA @calexpo @CAStateFair Not that great for all of the money being spent!	negative	negative

1705	"@Forbes: @CAHSRA's ""high speed rail idiocy, such as that boondoggle in California"" https://t.co/mMykyQCeUS @CAWater4All"	negative	negative
1712	@railLAorg @CaHSRA @Amtrak Why does the US still have mostly 1960s-era rail technology in 2016? When do we get to enter the modern world?	negative	negative
1759	#California Assemblyman #KevinMullin & Sen. #JimBeall support #Caltrain's illegal hijacking of #HighSpeedRail funds https://t.co/YNkBPpCxCk	negative	negative
1822	Kevin works past 8 pm 4 @CaHSRA as our @railLAorg posse heads back 2 Union Station in LA. https://t.co/3CipbBoyLD	positive	neutral
1862	railLA President @jeremytweet explains why its important to show LA the @CaHSRA construction in the Central Valley https://t.co/vgdK3jWtVo	positive	neutral
1890	"@RepJohnMica: Ã¢â‚¬Å“@CAHSRA mired in delays, doubled its budget and lowered its speed projectionsÃ¢â‚¬Â? https://t.co/ThqNCeAdq4 @CAWater4All"	negative	negative
1910	Gas leak caused by pre-construction work for @CaHSRA at McKinley.	negative	negative
1930	A nice piece of fiction. https://t.co/gd634UOEZQ	negative	negative
2064	@TransportiCA @ca_trans_agency @CaHSRA @cahsr thanks for sharing our work!	positive	positive
2066	"Instead of hyperloop dreams, I wish @MIT_alumni would focus on building real @CaHSRA infrastructure, or just streets that don't kill people."	negative	positive
2145	"The @Hyperloop is fucking stupid and classist and regionalist. The @CaHSRA will help connect the entirety of CA, not just SF to LA. Ughhh"	negative	positive
2147	"@CAHSRA's empty trains to move few people, gobble up #AB32 #CapAndTrade $, fails to solve transit needs https://t.co/QxZhzxBmCO @CAWater4All"	negative	negative
2150	@bradpomerance @CaHSRA @CalChannel about wasting money on it which could be used for decades promised Desert Wind?	negative	negative
2186	Are @CA_Bldg_Trades #Labor deals helping derail @CAHSRA? https://t.co/IVsYIuSwFd @CAWater4All	negative	negative
2211	Great to align with our partners at @SBAgov Los Angeles and @CaHSRA to support small business! @CAGoBiz https://t.co/f5jVdF5S4y	positive	positive
2235	"California High-Speed Rail July Construction Update: FRESNO, Calif. Ã¢â‚¬â€œ Hard work is paying off at th... https://t.co/RkBlETuS5r #railtube"	positive	positive
2296	"Register now! Learn about best practices, challenges related to California HSR @CaHSRA https://t.co/eHjiv0uyvV https://t.co/t9pVFO7AgL"	positive	positive
2314	California High Speed Rail - A Sustainable Transportation Solution https://t.co/dmPYfMbQS7 https://t.co/9PqmdJ12bL	positive	positive
2352	@JohnChiangCA Do you support California High Speed Rail?	neutral	neutral
2373	"@jake_bradford_1 I live in California. Booming economy, great diversity, budget surplus, high speed rail coming, new buildings everywhere.."	positive	positive
2387	"@LATimes: @CAHSRA ""running 15% over budget and has fallen about six months behind schedule"" https://t.co/cKUdAeFCPr @CAWater4All"	negative	negative
2388	"@CAHSRA """"investment' in @CalTrain jumps from $600 mil to $713 mil plus $84 mil more 'for other work'"" https://t.co/cKUdAeFCPr @CAWater4All"	negative	negative
2407	Who takes much of the credit for California High-Speed Rail initiation & progress? The unions that control the jobs. https://t.co/w1TVshydDv	neutral	neutral

2466	"The California High-Speed Rail Authority has begun work on the ""Fresno Trench"". The trench will go under 180, the... https://t.co/mhTy3SGgl8"	neutral	neutral
2515	Check out what @CaHSRA are doing in Fresno right now! @ca_trans_agency https://t.co/cuIzMpgPKA https://t.co/xMa0ejrgmr	positive	positive
2537	High-speed rail delays in Fresno area #California #hsr https://t.co/65CUoZavUL	negative	negative
2538	"@urbanlifesigns @CaHSRA @HSRail @CA4HSR @SPUR_Urbanist @burritojustice meanwhile, where is Desert Wind, promised for decades?"	negative	negative
2541	@VITCBOY @CaHSRA @HSRail @CA4HSR @SPUR_Urbanist @burritojustice the Desert Wind! XPressWest apparently died in June. https://t.co/FLGcKBNrFU	negative	negative
2561	Carbon futures drift well below #CapAndTrade auction minimum as #CARB seeks buyers - https://t.co/Bqsr2vcWUk @CAWater4All @CAHSRA	negative	negative
2583	@CAHSRA oversight & accountability bill receives unanimous vote of approval https://t.co/z9H7OA5jDS @CAWater4All	neutral	positive
2585	Learn why plans for the @CaHSRA is causing some concern in #AntelopeValley: https://t.co/poCybmcB7U	negative	negative
2586	Feline conservation center worried CA high-speed rail plans may threaten future of rare cats https://t.co/D4FnQFJ7D0 https://t.co/kGqO1rEnHe	negative	negative
2605	"@Caltrain @CaHSRA WTF! So to account for temp phase of 2 level platforms, we'll have trains with extra useless doors running next 40 years?"	negative	negative
2609	@Caltrain @CaHSRA why will Caltrain keep low level platforms then? Why not make all PF high level like HSR ones?	negative	negative
2613	@Caltrain @CaHSRA Sorry but that seems silly. Wasting space with 2 extra doors in ALL trains over decades < cost of raising 30 pf. Really?	negative	negative
2640	"A Fast Train Is Coming, Like It or Not.: The high-speed rail project in California continues to slog ahead an... https://t.co/Y63FWmL0nb"	positive	positive
2670	@sspencerthomas @smartunionworks @TrackSAFE @CaHSRA @RAILMag @UnionPacific you are all most welcome	positive	positive
2694	California's Cap-And-Trade Program Is Sick And Will Take High-Speed Rail Down With It https://t.co/CX6VXmBGw8 #Finance #Investments #ROIMeÃ¢â‚¬Â¦	negative	negative
2765	TradingStreet: California's Cap-And-Trade Program Is Sick And Will Take High-Speed Rail Down With It: Califor... https://t.co/c2mel9iV56	negative	negative
2769	HoerterFX's Notes: California's Cap-And-Trade Program Is Sick And Will Take High-Speed Rail Down With It: Cal... https://t.co/ZfMQw71SMQ	negative	negative
2825	"@RepJeffDenham's Transportation Committee to meet Monday in #SanFrancisco, evaluate progress of @CAHSRA https://t.co/ZfRtSmD1Bv @CAWater4All"	neutral	neutral
2836	High-speed rail is really happening in CA! Check out our video of @CaHSRA's progress! https://t.co/zNn9Vx1DQf https://t.co/ZQGOZ7lGUA	positive	positive
2885	california could really use a high speed rail train bc the Caltrain sucks Ass.	negative	negative
2888	California High-Speed Rail will be successful. It just requires a bit of government nudgery to get everyone on board. Start with the taxes.	positive	positive

2951	Congressional hearing today: @Caltrain and @CaHSRA admit 220 mph not achievable in urban areas. #CA #HighSpeedRail #Legal #Transit @GovTop	negative	negative
3032	"Canada Rx High-speed rail critics question the first route segment, which will end in an almond orchard: The ... https://t.co/YRStaOqHhP"	negative	negative
3102	"High-speed rail critics: 1st segment will end in almond orchard https://t.co/1nXUOsOK2I I think pecan, I think pecan https://t.co/x7nl1immi5"	negative	negative
3111	"@RepJeffDenham: ""All the money on @CAHSRA will be spent and you will be stuck somewhere in a field"" https://t.co/A6lWSooCEa @CAWater4All"	negative	negative
3113	Ya Think? @CAHSRA's Dan Richard: Ã¢â‚¬Å“It seems odd to be stopping in the middle of an almond orchard.Ã¢â‚¬Â? https://t.co/gQMfFuM1vJ @CAWater4All	negative	negative
3115	CA high speed rail critics question first route segment ending in almond orchard... https://t.co/qBg04mSW8t	negative	negative
3145	"California and the United States is well over do in having a High Speed Rail, like 20 years over do...""CaliforniaÃ¢â‚¬â„¢sÃ¢â‚¬Â¦https://t.co/XVNT6CNxnT"	negative	negative
3168	a #protest for those who eschew the choo https://t.co/BKmOfNPlMi @CaHSRA @CCHSRA	negative	negative
3178	"As citizens plan to protest CaliforniaÃ¢â‚¬â„¢s $64 million high-speed rail in San Fernando Wednesday night, members of a Ã¢â‚¬Â¦ https://t.co/i7Y6Ikd6Rb"	negative	negative
3180	California High Speed Rail protest at Lake View Terrace Library . https://t.co/577Z8i2tJS	negative	negative
3289	".@lvtia Yes, #Prop53 will require California High-Speed Rail to seek voter approval to sell $2+ billion in bonds to be paid back by revenue."	neutral	neutral
3303	"@Faqsicle: @CAHSRA #BulletTrain burns $3.6 million/day, sucks #Stimulus, #CapAndTrade $$ by the bushel https://t.co/JC1Owrc1GT @CAWater4All"	negative	negative
3324	"That was because the California High Speed Rail Authority (CHSRA), the autonomous state agency in charge of... https://t.co/ie9m3bbbUW"	neutral	neutral
3348	Texas' early success in building high speed rail could benefit California's planned #hsr line: https://t.co/WztNtfBMWB	positive	positive
3395	#HighSpeedRail: What's Good for Texas Is Good for California https://t.co/liLDjjrGe5	positive	positive
3424	Exciting developments for high speed rail in California & building of a 2.5 billion Union Station project. Thanks for sharing #APTAannual16	positive	positive
3426	@CaHSRA can't wait to hear what new BS you are selling this time.	negative	negative
3468	So @JerryBrownGov buried unfavorable reports on both @WaterFix #TwinTunnels and @CAHSRA bullet train? https://t.co/u5PxHzns05 @CAWater4All	negative	negative
3477	Fresno City Council today 9/15 votes on resolution to SUPPORT California High-Speed Rail. https://t.co/7ZaganhBoq Why? (People are nervous?)	neutral	positive
3500	Convenient sources of information about #CAHSR: https://t.co/OQKd2NUCIa #CALeg #Business #Economy #Env #Farm #Home #Legal #Transit @GovTop	positive	neutral
3551	Great Urban Growth Seminar today with Tony Mendoza of the @CaHSRA @METRANS_CENTER https://t.co/8DCLEOsIXD	positive	positive

3577	@CaHSRA Regional Dir. Gary Griggs happy to be back at #FresnoState! He discussed the $64.2 billion project & its 80Ã¢â‚¬Â¦ https://t.co/zjDjYv8mVz	positive	positive
3579	"Interesting... @CaHSRA Twitter page mentioned #SanFernando (#SunValley) event this evening, but not #Selma event https://t.co/In4iL9HCWY #CA"	neutral	neutral
3615	Will High-Speed Rail Development In Texas Benefit California? https://t.co/UKmYE75Slg by James Ayre #cleantech #energy	neutral	neutral
3660	Looks like #Texas will beat #California's @CAHSRA to finish line: Delays & cost overruns put CA behind https://t.co/KetamVrFDI @CAWater4All	negative	negative
3685	"@eparillon I still hold out faith that it will happen SOMETIME, but @CaHSRA is right to plan a ""temporary"" terminus at Fourth"	positive	positive
3696	Jerry Brown vetoes bipartisan California High Speed Rail Authority transparency bill. So much for trust but verify.. https://t.co/ItgaujYqo2	neutral	neutral
3697	"@quinnnorton On the other hand, itÃ¢â‚¬â„¢ll keep Musk distracted from undermining California high speed rail for a while."	negative	negative
3746	#Big_Government Jerry Brown Vetoes Bill to Improve High-Speed Rail Oversight https://t.co/9CTMNAwhst https://t.co/Q1IIdQKCgf	neutral	neutral
3760	.@JerryBrownGov Whatcha hiding Moonbeam? https://t.co/uxdjujpy1O Progressive Turd Vetoes Bill to Improve High-Speed Rail Oversight	negative	negative
3837	@CAHSRA: From unanimous vote to unreal @JerryBrownGov veto https://t.co/4gPBVBUoaf @CAWater4All	negative	negative
3841	"@JerryBrownGov vetoes @CAHSRA oversight bill that passed @CAAssembly 116-0, ignores #CALAO warning https://t.co/4gPBVBUoaf @CAWater4All"	negative	negative
3857	Jerry Brown's Train Wreck - CaliforniaÃ¢â‚¬â„¢s Governor doesnÃ¢â‚¬â„¢t want anyone looking under the high-speed rail track. https://t.co/TOhWyDGVyc	negative	negative
3859	Jerry Brown's Train Wreck. CaliforniaÃ¢â‚¬â„¢s Governor doesnÃ¢â‚¬â„¢t want anyone looking under the high-speed rail track. https://t.co/31TmlGYyZo	negative	negative
3885	Jerry BrownÃ¢â‚¬â„¢s Train Wreck - The California Governor doesnÃ¢â‚¬â„¢t want anyone looking under the high-speed rail track. https://t.co/vtvwQD8Z00	negative	negative
3891	"@HillaryClinton & @RealDonaldTrump ""fed $$ goes to questionable projects like CAÃ¢â‚¬â„¢s troubled @CAHSRA"" https://t.co/dh1UkZHaPr @CAWater4All"	negative	negative
3954	#GMFUrban Fellow supports #CAHSR: Bigger and Bolder: Preparing California Cities for High-Speed Rail https://t.co/374l9XXVSR @gmfus	positive	positive
3968	"@GavinNewsom @CAHSRA now: ""Not an opponent"" https://t.co/SWxy3arTSe; Then: ""More pressing problems"" https://t.co/gev2gza2ve @CAWater4All"	negative	negative
3971	"California High Speed Rail? This is 2016, can we just work on autonomous cars and mag-lev trains instead of a... https://t.co/sa9tyYvPBE"	negative	negative
3996	@LADailyNews: @RealDonaldTrump win could stop @CAHSRA #BulletTrain in tracks https://t.co/eJDsZ34kS6 @CAWater4All	neutral	negative
4012	@CaHSRA delays action on plans for #Fresno train station https://t.co/GKduDgYIsE #highspeedtrain	negative	negative
4045	MTI hosted tour of Diridon Station yesterday for Getting it Right on Governance and on the Station-Neighborhood IntÃ¢â‚¬Â¦ https://t.co/gbcUCWRtPa	positive	positive

4048	California Proposition 53 Bonds for big projects (Like high speed rail and Delta) Would need peopleÃ¢â‚¬â„¢s vote https://t.co/3p3wmfqiWC	neutral	neutral
4083	On the Nov. 8 ballot is a state proposition that could very well derail CaliforniaÃ¢â‚¬â„¢s high-speed rail project. https://t.co/nM7eT7SboI	negative	negative
4091	"@andybosselman @SFTRU @cahsr @CaHSRA @GavinNewsom @JohnChiangCA Forget HSR, it is time to focus on the Hyperloop. CA leads, not follows."	positive	positive
4092	Learn about the impact of new high speed rail systems in California from experts on issues from design to governancÃ¢â‚¬Â¦ https://t.co/wWMgwPf3ku	neutral	neutral
4096	FACT- @CaHSRA failed to respond to public questions about its $250M #CapandTrade shortfall during TuesdayÃ¢â‚¬â„¢s Board mtg. Was that transparent?	negative	negative
4124	"If there are no consequences for delay, delay, delay - guess what happens? This is regular occurrence for us withÃ¢â‚¬Â¦ https://t.co/e4wF1VYSbT"	negative	negative
4221	Just in! Trouble for high speed rail... https://t.co/uEl7gU1NeO	negative	negative
4245	"Jerry Brown, allies spend millions to kill measure that could doom high speed rail, Delta tunnel https://t.co/9xozVUEh5o #capolitics"	negative	negative
4319	"@GillMMcN @McMikeskywalker I've been called that, too. For not supporting our boondoggle California high speed rail."	negative	negative
4326	@TaupeAvenger Disagree. @CaHSRA will can only run a couple of trains during peak hours in their peak market (SF to LA). Each train counts.	negative	negative
4334	@TaupeAvenger @CALHSR yÃ¢â‚¬â„¢all talk like @CaHSRA will actually be competing and competitive with LA-SF air routes. It wonÃ¢â‚¬â„¢t.	negative	negative
4380	@CaHSRA Your heart is the best prototype for kindness. #TheNiceBot	positive	positive
4409	California @CaHSRA reduces size of HSR stations (RAIL: won't lose capacity if bi-level coaches used) | @latimes Ã¢â‚¬â€œ https://t.co/SHSXb9SkE1	negative	negative
4484	Slower speeds? Lower capacity? I'm sure this was entirely unforeseeable. https://t.co/lBeeHJVPer	negative	negative
4504	@nbroverman @CaHSRA @LA_mag [3/3] sound like huge de-scope; any reduction in budget?	negative	negative
4506	"Good news. @CaHSRA has cut planned capacity on HSR system by half, which will reduce costs and tunneling. https://t.co/95SiY7OGV1"	positive	positive
4511	Jerry Brown and company seem pretty worried that Prop. 53 (anti-Delta tunnels and high-speed rail) will pass: https://t.co/VUh2QtEkTa	negative	negative
4518	"@CAHSRA, @JerryBrownGov losing steam: cuts to peak speed, ridership; ""I know I can't,"" says train https://t.co/4rwxJwD72x @CAWater4All"	negative	negative
4519	"@CAHSRA, @JerryBrownGov: Like a second marriage, #BulletTrain is the triumph of hope over experience"" https://t.co/4rwxJwD72x @CAWater4All"	negative	negative
4541	"@CAHSRA ""plowing up some of best #California #farmland to build the first link of high-speed rail"" https://t.co/Xyt6z2We0s @CAWater4All"	negative	negative
4553	@CaHSRA Find that extra $42B so I can actually get Sf to LA 1.5 hrs slower and $10 more expensive than jet blue?	negative	negative

4579	Did you miss CV on Monday? Here is our interview with Dan Richard on high-speed rail in California. https://t.co/yye9zj6qSr	neutral	neutral
4598	".@CALawMama Some people will get free California High-Speed Rail rides, but the legislature will designate them. No need to jump turnstiles."	positive	neutral
4618	"Good News for California High-Speed Rail. Half the Capacity = Savings!!! What About Revenue? Don't Worry, Be Happy. https://t.co/K8dKaoO79V"	negative	negative
4644	@CAHSRA's pivot from #BulletTrain building to #RailModernization is illegal says #Prop1A's @QuentinKopp https://t.co/R4QJTDpOdr @CAWater4All	negative	negative
4697	"@MotherJones: @CAHSRA ""gross financial negligence in original plan, or else they're blowing smoke now"" https://t.co/jpAtHVfFF0 @CAWater4All"	negative	negative
4708	"Wish we had invested in this instead of disastrous money-pit @CaHSRA. Woulda, coulda, shoulda I guess... this is whÃ¢â‚¬Â¦ https://t.co/DWDiB6Z3Rs"	negative	negative
4744	"My brother, @jeremytweet, is on the front page of @CaHSRA talking about his work with @railLAorg! https://t.co/SEwXR2HlXL"	positive	positive
4761	@CaHSRA thanks for the great story about the work we do at @railLAorg!	positive	positive
4809	@4c4d @tdfischer_ @SFBART same thing hold for California high speed rail. Many lawsuits. Still building because it's right thing to build.	positive	positive
4815	"@CaHSRA This may be good news for CaHSR, Former mayor of LA, Antonio Villaraigosa is running for Governor of CA in 2018. What do you think?"	positive	positive
4848	This is an excellent question. Anyone know the answer? @CaHSRA https://t.co/7DCJdkgcch	positive	neutral
5227	Trump needs to look into the California High-Speed Rail and all the corruption it involved he needs to look into Diane Feinstein n her hubby	negative	negative
5245	@realDonaldTrump California is in urgent need of water infrastructure and a high speed rail train. Work with Nancy Pelosi & Gov Brown.	neutral	positive
5254	#CAHSRA is hosting a Free Small & Disadvantaged Biz Workshop in Fresno on 12/2. See the flier for details or visitÃ¢â‚¬Â¦ https://t.co/t5N2OcWfJN	neutral	positive
5295	"@LATimes, @JerryBrownGov, @CAHSRA fool no one: once railcar tooling is in China, work will stay there https://t.co/9ZkkfTCTOW @CAWater4All"	negative	negative
5326	"@SacBee @WaltersBee: @CAHSRA #Prop1A ""bait-and-switch ploy so voters finance local transit they otherwise would not support"" @CAWater4All"	neutral	neutral
5349	Connecting #California: High-Speed Rail to Enhance Statewide #Transportation Network https://t.co/txSepC3zyAÃ¢â‚¬Â¦ https://t.co/shuDJxHbGW	positive	positive
5354	"Could a Trump presidency hurt or help CaliforniaÃ¢â‚¬â„¢s High Speed Rail project? Ã¢â‚¬â€? Nov 17, 2016 by Take Two Show https://t.co/CkNZGwlWQF #trendÃ¢â‚¬Â¦"	neutral	neutral
5365	"If lack of ""Progressive"" support in Congress = dooming federal funding for California High-Speed Rail, so be it. https://t.co/dTZmazTh2s"	negative	negative
5411	We didn't realize Engineering & Construction were added to the curricula of young Master Edna's speedy come-up .Ã¢â‚¬Â¦ https://t.co/iqL68NZ248	positive	neutral
5472	Tapping @terplan @SPUR_Urbanist brain trust for Beyond the Track 2.0 @CaHSRA Land Use Committee https://t.co/eacQVQb2oF	neutral	neutral

5487	The only thing the California High-Speed Rail Authority Transit-Land Use Committee should be focusing on right now is naming the stations.	negative	negative
5490	Lou Correa came and went from the California High-Speed Rail Authority board so fast he didnÃ¢â‚¬â„¢t even get his bio up.Ã¢â‚¬Â¦ https://t.co/6xLkyqr4Yu	negative	negative
5501	"Credit to @CCHSRA (https://t.co/RG5zB7gHoW): ALWAYS vigilant, NEVER trusting California High-Speed Rail promises. https://t.co/1xrCAJmq7o"	negative	negative
5504	"Oh happy day! More cap and trade money from ""polluters"" for pollution-free California High-Speed Rail, built by hand to run with the wind."	positive	positive
5513	"@Wired: @CAHSRA #BulletTrain ""#SF-#LA plan hamstrung by bureaucracy, crippling land use issues"" https://t.co/uHmoEC2hPI @CAWater4All"	negative	negative
5518	Is Fresno Councilmember Brandau correct about the reduction in traffic @CaHSRA ? https://t.co/6AWe33YEic	neutral	neutral
5533	From all of us at #CAHSRA... Happy Thanksgiving! https://t.co/2srYvP8zWK	positive	positive
5542	"When California officials say high-speed rail will cost $64 billion between SF and LA to operate in 2029, is that an example of ""fake news?"""	negative	negative
5554	Will We Allow Technology to Rip America to Shreds? https://t.co/fvso3r96rh Not about California High-Speed Rail. ThÃ¢â‚¬Â¦ https://t.co/iWKxCpwROF	negative	negative
5560	Pls defund Fed $ to City of LA. CA is wealthy enuf to build $60B HSR we don't need Fed $ @realDonaldTrump @MayorOfLA @LAPDChiefBeck @CaHSRA	negative	negative
5594	"This week: Metro board, South LA plans, @LAGreatStreets Lankershim, faith diversity, @CaHSRA meetings, Cudahy + https://t.co/FS53YWECcY"	neutral	neutral
5650	Interactive map on the progress and plans for high-speed rail in California. https://t.co/e68bOmSvAc https://t.co/zf7rSz8k55	positive	neutral
5707	Fed Class 1 Railroad eminent domain authority trumps @CAHSRA taking of Union Pacific property https://t.co/yS2bapaBXj @CAWater4All	positive	positive
5779	@cspanwj @TimRyan No lying Californian... the high speed rail is critical to solving the transportation problems in California.	positive	positive
5840	186 MPH Commute: Here's what #rail #travel could look like once the @CaHSRA is complete (video taken in #Shanghai) https://t.co/XKjYeG0zKi	neutral	positive
5877	@CaHSRA says yes to spending bond funds. They also want $19.5 bil from highway funds 4 bond debt. My bill says no.Ã¢â‚¬Â¦ https://t.co/KOkfKbmAAQ	negative	negative
6052	Still in awe that California might get high speed rail before the northeast does.	positive	positive
6067	"California bullet train still barreling ahead, madly: The California High-Speed Rail AuthorityÃ¢â‚¬â„¢s decision this weekÃ¢â‚¬Â¦ https://t.co/w649QNsB6s"	negative	negative
6100	"Good news for @CaHSRA, good news for @cityoffresno! https://t.co/jgjCSZXHzE"	positive	positive
6104	@burberryant Supposed to but through the central valley & no of intersections will slow avg spd way down. Cost huge https://t.co/1S622y3c9U	negative	negative
6165	Exciting progress in the Central Valley on the @CaHSRA #transit #train #construction https://t.co/BNgu01ozWq	positive	positive
6182	"Every winter & holiday season with all the road traffic, airport congestion & flight disruptions it seems that we could use @CaHSRA today."	negative	positive

6227	"New post: ""Lame-Duck Obama Admin Rejects CA High-Speed Rail $15 Billion Loan"" https://t.co/4vdcKPpwfq"	negative	negative
6231	Kill this project ASAP-> Lame-Duck Obama Admin Rejects CA High-Speed Rail $15 Billion Loan https://t.co/KB11omhanh via @BreitbartNews	negative	positive
6241	"Check this article: https://t.co/zserAOV8MJ @CaHSR_Scam @CaHSRA Train to nowhere, voters voted approved $9B for $33B project. Flawed plan."	negative	negative
6379	LameDuck Obama Admin Rejects CA High-Speed Rail $15 Billion Loan.Collectivist Cal an embarassment https://t.co/Fq7CJUKFye via @BreitbartNews	negative	positive
6414	24 companies said they wouldnÃ¢â‚¬â„¢t put any money into @CAHSRA until CA proves #HSR would be profitable https://t.co/IoLPHZZZfB @CAWater4All	negative	negative
6558	@NancyPelosi's pet project dies. Hooray! https://t.co/sdBJVVyPSa	positive	negative
6561	What a mess at @flyLAXairport! - takes 45 minutes to taxi from gate to runway - can't wait for @CaHSRA to finish High Speed Rail #iwillride	negative	positive
6576	Wow! Amazing! Lame-Duck Obama Admin Rejects CA High-Speed Rail $15 Billion Loan https://t.co/zREhAnol6c via @BreitbartNews	negative	negative
6578	Even the Obama administration is refusing to bail out California's crazy & costly high speed rail social experiment https://t.co/UPvh0KErIH	negative	negative
6593	"Pres Obama refuses Jerry Brown 15 Bil Loan for Train to nowhere ,... https://t.co/roHBWjiQQK"	negative	negative
6601	#makeamericagreatagain No Money 4 CA Nothing - Lame-Duck Obama Admin Rejects CA High-Speed Rail $15 Billion Loan https://t.co/ZPagP7cFEg	negative	positive
6602	#Trump2016 No Money 4 CA Nothing - Lame-Duck Obama Admin Rejects CA High-Speed Rail $15 Billion Loan https://t.co/ZPagP7cFEg	negative	positive
6616	@kellymcorrigan @acocarpio @BurbankLeader I am really excited about the California High Speed Rail!	positive	positive
6640	@wadhwa aside from argument re:economics... do you also anticipate that 1000s of personal autos will be better for enviro than @CaHSRA?	negative	positive
6657	"@AllAboardOhio how's that 3-C ""high speed"" going? Oh that's right. Also see California passenger rail boondoggle. #Traincult"	negative	negative
6662	@steverichards83 Ã?Âºee‘â,¬ smiling.. Life is bÃ?ÂµÃ?Â°utiful in spite Ã?Â¾f evÃ?Âµrything! @RuckerKendall @CaHSRA @crashers23 @MichaelKCBento	positive	negative
6668	"With @SFBART failing falling apart, with endless traffic congestion in LA, why is @CaHSRA even still in existence? https://t.co/Y9DrmV0b57"	negative	negative
6694	"@KamalaHarris like alternative transportation options, i.e. @CaHSRA @metrolosangeles @SFBART and @sandiegometro !!!!!! #greenliving"	neutral	positive
6715	Maybe we shouldn't be blowing so much money on the unnecessary @CaHSRA project https://t.co/GC5UhOeqpu	negative	negative
6746	Reports from @USTreasury & @USDOT hail @CaHSRAÃ¢â‚¬â„¢s transformative economic impact for cities like @CityofFresno. MoreÃ¢â‚¬Â¦ https://t.co/sfPmkt3RFp	positive	negative
6760	No what ifs - just facts. #CAHSRA is spurring economic recovery & job creation https://t.co/w01blE7IrfÃ¢â‚¬Â¦ https://t.co/rAnVqs9ngt	positive	positive

6795	"@CAHSRA's ""#BulletTrain hurtling toward a multibillion-dollar cost overrun"" says Feds https://t.co/3aE0jndMaX @CAWater4All"	negative	negative
6797	"@CAHSRA's ""#BulletTrain mismanagement has gone too far to ignore any longer"" https://t.co/3aE0jndMaX @CAWater4All"	negative	negative
6861	@elizabethforma why R u not concerned about your liberal friends conflict of interest? https://t.co/cbsZhWe16s	negative	negative
6872	"The requirement should be that those corporations that win ""tax-payer"" funded infrastructure contracts be fully... https://t.co/hBLBGjA4mb"	negative	negative
6907	@latimes #JerryBrown is crazy the #bullet train is just a massive Union payout. https://t.co/rgT7OtoHaI	negative	negative
7066	It's inconceivable that the Project Labor Agreement on California High-Speed Rail could be 1 cause of inflated costÃ¢â‚¬Â¦ https://t.co/NQ6RbQU5na	negative	negative
7097	Worst of the Legacy Issues: Foolhardy California High-Speed Rail Promises to Voters in Proposition 1A (2008)Ã¢â‚¬Â¦ https://t.co/b3YYHHOElD	negative	negative
7133	.@JoeGruters @Reaganista California perspective: so odd to see Florida Democrats lambaste @FLGovScott for not buyinÃ¢â‚¬Â¦ https://t.co/y4lCziMrxV	negative	negative
7250	Can anyone say special official privilege? https://t.co/dti0GdK9sE	negative	negative
7510	@SenFeinstein Shameful and despicable! https://t.co/fbKSshiq2T	negative	negative
8329	"Another example of OUR tax dollars (not) at work! Thanx, BO CA High-Speed Rail: Over Budget, Behind Schedule - https://t.co/M8CT4gSmuq"	negative	negative
8337	@JayWeber3 Tell California not to worry the high speed rail will be a high priced trolley in 6 years...it will only go in circles...	negative	negative
8462	I spend half the year in Auburn CA I'm painfully aware. But I have a bunch of plastic bags I'll sell you for $0.05!Ã¢â‚¬Â¦ https://t.co/LOkebOOV2B	negative	negative
8519	CA Corruption: https://t.co/1wiyYXEBlJ	negative	negative
8750	Now wonder Hillary won California so big. They seem to aspire to ever increasing levels of corruption there. https://t.co/TbUPke9iZG	negative	negative
8859	"California ""is""a train wreck! Sure hope it's not a Omen for their high speed rail,that Feinstein and her family r going to get richer! From"	negative	negative
8874	"@Stuflash99: ""@CAHSRA is trying to muddy the waters so you canÃ¢â‚¬â„¢t see what is going on"" https://t.co/q6tQTwmPas @CAWater4All @CAHSR_Scam"	negative	negative
8924	Trump Trian makes California High Speed Rail look like a turtle!! https://t.co/9gEw6jWXHx	negative	negative
8930	"The more we learn about high-speed rail, the less we like it. https://t.co/4IQJzFyDld Its gettin kinda not good."	negative	negative
8948	"Me. Donald J. Trump, please kill the California high speed rail project with executive order, Please!!!!!"	negative	negative
8956	Expect an immediate aggressive lobbying offensive done by multi-national corporations & unions to keep California High-Speed Rail in the $$$	negative	negative
9050	@realDonaldTrump Want to win over California? Kill high speed rail for us!!!	neutral	negative
9054	@GovPdfs @CaHSRA I am very ashamed of those not attending the inauguration. Please set selfish pride aside and join the U.S. Community - GO!	negative	neutral

9076	.@CCHSRA publicly declared to @CaHSRA Board that #CAHSR Project would cost $500B. Board did not refute the projection. #CA #Legal #Transit	negative	negative
9082	.@CaHSRA Board failed to explain @ mtg what private parties were going to invest in #CAHSR. They were silent on matter. #CA #Legal #Transit	negative	negative
9086	.@CaHSRA Board also failed to discuss failing #CapandTrade funding @ #CAHSR meeting. #CA #Business #Env #Farm #Home #Legal #Transit @GovTop	negative	negative
9096	"@LCJandA @CaHSRA Thanks. I was listening in on the meeting; I heard the speakers address board, but appreciate the offer!"	positive	positive
9097	".@CaHSRA Board just declared at their meeting today: ""Public doesn't understand risk."" #CA #Business #HighSpeedRail #Legal #Transit @GovTop"	negative	negative
9209	dcexaminer NEW MichaelBarone: Infrastructure lessons from the California high-speed rail fiasco https://t.co/2MsbvAphHG Ã¢â‚¬Â¦	negative	negative
9233	What a joke. Reality meets #progressives #California High Speed Rail Faces 50 Percent Cost Overruns https://t.co/LbZzNJ528t	negative	negative
9314	". @realDonaldTrump promises, among other things, 'new railways' cc: @CaHSRA"	positive	positive
9322	"@USDOTFRA ""administered what may turn out to be a fatal blow to @CAHSRA"" #BulletTrain https://t.co/xAs078DSO3 @CAWater4All"	negative	negative
9370	Feinstein nepotism. https://t.co/oZHlMhN7fl	negative	negative
9413	Another interesting place along the California high speed rail route in the Central Valley #cahsra https://t.co/04IuQ8zzPJ	positive	positive
9433	"Any Ã¢â‚¬Å“museums, libraries and galleries around the worldÃ¢â‚¬Â? want my 2016 California High-Speed Rail Groundbreaking protÃ¢â‚¬Â¦ https://t.co/D5emLGPC5C"	negative	negative
9441	That isnÃ¢â‚¬â„¢t fair @CCHSRA. Your $500 billion claim for cost of California High-Speed Rail includes bond interest. https://t.co/BC6gmLYcRE	negative	negative
9462	"Looks like @CaHSRA is back to budget-busting elevated station at Diridon ""Intergalactic"". https://t.co/OLx5IlC95W https://t.co/B8x84PSQG1"	negative	negative
9475	"@jonahsachs @SenFeinstein You should drain the swamp while you're there, https://t.co/ZS6ZZN5oPj https://t.co/HxWSBtJXaJ"	negative	negative
9540	"@POTUS response to @JerryBrownGov insults: No @WaterFix, @CAHSRA funding https://t.co/WIyPcJKyWk @CAWaterAlliance @McClatchy"	negative	negative
9547	@elonmusk Hello Elon. Do you believe HyperLoop could be a better substitute to California High-Speed Rail System?	neutral	neutral
9554	"@DJdm67 i mean feel free to criticize california's massive expenditure on high speed rail, but that's not the same convo as sanctuary cities"	negative	negative
9569	"California's High-Speed Rail: Slow, Expensive, and Bound for Cancellation - National Review https://t.co/aymFMJljUQ"	negative	negative
9616	"@CALHSR @CaHSRA @Caltrain @cahsr @TransForm_Alert @SPUR_Urbanist which was already cut off by highways 280 & 87, at-grade would worsen.."	negative	negative
9624	"By 2029, a bullet train should run from #SF to #LA at speeds capable of over 200 miles per hour: https://t.co/3CEuqE235D"	positive	positive

9632	"@KQED @KQEDForum, @CaHSRA isn't going to be any better...We need some real solutions in CA, not an expensive mega-project no one will use."	negative	negative
9682	"@SPUR_Urbanist, @CaHSRA & @ca_trans_agency - TY for joining us for a great discussion @ucmerced today! #BuildtheFuture"	positive	positive
9704	"@chuckdevore: #CA High-Speed Rail: Slow, Expensive, and Bound for Cancellation: https://t.co/59SSvjvMO1 via @NRO #bullettrain #HSR"	negative	negative
9707	California High Speed Rail: Brilliant idea? https://t.co/U44Upn3FBr via @Youtube	negative	negative
9727	"@BRC4252 @darksecretplace @sacbee_news Same ppl paying now: TAXPAYERS! BHO said no, President Trump will say HECK no https://t.co/MJ5HnhpwrN"	negative	negative
9739	"Put a bullet in the bullet train https://t.co/xAs078DSO3 @CAHSRA's #HSR rail fiasco"" @CAWater4All"	negative	negative
9761	"You guys, if we had high speed rail in California like we do here, I'd meet you at Javier's once a month. Right nowÃ¢â‚¬Â¦ https://t.co/yz7SPbfCBv"	positive	positive
9798	"@KellyannePolls A free boarder wall, have the railroads pay for it, and install a high speed rail from California to Texas, thanks."	positive	neutral
9812	New California High-Speed Rail track will end up as faster Amtrak in San Joaquin Valley & Fresno Area Rapid Transit. https://t.co/rlt371jvED	negative	negative
9893	@SenFeinstein 's husband wins near-billion dollar California 'high speed rail' contract #MuslimBan Is #NotaMuslimBan https://t.co/OL4gdsZo00	neutral	negative
9982	"#TCOT CaliforniaÃ¢â‚¬â„¢s High-Speed Rail: Slow, Expensive, and Bound for Cancellation https://t.co/LaSYc6vyiB"	negative	negative
10001	Laying track Bakersfield-Palmdale poised to be one of most difficult & costly sections of @CAHSRA https://t.co/6RWRrrrg2D @CAWater4All	negative	negative
10055	Critical for @Caltrain and @CaHSRA. No delays! Call tomorrow! https://t.co/SXlYp2ud5v	neutral	positive
10078	"California GOP, possibly confusing Caltrain with Calif. high-speed rail, are putting $650M for upgrades at risk https://t.co/xP60BUXXRV"	negative	negative
10085	@MatierAndRoss: @USCongress may kill electrification of #CalTrain over audit of @CAHSRA https://t.co/P9B0ThiiSC @CAWater4All #SiliconValley	negative	negative
10492	"Retweeted Jerry Brown (@JerryBrownGov): .@realDonaldTrump, CaliforniaÃ¢â‚¬â„¢s ready. #CAHSRA Ã°Å¸Å¡â€ž https://t.co/NJS9aUhYKF https://t.co/MpFbP1YBjx"	positive	neutral
10547	".@MZanona @thehill @politicsreid the CA reps lied in their letter to #DOT - the grant is for @Caltrain, not @CaHSRA https://t.co/0okGdj3Jya"	negative	negative
10621	"Subtext: If Trump means what he says about investing in infrastructure, he'll press Congress to fund California higÃ¢â‚¬Â¦ https://t.co/aSwVvhG975"	neutral	neutral
10666	@DeletionMapping @JerryBrownGov we need real technology not vaporware. #CaHSRA gives us modern transit in next decade. Hyperloop is fantasy	positive	positive
10734	@ByRosenberg That's a steal compared to the High Speed Rail getting built in California lol	negative	negative
10804	.@NancyPelosi made a plug for CA's high-speed rail project (@GOPLeader McCarthy wants to scrap) during morning coffee today w/PM Abe #CAHSRA	negative	neutral

10818	@WashTimes Nov 2016 California high speed rail project update https://t.co/UJCxcGegrC awarded $3B from Federal Railroad Administration	positive	positive
10889	Trump laments lack of 'fast trains' in US during meeting with airline execs https://t.co/VNctw10SHY California isn't building a fast train.	negative	negative
10910	@CBSNews @MajorCBS The wall is money well spent. The high speed rail in california 3x more expensive https://t.co/dDpxeInrai	neutral	negative
10975	@GovPressOffice @CaHSRA @JerryBrownGov Have you requested Fed Funds? Have requested you protester-in-chief @KamalaHarris help secure funds?	negative	negative
10985	"@CaHSRA sorry but it needs to be derailed...now, if you're willing to put on hold and start exploring hyperloop you got my support"	negative	negative
10991	"@sullivanradio NOW would be a GREAT time for California to divert the 'High Speed' rail project money, to damn repair/ construction."	neutral	negative
11085	Right now it's really hard to believe that California could pull of a high speed rail project when they can't even maintain a dam.	negative	negative
11143	Construction proceeds on #California high-speed rail despite uncertain future. https://t.co/rHiGHdcmIT https://t.co/dtTHp6e5jD	negative	negative
11151	"https://t.co/DBGAQMiKjJ ""California Spent On High-Speed Rail And Illegal Immigrants, But Ignored Oroville Dam!"""	negative	negative
11261	U.S. high-speed rail projects might stand a chance if Breitbart is excited about it https://t.co/cE8x1bTBqX	positive	positive
11497	"California's ""High-Speed"" Rail boondoggle: Getting worse all the time | Washington Examiner https://t.co/dkvaPAMMbw"	negative	negative
11728	High-speed rail CEO says slower environmental reviews won't delay your first train trip https://t.co/NukCEvKelX via @svbizjournal	neutral	positive
11735	"Breaking news, California has decided to house illegals on high speed rail lines"	negative	negative
11762	"@POTUS Do not give California ONE RED CENT. They can use the High-Speed rail money. They can use ""welfare to illegals"" money. Calif. is rich"	negative	negative
11930	"@GovPressOffice @JerryBrownGov @fema WATCH CAREFULLY California what these slick magicians do with the $, like they did w/ high speed rail $"	negative	negative
12067	"@milguy23 @Thomas1774Paine @FoxNews The money sent to California was much better spent on High Speed Rail ""Oh that didn't work out so well"""	negative	negative
12116	CA chose Brown over sane alternative. Brown chose high-speed rail over Oreville Dam. No federal bailout! https://t.co/ahRC43SE29	negative	negative
12161	"Whether for @CAHSRA bullet train or for @WaterFix, #CADelegation's eyes are always on the wrong objective for helpiÃ¢â‚¬Â¦ https://t.co/2MRtWYutxr"	negative	negative
12185	"Great panel on transportation issues in Calif- thanks to @CalBCC & @MalcolmXdough , @CaHSRA Jeff Morales &Ã¢â‚¬Â¦ https://t.co/Nw7jXpt3t1"	positive	positive
12226	"@thecliffbar @kimmaicutler @Caltrain @CaHSRA @SVLeadershipGrp i think they know, but good to give face saving out"	negative	negative
12231	Feds delay decision on California's $650 million high-speed rail grant https://t.co/qQxYdORAUu https://t.co/ZBAvhUkBjF	negative	neutral
12279	@natehanco @tjon_t The plan for @CaHSRA has always envisioned a PPP to operate it. Maybe a PPP would work for electric express Caltrain?	neutral	neutral

12334	"Dam it Jerry Brown! Stop the stupid bullshit train, ... Oh I GET IT. You sold out California. You Fake! https://t.co/4PVzLAaPpO"	negative	negative
12347	I just initiated discussion on highlighting to public the formidable challenges of getting California High-Speed Rail through Pacheco Pass.	positive	neutral
12366	California's infrastructure is being demolished by the rain we've been getting. We're starting to sink! JB and that stupid high speed rail	negative	negative
12409	Wow Ã°Å¸ËœÂ® https://t.co/EZ5tJAuoK3	positive	positive
12427	I wonder how many people she's putting out of their homes for this. https://t.co/24eT6Gtqgf	negative	negative
12834	"Shutting down the California high speed rail project means also eliminating 20-50 thousand new jobs. But...like screw California, right?"	negative	negative
12843	California Republicans asks Trump admin to block new grant moneyÃ¢â‚¬â€?arguing it would likely benefit the high-speed raiÃ¢â‚¬Â¦ https://t.co/VArmrTaeJq	positive	positive
13175	"Doesn't ""progressive"" vision for San Joaquin Valley = building California High-Speed Rail AND restoring Tulare Lake? https://t.co/pLBX4VM1Z1"	positive	positive
13248	It's not about fast travel between cities. It's about changing how we live. Starting with YOU. California High-Speed Rail. #YouWillRide	positive	positive
13249	"We need roads, bridges, dams and other vital, and crumbling infrastructure fixed, not this stupid train that no... https://t.co/KeldxBc0qR"	negative	negative
13430	@ericgarland well we can start adding these kinds of anti-infrastructure.. https://t.co/JnxZiqyiyy	negative	negative
13553	".@GOP administration says infrastructure, clean energy, and public transportation is for losers. Take a limo! https://t.co/3omoqACnbe"	negative	negative
13560	@FoxNews and I was concerned about the California high speed rail project.	negative	negative
13585	@ElaineChao @realDonaldTrump is this story true? Promises of improved infrastructure & jobs & you nixed this? https://t.co/cSuKGd05mH	negative	negative
13638	@realDonaldTrump you're picking a fight you can't win here moron. #ImpeachTrumpTreason #LockHimUp https://t.co/4kIdBf9Lk5	negative	negative
13777	@SenFeinstein What is up with this...a bit of corruption? https://t.co/F64zTQtYQe	negative	negative
13794	I thought the bare minimum for fascism was making the trains run ontime --> Trump halts CA plans for high-speed rail https://t.co/q4espPAEyH	negative	negative
13836	"California ""High Speed"" Rail has $ from Feds held up. Question is, which insane agency even OK'd in first place?: https://t.co/ICxSPUfyGz"	negative	negative
13947	Backwards Trump administration halts California's plans for high-speed rail & infrastructure improvements - https://t.co/SxSBsBkQbj	negative	negative
14059	"Trump Administration Just Killed CaliforniaÃ¢â‚¬â„¢s High Speed Rail Project, ButÃ‚Â Why? https://t.co/qF0Ou7jP5a https://t.co/eTTkFd2ocw"	negative	negative
14089	Aww hell naw. Vote them ALL out https://t.co/ctzA3rJYc4	negative	negative
14383	.@AssemblyGOP @CAGOP Gas tax should be used to fund clean air protections & @CaHSRA to mitigate hidden costs of car culture. #TrainTwitter	positive	negative
14384	".@AssemblyGOP @CAGOP It's an embarrassment not an achievement. Cars are the problem - @CaHSRA is the solution, not more roads. #TrainTwitter"	negative	positive

14400	Did @CaHSRA just say the line can stop at San Jose? Don't just shrug and say @Caltrain is on its own - be an ally fÃ¢â‚¬Â¦ https://t.co/2yuLinxVQQ	positive	positive
14403	@SFTRU @CaHSRA @Caltrain But would it help Caltrain to have an ally the Republicans hate?	negative	negative
14425	Shame on California Republicans for sacrificing Caltrain electrification in their effort to stop high speed rail at any cost. Any.	negative	negative
14441	#CAHSRA is honoring the women engineers who are making high-speed rail a reality. #Eweek2017 #GirlDay2017Ã¢â‚¬Â¦ https://t.co/r9wVW8FTWM	positive	positive
14476	"In California's Commuter Rail Drama, Nobody's a Winner: Except, ironically, proponents of high-speed rail, theÃ¢â‚¬Â¦ https://t.co/NNKfmsrERL"	negative	negative
14550	@alevin HSR shows video including construction in district of CA Reps trying to kill the project @Caltrain @CaHSRA	negative	negative
14563	@alevin Bouchard as well as perturbation analysis showing how the scenarios would work with schedule disturbance @Caltrain @CaHSRA	neutral	neutral
14564	"@alevin CM Tanaka, new from Palo Alto, asks about a long tunnel. Tripousis says not feasible @Caltrain @CaHSRA"	negative	negative
14586	"Glad to see @CityLab making this important point: foolishness from @GOPLeader hurts @Caltrain commuters, not @CaHSRA https://t.co/rmfx56vU0h"	neutral	negative
14597	@alevin we've had massive support for this project - not only in Bay Area but where jobs created around country @Caltrain @CaHSRA	positive	positive
14610	Thank You @CaHSRA for giving @CordobaCorp's Melissa de la PeÃƒÂ±a a shout out! Let's celebrate #EngineersWeek2017 togeÃ¢â‚¬Â¦ https://t.co/4dYAWUptEK	positive	positive
14630	"High speed rail sucked up California's infrastructure spending for roads, levees, etc. ""The people chose high speed rail."" -Gov. Jerry Brown"	negative	negative
14664	Disappointed she signed letter with CA GOP halting improvements to California's High Speed Rail efforts and CalTraiÃ¢â‚¬Â¦ https://t.co/D6KiQR52Q2	negative	negative
14672	"I blame the Democrats running California for reckless spending on High Speed Rail, infrastructure fail https://t.co/2gEHIHL6hL #OrovilleDam"	negative	negative
14713	@CALHSR @CaHSRA Tiny number of freight trains on Caltrain line could be equipped with ERTMS equipment right?	negative	negative
14792	.@realDonaldTrump administration halts California's plans for high-speed rail and infrastructure improvements - https://t.co/89vpjcToi0	negative	negative
14880	@ScoJo760 @kgbveteran @saksivas_ Trump DOT denied $647 million grant to California High Speed Rail. https://t.co/9ksLFg1Mb9	negative	negative
14918	"California High Speed Rail, work in Silicon Valley but live on a cattle ranch in Merced, Madera and Fresno Counties. Only a matter of time."	neutral	positive
15032	TRUMP LIES 2 AMER WRKRS Trump Admin Halts California's plans 4 high-speed rail & Infrastructure Improvements https://t.co/uW524SIiqz	negative	negative
15035	"High Speed Rail services to flourish with focus on station area context, last-mile mobility and overall experience! https://t.co/WxnclsfgAR"	positive	positive
15041	.@GOP attack @GoCaltrain in an effort to kill California High Speed Rail. Our guest opinion from @alon_levyÃ¢â‚¬Â¦ https://t.co/fYXpVIfBrA	negative	negative

15048	Webcast starts soon! High Desert Corridor HSR Investment Study results: https://t.co/MBw6oUKS4B @CaHSRA @XpressWestÃ¢â‚¬Â¦ https://t.co/qadOplJ2zf	neutral	neutral
15084	"@KenCalvert @indivisible42 @GoRail @CaHSRA Mr. Calvert, it seems you ignored I.E. freeway gridlock issues and Nat'l Rail Day in DC. SAD!"	negative	negative
15089	"Bullet train suffers two big setbacks that could be fatal: Late Wednesday, the California High-Speed Rail AuthorityÃ¢â‚¬Â¦ https://t.co/WOQfwNXMeY"	negative	negative

Reference

AECOM. (2007). User guidebook on implementing public-private partnerships for transportation infrastructure projects in the United States. *Arlington, Virginia: Federal Highway Administration*.

Astaneh-Asl, A. (2008). Progressive collapse of steel truss bridges, the case of I-35W collapse. In *Proceedings of 7th International Conference on Steel Bridges, Guimar{\u{a}}es, Portugal*.

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on* (Vol. 1, pp. 492–499).

Aylien Ltd. (n.d.). Aylien Homepage. Retrieved from https://aylien.com/

Barnaghi, P., Ghaffari, P., & Breslin, J. G. (2016). Opinion Mining and Sentiment Polarity on Twitter and Correlation Between Events and Sentiment. In *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)* (pp. 52–57).

Bartolomé, L. J., De Wet, C., Mander, H., & Nagraj, V. K. (2000). Displacement, resettlement, rehabilitation, reparation and development. *WCD Thematic Review*, *1*.

Bennett, W. L., & Manheim, J. B. (2006). The one-step flow of communication. *The ANNALS of the American Academy of Political and Social Science*, *608*(1), 213–232.

Bird, Steven, Loper, E., & Klein, E. (2009). *Natural Language Processing with Python*.

Calais Guerra, P. H., Veloso, A., Meira Jr, W., & Almeida, V. (2011). From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 150–158).

California High-Speed Rail Authority. (2016a). *California High-Speed Rail Big Picture*.

California High-Speed Rail Authority. (2016b). *Connecting and Transforming California - 2016 Business Plan*.

California High-Speed Rail Authority. (2017). About California High-Speed Rail Authority.

Carnevale, A. P., & Smith, N. (2017). Trillion Dollar Infrastructure Proposals Could Create Millions of Jobs. *Center on Education and the Workforce, Georgetown University, Avail-Able at Https://cew. Georgetown. Edu/wp-Content/uploads/trillion-Dollar-Infrastructure. Pdf*.

CCHSRA. (n.d.). Our Story.

CCHSRA. (2016). COURT HEARS ARGUMENTS ON CALIFORNIA HIGH-SPEED RAIL AUTHORITY PROP 1A COMPLIANCE.

Cernea, M. M. (1988). *Involuntary resettlement in development projects: Policy guidelines in World Bank-financed projects* (Vol. 80). World Bank Publications.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring user influence in twitter: The million follower fallacy. *Icwsm*, *10*(10–17), 30.

Chadwick, A. (2011). Britain's first live televised party leaders' debate: From the news cycle to the political information cycle. *Parliamentary Affairs*, *64*(1), 24–44.

Checkoway, B. (1981). The politics of public hearings. *The Journal of Applied Behavioral Science*, *17*(4), 566–582.

Checkoway, B., & Van Til, J. (1978). What do we know about citizen participation? A selective review of research. *Citizen Participation in America*, 25–42.

Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 759–768).

Childers, T. L. (1986). Assessment of the psychometric properties of an opinion leadership scale. *Journal of Marketing Research*, 184–188.

Cho, J., & Garcia-Molina, H. (1999). *The evolution of the web and implications for an incremental crawler*.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, *9*(6), 811–824.

Chua, D. K. H., Kog, Y.-C., & Loh, P. K. (1999). Critical success factors for different project objectives. *Journal of Construction Engineering and Management*, *125*(3), 142–150.

Cohen, J. J., Reichl, J., & Schmidthaler, M. (2014). Re-focussing research efforts on the public acceptance of energy infrastructure: A critical review. *Energy*, *76*, 4–9.

Cole, R. L., & Caputo, D. A. (1984). The public hearing as an effective citizen participation mechanism: a case study of the general revenue sharing program. *American Political Science Review*, *78*(2), 404–416.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning* (pp. 160–167).

CPHA. (2015). *The Red Line: what now?*

Devine-Wright, P. (2007). Reconsidering public attitudes and public acceptance of renewable energy technologies: a critical review. *Beyond Nimbyism: A Multidisciplinary Investigation of Public Engagement with Renewable Energy Technologies*, *15*.

Diekmann, J. E., & Girard, M. J. (1995). Are contract disputes predictable? *Journal of Construction Engineering and Management*, *121*(4), 355–363.

Dimitropoulos, A., & Kontoleon, A. (2009). Assessing the determinants of local acceptability of wind-farm investment: A choice experiment in the Greek Aegean Islands. *Energy Policy*, *37*(5), 1842–1854.

Dincer, I. (2000). Renewable energy and sustainable development: a crucial review. *Renewable and Sustainable Energy Reviews*, *4*(2), 157–175.

Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 231–240).

Domingos, P., & Richardson, M. (2001). Mining the network value of customers. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '01*, 57–66. https://doi.org/10.1145/502512.502525

Dutzik, T., Bradford, A., Weissman, G., & others. (2017). Big Projects. Bigger Price Tags. Limited Benefits.

Engel, E., Fischer, R., & Galetovic, A. (2011). Public-private partnerships to revamp us infrastructure. *The Hamilton Project, Discussion Paper*, *2*, 8.

Erikson, R. S., & Tedin, K. L. (2015). *American public opinion: Its origins, content and impact*. Routledge.

Evans, B., Parks, J., & Theobald, K. (2011). Urban wind power and the private sector: community benefits, social acceptance and public engagement. *Journal of Environmental Planning and Management*, *54*(2), 227–244.

Federal Highway Administration. (1992). Examining Congestion Pricing Implementation Issues. In *Summary of Proceedings: Congestion Pricing Symposium. Searching for Solutions. A Policy Discussion Series.*

Forbes, L. P. (2013). Does social media influence consumer buying behavior? An investigation of recommendations and purchases. *Journal of Business & Economics Research (Online)*, *11*(2), 107.

Frilet, M. (1997). Some Universal Issues in BOT Projects for Public Infrastructures. *International Construction Law Review*, *14*, 499–512.

Galper, D. Z., Goyal, S., & Gilbertson, K. (2013). Method for automatic url shortening. Google Patents.

Gleick, P. H. (2012). China dams. In *The world's water* (pp. 127–142). Springer.

Gnip. (2017). Choosing a Historical API. Retrieved from http://support.gnip.com/articles/choosing-historical-api.html

Google. (n.d.). Geocoding API. Retrieved from https://developers.google.com/maps/documentation/geocoding/intro

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.

Hao, S. (2009). I-35W bridge collapse. *Journal of Bridge Engineering*, *15*(5), 608–614.

Hardcastle, C., Edwards, P. J., Akintoye, A., & Li, B. (2005). Critical success factors for PPP/PFI projects in the UK construction industry: A factor analysis approach. *Construction Management and Economics*, *23*(5), 459–471.

Hayes, D. J. (2014). Addressing the environmental impacts of large infrastructure projects: making "mitigation" matter. *Environmental Law Reporter*, *44*, 10016–10021.

Heberlein, T. A. (1976). Some observations on alternative mechanisms for public involvement: The hearing, public opinion poll, the workshop and the quasi-experiment. *Nat. Resources J.*, *16*, 197.

Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). Word cloud explorer: Text analytics based on word clouds. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on* (pp. 1833–1842).

Herrmann, A. W. (2013). Asce 2013 report card for america's infrastructure. In *IABSE symposium report* (Vol. 99, pp. 9–10).

Homeland Security. (2010). Aging Infrastructure: Issues, Research, and Technology. *Buildings and Infrastructure Protection Series*, *1*(December). Retrieved from https://www.dhs.gov/xlibrary/assets/st-aging-infrastructure-issues-research-technology.pdf

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168–177).

IFC. (1998). *Doing Better Business through Effective Public Consultation and Disclosure*.

James, J. (2014a). Data Never Sleeps 2.0.

James, J. (2014b). Data Never Sleeps 2.0. Retrieved from https://www.domo.com/blog/data-never-sleeps-2-0/

Jiang, H., Lin, P., & Qiang, M. (2015). Public-opinion sentiment analysis for large hydro projects. *Journal of Construction Engineering and Management*, *142*(2), 5015013.

Jiang, H., Qiang, M., & Lin, P. (2016). Assessment of online public opinions on large infrastructure projects: A case study of the Three Gorges Project in China. *Environmental Impact Assessment Review*, *61*, 38–51.

Jobert, A., Laborgne, P., & Mimler, S. (2007). Local acceptance of wind energy: Factors of success identified in French and German case studies. *Energy Policy*, *35*(5), 2751–2760.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68.

Katz, E. (1957). The two-step flow of communication: An up-to-date report on an hypothesis. *Public Opinion Quarterly*, *21*(1), 61–78.

Katz, E., & Lazarsfeld, P. F. (1966). *Personal Influence, The part played by people in the flow of mass communications*. Transaction Publishers.

Keller, E., & Berry, J. (2003). The influential. *J. Berry.--New York: The Free Press*.

Kellerman, B. (2007). What every leader needs to know about followers. *Harvard Business Review*, *85*(12), 84.

Kemp, R. (1985). Planning, public hearings, and the politics of discourse. *Critical Theory and Public Life*, 177–201.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, *54*(3), 241–251.

Koppel, M., & Schler, J. (2006). The importance of neutral examples for learning sentiment. *Computational Intelligence*, *22*(2), 100–109.

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., & Cebrian, M. (2016). Rapid assessment of disaster damage using social media activity. *Science Advances*, *2*(3), e1500779.

Larsson, A. O. (2013). Tweeting the viewer—Use of Twitter in a talk show context. *Journal of Broadcasting & Electronic Media*, *57*(2), 135–152.

Larsson, A. O., & Moe, H. (2012). Studying political microblogging: Twitter users in the 2010 Swedish election campaign. *New Media & Society*, *14*(5), 729–747.

Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1948). Thepeople's choice. *New York: Columbia UniversityPress, 1948LazarsfeldThe People's Choice1948*.

Lifson, T. (2016). Dianne Feinstein's Husband Wins Near-Billion Dollar California "High Speed Rail" Contract. Retrieved from http://conservative50.com/dianne-feinsteins-husband-wins-near-billion-dollar-california-high-speed-rail-contract/

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167.

MacSween-George, S. L. (2003). A Public Opinion Survey: Unmanned Aerial Vehicles for Cargo, Commercial, and Passenger Transportation. In *AIAA" Unmanned Unlimited" Systems, Technologies, and Operations Conference. San Diego, California*.

Maddex, W. E. (2012). *The Rigor of Negotiation; Why Public Private Partnerships are Effective*. Arizona State University.

Morrill, J. (2016). N.C. House passes bill to cancel I-77 toll contract. Retrieved from http://www.charlotteobserver.com/news/politics-government/article81344977.html

Mutz, D. C., & Young, L. (2011). Communication and public opinion: Plus ça change? *Public Opinion Quarterly*, *75*(5), 1018–1044.

NAACP. (n.d.). What does NAACP Stand for? Retrieved from http://www.naacp.org/about-us/

Newport, F., Saad, L., & Moore, D. (1997). How are polls conducted. *Where America Stands*.

Nisbet, M. C., & Kotcher, J. E. (2009). A two-step flow of influence? Opinion-leader campaigns on climate change. *Science Communication*, *30*(3), 328–354.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, *11*(122–129), 1–2.

Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques*. Springer Science & Business Media.

Palen, L. (2008). Online social media in crisis events. *Educause Quarterly*, *31*(3), 76–78.

Palen, L., Hiltz, S. R., & Liu, S. B. (2007). Online forums supporting grassroots participation in emergency preparedness and response. *Communications of the ACM*, *50*(3), 54–58.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (pp. 79–86).

Pennacchiotti, M., & Popescu, A.-M. (2011). A Machine Learning Approach to Twitter User Classification. *Icwsm*, *11*(1), 281–288.

Perrin, A. (2015). Social Media Usage: 2005-2015. Retrieved from http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/

Postel, J. (1994). Domain name system structure and delegation.

Postrel, V. (2016). California Hits the Brakes on High-Speed Rail Fiasco. Retrieved from https://www.bloomberg.com/view/articles/2016-06-28/california-hits-the-brakes-on-high-speed-rail-fiasco

Presidio Parkway. (2016). Presidio Parkway Overview. Retrieved from http://www.presidioparkway.org/about/

Qiao, L., Wang, S. Q., Tiong, R. L. K., & Chan, T.-S. (2001). Framework for critical success factors of BOT projects in China. *Journal of Project Finance*, *7*(1), 53–61.

RapidMiner. (2014). *RapidMiner Studio Manual*.

Rehurek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA.

Ritterman, J., Osborne, M., & Klein, E. (2009). Using prediction markets and Twitter to predict a swine flu pandemic. In *1st international workshop on mining social media* (Vol. 9, pp. 9–17).

Roberts, S. (2011). Appeals Court Rules Presidio Parkway Can Move Forward as P3. Retrieved from http://www.infrainsightblog.com/2011/08/articles/p3s/appeals-court-rules-presidio-parkway-can-move-forward-as-p3/

Robinson, P. (2008). The role of media and public opinion. *Foreign Policy: Theories, Actors, Cases*, 137–154.

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PloS One*, *6*(5), e19467.

Statistica. (2016). Most famous social network sites worldwide as of September 2016, ranked by number of active users (in millions). Retrieved from

https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267–307.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. In *ACL (1)* (pp. 1555–1565).

The Fresno Bee. (2015). Trial date set in Prop. 1A lawsuit over high-speed train plans.

The Fresno Bee. (2016). Delays from lawsuit cost high-speed rail cost $63 million, 17 months.

The White House. (2014). An Economic Analysis Of Transportation Infrastructure Investment, (July).

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, *63*(1), 163–173.

Thompson, D. (2017). California high-speed rail ready to lay some track. Retrieved from https://apnews.com/98f8e3bb3df94f94b4b06743f66ccc26/california-high-speed-rail-ready-lay-some-track

Tilt, B., Braun, Y., & He, D. (2009). Social impacts of large dam projects: A comparison of international case studies and implications for best practice. *Journal of Environmental Management*, *90*, S249--S257.

Turcotte, J., York, C., Irving, J., Scholl, R. M., & Pingree, R. J. (2015). News recommendations from social media opinion leaders: Effects on media trust and information seeking. *Journal of Computer-Mediated Communication*, *20*(5), 520–535.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 417–424).

Twitter. Inc. (2016). Twitter Developer Documentation. Retrieved from https://dev.twitter.com/overview/documentation

Ulberg, C. G. (1995). *An Evaluation of Public Opinion about Congestion Pricing and Tolls*. Transportation Northwest.

United States Government Accountability Office. (2012). *HIGH-SPEED PASSENGER RAIL Preliminary Assessment of California's Cost Estimates and Other Challenges*.

Update, A. G. C. (2017). January 2017. *Washington, DC: January*, *16*, 9.

Vaccari, C., Chadwick, A., & O'Loughlin, B. (2015). Dual screening the political: Media events, social media, and citizen engagement. *Journal of Communication*, *65*(6), 1041–1061.

Venables, W. N., Smith, D. M., Team, R. D. C., & others. (2004). An introduction to R. Network Theory Limited.

Vryniotis, V. (2013). The importance of Neutral Class in Sentiment Analysis. *Machine Learning Blog & Software Development News*.

Waldron, M. (2015). Building a Twitter Sentiment Analysis Process in RapidMiner. Retrieved from http://blog.aylien.com/building-a-twitter-sentiment-analysis-process-in/

Wang, X., Gerber, M. S., & Brown, D. E. (2012). Automatic crime prediction using events extracted from twitter posts. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (pp. 231–238).

Watts, D. J., & Dodds, P. S. (2007). Influentials, networks, and public opinion formation. *Journal of Consumer Research*, *34*(4), 441–458.

Weber, T. (2010). Why companies watch your every Facebook, YouTube, Twitter move. *Retrieved November*, *5*, 2010.

Weimann, G. (1994). *The influentials: People who influence people*. SUNY Press.

Weimann, G., Tustin, D. H., Van Vuuren, D., & Joubert, J. P. R. (2007). Looking for opinion leaders: Traditional vs. modern measures in traditional societies. *International Journal of Public Opinion Research*, *19*(2), 173–190.

Weisbrod, G., & Weisbrod, B. (1997). Assessing the economic impact of transportation projects: How to choose the appropriate technique for your project. *Transportation Research Circular*, (477).

Welch, M. (2016). The Political Class Knew California High-Speed Rail Was B.S., and Supported it Anyway. Retrieved from http://reason.com/blog/2016/06/28/the-political-class-knew-california-high

Widen I-77. (2016). Widen I-77. Retrieved from wideni77.org

WidenI77. (n.d.). Widen I-77. Retrieved from wideni77.org

Wiggins, O., & Turque, B. (2015). NAACP to challenge cancellation of Baltimore Red Line rail project. Retrieved from https://www.washingtonpost.com/local/md-politics/naacp-to-challenge-cancellation-of-baltimore-red-line-rail-project/2015/12/21/6cdb45aa-a7fc-11e5-8058-480b572b4aae_story.html

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level

sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing* (pp. 347–354).

Wüstenhagen, R., Wolsink, M., & Bürer, M. J. (2007). Social acceptance of renewable energy innovation: An introduction to the concept. *Energy Policy*, *35*(5), 2683–2691.

Yates, D., & Paquette, S. (2011). Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake. *International Journal of Information Management*, *31*(1), 6–13.

Zhang, W. R., WANG, S. Q., TIONG, R. L. K., Ting, S. K., & Ashley, D. (1998). Risk management of Shanghai's privately financed Yan'an Donglu tunnels. *Engineering, Construction and Architectural Management*, *5*(4), 399–409.

Zhang, X. (2005). Critical Success Factors for Public – Private Partnerships in Infrastructure Development. *Journal of Construction Engineering and Management*, *131*(1), 3–14. https://doi.org/10.1061/(ASCE)0733-9364(2005)131

Zoss, A. (2017). Introduction to Text Analysis: Analysis Methods and Tools. Retrieved from https://guides.library.duke.edu/text_analysis