2015

Agenda detector: labeling tweets with political policy agenda

Sheetal Kaul
Iowa State University

Agenda detector: Labeling tweets with political policy agenda

by

Sheetal Kaul

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Computer Science

Program of Study Committee:
Wallapak Tavanapong, Major Professor
Johnny S. Wong
David Peterson

Iowa State University

Ames, Iowa

2015

DEDICATION

I *dedicate* this thesis to my father and inspiration - Dr. G.L. Kaul, to my paternal grandfather and first love of my life - Mr. R.K. Kaul, and to my maternal aunt and a motherly figure - Ms. Sunanda Pushkar. I am sure had they been around to watch me complete this, they would have had the broadest smiles on their faces. You all are dearly missed every day.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank my major professor, Dr. Wallapak Tavanapong for her continued guidance, for helping me practice the art of "research and development" and keeping me motivated to work harder. I value her inputs and patience she has shown with me during our association.

Thank you to my POS committee member Dr. David Peterson for being very approachable and guiding me to problems and challenges in the political science domain. I thank him for his constructive feedback and for making this collaboration seamless and enjoyable.

A big thanks to my co-major professor Dr. Johnny Wong for his valued guidance and time for Q&A sessions. They were really encouraging and reassuring that research gets done my doing! I also want to thank him and Dr. Tavanapong for all the good food and fun we had in this association. A special thank you to Dr. Samik Basu for his unflinching support and guidance in my graduate life.

I want to acknowledge LAS Signature Research Initiative (SRI) for funding this project and thank them for their valuable support.Thank you to Eric Meyer, Jefferson Fink and my other friends for helping with manual labeling of tweets. Their efforts are greatly appreciated. A mighty thanks to the Cyride bus service and my sister Ms. Sunita Koul for gifting me her car. My gratitude to Daiana Coimbra for coming all the way from Brazil to become my roommate and filling this graduation journey with adventure and fun.

Most importantly, thanks to God for surrounding me with loving friends and an inspirational family. A special mention to my mother Ms. Veena Kaul, uncle Prof. O.N. Koul and brother Dr. Neeraj Koul for their continued words of encouragement.

ABSTRACT

In nearly one decade of Twitter's being it has witnessed an ever growing user base from various realms of the world, one of them being politics. In the political domain, Twitter is used as a vital tool for communication purposes, running effective e-campaigns, and mining and affecting public opinions to name a few. We study the problem of automatically detecting whether a tweet posted by a state's Senate's twitter handle in the US has a reference to policy agenda(s). Such a capability can help detect the policy agendas that a state focuses on and also capture the inception of ideas leading to framing of bill/law. Furthermore, analyzing the spatial and temporal dynamics of tweets carrying policy agendas can facilitate study of policy diffusion among states, and help in comprehending the changing aspects of states learning policy-making from each other.

Currently, no study has been carried out that analyzes Twitter data to detect whether or not a tweet refers to a policy agenda. We present our analysis on 122,965 tweets collected from verified Twitter handles of the US state's upper house – Senate. We present our high-level analysis on (a) how much Twitter has penetrated into state politics and (b) how states use the medium differently in terms of the messages they broadcast. Our proposed approach aims to automate classification of a tweet based on having a reference to policy agenda (Has Agenda) or not (No Agenda). We accomplish this by leveraging existing text classification methodology and achieve a recall of 89.1% and precision of 77.2% for the "Has Agenda" class. We investigate several machine learning algorithms to determine the best performing one for our binary classification problem. We conclude that support vector machine using linear kernel was the most efficient algorithm to use for our dataset. Lastly, we propose a set of hand-crafted features that together with feature selection and stemming improved our classifier's performance. Prior to including these features the classifier was developed using, basic preprocessing techniques, and term occurrence (for feature extraction). An overall improvement of 5.187 % at a significance level of $\alpha = 0.05$ was achieved.

CHAPTER 1 INTRODUCTION

Poli-informatics is an interdisciplinary field that promotes diverse methodological approaches to the study politics and government[1]. Publicly available high volumes of government datasets [2] together with advances in computational linguistics, machine learning, data visualization, and high performance computing, facilitate innovation in perspectives related to governance. These government datasets are vast and vary from data on agriculture, business, climate, health, finance, local government, education, energy amongst many others.

In our study, we focus on state governments and their respective policy agendas. A policy agenda is a set of issues viewed as important by people in policymaking (e.g., government officials, government decision-makers). Since policy agendas can vary in terms of the issues they address, Policy Agendas Project [3] divides them into 20 main topics and 220 subtopics. The main topics are shown in Table 1.1.

Table 1.1. List of major topic codes

1. Macroeconomics
2. Civil Rights, Minority Issues, and Civil Liberties
3. Health
4. Agriculture
5. Labor and Employment
6. Education
7. Environment
8. Energy
9. Immigration
10. Transportation
12. Law, Crime, and Family Issues
13. Social Welfare
14. Community Development and Housing Issues
15. Banking, Finance, and Domestic Commerce
16. Defense
17. Space, Science, Technology and Communications
18. Foreign Trade
19. International Affairs and Foreign Aid
20. Government Operations
21. Public Lands and Water Management

Another key aspect of our study is based on the rapid mileage gained by online social networking sites in the current digital world. One of the prominent platforms in this domain is Twitter, which in less than 10 years of its being has gathered a user base of more than 302 million

monthly active users across the globe [4]. Twitter enables its registered users to send short 140 characters messages called "tweets", and follow other users' twitter posts. As of May 2015, US President Barack Obama is the most followed politician in the world on Twitter with 58.1 million followers [5]. The use of Twitter in the political domain is on the rise. The 115 studies compiled in the survey [6] reveal that Twitter is used in the political domain by politicians for running effective e-campaigns, sharing their political strategies, their vision for the future and discussions on various policy agendas. The medium is also used by constituents to voice their opinions on political matters to name a few.

Our work aims at automating the detection of policy agendas at the state-level in the US that are under discussion, or framed into a bill, or passed as a law. The work marks a first attempt in analyzing this domain on Twitter. We study the tweets posted from State Legislature twitter handles to ascertain whether or not a tweet contains a policy agenda. Tweets that mention about passing of bills and state the agenda topic it is related to are considered in our class of "Has Agenda". We also observed tweets that do not mention passing of bills directly but discuss action items on agenda topics. Such discussions bear a high chance of formulation of bills related to the corresponding agenda topics, hence, we consider them in our "Has Agenda" class as well. Keeping these guidelines in mind Table 1.2 exemplifies performing the binary classification manually.

Table 1.2. Manual classification of tweets

Tweet text	**Label**
The Senate passes H.B. 2036, adding and modifies certain statutes related to the regulation of abortion and abortion clinics. Vote 20 – 10	Has Agenda
Senate passes HB 2601, increasing maximum amount of unpaid wages that enable an employee to file a claim with the ICA. Vote 29-0	Has Agenda
GOP puts minimum wage bill on fast track to help boost prospects for oil tax law.	Has Agenda
9:30 this morning on the Senate Floor....Marine Corps celebrates its 237th birthday.	No Agenda
Good morning Alabamians!!!!! Today is the last day of the 2013 Legislative Session!!!! Stay tuned for updates throughout the day :22)	No Agenda

Policy-making has been studied to understand the role media plays in formulating policies [7]. The study in [8] studies open government datasets and analyzes how policy agendas have been addressed/ignored in the history of American politics. Congressional speeches have been analyzed in [9] and mapped to topic codes as per Table 1.1. Twitter data has been analyzed to study different

aspects of the political world [10-14] as well as outside of it [15-21]. The work carried out in [22-24] analyzes the underlying sentiment of tweets and [25-29] addresses the privacy concerns that come along with the medium's growing popularity and usage. Besides using Twitter as the data source, there are studies that analyze data from YouTube to determine the impact of political video-campaigns on online communities [30]. The Lydia project in [31] uses traditional online news sources- "New York Times" and "Time Magazine" to gauge how political entities are framed in the media.

To the best of our knowledge, we do not find any existing study that analyzes Twitter data to detect policy agendas in the US. In this paper, we propose an approach to automatically label a tweet as "Has Agenda" or "No Agenda" based on the existence of a policy agenda in it as discussed earlier. Such an ability would facilitate political scientists to (a) study intra-state politics in the new-age digital world, (b) understand which policy agendas are communicated to public using online social media, (c) study the course of policy framing at the state level from the inception of its idea to becoming a law, and (d) study dynamics of policy diffusion among states by analyzing how states impact one another in terms of policy making, if at all. In order to lend credence that we analyze data that bears a direct impact in policy making, we manually collected the verified State Legislatures' twitter handles. After initial investigations, we found that the Twitter handles of Senate are more prevalent when compared to the twitter handles of House, hence we collected the Senate twitter handles for all the 50 states in the US.

1.1 Contribution

We make three main contributions in our study. Firstly, in terms of the text classification problem that we study, (a) we present how each processing step in the process impacts the performance of the classifier for this problem; (b) we examine several machine learning algorithms to determine the best performing algorithm that achieves a recall of 89.1% and precision of 77.2 % for the "Has Agenda" class; and (c) we propose a set of new hand-crafted features which we include in our data sets' feature set. Including these features enhances our support vector machine based binary classifier's performance significantly by 5.187% at a significance level of $\alpha = 0.05$. We validate our results on a vast dataset of over 4,000 manually classified tweets of 5 states.

Secondly, we present our findings on (a) how much Twitter usage has penetrated into the world of state politics, (b) how various states make use of the medium in terms of communicating messages with or without policy agendas. Thirdly, we make our collected data (manually collected and using Twitter API) publicly available for further analysis. We manually collected verified Twitter handles representing Senate and House of states in the US. Senate representation of states was far more prevalent when compared to House representation. As of 12-16-2014, 30 states had a verified Senate handle on Twitter and some of the states had a separate representation of Senate Republicans and Senate Democrats. Overall 42 handles representing state's Senate were collected and a subset of 122,965 tweets posted by them was stored in our tweet repository. We also developed a robust web application for manual classification of tweets which can be modified for similar n-class classification problems. As a final point, our study marks a maiden attempt to analyze twitter data for detecting policy agendas at the state-level.

1.2 Organization

The rest of the thesis is organized as follows. Chapter 2 presents related work in political science and social media analysis. Chapter 3 discusses the approach of our proposed work and Chapter 4 illustrates the experiment design and results. Chapter 5 presents our conclusion and avenues of future work.

CHAPTER 2 RELATED WORK

We consider (i) studies in the Political Science domain that relate to our work, (ii) studies that employ machine learning techniques in text classification, and (iii) studies that analyze data mined from social media. We found Twitter was used as a data source for majority of the studies along with other sources like Pinterest, YouTube and online News sources. As of May, 2015, Twitter has 302 million monthly active users and 500 million tweets being posted per day on an average. Hence, it is unsurprising that a significant amount of work has been carried out on analyzing Twitter data among varied domains, not just in politics. Since we encountered a variety of studies in our survey, we organize our related work into different classes and subclasses as shown in Table 2.1.

Table 2.1. Categorization of related work

Class	**Subclass**
Research Theme	**Poli-informatics**: Analyzes publicly available data in the realm of politics [6, 7, 10-14, 31, 32] **Social Network Analysis**: Studies the flow of information within online social networks [15, 17, 18] **Information Mining**: Extracts information relevant to a research topic from the text content [16, 19-21, 33-35] **Sentiment Analysis**: Gauges sentiment of the text towards a relevant topic [22-24, 30] **Privacy Concerns**: Addresses privacy hazards on online social media [25-29, 36] **Text Classification**: Uses machine learning techniques for classifying text [26, 32, 37-39]
Data Source	Policy Agendas Project [3] Congressional Bills [40] Congressional speeches Social Media : Twitter, YouTube, Pinterest Online news sources: New York Times, Time Magazine

Table 2.1. (continued)

Data Collection Methods	Twitter APIs (REST API, Search API and Streaming API) YouTube API Web crawler Direct download
Technical Approach	Supervised Machine Learning Content Analysis Sentiment Analysis Graph-based Frameworks Statistical Modeling Manual Classification
Results Visualization	Graphs, Tables, Pie Charts, Heat Maps

In our survey, we found that Twitter is a powerful medium in the domain of Politics. The claim can be substantiated by the comprehensive literature survey carried out by [6]. This literature survey by Jungherr is a compilation of 115 studies carried out at the intersection of political and Twitter domains. The survey grouped the studies into three categories: (a) usage of Twitter by politicians, (b) usage of Twitter by constituents during elections, (c) usage of Twitter by various actors in reaction to facilitated campaign events, e.g., televised debates or coverage of election-day. None of the studies mentioned in this survey present an approach to detect policy agendas mentioned by State Legislatures on the social networking medium.

Outside of the research work cited in this survey, we present related studies in the domain of policy agendas. The study presented in [7] showcases the effect of traditional media sources on policy-making in U.S Congress. It bases its findings on data collected from New York Times and agenda topics of bills introduced in Congress (made available by Policy Agendas Project). The study also correlates media coverage on crime related issues to the money invested by state budget in corrections (punishment, treatment, and supervision of people convicted of crimes).

The authors of [32] work with a dataset of federal public bills introduced since 1947, referred to as the Congressional Bills Project [40]. They present a supervised machine learning approach to develop a multi-class classifier which can annotate the subtopic (based on 226 subtopics in Policy Agendas Project) of a given bill. Our work employs a similar supervised machine learning approach, but we assert that our dataset extracted from Twitter comes with a lot of noise, is not formal in structure as the bills and has a much higher volume and veracity in comparison to the Congressional Bills dataset. In our study, we also present an overview of how various processing steps in text classification impact classifier's performance. Finally, we also use hand-crafted features that significantly improve classifier's performance on our dataset and can be re-used in similar Twitter-based studies.

In the realm of social media and politics, the research work in [11] uses speech acts methodology on Twitter data. Speech acts refers to an attempt made in speech to get someone to do something regarding the topic mentioned in the speech. This study manually categorizes these speech acts into 16 different categories implying the linguistic approach in which constituents lobby Congress using the medium. The authors identified 4 prominent political topics and the hashtags used for tweeting about such topics. Using Twitter's streaming and search APIs, they collected 76,454 tweets that used the identified hashtags and were addressed to Twitter handles owned by members of the 112th and/or 113th Congress. They filtered out 42,398 retweets and manually classified a random subsample of 925 tweets from the remaining ones to accomplish the grouping into the 16 different categories.

The study in [10] presents a 6-class categorization of the speech acts of Congress on Twitter. The study develops an automated classifier using supervised machine learning techniques to label the speech category of a tweet. It uses 526 manually classified tweets to train the classifier and overall it collects 30,373 tweets to present analysis on tweeting frequency by gender, party and chamber (U.S Senate and House.) The study in [41] and several other studies in the literature survey in [6] confirm that Twitter can be used as a significant tool when it comes to election campaigning. Moreover, the medium can help determine political affiliations of a citizen by performing sentiment analysis on their tweets [12]. This study uses existing naïve lexicon based approaches (Subjectivity Lexicon [42] and SentiWordNet 3.0 [43]) to determine the sentiment

polarity of words in text. The study enhances the sentiment analysis' performance by incorporating support vector machine in a 5-fold cross validation setup.

The authors of [12] develop a 3-class classifier of tweets into positive, negative, and neutral towards a political entity by using a manually classified set of 2,624 tweets as ground truth. Lastly, in the political domain, Twitter has also been used to predict election results by using the sentiment analysis approach [14]. This study reports the analysis of 104,003 tweets that appeared few weeks prior to the election of the national parliament in Germany, 2009. Sentiment Analysis of the tweets was carried out using the tool LIWC2007 [44] and it was concluded that the Twitter activity prior to elections was a valid reflection of the election outcome. Outside the strict domain of politics, [22] uses psychometric instruments to automatically models the moods and emotions of public by mapping text to 6 mood states (tension, depression, anger, vigor, fatigue, confusion.) and establishes a significant correlation between worldly events and public temperament on Twitter.

As mentioned earlier in Table 2.1, the data on Twitter has been used to analyze information flow on the social networking site. To understand the spatial and temporal mechanics of flow of hashtags at a global scale, study in [15] analyzes the geolocation and time properties of 27 million unique hashtags extracted from Twitter using the Streaming API. The study quantifies the global footprint of the hashtag, analyzes spatial properties of hashtag propagation, and measures the spatial impact of a location on hashtag propagation dynamics. Another research work carried out in [17] detects a cluster of messages on Twitter that are bound by a common theme referred to as campaigns. In this study, 1912 tweets were manually examined to extract such campaigns that share a common talking point. Using these groupings as the ground truth, the authors identified Shingling [45] as the best existing near-duplicate detection algorithm to find relatedness between texts of two tweets. Founded on this approach, authors of [17] constructed message graphs over large tweet datasets (~1.5 million tweets), where tweets become nodes and edges reflect the relatedness between them. Based on stated formulae in the paper, they were able to extract loosely /tightly bound campaigns. And upon manual inspection of certain attributes of these campaigns, they were able to categorize them into legitimate/spam campaigns.

Lastly, the work carried out in [29] quantifies the influence of a user on Twitter by analyzing attributes like, user's followers' network and its dynamics of retweeting a user's tweet that contains a URL. Retweeting such tweets clearly implies propagating information in the URL

to their respective follower networks. Quantifying influence can prove to be a strong component in laying out strategies to craft public opinions or obtain word of mouth publicity.

With the variability of purviews these tweets can fall into, different studies focus at mining variegated information from Twitter. For example, the occurrence of diseases like Influenza is mined from Twitter in [19]. The study evaluates the ability of statistical models like linear regression, multivariable regression, and support vector machine regression (SVMRegression) to accurately assess the prevalence of the disease. In another study, a probabilistic framework is proposed to narrow down the user's whereabouts to the city level by only analyzing the content of tweets by a user [16]. The dataset used in this study is the 5 million tweets posted from 5,190 users spread across the continental United States. The study uses maximum likelihood estimation to identify probability distribution of unique words in their dataset over cities and employs classification algorithms available in Weka Toolkit [46] to classify words as being local to a city. It uses various techniques to smoothen the probability distributions of words over cities and proves to be a promising method in estimating content-driven locations in the future.

The study in [20] is centered around journalism and confirms that the death of Osama Bin Laden first broke on Twitter before it reached traditional media. This was established by manually backward tracing the tweets containing the keyword "laden" and verifying the user as @keithurbahn who broke the news on Twitter prior to US President Barack Obama stating it in his official address.

From the visible outburst of information on Twitter, it was found that people tend to give away vital personal details that could have serious ramifications by attackers. For example, the study in [36] ascertains that online social media divulges our personally identifiable information (PII) on the internet. Study in [25] presents an approach to combat privacy hazards on online social networks.

Apart from PII, study in [26] states that users tend to reveal information that may belong to three subjects of: divulging vacation plans, tweeting about driving in a drunk state, or tweeting information regarding diseases contracted. This study aims at developing an automated binary classifier to label tweets as "sensitive" if they belong to any of the aforementioned three subjects otherwise label as "insensitive." The study performs a primary filtering on tweets by extracting those tweets that have matching subject-specific-keywords in it. It uses a set of 600 tweets as

ground truth and has a 70:30 distribution of the size of the training set to testing set. The study uses machine learning techniques - Naïve Bayes and SVM to train the binary classifier. Our problem under consideration is similar in nature being a binary classification problem, but we emphasize that we do not use any keyword matching in data collection which is bound to add a positive bias in the classifiers performance. Also, we examine performance of additional machine learning algorithms like Decision Trees (J48,) Cost Sensitive Classifier, Attribute Selected Classifier, and CV Parameter Selection. Our study is based on a vaster ground truth of over 4,000 tweets with a 50:50 distribution of training set to testing set. Lastly, we also contribute towards employing Twitter specific and Policy Agendas specific feature extraction from tweets in our classification process which significantly enhances classifier's performance.

Analyzing data from social media platforms is not limited to Twitter but extends to other such platforms like Pinterest [35] and YouTube [30]. Other than using social media platforms as the data source, the Lydia Project [31] uses data from online news sources like New York Times and Time Magazine. Lydia enables a user to visualize media trends on political entities encompassing (a) reference classification by type (News/ Business/ Entertainment/ Sports/ Other), (b) polarity (positive/negative sentiment) and subjectivity (number of sentiment references) trends for an entity in the media, (c) analysis of the words co-occurring with an entity in media referred to as juxtaposition analysis, (d) spatio-temporal analysis on the entity i.e., monitoring trends over time and geography.

After careful research, we can affirm that we did not find any existing study that presents an approach to detect policy agendas in tweets posted by State Legislature handles. Since our problem falls under the purview of text classification, we leverage the basic schematics of text classification from [26, 32, 37-39]. We utilize the text classification framework which involves Stemming, Data Cleansing, Feature Extraction, and Machine Learning along with our proposed set of features (mentioned in upcoming sections), which improves our binary classifier's performance significantly.

CHAPTER 3 PROPOSED POLICY AGENDA DETECTOR

In this chapter, we present our approach to classify a given tweet into one of the two classes---"Has Agenda" or "No Agenda." We analyzed the tweets posted from state's Senate Handles in USA to develop an algorithm to automate the categorization of these tweets into the aforementioned two classes. Manual collection of twitter handles and formulating the ground truth (discussed in upcoming sections) proved to be a very time-consuming and painstaking effort.

In the following sections, we discuss the step by step process carried out to achieve our goal. Section 3.1 defines a "policy agenda". In Section 3.2, we discuss the overview of our approach. In 3.3, we detail the data collection process and in Section 3.4 we present the process of ground truth formulation. In Section 3.5, we discuss steps to prepare data for training and in Sections 3.6 and 3.7, we articulate the process of feature extraction and feature selection respectively. Finally, in Section 3.8, we present the machine learning techniques used to develop an efficient binary classifier.

3.1 Definition of a policy agenda

A policy agenda by definition refers to a set of issues and policies laid out by political groups which may possess the credibility to be formulated into a bill and later law. Also considered as policy agenda are topics under discussion by a governmental executive, or a cabinet in government that tries to influence current and near-future political news and debate.

3.2 Approach overview

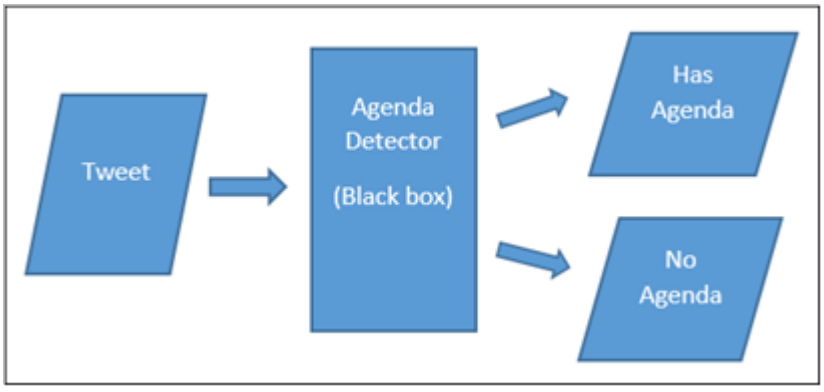

Figure 3.1. Overview of the proposed approach

As depicted in Figure 3.1, the goal of our proposed approach is to develop a binary classifier for tweets in order to label them into classes- Has Agenda/No Agenda. This goal in achieved in several steps by using supervised machine learning techniques and following architecture of text classification process depicted in Figure 3.2. The processing steps include – data collection and storage, formulation of ground truth, data cleansing, feature extraction, feature selection, and application of machine learning algorithms.

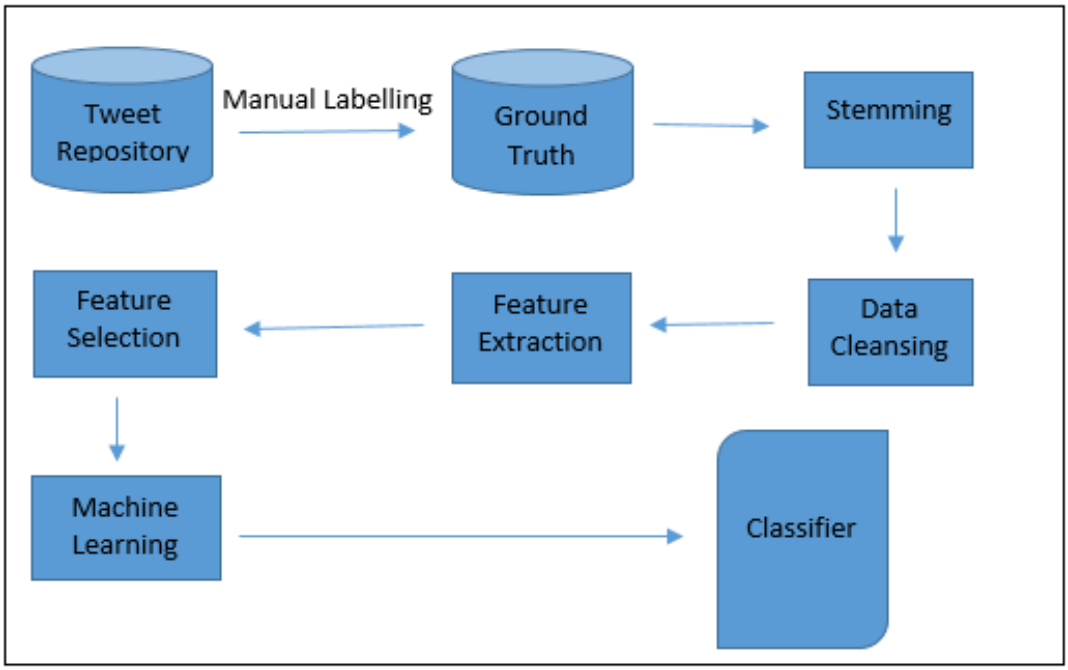

Figure 3.2. Application of supervised machine learning to determine if a tweet contains a policy agenda.

3.3 Data collection and storage

We began data collection by manually collecting the Twitter handles for State Legislatures in the US. After initial investigations, we found that the user handles for Senate outnumbered the House and we proceeded with collecting the Senate twitter handles for each state. As of December 16, 2014, 31 states out of 50 had Senate Twitter handles and majority of these had two Senate

handles, one for Senate Democrats and another for Senate GOP (Grand Old Party or Republicans). Overall, we collected 42 twitter handles and gathered tweets posted from by them.

In order to collect data from Twitter, we developed a standalone Java based application that used the Twitter4J library, and Twitter's REST API for collecting twitter data. We used the "GET statuses" method from the Twitter API, which returns 3,200 most recent tweets for each user handle for which data is requested. Due to this limitation of the API, the "date of creation of tweets" that our collected tweets span across varies as per the tweeting frequency of these user handles. Overall, we had tweets from as early as September 13, 2007 until December 16, 2014 (the date we requested the tweets). From these 42 handles, we gathered a total of 122,965 tweets and present our findings in section 4.3. Based on our literature survey, this is the largest dataset analyzed upon in the domain of politics and twitter. Since we work with tweets that are posted by user handles and are not based on keyword search we did remove any duplicate tweets.

We stored this Twitter data in a Relational database using MySQL Community Server 5.5. Apart from storing the text in the tweet, we also stored tweets' other attributes and data related to twitter handles as shown in the database schema in Figure 3.3.

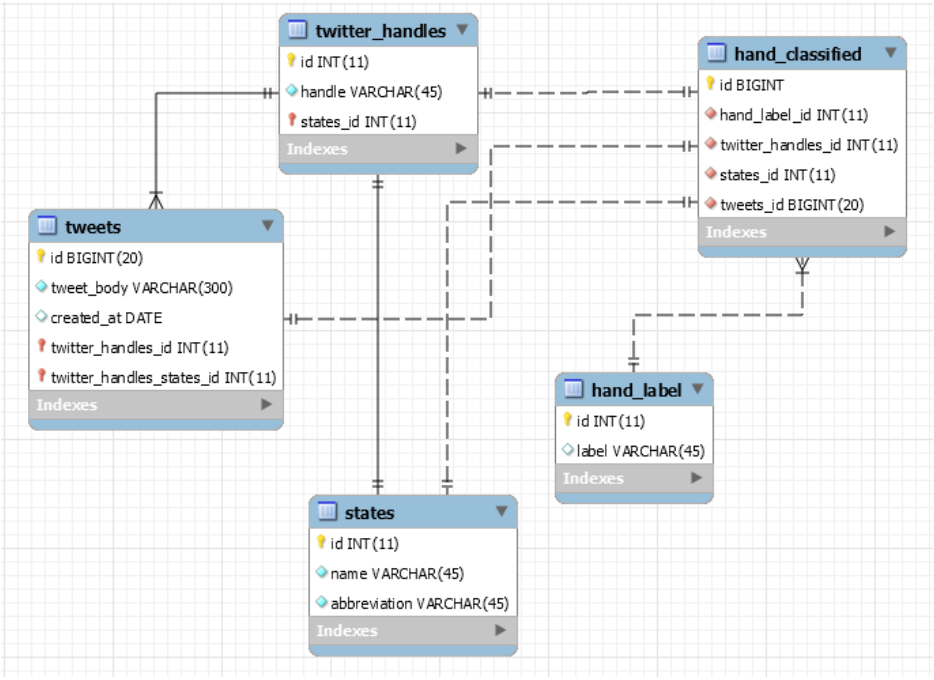

Figure 3.3. Snapshot of DB schema of our experiment.

3.4 Formulation of ground truth

In text classification problems that use supervised machine learning techniques, the term "ground truth" refers to a set of accurately labeled documents which can be used to train and later test classifier's performance. Since no such set is publicly available for our classification problem, we manually classified a set of tweets to serve as ground truth. We shortlisted states that had both Democrat and Republican verified Twitter handles. From these we selected 5 states such that we could work with different data dimensionalities and at the same time the manual labelling for which would be under reasonable human capacity. In this study we worked with states namely, Alabama, Arizona, Minnesota, Pennsylvania, and Oklahoma.

We extracted the respective tweets for each state which totaled to 4,790 tweets. Five human coders from different educational backgrounds (Computer Science, Political Science, Journalism and Agronomy) labeled different data sets according to the criteria for "Has Agenda" class as described in Table 3.1. The criteria was developed under consultation with the domain expert in Political Science and relies on Policy Agendas Project [3] that explains 20 political agenda topics and 220 subtopics in detail. Tweets that do not match the criteria were labeled as "No Agenda."

Table 3.1. Criteria for manually labeling a tweet as "Has Agenda"

Aspects in Tweet	**Labeled Class**
Mention of passing/proposing a bill containing/referring to policy agenda(s)	Has Agenda
Laying out political strategies centered on policy agenda(s)	Has Agenda
Expressing an interest in bringing forth a reform centered on policy agenda(s)	Has Agenda
Criticizing existing scenario on policy agenda(s)	Has Agenda
Sharing inputs/reporting progress in the state regarding policy agenda(s)	Has Agenda
Any other direct reference to policy agenda(s) that has political relevance	Has Agenda
Any reference to state's budget	Has Agenda

To facilitate the ground truth manual labeling, we developed a PHP web-application that multiple people could access the tweets in our central data repository. In the application, tweets

are populated on screen based on the selected state. The human coder assigns a label to each tweet accordingly.

3.5 Data preprocessing

The collected data have noise (unwanted data). Preprocessing techniques ensure that all the noise is disregarded before processing to next steps in text classification. We employ two techniques: Stemming and Data Cleansing. Stemming is a common term used in Natural Language Processing (NLP) which refers to a method of reducing a word to its root word or stem. The technique drastically reduces the number of unique words in the dataset while still maintaining the word's basic intent of usage. There are various existing stemming algorithms that are used in NLP today namely, Porter Stemmer, K-Stemmer and Hunspell Stemmer among few others [47]. Hunspell Stemmer can be applied to languages other than English. Different stemmers over stem and under stem to a different degree. As exemplified in Table 3.2, Porter stemmer performs aggressive stemming whereas K-Stem algorithm is known to under stem words [48]. Since the tweet content is already limited to 140 characters, we prefer using K-Stem to avoid any data loss. For Data Cleansing, we remove the standard stop words in English dictionary and also eliminate URLs in the tweets.

Table 3.2. Examples of stemming results using Porter and K-Stem algorithms

Original word	**Over stemming (Porter)**	**Under stemming (K-Stem)**
Recession	Recess	Recession
Importance	Import	Importance
Import	Import	Import
Namely	Name	Namely
Addicting	Addict	Addict
Political	Polit	Politics
Policy	Polic	Policy
Police	Polic	Police
Educating	Educ	Educate

3.6 Feature extraction

We observed the preprocessed data and introduced additional features listed in Table 3.3. These features include three Twitter specific features and one policy agenda specific feature. The latter, HasTopicKeyword, is to extend the capability of the classifier to classify tweets not seen in the training set. We created a dictionary of representative words of 20 policy agenda topics from the description of the topics provided by Policy Agendas Project [3]. A value 1 or 0 is assigned as the value of this feature based on whether or not the tweet includes any of the words in the dictionary.

Table 3.3. Description of Twitter specific/policy agenda specific feature(s)

Feature Name	Allowed Feature Values	Description
Twitter specific		
HasHashtag	0,1	Whether or not a tweet contains a hashtag
HasURL	0,1	Whether or not a tweet contains a URL
TweetStrength	0-8, 9-13, 14-18, 19-24, >25	The histogram bin range of word counts in a tweet.
Policy agenda specific		
HasTopicKeyword	0,1	Whether or not a tweet contains a word in the policy agenda dictionary

After the data were preprocessed, we segregated them based on states because we observed several differences in each state's pattern of tweeting frequency, choice of words used in tweets, structure of framing sentences, and policy agendas referred to in the tweets. For any given state's dataset, we divided it into equal halves of training set and testing set. We developed a program to process the training set to generate a set of unique terms and a feature vector to represent each tweet. The basic features are Term Occurrence (0 when the term is not present in the tweet or 1 otherwise) of the unique terms. The Twitter specific features and policy agenda specific feature are also extracted. In Section 4.3.2, we discuss the effectiveness of these features.

3.7 Feature selection

Feature selection refers to the process of selecting the optimal set of features/attributes from the given list such that the classifier yields best results. We used Weka Toolkit [46] to perform this study. We experimented with "InfoGainAttributeEval" Attribute Evaluator and the "Ranker" search method. InfoGainAttributeEval evaluates the worth of an attribute by measuring the information gain with respect to the class. The Ranker search method ranks attributes by their individual evaluations. We also experimented with "CfsSubsetEval" Attribute Evaluator and the "BestFirst" CfsSubsetEval evaluates the worth of a subset of attributes by considering the individual predictive ability of each attribute along with the degree of redundancy between them. The "BestFirst" search method searches the space of attribute subsets by greedy hill-climbing augmented with a backtracking facility.

3.8 Machine learning

Machine learning (ML) explores the development and study of algorithms that can artificially learn from existential data (in a certain realm) and based on the learning predict outcomes on unseen data (in the same realm) without being explicitly programmed to do so.

In our study, after formulating ground truth, preprocessing the data, adding new features, and extracting an optimal set of features to work with, we applied ML algorithms to develop a binary classifier. Although, performance of ML algorithms is largely dependent on the nature of datasets that are worked upon, but existing literature [26, 32, 37-39] claims that support vector machine (SVM) is the better algorithm for text classification. We explore the performance of SVM along with other ML algorithms for our classification of tweets.

CHAPTER 4 EXPERIMENTAL DESIGN AND RESULTS

4.1 Design of experiments

We worked with 5 datasets from our tweet repository where each dataset comprises of tweets belonging to each of these states – Alabama, Arizona, Minnesota, Pennsylvania and Oklahoma, respectively. We manually classified 4,790 of these tweets to serve as the ground truth and for each dataset we divide tweets labelled as "Has Agenda" and "No Agenda" into respective equal subsets. We used one subset of each class for training the binary classifier using machine learning techniques and the other subset for testing its performance.

We aim to:

- find the best performing classification algorithm for our datasets
- employ combination of processing steps discussed in Section 3.5 through Section 3.8 to gauge the impact of these steps in the classification process
- and finally, to establish whether or not our hand-crafted features enhance the classifier's performance

We refer to these different combination of processing steps as a Method Type and details are shown in Table 4.1 (underlined steps in the table are a part of feature extraction step).

Table 4.1. Method types based on combination of processing steps

Method Type	**Combination of processing steps**
DT	Data Cleansing + Term Occurrence
DTF	Data Cleansing + Term Occurrence + Feature Selection
SDTF	Stemming + Data Cleansing + Term Occurrence + Feature Selection
SDTTF	Stemming+ Data Cleansing+ Term Occurrence + Twitter Features + Feature Selection
SDTTAF	Stemming + Data Cleansing + Term Occurrence + Twitter Features + Agenda Topic Keywords + Feature Selection

DT applies the Data Cleansing technique and use Term Occurrence as the feature set. Learning about successful machine learning algorithms in text classification from our literature survey, we chose to analyze performance of three algorithms on our dataset namely, (a) Decision Trees (J48), (b) Naïve Bayes, and (c) Support Vector Machine (SVM) with linear kernel. We used Weka 3.6 [49], an open source data mining software in Java to employ aforementioned algorithms and perform 5-fold cross validation on our training set.

Among these 3 algorithms, SVM performed the best for four out of five datasets, hence we used it for the remainder Method Types. Using Weka, we employ CV Parameter Selection algorithm on SVM to find the optimal values for two parameters: Cost (C) and Kernel function. For all five datasets optimal values were C = 1 and Kernel function = Linear Kernel. In the rest of the study, we use this algorithm and SVM refers to SVM with linear kernel and C=1.

DTF uses Attribute Selected classifier and feature selection algorithms as discussed in Section 3.6. We use SVM on the reduced feature set and analyze its performance on all five datasets. SDTF applies the K-Stem algorithm first for stemming and then follow the steps as in Method 2.

SDTTF differs from SDTF by not limiting the feature extraction step to Term Occurrence and it uses Twitter specific features: HasHashtag, HasURL, and TweetStrength. As discussed in Section 3.6, SDTTAF adds HasTopicKeyword as another feature in this feature extraction step. Table 4.2 shows dimensionality of our dataset, and Table 4.3 shows the number of training instances used in this study and the reduced length of feature vectors after various processing steps.

Table 4.2. Count of total tweets in our datasets

Dataset Name	Total Number of Tweets
Minnesota	953
Alabama	938
Oklahoma	364
Pennsylvania	800
Arizona	1735

Table 4.3. Count of training instances and length of feature vector in our datasets

Dataset Name	Training Instances	Total number of features in a feature vector		
		DTF	**SDTF**	**SDTTAF**
Alabama	468	1404	1216	1219
Arizona	850	1890	1657	1660
Minnesota	476	1729	1506	1509
Oklahoma	183	862	783	786
Pennsylvania	413	1139	1004	1007

We use the common metrics of recall and precision to evaluate our classifier's performance. In pattern recognition and information retrieval with binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved [50].

4.2 Findings

In the subsequent sections, we discuss the findings of our study. The first subsection includes the observations on characteristics of the tweets that we collected. In our second subsection, we share insights on the best performing machine learning algorithm for our classification problem and report the features from the feature extraction step (described in Section 3.6) that contribute the most in enhancing our classification performance.

4.2.1 Characteristics of political agendas tweets

As mentioned in Section 3.3, we articulate a list of user handles for which we extracted the tweets. Post the tweets collection phase, we found that Senate handles from states like Washington, Colorado, Florida, North Carolina and Alaska tweet heavily; Senate handles from states like Oklahoma, North Dakota use Twitter very judiciously. Senate handles from states like Arkansas, Georgia, Texas, Wyoming and Vermont, are yet to enter the online social media platform. Figure

4.1 depicts an assessment of the penetration of the microblogging platform into the Senate of states in USA.

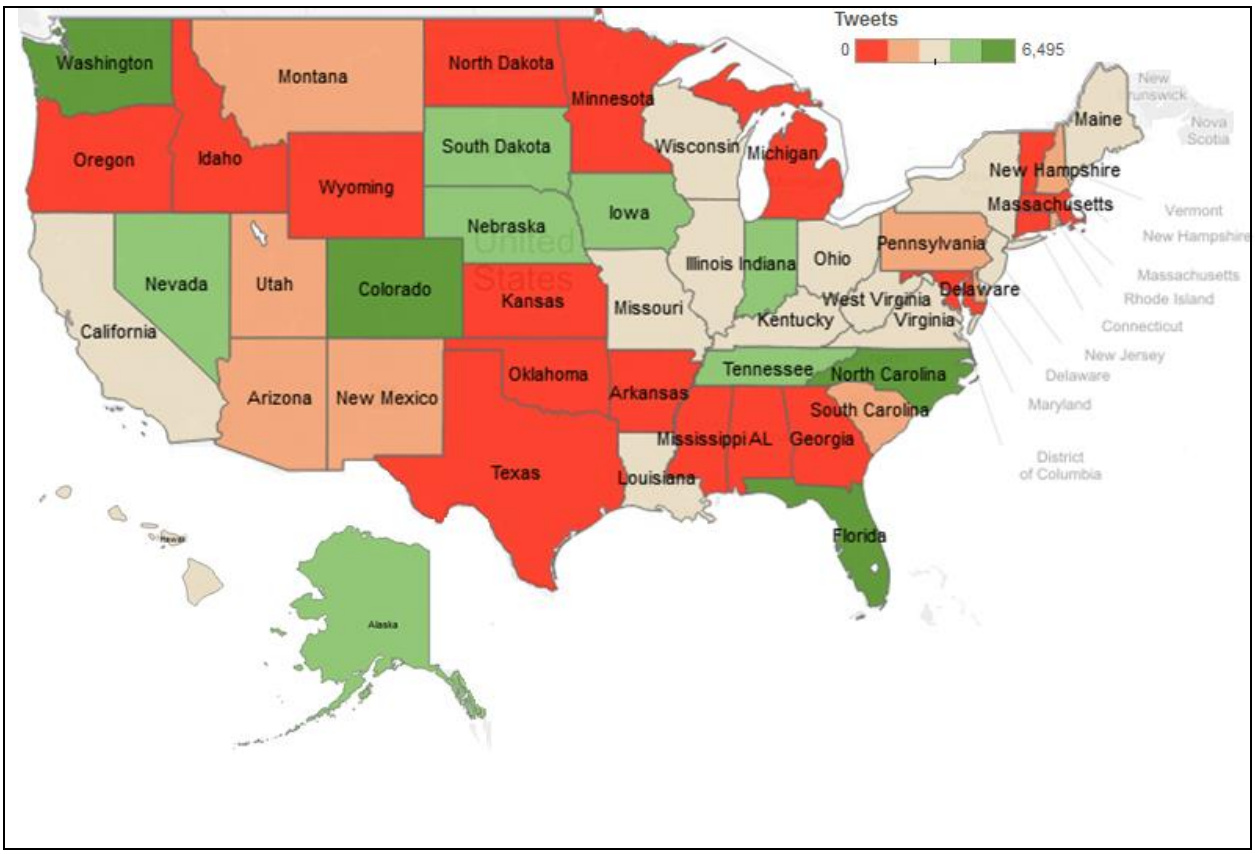

Figure 4.1. Heat map of number of tweets posted from the Senate handles in the US as of December 16, 2014

From the collected tweets, we observed that the social media medium was used for purposes of telling the masses about - bills that are proposed, bills being discussed on the floor, issues that concern the state, new senators that join the Senate, the details of Senators' public appearances, alerts as to when the session goes into recess, resumes or adjourns, the strategies that get condemned by the minority, and for sending wishes on special occasions to its constituents.

Also observed was that two states could have an entirely contrasting pattern of tweeting and usage of words. For instance, tweets by one of the Arizona Senate's handle were very formal and followed a strict pattern of stating the bill number that was being read in the Senate and 2-3 keywords that best related to it. Whereas Alabama followed no such pattern and informally mentioned the reforms they would want to bring into the state. The Alabama Senate handle made

maximal use of the platform to apprise its constituents of the media appearances of its senators and other such events held with the mainstream media.

A common trait that was observed for majority of tweets was that the usage of misspelt words, internet slang and political jargon was kept to the minimum. Also common was a frequent use of hashtags and of attaching shortened URLs (with the help of shortening devices like – bitly [51]and TinyURL [52]).

As discussed in Section 3.4, we worked on formulating the ground truth for our proposed approach. Post manual labelling, we analyzed the distribution of tweets between the two classes for each state. As can be seen in Figure 4.2, Arizona has the highest percentage of tweets with a Policy Agenda at 76%, with Pennsylvania as a close second at 74%. Minnesota has the least percentage of tweets with Policy Agenda at 25%.

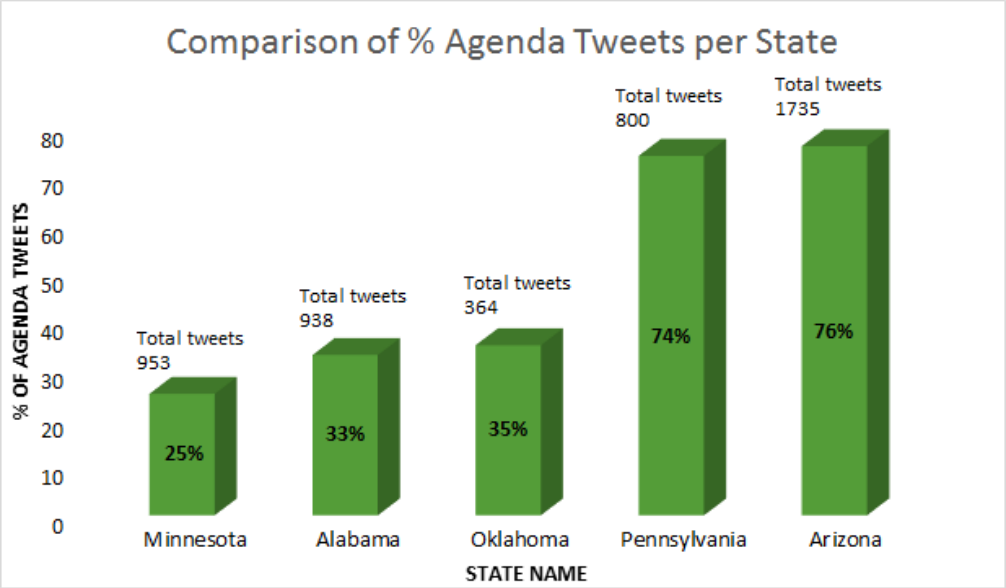

Figure 4.2. Comparison of count of tweets labeled as "Has Agenda" per state

4.2.2 Classification results

We mainly worked with three machine learning algorithms, (a) Support Vector Machine (SVM) using linear kernel and Gaussian kernel, (b) Decision Trees (J48), and (c) Naïve Bayes. SVM with Gaussian kernel did not yield satisfactory results whereas SVM with linear kernel

yielded the best overall classification performance on our 5 datasets. SVM yielded adequate results on states which had at least 465 tweets in the training set and 1,215 features after stemming. In case of Oklahoma, the dataset was small with only 183 tweets available for training and the number of features after stemming at 783. In such a scenario, the Naïve Bayes algorithm performed best in comparison to the other algorithms. Table 4.4 compares the recall of these algorithms on our 5 datasets.

Table 4.4. Performance comparison of machine learning algorithms on 5 independent datasets for Has Agenda class

	Recall for Has Agenda class		
State Name	**Naïve Bayes**	**Decision Tree (J48)**	**SVM with linear kernel**
Minnesota	0.109	0.773	0.782
Alabama	0.431	0.549	0.822
Oklahoma	0.641	0.219	0.469
Pennsylvania	0.907	0.923	0.891
Arizona	0.734	0.207	0.981

As discussed in Section 3.6, we extracted four new features from the dataset (3 Twitter specific and 1 Policy Agenda specific). As discussed in Section 3.7, we implemented "Attribute Selection" algorithms in Weka to find the optimal set of features that yielded best classification results. We found that all these hand-crafted features appeared among the top fifty features with highest information gain. This finding was consistent across all datasets wherein each dataset consisted of 1,400 features each on an average. Based on information gain, Table 4.5 lists top eight features of all datasets and the feature – "TweetStrength" consistently showed up in all 5 datasets.

Table 4.5. Top eight features of each state with highest information gain

Minnesota	Alabama	Oklahoma	Pennsylvania	Arizona
mnsure	Act	TweetStrength	Senate	Pass
tax	accountable	HasURL	TweetStrength	vote
#mnsots	budget	ivester	recess	TweetStrength
TweetStrength	TweetStrength	HasHashtag	HasHashtag	HasURL
politician	recess	approve	#voterid	hb
HasTopicKeyword	HasTopicKeyword	oeta	people	HasTopicKeyword
higher	@ltgovivey	education	amendment	renewable
obamacare	gun	bond	economic	final

Besides using the three machine learning algorithms, we also worked with Cost Sensitive Classifier [53] to be able to achieve high recall of tweets that have a Policy Agenda. Attaining a high recall for "Has Agenda" class was our top priority because the future work of our study envisions extracting specific policy agenda major topics for each state. Although we achieved a high recall of 82.3% for Has Agenda class but an attempt to improve it further drastically affected the precision. We wanted to keep the precision for both the classes at least 70% hence we did not investigate further along this direction.

Table 4.1 illustrates description of several Method Types. The performance metrics for all these Method Types averaged over all datasets are reported in Table 4.6. As illustrated in Fig 4.3, the recall for "Has Agenda" class follows a non-decreasing curve corresponding to application of each Method Type. We report that attribute selection did not improve the performance of the classifier. With the help of SVM and feature extraction techniques the best recall achieved for the "Has Agenda" class was 0.891 and the worst recall achieved for the same class was .769.

Table 4.6. Performance metrics of the binary classifier using different method types

Method	Recall (Has Agenda)	Recall (No Agenda)	Precision (Has Agenda)	Precision (No Agenda)
DT	0.769	0.769	0.769	0.834
DTF	0.769	0.878	0.802	0.839
SDTF	0.823	0.822	0.748	0.839
SDTTF	0.859	0.835	0.763	0.868
SDTTAF	0.891*	0.836	0.772	0.873

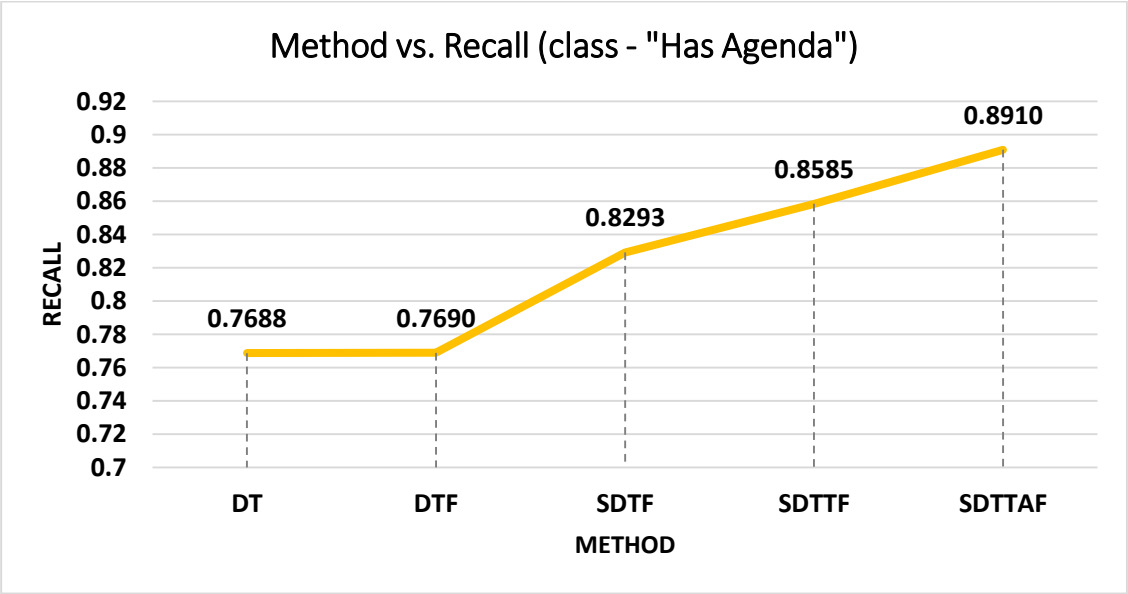

Figure 4.3. Impact of data processing techniques on recall for class -"Has Agenda"

In order to validate that our findings of adding features and using topic keywords actually enhanced the classifiers performance and was not just a mere coincidence, we performed paired t-tests on the recall values for these four states – Alabama, Arizona, Minnesota, Pennsylvania. We omit Oklahoma because the best performing algorithm for this dataset was Naïve Bayes and not SVM like the other states. We verify that the difference between the recall values for these states before and after applying feature extraction is normally distributed by using the Shapiro-Wilk Normality test [54]. The paired t-tests executed using R software [55] confirms that the difference between the mean value of "Recall" for class "Has-Agenda" before and after feature extraction is statistically significant at $\alpha = 0.05$. The overall gain in recall was that of 5.187%.

4.3 Limitation of the study

In labeling the ground truth, each dataset was manually coded by one person. It is possible that different human coders may label the same tweet differently. However, we do not expect high inter-coder disagreement because each human coder was briefed about the coding guideline; the guideline is clearly presented in the web application used for ground truth labeling. Secondly, we did not consider images embedded in tweets, tweet's spatial and temporal attributes, and number of times a tweet has been retweeted. Including these features in our hand-crafted twitter specific features may impact the classifier's performance.

CHAPTER 5 CONCLUSIONS AND FUTURE WORK

Twitter has emerged as a new medium for politics. No prior work has studied this medium for state politics, perhaps due to the large volume of the data and lack of effective methods to process them. Official tweets from State Senate and House may contain information regarding policy topics under discussion which possess a high chance of being formulated into a bill. Hence, collecting and analyzing this type of data adds a new dimension to the study of state politics.

We collected 122,965 tweets from various State Legislature handles of 50 states. Our finding from manual inspection of over 4500 tweets from 5 states shows a large variation on the percentage of agenda tweets (those with reference to Policy Agendas topics) between 25% (Minnesota) and 76% (Arizona). Since manual analysis is time consuming and does not allow for the study to cover 50 states, we proposed an approach to automate the classification of tweets. This approach uses our Twitter specific and policy agenda specific features that were shown to significantly enhance the classification performance. Among several supervised machine learning algorithms investigated in this study, SVM with linear kernel gave the best performance with 89.1% recall of agenda tweets.

Our future work includes (a) automatic detection of which policy agenda topic and subtopic is referred to in the tweet, (b) study of policy diffusion among states by analyzing spatio-temporal characteristics of these topics/subtopics of states in the US, and (c) hypothesis testing of whether or not states learn policy-making from each other.

BIBLIOGRAPHY

[1] *PI-Net - Poliinformatics*. Available: http://poliinformatics.org/

[2] *Government Data*. Available: https://www.data.gov/

[3] The data used here were originally collected by Frank R. Baumgartner and Bryan D. Jones, with the support of National Science Foundation grant numbers SBR 9320922 and 0111611, and were distributed through the Department of Government at the University of Texas at Austin. Neither NSF nor the original collectors of the data bear any responsibility for the analysis reported here. [Online].

[4] *About: Twitter*. Available: https://about.twitter.com/company

[5] *Twitter Statistics*. Available: http://www.statista.com/statistics/273172/twitter-accounts-with-the-most-followers-worldwide/

[6] A. Jungherr. (2014). *Twitter in politics: a comprehensive literature review*. Available: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2402443

[7] B. D. Jones, H. F. Thomas III, and M. Wolfe, "The Role of the Media in Policy Bubbles," *Talk at the Department of Communication, Stanford University.* *http://www.policyagendas.org/document/jones-role-media-policy-bubbles-talk-stanford,* 2013.

[8] B. D. Jones, "The dynamics of agenda expansion and contraction in the US," *Perspectives on Europe,* vol. 42, pp. 22-28, 2012.

[9] K. M. Quinn, B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev, "An automated method of topic-coding legislative speech over time with application to the 105th-108th US senate," in *Midwest Political Science Association Meeting*, 2006, pp. 1-61.

[10] L. Hemphill, J. Otterbacher, and M. Shapiro, "What's congress doing on twitter?," in *Proceedings of the 2013 conference on Computer supported cooperative work*, San Antonio, TX, USA, 2013, pp. 877-886.

[11] L. Hemphill and A. J. Roback, "Tweet acts: how constituents lobby congress via Twitter," in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, Baltimore, MD, USA, 2014, pp. 1200-1210.

[12] A. Bakliwal, J. Foster, J. van der Puil, R. O'Brien, L. Tounsi, and M. Hughes, "Sentiment analysis of political tweets: Towards an accurate classifier," in *Proceedings of the NAACL Workshop on Language Analysis in Social Media, Association for Computational Linguistics*, Sofia, Bulgaria, 2013.

[13] T. Mullen and R. Malouf, "A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse," in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, Palo Alto, California, 2006, pp. 159-162.

[14] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," *ICWSM,* vol. 10, pp. 178-185, 2010.

[15] K. Y. Kamath, J. Caverlee, K. Lee, and Z. Cheng, "Spatio-temporal dynamics of online memes: a study of geo-tagged tweets," in *Proceedings of the 22nd international conference on World Wide Web*, Rio de Janeiro, Brazil, 2013, pp. 667-678.

[16] Z. Cheng, J. Caverlee, and K. Lee, "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*, Toronto , ON, Canada, 2010, pp. 759-768.

[17] K. Lee, J. Caverlee, Z. Cheng, and D. Z. Sui, "Content-driven detection of campaigns in social media," in *Proceedings of the 20th ACM Int'l Conf. on Information and Knowledge Management*, Glasgow, Scotland, UK, 2011, pp. 551-556.

[18] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: quantifying influence on twitter," in *Proceedings of the fourth ACM international conference on Web search and data mining*, Hong Kong, 2011, pp. 65-74.

[19] T. Bodnar and M. Salathé, "Validating models for disease detection using twitter," in *Proceedings of the 22nd international conference on World Wide Web companion*, Rio de Janeiro, Brazil, 2013, pp. 699-702.

[20] M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko, and K.-L. Ma, "Breaking news on twitter," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Austin, TX, USA 2012, pp. 2751-2754.

[21] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy," in *Proceedings of the 22nd international conference on World Wide Web companion*, Rio de Janeiro, Brazil, 2013, pp. 729-736.

[22] J. Bollen, A. Pepe, and H. Mao, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, Barcelona, Spain, 2011.

[23] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, Portland, Oregon, USA, 2011, pp. 151-160.

[24] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, Stroudsburg, PA, USA, 2002, pp. 79-86.

[25] S. Jahid, S. Nilizadeh, P. Mittal, N. Borisov, and A. Kapadia, "DECENT: A decentralized architecture for enforcing privacy in online social networks," in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*, Lugano, Switzerland, 2012, pp. 326-332.

[26] H. Mao, X. Shuai, and A. Kapadia, "Loose tweets: an analysis of privacy leaks on twitter," in *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, Chicago, IL, USA, 2011, pp. 1-12.

[27] S. Patil, Y. Le Gall, A. J. Lee, and A. Kapadia, "My privacy policy: exploring end-user specification of free-form location access rules," in *Financial Cryptography and Data Security*, ed: Springer, 2012, pp. 86-97.

[28] S. Patil, G. Norcie, A. Kapadia, and A. J. Lee, "Reasons, rewards, regrets: privacy considerations in location sharing as an interactive practice," in *Proceedings of the Eighth Symposium on Usable Privacy and Security*, Washington, DC, USA, 2012, p. 5.

[29] R. Schlegel, A. Kapadia, and A. J. Lee, "Eyeing your exposure: quantifying and controlling information sharing for improved privacy," in *Proceedings of the Seventh Symposium on Usable Privacy and Security*, Pittsburgh, PA, USA, 2011, p. 14.

[30] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces*, Alicante, Spain, 2011, pp. 169-176.

[31] L. Lloyd, D. Kechagias, and S. Skiena, "Lydia: A system for large-scale news analysis," in *String Processing and Information Retrieval*, ed: Springer Berlin Heidelberg, 2005, pp. 161-166.

[32] S. Purpura and D. Hillard, "Automated classification of congressional legislation," in *Proceedings of the 2006 international conference on Digital government research*, San Diego, CA, USA, 2006, pp. 219-225.

[33] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," *Journal of the American society for information science and technology,* vol. 60, pp. 2169-2188, 2009.

[34] Y.-R. Lin, D. Margolin, B. Keegan, and D. Lazer, "Voices of victory: A computational focus group framework for tracking opinion shift in real time," in *Proceedings of the 22nd international conference on World Wide Web*, Rio de Janeiro, Brazil, 2013, pp. 737-748.
[35] K. Y. Kamath, A.-M. Popescu, and J. Caverlee, "Board Recommendation in Pinterest," in *UMAP Workshops*, Aalborg, Denmark, 2013.
[36] B. Krishnamurthy and C. E. Wills, "On the leakage of personally identifiable information via online social networks," in *Proceedings of the 2nd ACM workshop on Online social networks*, Spain, Barcelona, 2009, pp. 7-12.
[37] C. C. Aggarwal and C. Zhai, "A survey of text classification algorithms," in *Mining text data*, ed: Springer, 2012, pp. 163-222.
[38] M. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text classification using machine learning techniques," *WSEAS Transactions on Computers,* vol. 4, pp. 966-974, 2005.
[39] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR),* vol. 34, pp. 1-47, 2002.
[40] E. S. Adler and J. Wilkerson, "Congressional Bills Project," *NSF,* vol. 880066, p. 00880061, 2006.
[41] O. O. İkiz, M. Z. Sobaci, N. Yavuz, and N. Karkin, "Political use of twitter: the case of metropolitan mayor candidates in 2014 local elections in Turkey," in *Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance*, Guimaraes, Portugal, 2014, pp. 41-50.
[42] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics,* vol. 37, pp. 267-307, 2011.
[43] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the conference on human language technology and empirical methods in natural language processing*, Vancouver, Canada, 2005, pp. 347-354.
[44] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, "The development and psychometric properties of LIWC2007," ed. Austin, TX, 2007.
[45] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, "Syntactic clustering of the web," *Computer Networks and ISDN Systems,* vol. 29, pp. 1157-1166, 1997.
[46] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1 (2009) 11," *Affendey, LS, Paris, IHM, Mustapha, N., Sulaiman, MN, Muda, Z.: Ranking of Influencing Factors in Predicting Student Academic Performance. Information Technology Journal,* vol. 9, pp. 832-837, 2010.
[47] *Stemming Algorithms*. Available: http://www.elastic.co/guide/en/elasticsearch/guide/master/choosing-a-stemmer.html
[48] A. Wiese, V. Ho, and E. Hill, "A comparison of stemmers on source code identifiers for software search," in *Software Maintenance (ICSM), 2011 27th IEEE International Conference on*, Williamsburg, VA,USA, 2011, pp. 496-499.
[49] *Weka 3.6*. Available: http://www.cs.waikato.ac.nz/ml/index.html
[50] *Precision and Recall*. Available: http://en.wikipedia.org/wiki/Precision_and_recall
[51] *Bitly*. Available: https://bitly.com/
[52] *TinyURL*. Available: http://tinyurl.com/
[53] *CVParameter Selection*. Available: https://weka.wikispaces.com/Optimizing+parameters
[54] *Shapiro-Wilk Normality Test*. Available: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/shapiro.test.html
[55] R. C. Team, "R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012," ed: ISBN 3-900051-07-0, 2012.