NUMBERS IN POLITICS: COMPARATIVE QUANTITATIVE ANALYSIS & MODELING IN FOREIGN POLICY ORIENTATION AND ELECTION FORECASTING

A Master ‘s Thesis

by

ENES TAYLAN

Department of
International Relations
İhsan Doğramacı Bilkent University
Ankara
May 2017

To Ş.Y.

NUMBERS IN POLITICS: COMPARATIVE QUANTITATIVE ANALYSIS & MODELING IN FOREIGN POLICY ORIENTATION AND ELECTION FORECASTING

Graduate School of Economics and Social Sciences
of
İhsan Doğramacı Bilkent University

by

ENES TAYLAN

In Partial Fulfillment of the Requirements for the Degree of
MASTER OF ARTS

THE DEPARTMENT OF
INTERNATIONAL RELATIONS
İHSAN DOĞRAMACI BILKENT UNIVERSITY
ANKARA
May 2017

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in International Relations.

Assistant Prof. Dr. İ. Özgür Özdamar
Supervisor

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in International Relations.

Prof. Dr. Halil Altay Güvenir
Co-Supervisor

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in International Relations.

Assistant Prof. Dr. Seçkin Köstem
Examining Committee Member

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in International Relations.

Assistant Prof. Dr. Nihat Ali Özcan
Examining Committee Member

Approval of the Institute of Economics and Social Sciences

Prof. Dr. Halime Demirkan
Director

ABSTRACT

NUMBERS IN POLITICS: COMPARATIVE QUANTITATIVE ANALYSIS & MODELING IN FOREIGN POLICY ORIENTATION AND ELECTION FORECASTING

Taylan, Enes

MA, Department of International Relations

Supervisor: Assist. Prof. Dr. Özgür Özdamar

Co-Supervisor: Prof. Dr. Altay Güvenir

May 2017

To advance social science in the direction of accurate and reliable quantitative models, especially in the fields of International Relations and Political Science, new novel methodologies borrowed from the Computer Science and Statistics should be employed. In International Relations, quantitative analysis can be carried out to understand foreign policy topic relations in public discourse of decision makers. In domestic politics, Election Forecasting is a suitable area, because of its offering of already quantified vote results and its importance in the decision-making process in domestic politics and foreign policy. This work embarks upon a computational-

statistical model built on social media (Twitter) data and texts' meaning extracted, analyzed and modeled with the state-of-the-art methodologies from Computer Science (Machine Learning and Natural Language Processing) and statistics to forecast election results and foreign policy orientation. To verify the model, Turkish General Election 2015, US Presidential Election 2016 and campaign period of Donald Trump are analyzed. This work shows that, sentiment of political tweets can be captured with high predictive accuracy (92% in Turkish, 96% in English) and using opinion poll results for a given period of time, vote percentage fluctuations can be predicted. Furthermore, it is possible to capture the foreign policy orientation of a candidate by his and his team's tweets in the campaign period.

Keywords: Election Forecasting, Foreign Policy Analysis, Machine Learning, Natural Language Processing, Sentiment Analysis

ÖZET

POLİTİKADA SAYILAR: DIŞ POLİTİKA YÖNELİMİ VE SEÇİM TAHMİNİNDE KARŞILAŞTIRMALI SAYISAL ANALİZ VE MODELLEME

Taylan, Enes

Yüksek Lisans, Uluslararası İlişkiler Bölümü

Tez Danışmanı: Yrd. Doç. Dr. İbrahim Özgür Özdamar

2. Tez Danışmanı: Prof. Dr. Halil Altay Güvenir

Mayıs 2017

Sosyal bilimleri, özellikle Uluslararası İlişkiler ve Siyaset Bilimi alanlarında, doğru ve güvenilir niceliksel modeller doğrultusunda ilerletmek için, Bilgisayar Bilimleri ve İstatistik'ten alınan yeni metodolojiler kullanılmalıdır. Uluslararası İlişkiler'de, karar alıcıların kamusal söhylemlerinde, dış politika konuları arasındaki ilişkileri anlamak için niceliksel analiz yapılabilir. İç siyasette, Seçim Tahmini, nicelenmiş oy sonuçlarının varlığı ve iç politika ve dış politika karar alma sürecindeki öneminden dolayı uygun bir alandır. Bu çalışma, seçim sonuçlarını ve dış politika

yönelimlerini, istatistikten ve bilgisayar bilimlerinden (Makine Öğrenimi ve Doğal Dil İşleme) en yeni metodolojilerle analiz edilen ve modellenen, sosyal medya (Twitter) verileri ve metinlerin anlamı üzerine kurulu sayısal bir modelle tahmin etmektedir. Modeli doğrulamak için, 2015 Türkiye Genel Seçimi, 2016 ABD Başkanlık Seçimi ve Donald Trump'in seçim kampanyası dönemi analiz edildi. Bu çalışma, siyasi yönelimlerin yüksek doğrulukla (%92 Türkçe, %96 İngilizce) yakalanabildiğini ve anket sonuçlarındaki dalgalanmaların tahmin edilebileceğini gösteriyor. Ayrıca, bir adayın dış politika yönelimlerinin, kendisinin ve ekibinin kampanya döneminde attığı tweetlerle yakalanabildiği gösterilmiştir.

Anahtar Kelimeler: Dış Politika Analizi, Doğal Dil Işleme, Duygu Analizi, Makine Öğrenmesi, Seçim Tahmini

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1:

INTRODUCTION

Social science research community has a growing interest in the direction of quantitative models. Besides economics, there is a lack of reliable and accurate models in social science fields, and this is certainly true in International Relations and Political Science. There are several reasons behind it, but an important factor is the nature of the Social Sciences, it is difficult to quantify information that can be used in the state-of-the-art modeling techniques employed in natural sciences.

Today, we have much more data available than the social scientists in the past and much more opportunities to quantify and analyze social content and interactions. For example, hundreds of millions of people express their opinion freely and publicly in social media channels and these data and interactions are available to researches as raw data, regressable and analyzable linguistic features or network structures. With Machine Learning and Natural Language Processing algorithms, sentiment polarity of a post of a citizen or a policy decision maker can be analyzed probabilistically. Also, a decision maker's network structure can be analyzed through her connections in a

social network to understand that decision maker's policy orientation through not only her own data but with the data of her close environment as a whole.

Election forecasting is an interesting area which can be modeled using advanced data techniques due to availability of dependent and independent certain numerical measures (election results or opinion polls) that social scientist can crunch to produce reliable, accurate models and therefore predictions. Availability of social media medium data offers the chance of capturing voting intention on a very large scale, for example with millions of tweets from Twitter.

Being able to build quantitative models in elections is academically and practically valuable not just for its usefulness in the area of election forecasting but also for the advancement of social sciences in general, in terms of its reliability, accuracy and methodological plurality.

In International Relations, public opinion is a key element to scholarly analysis especially in some theoretical approaches and schools of thought. This work stipulates that decision making in international arena and conduct of foreign policy are directly dependent on the key people in the governments, which are elected by electorates' votes in ballot box in democratic countries, and their strategic interactions. Therefore, leaders and governments those leaders head are very important parameters in deciding how international relations evolve in time for the research community and politicians. Also, after the election and key decision makers assume take their governmental positions, public opinion continues to exert its influence on decision making process. Decision makers, because of the audience cost and their prospective success in the next elections, take into account how public reacts to their conduct of

domestic and foreign policy. Therefore, being able to predict and forecast which political party (or parties) will run a country or which leader will be the next president, or more generally keeping track of the public opinion are fundamental to an accurate foreign policy analysis. To achieve the research objectives, this work focuses on both sides of the international - domestic nested game, the mutual dependency between foreign and domestic policies.

Domestic Politics: Election Forecasting through citizens voting intentions of 1 November 2015 Turkish general election and US presidential primary elections in 2016 using Twitter data

International Relations: Foreign Policy Topic Analysis through foreign policy orientation of US presidential election winner, president-elect Donald Trump's campaign teams' Twitter data during the campaign season

1.1 Domestic Politics, Election Forecasting

Current election forecasting methodologies differ in terms of their data and their emphasis on qualitative or quantitative models. There are four major types of election forecasting approaches according to (Lewis-Beck and Stegmaier, 2014):

1. Structuralist: Static, usually single equation models
2. Aggregative: Aggregate results from polls
3. Expertise on the domain: Depends on expert qualitative knowledge and intuition
4. Synthesizers: Combining polls data with structural features

We can define social media analysts as the fifth type which captures voting intentions of electorates from social media data using Facebook, Twitter, etc. This thesis does social media analysis and proposes a novel approach which combines Natural Language Processing and Machine Learning approaches to determine the voting intentions of electorate probabilistically in a statistically confident interval. The cases for election forecast methodology are Turkish 1 November 2015 general election and US primary elections in 2016.

1.2 International Relations, President-Elect D. Trump's Policy Orientation

In the US Presidential Election of November 8, 2016, Trump was elected the president for 2017-2021 period. By assuming campaign team members hold key positions in administrative and governmental positions in successive presidencies, this work analyzes the twitter timelines of key members of Trump campaign team and Trump himself, in several important foreign policy issues for USA:

- Russia
- China
- Syria – Iraq - Terrorism – ISIS
- Iran

1.3 Main Research Objectives and Questions

1) Can we create a quantified model which can capture the voting intentions of electorate and over performs current methodologies used in the research community? How reliable are these current applied methodologies?

2) Can we create a model capable of explaining a decision maker's policy orientation on specific issues during the campaign through their own twitter posts and network structure?

To be able to answer the questions given above from a collect ion of twitter data, we should be able to capture a tweet's meaning in terms of sentimental polarity (positive vs. negative) for a specific policy issue accurately in a probabilistic model. Therefore, our third research question is:

3) In how much accuracy a tweet's sentimental polarity (positive vs. negative) can be captured?

1.4 Research Approach

This work assumes social science issues can be analyzed and predicted by increasingly advanced computational and mathematical approaches applied on relevant datasets. Therefore, it focuses on machine learning, natural language processing, statistical algorithms and approaches to build models for forecasting elections and foreign policy choices.

1.5 Thesis Outline

Because this work is transdisciplinary and makes predictive analysis for international relations - foreign policy and to achieve this aim creates models using machine learning and natural processing from computer science, first, an extensive review of related literatures from these fields are provided. Methodological choices are represented afterwards. Because sentiment analysis is at the core of the underlying model, results are discussed in detail especially for the Turkish sentiment polarity analysis to represent the process of algorithmic development. Therefore, English sentiment analysis will be given in less detail. After creating models, Turkish 2015 General Election and US 2016 Presidential Primaries will be analyzed and discussed. In the last case, Donald Trump's campaign team's discussion of several foreign policy issues will be discussed by analyzing his key campaign team members' tweets.

CHAPTER 2:

LITERATURE REVIEW

In this section, election forecasting literature, leadership public content analysis literature and related sentiment analysis literature will be reviewed.

2.1 Political Science Election Forecasting Literature

In the current literature, election forecasting can be categorized into five different labels: Structuralists, Aggregators, Synthesizers, Experts (Lewis-Beck and Stegmaier, 2014), and Social media analysts. Because experts rely on their private knowledge, intuition and not on modeling this work does not discuss on them.

The Structuralists (Abramowitz, 2012; Campbell, 2014; Lewis-Beck and Tien, 2014) present a theoretical model of the election outcome with generally core political and economic parameters such as GDP growth, incumbency or attrition amount a running party burdens. The unit of analysis is either nation or district, based on the available granularity of the data. Estimation is done through regression and resulting function is static rather than dynamic.

The Aggregators, on the other hand, (Blumenthal, 2014; Traugott, 2014), aggregate vote intentions in opinion polls like Real Clear Politics. Voter preferences are combined over multiple polls and weighted according to their past historic performance and recency. Aggregators do not offer a theory which means not causal streams but correlation among independent and dependent variables important. Here accuracy of the forecast is the key not underlying theory. Because aggregators have access to polls result over a period they have the ability to create time series and update their forecasts.

The Synthesizers (e.g. Bafumi, Erikson, et al., 2014; Linzer, 2014) combines methods of the Structuralists and the Aggregators. Their work begins with a political economy theory of the vote and use aggregated and ongoing polling preferences as well. Analysis may include multiple equations and done on national or district level.

Social media analysts, in comparison, use publicly available user intentions and opinions, create metrics and crunch on the numbers. These kinds of works claimed social media data allows a reliable forecast of the final result (Sang & Bos, 2012). Some of these works rely on very simple naïve techniques, focusing on the volume share of the data related to parties or candidates. Along the same line, some scholars claimed that the number of Facebook supporters could be a good indicator of election results (Williams & Gulati, 2008), while Tumasjan, et al., 2011) compared party mentions on Twitter with the results of the 2009 German election and argued that the relative number of tweets related to each party is a good predictor for its vote share.

Some scholar criticized this kind of mere volume works and argued number of mentions or retweets, or the number of "like"s are crude ways of trying to forecast future (Gayo Avallo, et al., 2011). Some studies tried to improve this simple analysis by conducting sentiment analysis. Lindsay (2008), for example, built a sentiment classifier based on lexical induction and found correlations between several polls conducted during the 2008 presidential election and the content of wall posts available on Facebook. (O'Connor, et al., 2010) show similar results displaying correlation between Obama's approval rate and the sentiment expressed by Twitter users. In addition, sentiment analysis of tweets proved to perform as well as polls in predicting the results of both the 2011 (Sang, et al., 2012) and the 2012 legislative elections in the Netherlands (Sanders, den Bosch, 2013), while the analysis of multiple social media (Facebook, Twitter, Google, and YouTube) was able to outperform traditional surveys in estimating the results of the 2010 U.K. Election (Franch, 2013).

However, other scholar argued that because just successful works are published, predictions from social media should not be taken reliable (GayoAvello, et al., 2011; Goldstein & Rainey, 2010; Huberty, 2015). For instance, it has been shown that the share of campaign weblogs prior to the 2005 federal election in Germany was not a good predictor of the relative strength of the parties insofar as small parties were overrepresented (Albrecht, et al., 2007). In a study on Canadian elections, Jansen and Koop (2006) failed in estimating the positions of the two largest parties. Finally, Jugherr, Ju¨rgens, and Schoen (2012) criticized the work of Tumasjan et al. (2011), arguing that including the small German Pirate Party into the analysis would have yielded a negative effect on the accuracy of the predictions. Gayo-Avello (2011)

argues that several theoretical problems with predicting elections based on tweets. First, he stresses how several of the quoted works are not predictions at all, given that they generally present post hoc analysis after an election has already occurred. This also increases the chances that only good results are published, inflating the perceived ability of using social media to correctly forecast election. Second, he underlines the difficulty to catch the real meaning of the texts analyzed, given that political discourse is plagued with humor, double meanings, and sarcasm. Third, he highlights the risk of a spamming effect: Given the presence of rumors and misleading information, not all the Internet posts are necessary trustworthy. Finally, in most of the previous studies, demographics are neglected: Not every age, gender, social, or racial group is in fact equally represented in social media. This study responses to the criticisms made by Gayo-Avello in the limitations chapter.

There are also other papers in the Turkish social media analysis literature which do not focus on election forecast in general, but touches upon special cases in political matters in Turkey. Among this, there is Yenigun, G. E., 2013 that analyzes political mobilization and alliance structures in social networks. There is also reports by ORSAM like the one which works on tweets of terrorists in Turkey[1]. However, in current Turkish social media literature focusing on social media political matters, there is no study which works millions of tweets gathered via Twitter stream API and applies Machine Learning and Natural Language Processing algorithms.

[1] http://www.orsam.org.tr/files/Raporlar/205/205_ENG.pdf

2.2 International Relations FPA Through Decision Makers

This work focuses primarily on specific foreign policy issues discussed by the President Elect Donald Trump himself and his campaign team during the 2016 US Presidential Election by analyzing their tweets. There are "personality at distance" approaches in Foreign Policy Analysis literature, which, like this work, analyze leaders' speech or interview contents although they model personalities of leaders in general not their orientation for specific foreign policy issues.

In Foreign Policy Analysis, "personality at a distance" analytical approaches to determine leadership styles, in decision making process occupy an important place (Hermann, 1977). Leaders' interviews, speeches, behavior and past experiences, biographies and autobiographies are all analyzed to make this kind of analysis.

Of these several approaches, Operational Code (OpCode) ((George, 1969) and Leadership Trait Analysis (LTA) (Hermann, 1977). uses decision makers' speeches and answers to interview questions, respectively, as their working material. They are closely related, share some parameters and differ in some aspects. This work enhances prediction capabilities of both of these statistical, related approaches via new channels of information we have in today's world (social media data).

LTA and OpCode focuses on grand personality features, an approach which does not focus on specific foreign policy issues. Besides using twitter data of decision makers and their campaign teams as new sources, another novelty of this work is, it directly analyzes specific foreign policy issues instead of personalities. Because there

is no relevant literature focusing on these two aspects, above mentioned two most well-known approaches, LTA and OpCode will be reviewed briefly.

2.2.1 Leadership Traits Analysis

Leadership Traits Analysis, models leaders' propensities analyzing their answers to the interview questions and calculate values for the parameters: Control over Events, Need for Power, Conceptual Complexity, Self-Confidence, Task Orientation, Distrust of Others and In-Group Bias (Hermann, 1977). LTA focuses on interviews because it sees them more spontaneous than public speeches which may be written by leaders' advisers or speech writers (Hermann, 1977)

LTA uses verbs or other words (adjectives, pronouns etc.) to calculate the measures for the parameters. For accurate LTA analysis, at least 50 interview responses with at least 100 words are needed. These interviews should span through leaders' the term in the office and should be on variety of topics and spontaneity levels.

2.2.2 Operational Code

Operational Code first developed by Leites in 1951 in Rand and then redeveloped (George, 1969). It models typology and personality of leader (Walker, 1998). Operational Code Approach models leaders' predispositions by doing mathematical

calculations on the speech material of those decision makers and “asks what the individual knows, feels, and wants regarding the exercise of power in human affairs” (Schafer & Walker, 2006). To do so it focuses on two sets of variables sets, philosophical ones and instrumental ones (George, 1969) and (Holsti, Rosenau, 1979):

“P-1. the fundamental nature of politics, political conflict, and the image of the opponent;

P-2. the general prospects for realizing one's fundamental political values

P-3. the extent to which the political future is predictable

P-4. the extent to which political leaders can influence historical developments and control outcomes

P-5. the role of chance? What are the leader's instrumental propensities for choosing?

I-1. the best approach for selecting goals for political action, i.e., strategy

I-2. how such goals and objectives can be pursued most effectively, i.e., tactics

I-3. the best approach to calculation, control, and acceptance of the risks of political action

I-4. the "timing" of action"

1-5. the utility and role of different means?” (Walker, et al., 1998)

In calculating these propensities, our main parameters are leaders’ attributions to themselves (to their individual beings) and to the others, positively or negatively. To understand whether a verb (which is what we focus) is positive or negative, Verbs and Context System (VCIS) is used. Verbs in context system codes verbs according to verbs’ subject, time and category. It also looks at domain and the context.

2.3 Sentiment Analysis Literature

Sentiment Analysis, also known as Opinion Mining, is the use of Natural Language Processing, text analysis and computational linguistic techniques with the help of Data Mining and possibly Machine Learning to extract meaning from text data. Because of much more available public and free data on the Internet, sentiment Analysis has gained a lot of interest in the recent years.

Sentiment Analysis works have been done on a wide range of fields such as movie reviews (Pang et al., 2002), product reviews and news and blogs (Bautin et al., 2008) and they can be broadly categorized into three classes:

Polarity: Documents are categorized into positive, negative and neutral.

Emotion: This works focus on emotions or mood states expressed in documents, such as happiness, joy, surprise, among others.

Strength: As a cross-cutting feature, polarity and emotion based classification can be carried out with strength values attached, like numerical scores showing the level of positivity or joy.

Although in Sentiment Analysis, there are less categories than topic classification, it is considered as a harder field of Natural Language Processing because usually sentiment is represented on at least sentence level instead of topic classifiers dependence on words. Therefore, for a successful sentiment classification, it is necessary to analyze words in their order in domain also considering their potential expression with irony, sarcasm and negation. (Pang and Lee, 2008)

Although in polarity based sentiment analysis, texts can be categorized into subjective (positive, negative) and objective (neutral) classes in terms polarity, in Tweet sentiment analysis, most studies assume the subjectivity of the text, therefore focuses on positive and negative labels.

In terms of classification methodology, there exists two different paradigms: Sentiment Analysis (SA) with Lexicons [aka Knowledge Based] and SA with Machine Learning. The lexical approach utilizes a dictionary of words tagged with sentimental polarity; positive, negative or neutral. Some lexicons use has been built on word's emotional context such as happiness, excitement etc. Works using lexicons, compare words in their corpus to the words in the lexicons. sentiment meaning, then, is computed by counting positive and negative words in a sentence (or in a document, depending on the level of granularity of the classification) and if the number of positive words is greater than the number of negative ones, then that sentence is labeled is positive otherwise negative. Some lexicons have also strength value in a scale attached to the words. In this case, sentiment labeling of that sentence is done through the summation of weighted word polarity values by their strength value. It is also possible to use supervised machine learning for modeling sentence level aggregated polarity values by taking them features and annotated sentence polarities as the labels.

There are several lexicons used in sentiment analysis such as OpinionFinder (OPF), Affective Norms for English Words (ANEW), AFINN, SentiWordNet, WordNet, SentiStrength, NRC, and others. The problem with lexical sentiment analysis is the fact that it does not take into account of twitter jargon. Because words

used in Twitter do not adhere the rules of formal language, it is possible that many words in Twitter corpus escape the matching process between tweets and a lexicon. These types of words include Twitter slang, informal abbreviations, grammatical mistakes etc. Although with extensive preprocessing and stemming (in Turkish for example), this problem can be solved in some degree, this solution can't catch the sentiment meaning of that words in the context of their usage. To be able to capture the sentiment meaning of a word in relation to other words in a sentence (or a document), sentence (or document) level analysis is required.

The other paradigm in sentiment analysis, namely using machine learning, requires supervised and unsupervised approaches, however unsupervised algorithms only capture the sentiment meaning to some degree and is used for some auxiliary work such as topic classification by Latent Dirichlet allocation (LDA), Latent Semantic Analysis (LSA) or vector representation by Word2Vec. In supervised approach, on the other hand, labeled data is needed and that can be built by manual annotation or by distant supervised methods such as labeling from emoticons or hashtags. A commonly used method is labeling tweets containing ":)" as positive and ":(" as negative.

2.3.1 Sentiment Analysis in Twitter

Sentiment analysis of tweets are especially more difficult than long, informal texts such as blog posts because: (Martinez-Camara et al, 2014)

1) Tweets are informal in the sense that they have lots of abbreviations, internal slang and its has its own style of jargon.

2) Grammar of tweets is problematic. Many users do not care about the grammatical correctness of text. Twitter's policy of maximum text length of 140 plays important role here.

3) Users refer to the same concept via different words, abbreviations and irregular forms. Because of this, same concepts usually get represented by different words, which creates data sparsity problem. If not handled, this problem creates performance issues in sentiment analysis classifiers.

4) Because tweets are short, identifying their context is difficult conceptually and computationally, meaning a token's previous and successive tokens are few.

One of the earliest studies of tweet sentiment analysis was a classification of tweets, in English, by supervised machine learning algorithms (Go, Bhayani, Huang, 2009). Their work was on a dataset collected through Twitter Search API and labeled through the existence of emoticons (":" for positive, ":(" for negative). That work's results showed that, supervised algorithms are the best approach for sentiment analysis, POS (Part of Speech) tags are irrelevant in tweet classification and unigrams works quite well in short text as opposed to long documents. Furthermore, they found that predictive accuracy can be improved slightly by using a combination of unigrams and bigrams. Another work (Kim et al. 2009), uses not machine learning but uses the Affective Norms for English Words (ANEW) lexicon to capture emotion sentiment.

Pak and Paroubek (2010), in their work, annotates tweets in their dataset with emoticons for positive and negative labels. Also, they collect neutral tweets from the

official twitter accounts of newspapers and magazines. At the modeling step, they use supervised machine learning algorithms: Conditional Random Fields (CRF), Support Vector Machines and Naive Bayes. They conclude that n-grams, POS tags and Naive Bayes are the best the configuration for accurate classification.

Thelwall, Buckley and Paltoglou (2011) focuses on the intensity of tweet sentiments. For this purpose, they use SentiStrength lexicon which assigns a score for positivity and negativity on a strength scale of 1 to 5. Their conclusion is sentiment strength is a useful predictor for learning the behaviors of twitter users.

Khan et al. (2015) uses a hybrid approach and uses a machine learning algorithm, support vector machine, after assigning each tweet using the lexicons. They found that turning abbreviations into whole words, removal of links, POS tagging, deleting retweets and modeling with SVM results in accurate models.

Barbosa and Feng (2010), theorized that just using n-grams can decrease predictive power due to existence of large number of infrequent words in Twitter because of grammatical mistakes. They proposed, instead, inclusion of Twitter and microblogging specific features such as mentions, emoticons, punctuations, retweets and hashtags. In their work, using this enlarges set of features improved accuracy 2% in Support Vector Machine classification. A similar approach was advised by Kouloumpis, et al, (2011) with the addition of abbreviations and intensifier words as features. They showed that best performance in their work was achieved by a combination of prior probability of words, n-grams, Twitter specific and lexicon polarity features. However, POS tags decreased the predictive accuracy.

Davidov, Tsur and Rappoport (2010) uses hashtags and emoticons as instance features and models them with K-Nearest Neighbors (KNN) algorithm with promising results. In a comparative study by Agarwal et al. (2011), using unigrams, tree-based models, partial tree kernels and several combinations of them. They assign polarities to words in tweets using DAL dictionary. They conclude that, based on comparison of their results with other research, numerical word polarity based methods are superior to the others. Bifet and Frank (2010) uses data flow algorithms and Kappa evaluation measure on sentiment classification besides three machine learning algorithms, Multinomial Naive Bayes, Stochastic Gradient Descent (SGD) and Hoeffding tree. Their results show that accuracy of SGD is better than the others.

Hernández and Sallis (2011) uses an unsupervised algorithm Latent Dirichlet allocation (LDA) although they do not classify tweets based on sentiments. They just show that vector representation of tweets after LDA analysis and Tf-Idf weighting has less entropy therefore these methods can give better results when combined with classification algorithms. Aisopos et al. (2012) uses character n-grams to solve the problems with Twitter data mentioned above. Because on character level every text is independent of language, sarcasm and irony, grammatical errors and language style becomes irrelevant. They found their supervised model shows good performance. Martinez-Camara et al, (2014) gives a good summary of existing works on Twitter sentiment analysis with their accuracies when available.

CHAPTER 3:

BACKGROUND THEORY AND METHODOLOGY

3.1 Background Theory

This work, to forecast elections and foreign policy, creates models by building upon novel machine learning (ML) and natural language processing (NLP) algorithms and techniques. In this section, related background ML and NLP definitions, concepts, theories, algorithms and approaches are discussed.

3.1.1 Machine Learning

"The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience." (Mitchell, 1997)
As a shortly formalized version:
"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." (Mitchell, 1997)

Machine Learning algorithms are data driven, meaning they are algorithms which create other algorithms' steps through their dataset. For example, in this work's case: sentiment analysis is not done by writing the rules of extracting sentiment meaning from sentence but modeled through the dataset which is annotated for this work. The main aim of a machine learning algorithm is building a probabilistic model which in essence an algorithm.

Machine Learning algorithms has two main approaches in terms of learning style:

Supervised

In supervised machine learning, instances in the dataset are labeled to give directions to the algorithms. In this work's case, sentences in the dataset are labeled as positive or negative manually by the humans (or neutral which are filtered out before training stage).

Unsupervised

Unsupervised machine learning algorithms create models from a dataset of unlabeled instances. They do not need human assistance for classification. Main uses of unsupervised methods are clustering and feature engineering.

Supervised and unsupervised approaches can be mixed to create intermediate forms and they are called semi-supervised algorithms. In this work, classification through Doc2Vec, is that kind of machine learning in which first vector representations of sentences in the training dataset are built using Doc2Vec (internally

using Word2Vec on word level, another unsupervised algorithm) in unsupervised fashion, and then these vectors are fed into supervised algorithms.

3.1.1.1 Machine Learning Terminology

In this chapter, fundamental machine learning terminology and concepts will be defined.

Training data

The data we have (labeled or unlabeled), to feed into the machine learning algorithms. The algorithms learn models from the training data.

Test data

The models leaned by the algorithms are, tested against the test data. Test data and training data are assumed to be disjoint.

Instances and their features

Each member of the training and testing dataset is called an instance. So, for a supervised algorithm, a row in the training set is composed of instance with its label. The properties of an instance are its features.

Sparse Matrix

In machine learning, instances and their features are represented as matrices whose rows are instances and columns as features. In NLP, due to high number of features because of high number of words and misspellings, many (instance, feature) pairs are just empty in the matrix representation. Therefore, these empty pairs are dropped from the matrix to decrease memory usage. This representation is called the sparse matrix.

Classes

Classes, aka labels, are the categories which training instances belong to. In sentiment analysis, classes can be positive, negative, neutral, or some emotional categories such as happiness, excitement etc.

Table 1. Sentiment Examples

Instance	Class
Today is Monday.	Neutral
Weather is excellent today.	Positive
What a bad country.	Negative

Training

A classifier said to be trained on the training dataset when it creates a model from the training instances.

Classification

Classification means assigning an instance of a test dataset to a class label such as assignment of a tweet to positive class.

Classifier

The model, built upon the training data, which assigns class labels to instances are called classifier or estimator because of its probabilistic nature.

Feature Engineering and Feature Selection

Although instances may have many features, just a subset of them may be relevant or sufficiently predictive for an accurate and computationally efficient classification task. Determining which subset is optimal for an accurate modeling is

called feature selection. Sometimes, new features may need to be defined from existing ones by combining them. The process of feature selection and creating new features from existing ones is called Feature Engineering.

Learning Curve

Learning curves are used to plot the relationship between number of training instances and the performance of a classifier. It is a great way to show how an information processing algorithm learns with experience.

3.1.1.2 Machine Learning Algorithms

In subsequent sections; the terms “algorithm”, “classifier” and “model” are used interchangeably. Also, “label” and “class” have same meanings.

3.1.1.2.1 SVM

Support Vector Machines (SVM), (Cortes and Vapnik, 1995), builds a decision surface with a set of hyperplanes on the borders of training instances (support vectors) whose locations in the instance space gives the optimum margin to differentiate members of different classes.

The key idea behind SVM’s are if training instances are not linearly separable in their n dimensional space, then in a hyperspace with higher than n dimension, there can be found hyperplanes which separates the instances optimally. This approach is called the kernel trick. Popular kernel functions are Radial Basis Function (RBF),

Sigmoidal, linear and Polynomial Kernel. Furthermore, an SVM model with a kernel (similarity function) is analogous to two-layer Neural Network with a different cost function (as the first layer to project data into another space, and second layer to classify). SVM algorithm can also be used for classification problems with more than two classes. In these cases, usually, one-again-rest or one-against-all methods are used.

Advantages of SVM's are:

- They are effective in high dimensional spaces which is characteristic in text classification.
- Uses a subset of training instances (support vectors) in the decision function, therefore it is memory efficient.
- It can use custom kernel functions (similarity measure) specific to needs (this kernel function determines the distance among instances in the space)

3.1.1.2.2 Logistic Regression

Logistic Regression classifiers regress the training data but by using a threshold they assign classes (discrete values) to instances instead of regression algorithm's continuous outputs, their dependent variables are categorical.

Although logistic regression is special kind of generalized linear regression, it assumes conditional probabilities have Bernoulli distribution instead of Gaussian distribution. In addition, Logistic Regression is equivalent to Maximum Entropy

modeling (Mount, 2011), their decision functions can be derived from each other. However, there are several implementations of Maximum Entropy classifiers using different underlying algorithms, in turn, have different accuracies on their classification problem.

The logistic regression algorithms are a widely successful algorithm in a wide range of domains such as topic classification, sentiment analysis and language detection.

3.1.1.2.3 Decision Trees

Decision trees classify instances by hierarchically sorting them through their feature values. Nodes refer to features while leaves are classes. At each tree level, best feature in terms of its ability to differentiate instances, is selected for the nodes. Most commonly used methods for selecting best features are entropy and information gain. Feature selection continues until a level is found where all instances have the same labels (stop conditions). After the training step, a new instance is classified by going through the root of three till a class label is found.

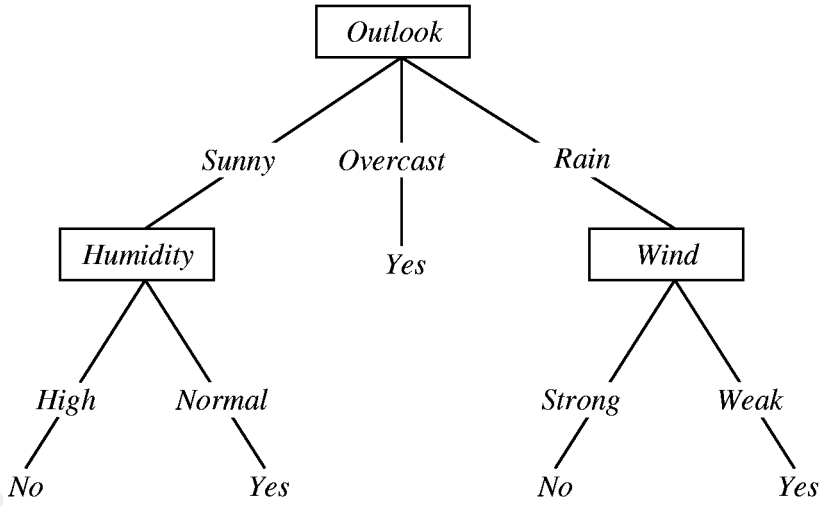

Figure 1. A sample decision tree (Mitchell, 1997)

A decision tree can be seen as a condensed version of a rule table. Each path from root to the leaves is a rule of the model and leaves are the categories in which training instances fall into.

3.1.1.2.4 Stochastic Gradient Descent

Gradient descent (or ascent) algorithms try to approximate maximum or minimum of an objective function (aka decision, loss, error or cost function) in a large problem space which is computationally very expensive to trace every corner of it. Stochastic gradient descent, on the other hand, is the stochastic approximation of the Gradient Descent, which optimizes the parameters of objective function by one training sample at a time, in each iteration. It is widely used in text classification problem which requires lots of features, and is very efficient in terms of computational cost. Another

advantage of Stochastic gradient descent algorithm in Natural Language Processing tasks is its capability in sparse data problems.

3.1.1.2.5 Maximum Entropy Classifier

Another well-known technique for Natural Language Processing is Maximum Entropy algorithm (Nigam et al. 1999) and its varieties that produce a discriminative model by learning soft or hard boundaries between classes in the instance space (similar to SVM and decision trees) as opposed to generative models which model the underlying distribution of training data (such as Naive Bayes algorithm). Maximum Entropy Classifiers do not assume, unlike Naive Bayes classifiers, features are conditionally independent of each other. Its underlying assumption is the Principle of Maximum Entropy which states that maximum information gain can be realized through the model that has the maximum entropy (entropy means the unpredictability of an information content) if we do not know the underlying distribution of an event.

Maximum entropy classifiers need more time than Naive Bayes classifiers because they first need to solve internal optimization problem to find parameters of the model. There are several training methods for Maximum Entropy classifiers such as GIS, IIS and Megam (Daumé III, Hal., 2004)

3.1.1.2.6 Naive Bayes

Naive Bayes classification algorithm rests upon the Bayesian Theorem which can be, in our case:

P(class | feature) = P(class) * P(feature | class) / P(feature)

with the meaning that probability of a class (positive or negative) of a given the feature is dependent on prior probability of that class, probability of seeing that feature given the class and probability of that feature in the corpus. Naive Bayes classifier builds upon this idea and considers every feature independent of other features.

In a document d, the assignment of class c to d is done by calculating the probabilities of each class given the document, then selecting the Maximum Posterior Probability (MAP) estimate:

$$c\ MAP = \text{argmax}\ Pc \in C\ (c|d)$$

using the Bayesian Formula, then

$$c\ MAP = \text{argmax}c \in C\ P\ (d|c) \times P\ (c) / P\ (d)$$

because P(d) is independent of and same in each class

$$c\ MAP = \text{argmax}c \in C\ P\ (d|c) \times P\ (c)$$

if we show a document with feature set,

$$d = f1,f2,f3,f4,,,fn$$

which makes our Naive Bayes classifier formula as:

$$c\ NB = \text{argmax}c \in C\ P\ (c) \prod a\ P\ (a|c)$$

Naive Bayes classifiers can be differentiated according to underlying probability distribution of p(feature | class) such as Multi-variate Bernoulli Naive Bayes, Multinomial Naive Bayes or Gaussian Naive Bayes.

In certain domains, it has been showed that Naive Bayes algorithm can perform better than more complex algorithms such as Neural Networks (Mitchell, 1997).

3.1.1.2.7 Ensemble Methods

Ensemble learning use multiple classifiers and combines their results for better predictive accuracy. The aggregate result can be calculated by majority rule voting or sum of the predicted probabilities for each of the class label. Voting and summation can be done using a weighted schema by giving not uniform weights for classifiers.

3.1.1.2.7.1 Random Forests

Random Forests classifiers, as an ensemble method and meta estimator, use a set of decision trees and combines their results for a given example to predict. Its assumption is by building n number of decision trees each trained with a subset of the training data set (bagging – bootstrap aggregating), totality of all trees can have higher accuracy than a single decision tree trained on the whole data set.

3.1.1.2.7.2 Gradient Boosting Classifier

"With excellent performance on all eight metrics, calibrated boosted trees were the best learning algorithm overall. Random forests are close second." (Caruana, 2006)

Gradient Boosting Classifiers, like Random Forest models, uses decision trees. Instead of Random Forest's approach of estimating with full blown trees, Gradient Boosting Trees are weak learners with high bias and low variance. The algorithm, in a forward-stage-wise fashion incrementally improves the trees by mainly reducing bias but also, to some extent, variance by aggregating the predictions of weak decision trees.

3.1.1.2.7.3 Extremely Randomized Trees Classification

Extremely Randomized Trees classification first fits randomized decision trees (in other word extra trees) on sub samples of whole dataset. Then, it aggregates results of sub trees to improve predictive power and reduce bias. Although similar, there are two main differences between Extra Trees and Random Forests which are:

- Extra Trees use sub samples from the whole training data set instead of bootstrap of the training data in each split stage while in a Random Forest meta estimator, each decision tree works specifically on a subset of the data (bagging – bootstrap aggregating)

- In each component tree, splits are chosen randomly which involves extra randomness in the ensemble, meaning components trees' mistakes are less correlated to each other.

3.1.1.2.7.4 Adaptive Boosting Classifier

Adaptive Boosting algorithm (Freund et al, 1999) is an meta estimator as an ensemble method, begins by fitting a classifier on the original whole training dataset. Then, each cycle by increasing the weights of the wrongly classified instances, AdaBoost fits the base classifier. By focusing on the wrongly classified instances in successive steps, it tries to reduce error rate.

3.1.1.3 Classification Evaluation Criteria

In this work, performance of Machine Learning algorithms will be presented comparatively by using the evaluation criteria mentioned below.

3.1.1.3.1 K-fold Cross Validation

Cross validation (aka rotation estimation) is a model validation technique to measure the generalization capability of a model learned from training dataset to out-of-sample instances. In K-fold Cross Validation, training dataset is divided into k separate subsets and with each (k-1) of them, model is trained to predict the remaining subset. Predictive accuracy for this configuration is calculated on that remaining one.

Therefore, there are k possible subsets and in each k iteration, a different (k-1) combination is selected for training. To calculate overall model predictive accuracy, k accuracies are averaged to get a smooth precision and decrease the weight of possible outliers. In this work, k is chosen as 10 which is the most common value in the literature.

3.1.1.3.2 Precision

Precision, aka positive predictive value, represents the exactness of a classification system. It is defined as (True Positives) / (True Positives + False Positives), in other words, number of correct positive results divided by the number of all positive results.

3.1.1.3.3 Recall

Recall, aka sensitivity, represents the completeness of a classification system. It is defined as (True Positives) / (True Positives + False Negatives), in other words, number of correct positive results divided by the number of positive results that should have been returned (all positive results in the dataset but not returned by the classifier).

3.1.1.3.4 F1-Score

F1-Score incorporates both precision and recall by using both in its formula. F1 is the harmonic mean of precision and recall:

F1 = (2 * precision * recall) / (precision + recall)

3.1.1.3.5 Roc and Auc

Receiving Operator Characteristic (ROC) Curve and Area under Receiving Operator Characteristic Curve (AUC) are used to illustrate the performance of a classifier ranking function. ROC is plotted as True Positive Rate (recall, sensitivity) against False Positive Rate (1 – specificity).

3.1.1.3.6 Training Score and Testing Score

Training score is the accuracy which the classifier performs in the training data after it is fitted. Testing score is the accuracy of the classifier in the testing data. When precision is mentioned for a classifier, it corresponds to the testing score. However, training score is also very important, because a classifier with a very high training score is in the danger of overfit. Therefore, classifiers with high and similar testing and training scores can be said to have optimal configuration, probabilistically.

3.1.2 Natural Language Processing (NLP)

In this work, several NLP methods and text representations were used to be able to decide which ones are suitable for Turkish and English in terms of their contribution to the sentiment analysis accuracy.

3.1.2.1 N-gram model

N gram model of text is probabilistic language model to predict the next token given the tokens specified. It is highly used in NLP and Information Retrieval fields. N can be 1,2, 3... etc. As an example, the sentence “the United States is a member of United Nations” can be represented in:

Unigram (1-gram): “the”, “united”, “states”, “is”, “a”, “member”, “of”, “united”, “nations”.

Bigram (2-gram): “the united”, “united states”, “states is”, “is a”, “a member”, “member of”, “of united”, “united nations”.

3.1.2.2 Bag of words (bow)

A bag of words can be considered as a set in which each member is attached with its frequency. In our example “the United States is a member of United Nations” corresponding bow representation is:

“the:1, united:2, states:1, is:1, a:1, member:1, of:1, nations:1”

3.1.2.3 Tf-Idf

Tf–Idf, term frequency–inverse document frequency, is a widely used weighting method in Information Retrieval and NLP communities. It gives more weights proportionally to terms in a dataset which occurs in a document more frequently but offset by term’s occurrences in the whole corpus. Therefore, a term which is seen frequently in small number of documents has higher Tf-Idf value than a term which is

sprawled over entire corpus. As a result, terms with higher discriminating value for document or sentence classification is rewarded by Tf-Idf. Although there are several implementations of Tf-Idf a classical formula of it is:

$$w(i,j) = tf(I,j) * \log(N / dfi)$$

where

tf j,j = number of occurences of I in document j

df I = number of documents containing term I

N = total number of documents in the corpus

3.1.2.4 Word2Vec and Doc2Vec

Word2vec (Goldberg et al, 2014) is a shallow neural network to produce vector outputs for given words. This is achieved through transforming each word in the text into n-dimensional (a parametric variable number) continuous vector space by analyzing distributional properties of that token in the document. As an unsupervised algorithm, word2vec does not need labeled data, creates word embedding model by analyzing corpus (set of documents) and extracting token distribution through cbow and skipgram (Levy et al, 2015) as explained below.

3.1.2.5 Cbow (continuous bag of words)

Uses bag of words approach mentioned above. The goal of the word2vec shallow neural network which uses cbow is to maximize P(token|context), meaning predicting the word given the context.

3.1.2.6 Skip-gram

It is similar to n-gram but skips some tokens so its components are not consecutive in the text. For example, in the text “the United States is a member of United Nations”, 1 skip 2grams are: “the states, united is, states a, is member, a of, member united, of nations”. When word2vec model initialized with skip-gram, it maximizes P(context|token), in other words it predicts the context given the word.
“The skip-gram method weights nearby context words more heavily than more distant context words.” (Mikolov, et al, 2013). Therefore, although Cbow is faster than skip-gram, skip-gram gives higher accuracy for infrequent words.

In Word2vec, the model can be trained via negative sampling or hierarchical softmax are used. Negative sampling, minimizes the log-likelihood of sampled instances of PMI (Pointwise Mutual Information). On the other hand, hierarchical softmax algorithm uses Huffman trees to reduce calculation so it does not need compute the conditional probabilities of all vocab words. According to (Mnih et al, 2009) hierarchical softmax works better for infrequent words while negative sampling works better for frequent words.

Therefore, there are four combinations in a word2vec model, with two representation methods (cbow and skip-gram) and two training algorithms (negative sampling and hierarchical softmax). Doc2vec uses word2vec algorithm to vectorize not just words but also text chunks, sentences, paragraphs and documents by combining text’s word2vec representations.

3.2 Methodology

To build the election and foreign policy models, an advanced programming system is created. In this section, technical details of this system, data collection step, preprocessing and annotation of tweets, high level statistical analysis of the dataset will be discussed.

3.2.1 System Design and Programming Environment

All data analysis and modeling were done using Python language using PyCharm as the IDE. In data collection step, tweets are stored in a text file each row corresponding to a tweet in json format (the format of Twitter API's). In analysis stage, tweets were preprocessed and resulting reduced json format were stored in MongoDb because of its inherent support for json format and Python dictionary data structure. For Natural Language Processing, Gensim and NLTK libraries of Python were used. In the modeling step, for machine learning, Scikit-Learn (sklearn) was used (Pedregosa, et al, 2011).

Rationale behind System Design and Programming Environment

In all computational steps Python was used as the programming language. There are other alternative tools: R, Microsoft Office Excel, SaS, Wecka etc. However, Python combines two very important advantages:

- Availability of excellent statistical, machine learning and natural language processing packages with multi-core support (sklearn, gensim, nltk)

- Easiness of coding as a very high level language with higher abstraction compared to lots of other programming languages

As the database management system, NoSql MongoDb was used because of its excellent support for Twitter API's json results and their corresponding Python dictionary data structure. Another reason for a NoSql Db was the Twitter API's json results' self-referential structure which is hard to represent and not natural in classical Sql Db's such as MySql.

3.2.2 Data Collection

Tweets were collected from Twitter Streaming API with selected keywords. Data collection is done via Tweepy package (a wrapper for the rest based Twitter api, json based) and Python programming language. Twitter streaming API provides maximum 1% of all tweets randomly in real time, filtered by the set keywords. Although Twitter Gardenhouse API (10% of all tweets) and Firehouse API (100% of all tweets), these options are not free and for the work's purposes randomly selected 1% of all tweets is suitable. There is also Twitter Search API, which gives only a few days of data, therefore is not practical for the purposes of this work.

3.2.2.1 Keywords

3.2.2.1.1 TURKEY

akparti = ["adalet ve kalkınma", "akparti", "akp", "erdoğan", "davutoğlu", "@Akparti", "@RT_Erdogan", "@Ahmet_Davutoglu"]

chp = ["cumhuriyet halk", "chp", "kılıçdaroğlu", "@herkesicinCHP", "@kilicdarogluk"]

mhp = ["milliyetçi hareket", "mhp", "bahçeli", "@dbdevletbahceli"]

hdp = ["halkların demokratik", "hdp", "demirtaş", "@HDPgenelmerkezi", "@hdpdemirtas"]

To collect tweets from Twitter API, keywords were selected carefully to not to introduce topic bias into the dataset. For each political party; name of the party, its abbreviation, surname of the leader of the party, official twitter accounts of both of the party and the leader were used. In Justice and Development Party (Akparty), President Recep Tayyip Erdoğan and Prime Minister Ahmet Davutoğlu were both considered as the party leaders because of electorates obvious perception of that. Also, due to widespread usages of akparti and akp as abbreviations, both of them were used. Furthermore, in Nationalist Movement Party, there is no official party Twitter account.

3.2.2.1.2 USA

Republican Presidential Candidates

trump = ["Donald Trump", "@realDonaldTrump"]

cruz = ["Ted Cruz", "@tedcruz"]

rubio = ["Marco Rubio", "@marcorubio"]

carson = ["Ben Carson", "@RealBenCarson"]

kasich = ["John Kasich", "@JohnKasich"]

bush = ["Jeb Bush", "@JebBush"]

Democratic Presidential Candidates

sanders = ["Bernie Sanders", "@BernieSanders"]

clinton = ["Hillary Clinton", "@HillaryClinton"]

In US Presidential Primary Elections, again a formal procedure was used to select keywords for Twitter Streaming API. For each of the candidates, only their names and their official twitter account names were used. However, in this dataset, after the primary election results, only Trump's and Clinton's tweets are analyzed.

3.2.3 Preprocessing

Several steps for taken to modify tweet texts before modeling their sentimental polarity.

Lowercase characters

All tweets were lowercased. In Turkish example because of special characters (İ,Ü,Ç...etc.) first all characters were capitalized then lowercased to get uniform representation.

Json reduction

All fields from Twitter json data which are not needed were filtered out such as entities, places, url tags. A twitter json example:

{
"text": In preparation for the NFL lockout, I will be spending twice as much time analyzing my fantasy baseball team",

```
    "favorited": false,
    "source": "<a href=\"http://twitter.com/\" rel=\"nofollow\">Twitter for
iPhone</a>",
    "in_reply_to_screen_name": null,
    "in_reply_to_status_id_str": null,
    "id_str": "5469180224328",
    "entities": {
        "user_mentions": [
            {
                "indices": [
                    3,
                    19
                ],
                "screen_name": "X",
                "id_str": "271572434",
                "name": "X",
                "id": 271572434
            }
        ],
        "urls": [ ],
        "hashtags": [ ]
    },
    "contributors": null,
    "retweeted": false,
    "in_reply_to_user_id_str": null,
    "place": null,
    "retweet_count": 4,
    "created_at": "Sun Apr 03 23:48:36 +0000 2011"
}
```


Valid, invalid, spam types

Twitter API returns true tweets and sometimes HTTP status messages both in json content. All json messages labeled with either valid (true tweet), invalid (HTTP status message or broken messages due to network issues) or as spam.

Spam detection

In twitter, there are lots of spam accounts posting junk tweets. All tweets were analyzed for their posting user. Spam users' tweets were labeled as spam besides tweets from the users who are very ineffective in terms of their network features or posting behavior. Criteria for spam filtering are listed below.

Default profile image

Accounts with default profile image (egg) were labeled as spam.

Followers count

Accounts with less than or equal to five followers were labeled as spam (in this case not valuable) because their owner does not enough say in Twitter network structure (ineffective users) or actual spam due to spam accounts do not attract large numbers of followers.

Friends count / followers count

If a user has many friends (accounts which that user is following in Twitter terminology) but very few followers, it is highly likely that user is a spam. To be on the safe side, 100 was selected as the ratio threshold.

Statuses count

Accounts whose total statuses count in their account is less than 50 (including Rts) were considered as very ineffective.

Account creation date

Because spam accounts are created to affect a specific issue on Twitter, accounts whose creation date is not at least 30 days further away from the related tweet were considered as spam.

Daily number of posted tweets

One of the characteristic features of spam accounts are they post tweets very frequently because of their automated nature. If an account posted on daily average at least 30 tweets after its creation date, it was considered as spam.

Tokenization

All twitter texts were tokenized into constituent parts.

Punctuation

Punctuation symbols were removed except the ones included in emoticons.

Stopwords

In current sentiment analysis literature, there are several examples which filter stopwords but in this work's dataset it reduced the accuracy. So, stopwords were preserved.

Emoticons

Emoticons were preserved because they have sentiment meaning such as ":)" smile and sadness ":(". When emoticons' characters are separately written with white spaces between them, they were combined. For example, ":)" was modified to ":)".

RT texts

In retweets (definition), this work's assumption was that twitter users usually retweet other's tweets to show their agreement with the content. Because Twitter API gives shortened texts for RT's, original text was restored.

Repetitive chars

Refpetitive characters in tweets were reduced to two characters. For example, in the token "heeeeeey", multiple e's were filtered out, and "heey" was used. Therefore, while noise was decreased by reducing both "heeeeeeeeeey" and "heeey" to "heey", their emphasis and stress were preserved by differentiating them "hey".

Urls and mentions

All url's were deleted. Mentions other than the ones used as filter keywords in Twitter Streaming API (official party and leader twitter accounts) were deleted.

3.2.4 Setting party labels

A tweet's party label was determined through how many its tokens contain filtering keywords used in data collection process. If it contains keywords from one party, that

party was assigned to the tweet. If contains more than one party's keywords it was labeled as "multiple party" tweet and filtered out.

3.2.5 Annotation

In sentiment analysis step, supervised machine learning algorithms were used. Therefore, randomly sampled 5000 tweets were sentimentally and manually annotated (positive, negative and neutral), and neutral tweets were filtered out. Positive and negative ones were used as the training data

3.2.6 Sentiment Analysis of Tweets

This work assumes that for best modeling and predictive accuracy, a machine learning approach is necessary as opposed to lexical methods as indicated by the relevant literature. Sentiment Analysis of tweets were done in this work on Turkish tweet dataset for Turkish 2015 November General Election and English tweet dataset for 2016 US Primaries. Due to underlying differences between Turkish and English, their respective semantic analysis will be represented in two separate sections. In each section, natural language processing and machine learning approaches and algorithms will be discussed comparatively. In the third section of this chapter, the differences between both languages and their corresponding semantic analysis will be compared.

Comparisons will be discussed on the combinations of several NLP methods and Machine learning algorithms:

NLP methods:

- Tf-Idf
- Bag of Words
- Unigrams
- Bigrams
- Unigrams + Bigrams
- word2vec - doc2vec

Machine Learning Algorithms:

- Support Vector Machines
- Linear SVC (although uses same logic with SVM's, a variation because of underlying implementation will be used)
- Logistic Regression
- Decision Trees
- Stochastic Gradient Descent
- Maximum Entropy Algorithms
- IIS
- GIS
- MEGAM
- Naive Bayes Algorithms
    - Multinomial Naive Bayes
    - Gaussian Naive Bayes
    - Bernoulli Naive Bayes
- Random Forests
- Gradient Boosting

- Extremely Randomized Trees
- Adaptive Boosting

3.2.6.1 Turkish Sentiment Analysis

After the preprocessing step was carried out, all tweets were stored in MongoDb. For the training set of Sentiment Analysis, 10000 tweets were sampled out among all tweets. Out of these 10000 tweets, 7265 of them were manually labeled in polarity categories of 1 (positive), 0 (negative) and X (neutral). After annotation, polarity distribution is:

Table 2. Turkish Tweet Dataset Polarity Distribution

Polarity Category	Count
0	3042
1	2513
X	1710

X(neutral) tweets were filtered out, in the final training dataset, there are 3042 negative, 2513 positive tweets.

3.2.6.1.1 Feature Engineering and Selection of Features

In NLP programs, there are many features because of the complexity of human language. In mediums, such as Twitter, due to informality and inattention to

grammatical rules, there is also high noise in the feature set. This work handles this problem through a statistical test on the best feature extraction algorithm, in terms of accuracy, unigrams. Hence, a chi-square test on the whole Twitter training dataset is applied to find the best unigram features which separates instances well, on the given statistical conditions. Instead of using all 22864 possible unigram features, this work will use 5000 best of them and results for this feature set will be given separately.

As a result, features will be extracted from each tweet using Bag of Words (BOW), unigrams (all 22864 tokens), bigrams, unigrams + bigrams, TfIdf, Doc2Vec and best unigrams (best 5000 tokens selected with chi square test) approaches for each of the algorithm.

3.2.6.1.2 Comparative Predictive Accuracies of Algorithms in Tweet Sentiment Analysis

In this section, for the selected features, parametrically optimized machine learning algorithms will be discussed comparatively. To achieve this, for each algorithm, 10-fold cross validation scores with means and standard deviations, classification report (with train – test split 80% - 20%), learning curve and ROC curve with AUC value will be presented. For the first algorithm, tables are discussed in detail. Classifier instances are implemented with optimum parameters which will be shown under the section of the relevant classifier for purposes of reproducibility. Optimum parameters are found with Scikit-Learn GridSearchCv and RandomizedSearchCV.

3.2.6.1.2.1 Support Vector Machines

Classifier Parameters:

(C=0.5, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape=None, degree=3, gamma='auto', kernel='linear', max_iter=-1, probability=True,random_state=None, shrinking=True, tol=0.1, verbose=False)

(parameters are determined by using GridSearchCv and RandomizedSearchCV)

This work uses Svm with linear kernel because others such as Radial Basis Function (RBF), Sigmoidal or Polynomial kernels are not very suitable for text classification tasks and performed poorly in experiments.

Table 3. SVC 10-Fold Cross Validation Accuracies

	Accuracy
BOW	0.80648
Unigram	0.80485
Bigram	0.73373
Unigram + Bigram	0.81475
Tf-Idf	0.81421
Best Unigrams	0.84986

Best Unigrams - Standard Deviation of Precision: 0.01781

Best Unigrams - Training Score: 0.95344

In the above table, SVC accuracies, for each of the NLP feature type, are given. For example, in the first row, it can be seen that SVC algorithm fed with BOW features of training tweets, can predict with 80% accuracy of the sentiments of test tweets. This result can also be interpreted as, for a given single tweet, it is expected, that tweet's polarity can be predicted correctly in 80% of time. The best features are Best Unigrams (unigrams which passed the statistical test). On the training data set, SVC with Best Unigrams features has 0.01 standard deviation in 10-Fold Cross validation, which makes the SVC-Best Unigrams combination a very stable algorithm – feature set pair.

Table 4. SVC Classification Report with Best Unigram Features

	Precision	Recall	F1-Score
0	0.84448	0.85884	0.85160
1	0.83821	0.82218	0.83012
Avg / total	0.84153	0.84158	0.84149

In the above table, SVC performance metrics (precision, recall and F1-Score) are given for class 0 (negative tweets) and class 1 (positive class). The last row is the average of the total set (positive tweets + negative tweets).

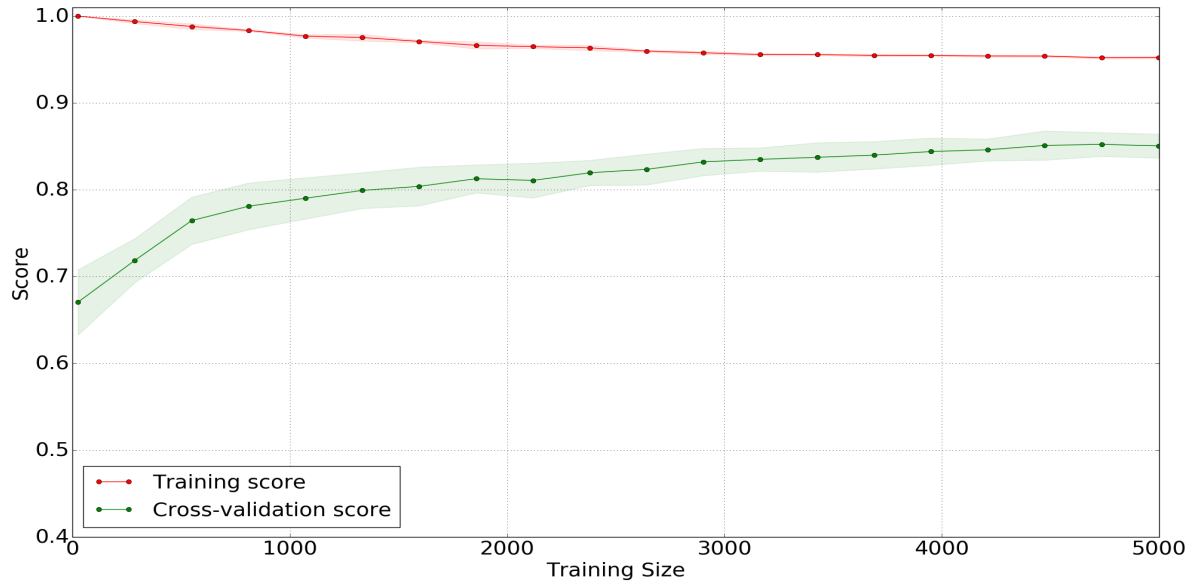

Figure 2. SVC Learning Curve with Best Unigram Features

Red line shows training score, while green is the 10-fold-cross validation testing score. The standard deviation of a test accuracy on a given point, is showed by the wideness of the area wrapped around that point. For the test accuracy of 80% in the task of predicting tweet sentiments, SVC needs about 1200 training instances and with 5000 instances it reaches 85% test accuracy.

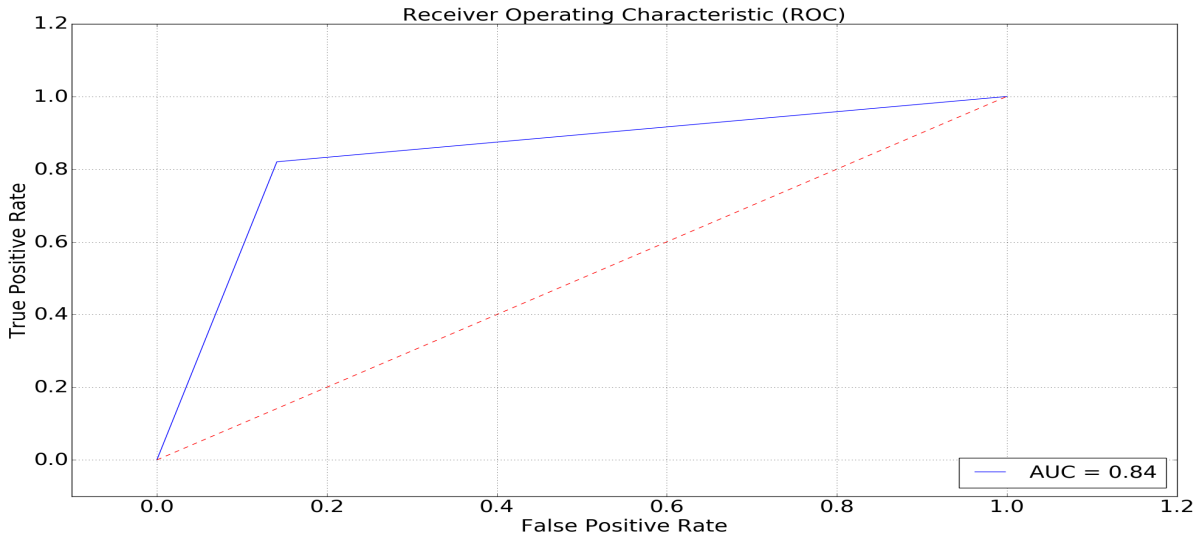

Figure 3. SVC ROC Curve with Best Unigram Features

In the above figure, SVC False Positive Rate (FPR) vs. True Positive Rate (TPR) is given for tweet sentiment analysis task. Area Under the Curve (AUC) is 0.84.

3.2.6.1.2.2 Linear SVC

Classifier Parameters:

(C=0.8, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, loss='squared_hinge', max_iter=1000, multi_class='ovr', penalty='l2', random_state=None, tol=0.1, verbose=0)

Table 5. LinearSVC 10-Fold Cross Validation Accuracies

	Accuracy

BOW	0.80486
Unigram	0.80575
Bigram	0.73156
Unigram + Bigram	0.81367
Tf-Idf	0.79441
Best Unigrams	0.85993

Table 6. LinearSVC Classification Report with Best Unigram Features

	Precision	Recall	F1-Score
0	0.85522	0.86395	0.85956
1	0.84526	0.83556	0.84038
Avg / total	0.85053	0.85059	0.85053

Best Unigrams - Standard Deviation of Precision: 0.01317

Best Unigrams - Training Score: 0.97120

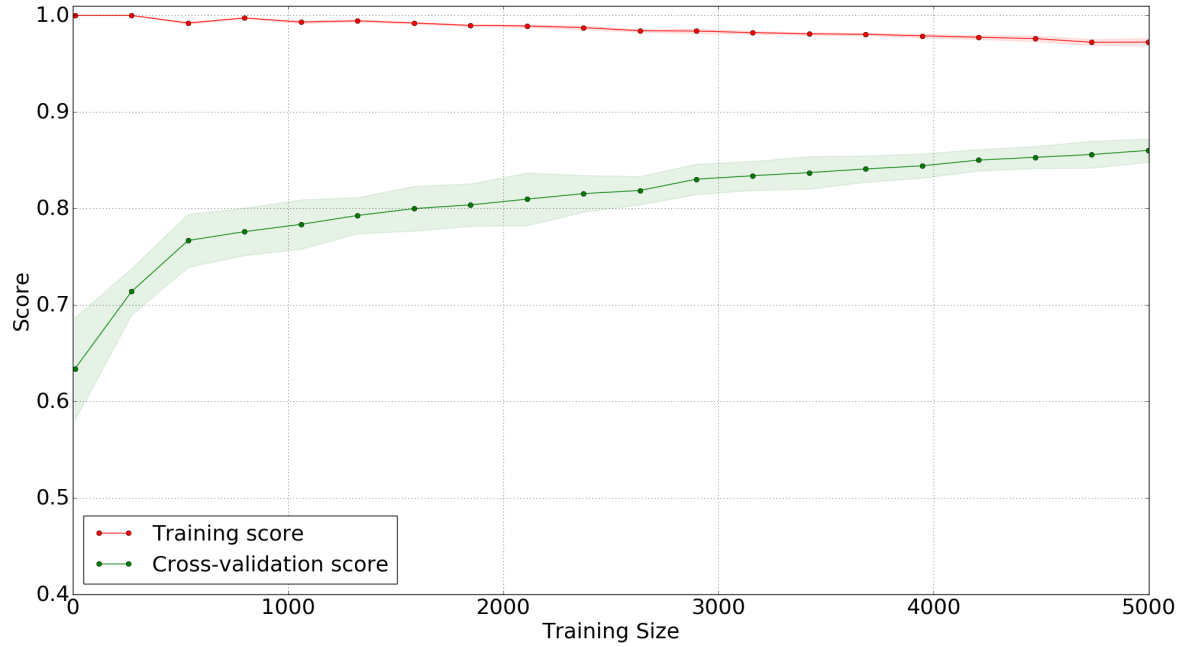

Figure 4. LinearSVC Learning Curve with Best Unigram Features

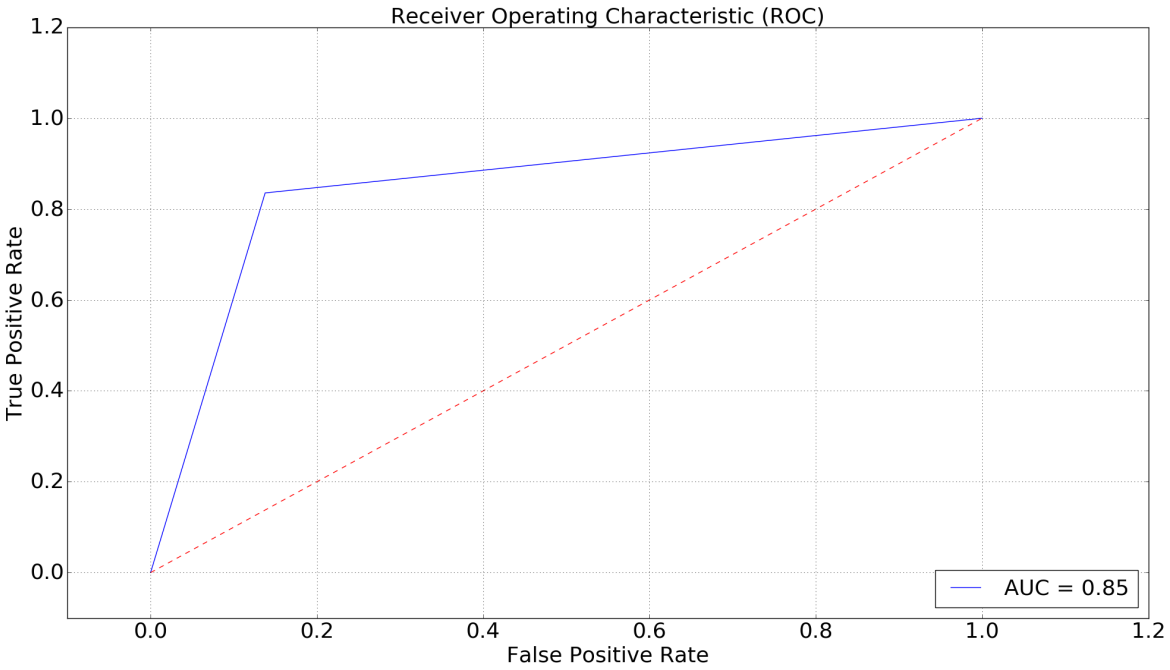

Figure 5. LinearSVC ROC Curve with Best Unigram Features

3.2.6.1.2.3 Logistic Regression

Classifier Parameters

(C=2.5, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=-1, penalty='l2', random_state=None, solver='sag', tol=0.1,verbose=0, warm_start=False)

Table 7. Logistic Regression 10-Fold Cross Validation Accuracies

	Accuracy
BOW	0.80899
Unigram	0.81331
Bigram	0.72581
Unigram + Bigram	0.81942
Tf-Idf	0.81295
Best Unigrams	0.85129

Table 8. Logistic Regression Classification Report with Best Unigram Features

	Precision	Recall	F1-Score
0	0.85025	0.86905	0.85955
1	0.84902	0.82792	0.83833
Avg / total	0.84967	0.84968	0.84956

Best Unigrams - Standard Deviation of Precision: 0.01613

Best Unigrams - Training Score: 0.94057

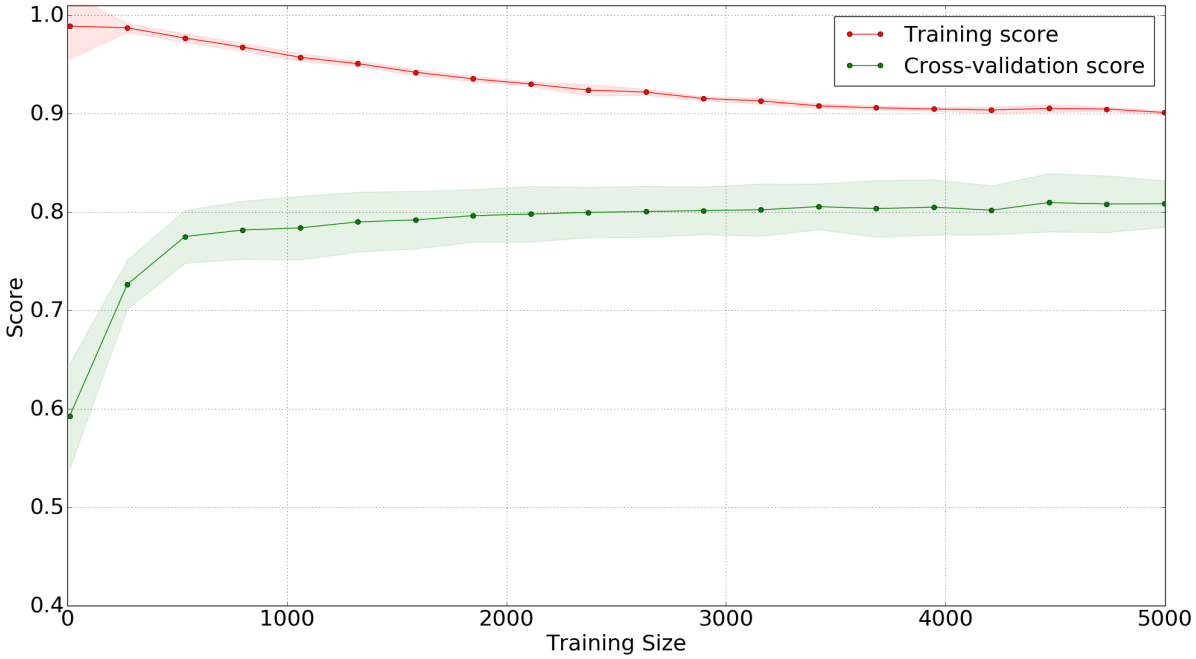

Figure 6. Logistic Regression Learning Curve with Best Unigram Features

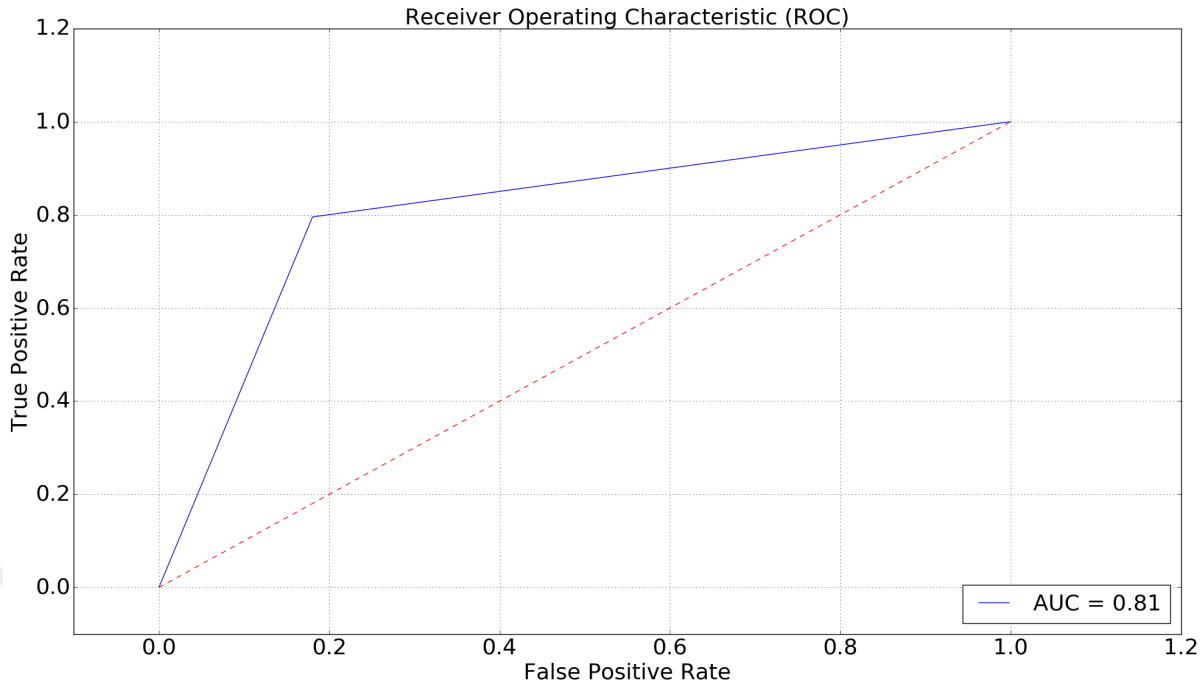

Figure 7. Logistic Regression ROC Curve with Best Unigram Features

3.2.6.1.2.4 Decision Trees

Classifier Parameters

DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=None, max_features=None, max_leaf_nodes=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort=False, random_state=None, splitter='best')

Table 9. Decision Trees 10-Fold Cross Validation Accuracies

	Accuracy
BOW	0.76398
Unigram	0.75967

Bigram	0.72023
Unigram + Bigram	0.75517
Best Unigrams	0.74022

Table 10. Decision Trees Classification Report with Best Unigram Features

	Precision	Recall	F1-Score
0	0.77759	0.80272	0.78996
1	0.76984	0.74187	0.75560
Avg / total	0.77394	0.77408	0.77378

Best Unigrams - Standard Deviation of Precision: 0.02184

Best Unigrams - Training Score: 0.99664

Figure 8. Decision Trees Learning Curve with Best Unigram Features

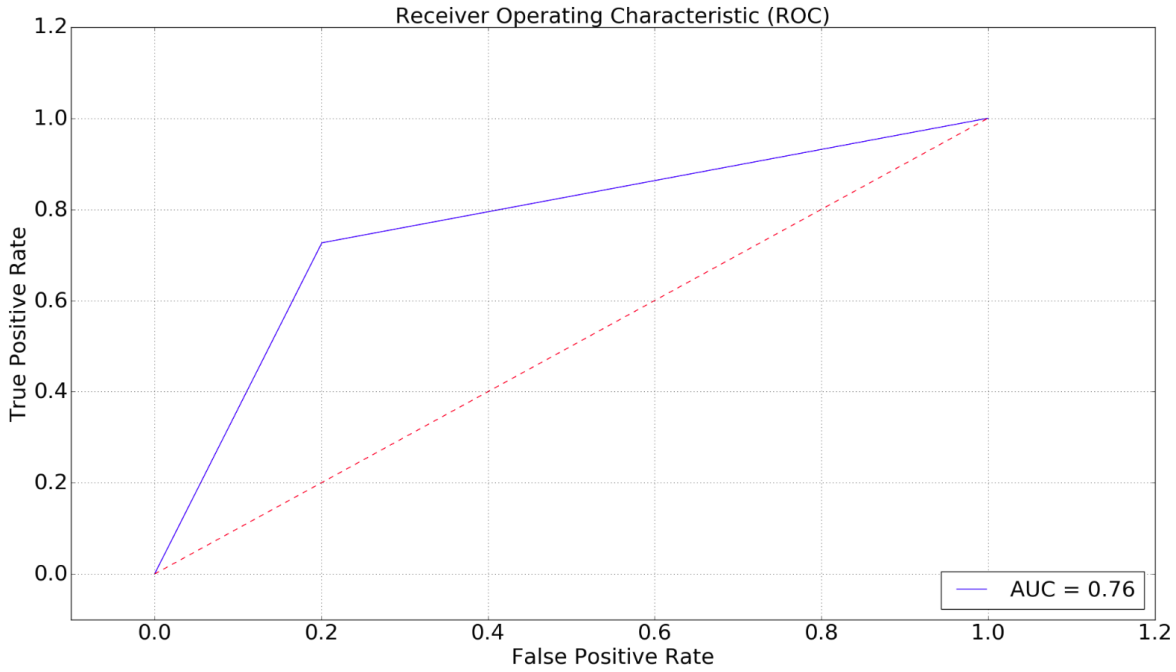

Figure 9. Decision Trees ROC Curve with Best Unigram Features

3.2.6.1.2.5 Stochastic Gradient Descent

Classifier Parameters

SGDClassifier(alpha=0.0001, average=False, class_weight=None, epsilon=0.1, eta0=0.0, fit_intercept=True, l1_ratio=0.15, learning_rate='optimal', loss='hinge', n_iter=5, n_jobs=-1, penalty='l2', power_t=0.5, random_state=None, shuffle=True, verbose=0, warm_start=False)

Table 11. SGD 10-Fold Cross Validation Accuracies

	Accuracy
BOW	0.79387
Unigram	0.79620
Bigram	0.73264
Unigram + Bigram	0.80394
Tf-Idf	0.79873
Best Unigrams	0.83635

Table 12. SGD Classification Report with Best Unigram Features

	Precision	Recall	F1-Score
0	0.83361	0.86054	0.84686
1	0.83730	0.80688	0.82181

Avg / total	0.83535	0.83528	0.83507

Best Unigrams - Standard Deviation of Precision: 0.02123

Best Unigrams - Training Score: 0.94164s

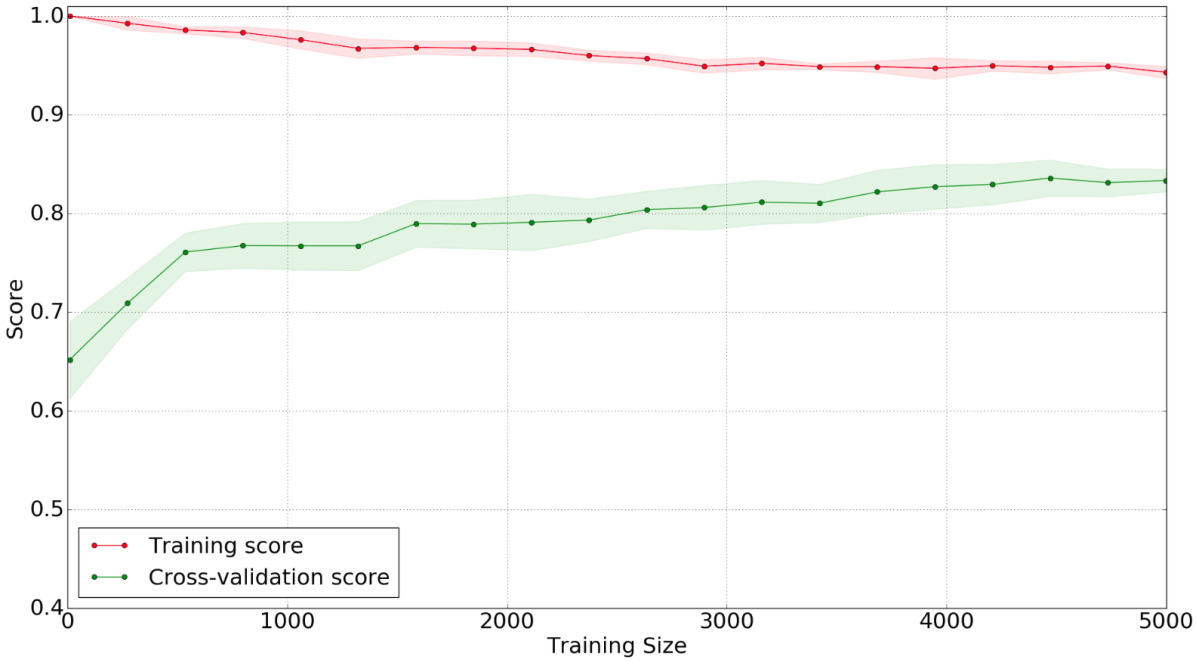

Figure 10. SGD Learning Curve with Best Unigram Features

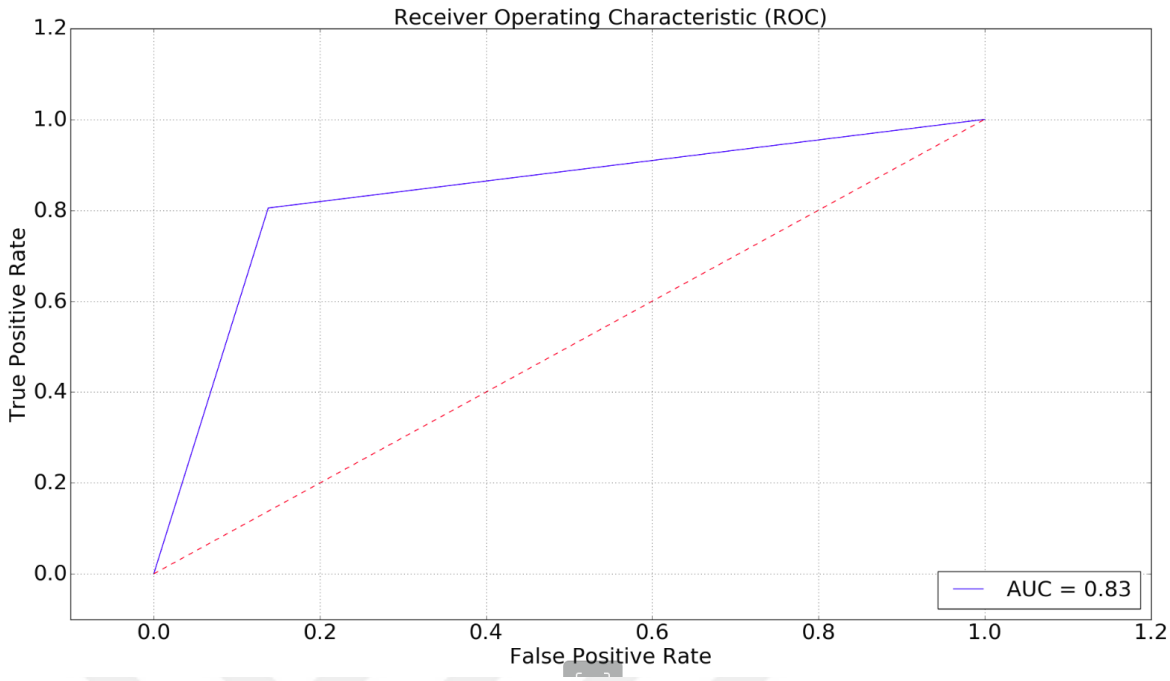

Figure 11. SGD ROC Curve with Best Unigram Features

3.2.6.1.2.6 Naive Bayes Algorithms

3.2.6.1.2.6.1 Multinomial Naive Bayes

Classifier Parameters

MultinomialNB(alpha=1e-10, class_prior=None, fit_prior=True)

Table 13. MNB 10-Fold Cross Validation Accuracies

	Accuracy
BOW	0.75840
Unigram	0.74993
Bigram	0.63633
Unigram + Bigram	0.63633

Tf-Idf	0.78630
Best Unigrams	0.90836

Table 14. MNB Classification Report with Best Unigram Features

	Precision	Recall	F1-Score
0	0.89644	0.94218	0.91874
1	0.93103	0.87763	0.90354
Avg / total	0.91273	0.91179	0.91159

Best Unigrams - Standard Deviation of Precision: 0.01719

Best Unigrams - Training Score: 0.95296

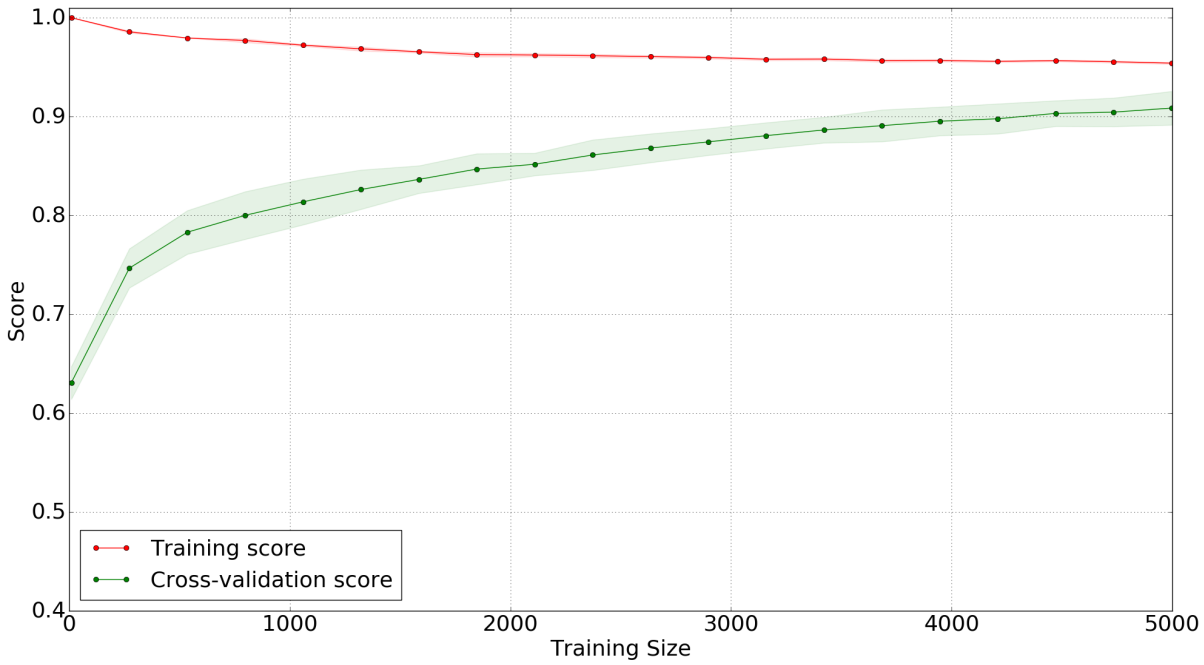

Figure 12. MNB Learning Curve with Best Unigram Features

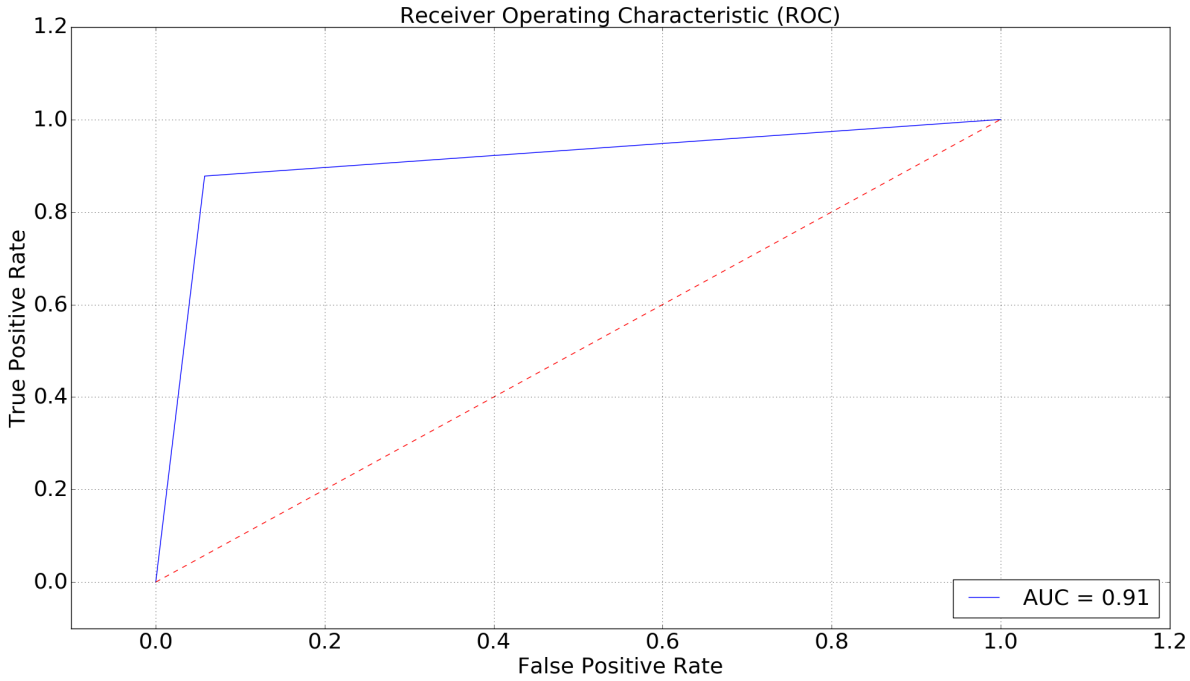

Figure 13. MNB ROC Curve with Best Unigram Features

3.2.6.1.2.6.2 Gauissian Naive Bayes

Table 15. GNB 10-Fold Cross Validation Accuracies

	Accuracy
BOW	0.69053
Unigram	0.69251
Bigram	0.59834
Unigram + Bigram	0.71807
Tf-Idf	0.68297
Doc2Vec	0.72793
Best Unigrams	0.89827

Table 16. GNB Classification Report with Best Unigram Features

	Precision	Recall	F1-Score
0	0.94073	0.83673	0.88569
1	0.83673	0.94073	0.88569
Avg / total	0.89177	0.88569	0.88569

Best Unigrams - Standard Deviation of Precision: 0.01793

Best Unigrams - Training Score: 0.93373

3.2.6.1.2.6.3 Bernoulli Naive Bayes

Classifier Parameters

BernoulliNB(alpha=0.001, binarize=0.0, class_prior=None, fit_prior=True)

Table 17. BNB 10-Fold Cross Validation Accuracies

	Accuracy
BOW	0.77064
Unigram	0.76290
Bigram	0.63759
Unigram + Bigram	0.77369

Best Unigrams	0.90800

Table 18. BNB Classification Report with Best Unigram Features

	Precision	Recall	F1-Score
0	0.89773	0.94048	0.91860
1	0.92929	0.87954	0.90373
Avg / total	0.92159	0.91179	0.91160

Best Unigrams - Standard Deviation of Precision: 0.01745

Best Unigrams - Training Score: 0.95679

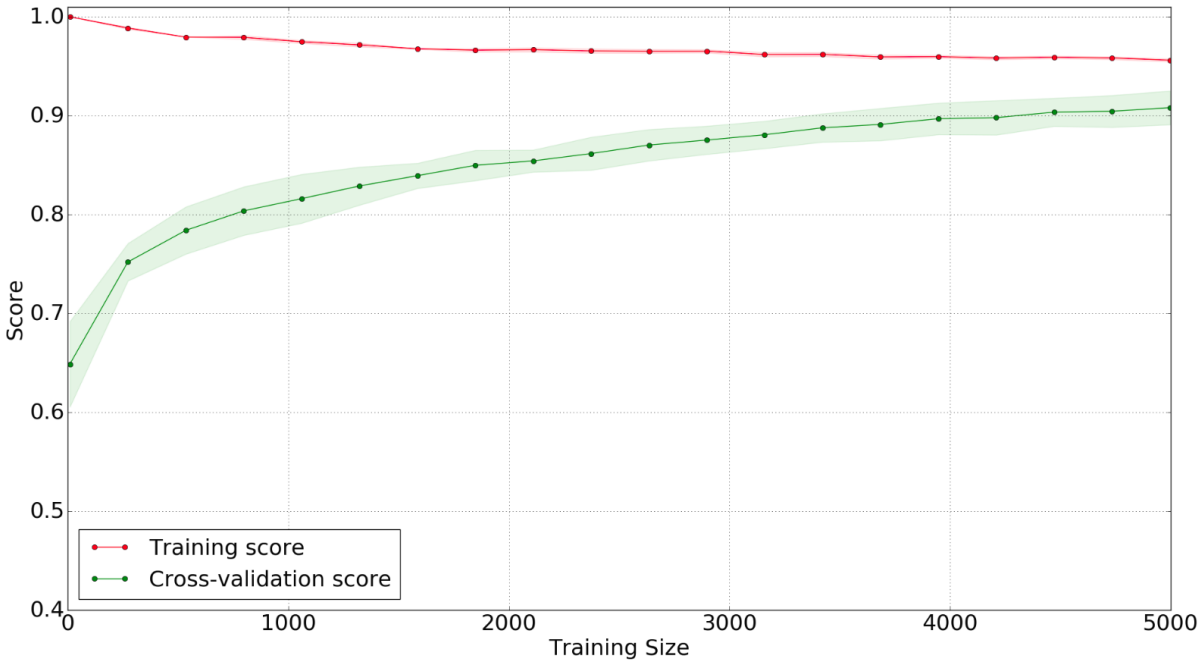

Figure 14. BNB Learning Curve with Best Unigram Features

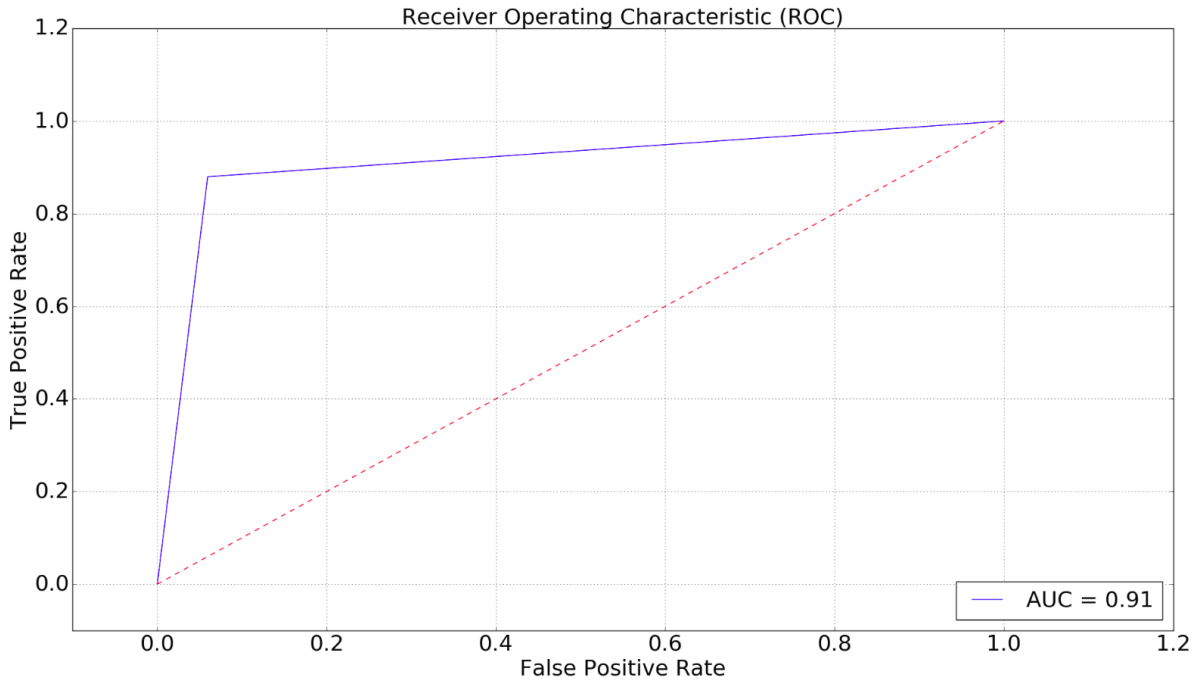

Figure 15. BNB ROC Curve with Best Unigram Features

3.2.6.1.2.7 Maximum Entropy (IIS, GIS, MEGAM)

Classifier Parameters

trace=3, labels=None, gaussian_prior_sigma=0, max_iter=15

Table 19. Maximum Entropy 10-Fold Cross Validation Accuracies

	Accuracy (IIS)	Accuracy (GIS)	Accuracy (MEGAM)
Unigrams	0.79946	0.79316	0.78507
Bigrams	0.63579	0.63758	0.71043
Unigrams + Bigrams	0.80395	0.796762	0.79766

Best Unigrams	0.87320	0.85341	0.82284

3.2.6.1.2.8 Ensemble Methods

3.2.6.1.2.8.1 Random Forests

Classifier Parameters

RandomForestClassifier(bootstrap=True, class_weight=None, criterion='entropy', max_depth=None, max_features='auto', max_leaf_nodes=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=-1, oob_score=False, random_state=None, verbose=0, warm_start=False)

Table 20. Random Forests 10-Fold Cross Validation Accuracies

	Accuracy
BOW	0.81043
Unigram	0.80017
Bigram	0.72401
Unigram + Bigram	0.79872
Best Unigrams	0.81114

Table 21. Random Forests Classification Report with Best Unigram Features

	Precision	Recall	F1-Score
0	0.81633	0.81633	0.81633
1	0.79350	0.79350	0.79350
Avg / total	0.80558	0.80558	0.80558

Best Unigrams - Standard Deviation of Precision: 0.02341

Best Unigrams - Training Score: 0.99910

Figure 16. Random Forests Learning Curve with Best Unigram Features

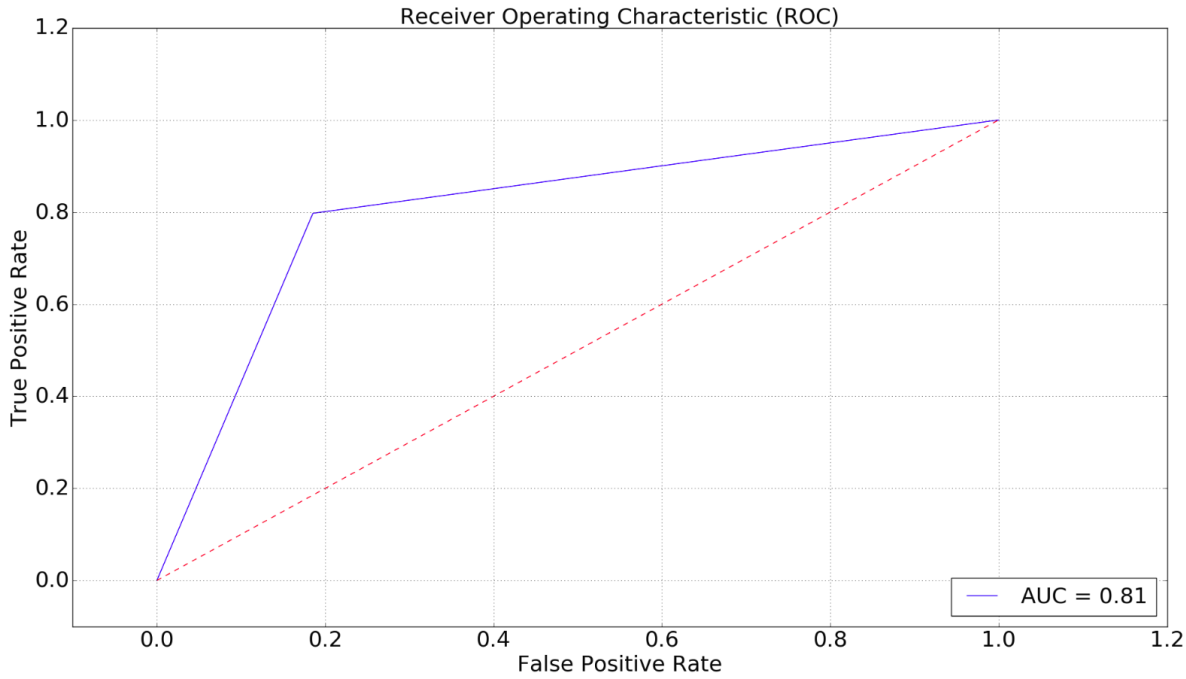

Figure 17. Random Forests ROC Curve with Best Unigram Features

3.2.6.1.2.8.2 Gradient Boosting

Classifier Parameters

GradientBoostingClassifier(init=None, learning_rate=0.1, loss='deviance',

max_depth=3, max_features=None, max_leaf_nodes=None, min_samples_leaf=1,

min_samples_split=2,

min_weight_fraction_leaf=0.0, n_estimators=100, presort='auto',

random_state=None, subsample=1.0, verbose=0, warm_start=False)

Table 22. Gradient Boosting 10-Fold Cross Validation Accuracies

	Accuracy
Best Unigrams	0.70639

Table 23. Gradient Boosting Classification Report with Best Unigram Features

	Precision	Recall	F1-Score
0	0.67932	0.88265	0.76775
1	0.80115	0.53155	0.63908
Avg / total	0.73667	0.71737	0.70718

Best Unigrams - Standard Deviation of Precision: 0.021849

Best Unigrams - Training Score: 0.72659

Figure 18. Gradient Boosting Learning Curve with Best Unigram Features

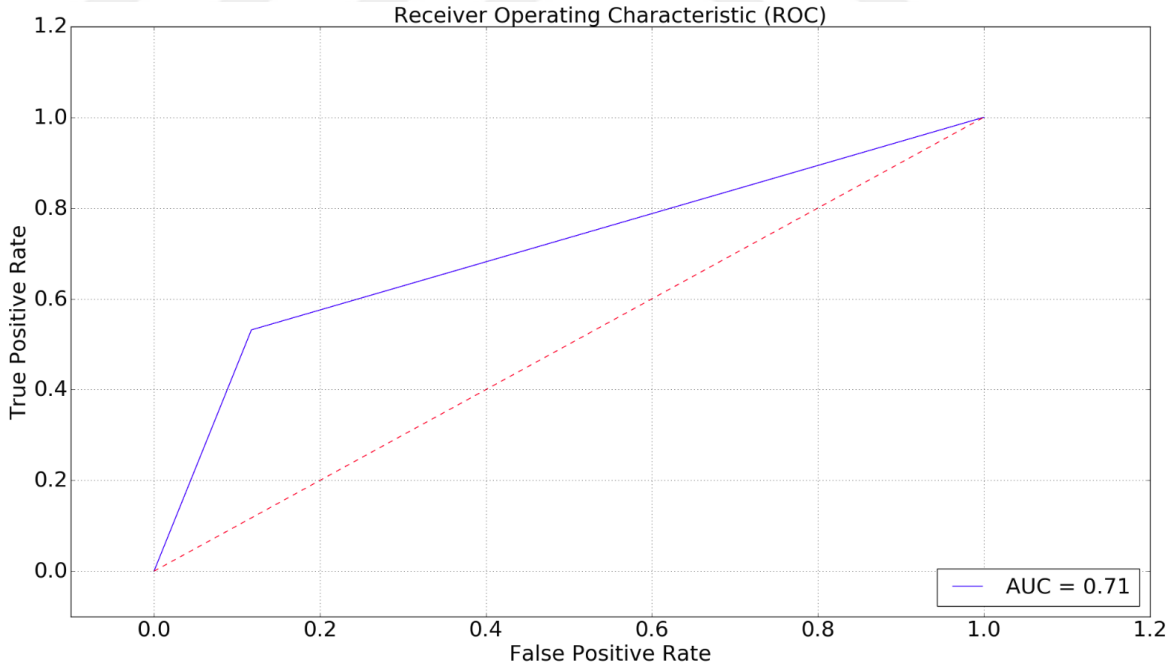

Figure 19. Gradient Boosting ROC Curve with Best Unigram Features

3.2.6.1.2.8.3 Adaptive Boosting (AdaBoost)

Classifier Parameters

AdaBoostClassifier(algorithm='SAMME.R',

base_estimator=BernoulliNB(alpha=0.001, binarize=0.0, class_prior=None,

fit_prior=True), learning_rate=0.001, n_estimators=1500, random_state=None)

Table 24. Adaptive Boosting 10-Fold Cross Validation Accuracies

	Accuracy
Best Unigrams	0.86965

Table 25. Adaptive Boosting Classification Report with Best Unigram Features

	Precision	Recall	F1-Score
0	0.85762	0.88095	0.86913
1	0.86193	0.83556	0.84854
Avg / total	0.85965	0.85959	0.85944

Best Unigrams - Standard Deviation of Precision: 0.01698

Best Unigrams - Training Score: 0.94081

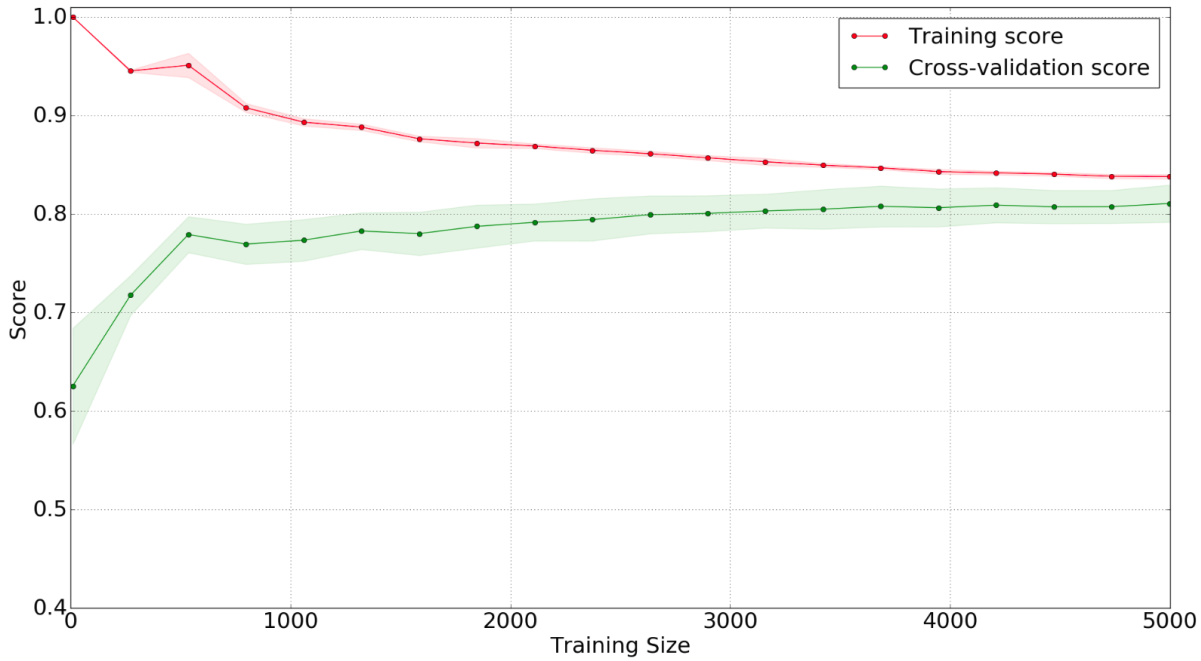

Figure 20. AdaBoost Learning Curve with Best Unigram Features

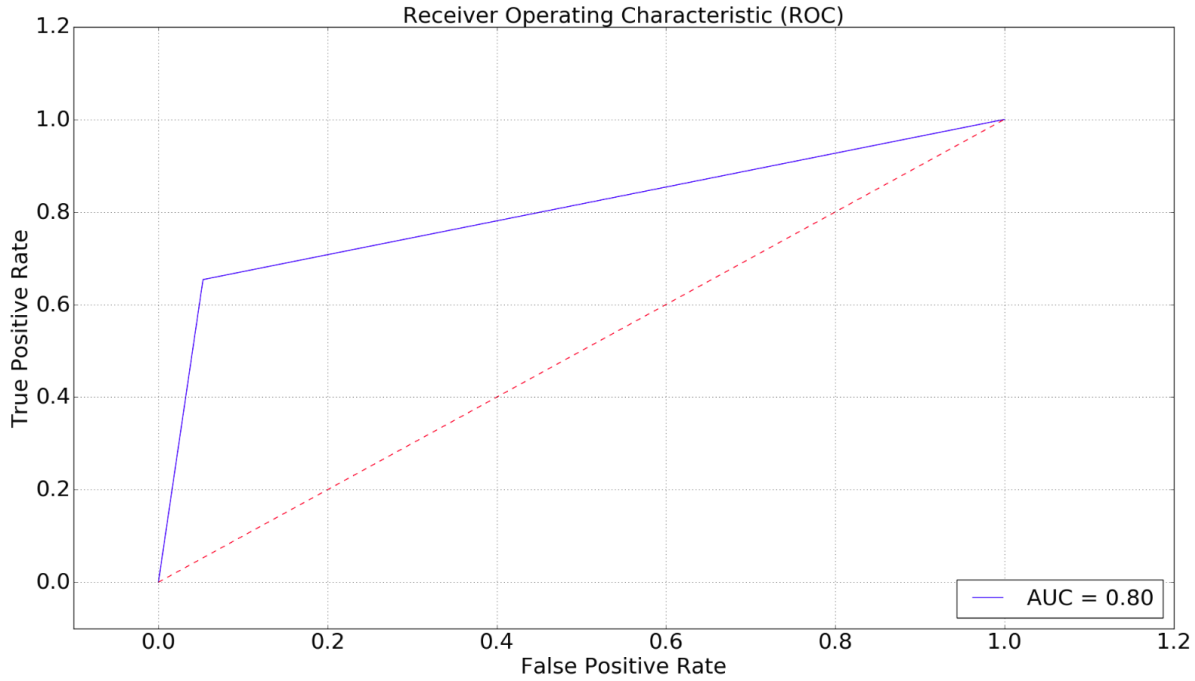

Figure 21. AdaBoost ROC Curve with Best Unigram Features

3.2.6.1.2.8.4 Extremely Randomized Trees

Classifier Parameters

ExtraTreesClassifier(bootstrap=False, class_weight=None, criterion='entropy', max_depth=None, max_features='auto', max_leaf_nodes=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=-1, oob_score=False, random_state=None, verbose=0, warm_start=False)

Table 26. Extremely Randomized Trees 10-Fold Cross Validation Accuracies

	Accuracy
Best Unigrams	0.793682

Table 27. Extremely Randomized Trees Classification Report with Best Unigram Features

	Precision	Recall	F1-Score
0	0.80801	0.82313	0.81550
1	0.79688	0.78011	0.78841
Avg / total	0.80277	0.80288	0.80275

Best Unigrams - Standard Deviation of Precision: 0.02221

Best Unigrams - Training Score: 0.99709

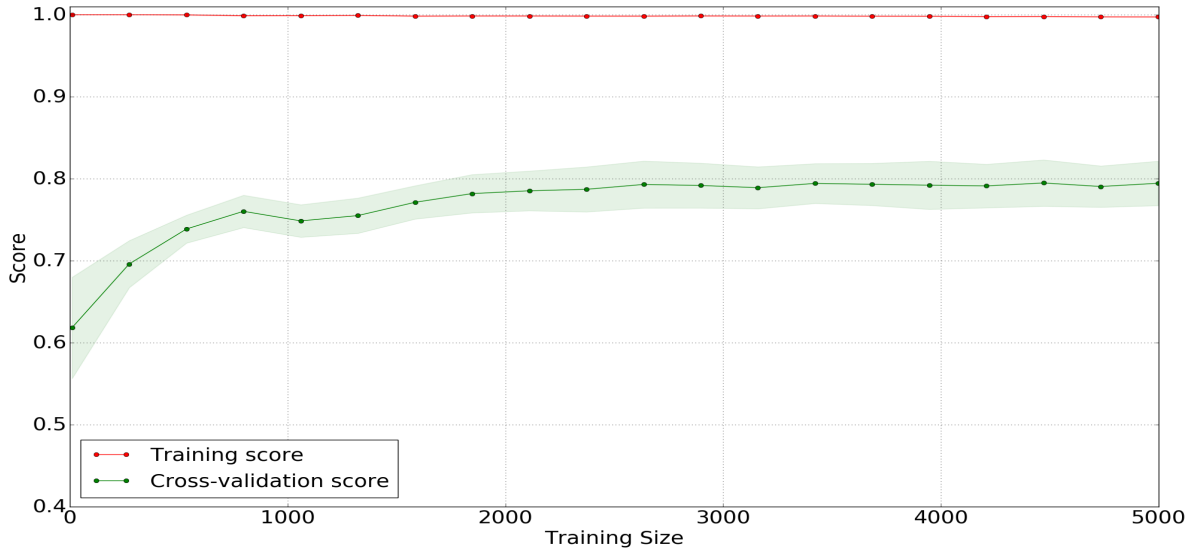

Figure 22. ERT Learning Curve with Best Unigram Features

3.2.6.1.3 Turkish Sentiment Analysis Conclusion

According to above results, in Turkish sentiment analysis, the best method is Bernoulli Naive Bayes classifier with Unigram features (statistically selected) with results:

Accuracy: 0.92159

Standard Deviation of Precision: 0.01745

Area under ROC: 0.91

Training Score: 0.95679

With the accuracy of 92%, Bernoulli Naïve Bayes classifier, when fed with statistically selected unigram features of tweets, predicts sentiments of tweets more precise than the other algorithm feature set combinations. Standard deviation of just 0.01 in 10-fold cross validation experiment gives the algorithm the ability to have a stable prediction capability in different datasets. Furthermore, because training score

(0.95) and accuracy in the test set (0.92) are close, the model does not show overfit. Therefore, in election prediction cases this model will be used.

3.2.6.2 English Tweets Sentiment Analysis

Sentiment analysis for 82 million English tweets will be discussed by giving summary results below (for detailed sentiment analysis methodology, look Turkish case).

3.2.6.2.1 Algorithm Comparison Results

Table 28. English Sentiment Analysis Accuracy Comparisons

Algorithm	10-Fold Accuracy (Best Unigram)
Support Vector Machines	0.91494
Linear SVC	0.91217
Logistic Regression	0.90506
Decision Trees	0.79316
Stochastic Gradient Descent	0.87896
Multinomial Naive Bayes	0.96319
Gauissian Naive Bayes	0.95910
Bernoulli Naive Bayes (best classifier)	**0.96675**
Maximum Entropy	0.86021

Random Forests	0.86985
Gradient Boosting	0.82058
Adaptive Boosting (AdaBoost)	0.89755
Extremely Randomized Trees	0.87619

As in the Turkish sentiment analysis best predictive capability is shown with Bernoulli Naïve Bayes classifier with statistically selected unigram features.

3.2.6.3 Comparison of Turkish and English Sentiment Analysis

For sentiment analysis both Turkish and English languages show similar results in their classifier and feature selection algorithms. Both have the highest accuracy with Bernoulli Naive Bayes Classifier with unigram features selected with statistical significance test. However, English sentiment accuracy has higher precision 0.96 compared to 0.92 although both are very accurate in tweet texts with maximum 140-character length.

CHAPTER 4:

ELECTION FORECASTING

The hypothesis that this work test out is: whether real election results in Turkish General Election November 1, 2015 can be predicted through twitter data or not. For the US primary presidential elections 2016, it employs a more complicated approach by creating a time series out of twitter data polarity analysis and takes national polls as the grand truth to forecast.

4.1 Case1: Turkish Election

In Turkish General Election of November 1, 2015, real results are:

- Akparti: 49,50
- Chp: 25,32
- Mhp: 11,90
- Hdp: 10,76

For this election, all tweets are tagged with their polarity values using the model discussed above. Then only tweets with positive polarity values are filtered out

because only that tweets correspond to party support. Two approaches were tested against the hypothesis that election results can be predicted by analyzing relevant twitter data: tweet based and user based. In each approach, two different results are produced, whole relevant tweets and non-spam tweets. As mentioned above, spam analysis is for testing purposes and rely on naive if-then-else production rules.

4.1.1 Tweet based Forecast

All tweets with positive polarity with a related party label were used to forecast election results without doing any filtering on user (twitter account) level. Therefore, one user account has the possibility to contribute to the voting with more than tweet. This approach tests the hypothesis that whole twitter medium can reflect general public opinion.

Predictions from whole relevant tweet dataset (spam + unspam):

Akparty: 49,29

Chp: 18,28

Mhp: 14,27

Hdp: 19,57

R-Squared value of the prediction against real election results: 0.8634

Prediction from only non-spam tweet dataset:

Akparty: 48,11

Chp: 18,03

Mhp: 14,27

Hdp: 19,57

R-Squared value of the prediction against real election results: 0.8578

4.1.2 User based Forecast

To give each user (Twitter account) the right of one tweet, meaning one vote, among all tweets with positive polarity, each user's only last tweet is selected as the vote. This approach test the hypothesis that not whole tweet medium but individual twitter accounts' tendencies may correspond to general public opinion.

Predictions from whole relevant tweet dataset (spam + unspam):

Akparty: 42,31

Chp: 20,89

Mhp: 15,11

Hdp: 21,66

R-Squared value of the prediction against real election results: 0.7940

Predictions from only non-spam tweet dataset

Akparty: 42,01

Chp: 21,17

Mhp: 14,89

Hdp: 21,90

R-Squared value of the prediction against real election results: 0.7879

4.1.3 Election Forecast Discussion

Table 29. Election Forecast Results

	Tweet Based Forecast R-Square Values	User Based Forecast R-Square Values
Whole Dataset	0.8634	0.7940
Non-Spam Dataset	0.8578	0.7879

In the above table of R-Square values of Tweet based forecast and User based forecast against Turkish General Election November 1, 2015, best R-Squared result is obtained by taking whole dataset as the input. Therefore, for the Turkish General Election of November 1, 2015 best predictor for the election result is not the approach of user level analysis but overall tendency in relevant tweets. Our naive spam filtering mechanism, by the way, does not improve predictive accuracy both on tweet and user level analysis.

Against the claims of some relevant papers discussed above in the literature review section, tweet volume and user preferences based on twitter account do not predict election results well even they are employed with state-of-the-art machine learning and natural language processing algorithms as opposed to simple word lexicon based approaches. Although R-Squared values can be considered as somewhat high, predictions cannot capture the relationship between MHP and HDP vote percentages. In all predictions, HDP gets higher vote than MHP, but in the real

election MHP gets more vote than HDP. This problem can be considered as the biggest deficiency of the predictive model.

Therefore, another important insight for election prediction literature is a successful approach should prove itself on several elections to be statistically important. A novel approach can be carried on a single election and can give good predictive results, but to be able to consider that approach as successful it should forecast several elections accurately.

4.2 Case2: US Primary Elections

In US Presidential primary elections, 2016, modeling results created by daily polarity distribution for two presidential candidates (Hillary Clinton and Donald Trump), are analyzed in terms of daily fluctuations. In the Real Clear Politics (RCP) aggregated poll results gathered via link$_2$, there are X fluctuations between 2016-02-13 and 2016-05-12 (90 days' time frame in which tweets are collected for this work). This work models tweet dataset on tweet and user basis like in Turkish case and compares fluctuations in the model and RCP dataset.Between 2016-02-13 and 2016-05-12, RCP datasets 24 fluctuations (increase or decrease).

Accuracies:

Tweet based: 23 / 24 = 0.95833

User based = 14 / 24 = 0.58333

1 http://www.realclearpolitics.com/epolls/json/5491_historical.js

As in the case of Turkish election, Tweet based modeling highly over performs user based modeling.

4.3 Comparisons of Case1 & Case2 methodologies

In Turkey because of the availability of twitter data set before the real elections, dependent variable is the real election results. However, for US presidential primaries, we have a time series of national poll results day by day which creates a time series. To model national polls time series, this work creates a time series out of its daily twitter dataset polarity model and analyzes daily fluctuations prediction rate.

CHAPTER 5:

FOREIGN POLICY ORIENTATION ANALYSIS

5.1 Dataset

22 prominent members of Donald Trump's campaign team members' tweets are collected via Twitter API:

- Donald Trump - @realDonaldTrump
- Melania Trump - @MELANIATRUMP
- Mike Pence - @GovPenceIN
- Paul Manafort - @PaulManafort
- Michael Glassner - @michaelglassner
- Ken McKay - @KenKMcKay
- Barry Bennett - @BBennett152
- Brian Jack - @briantjack
- Daniel Scavino - @DanScavino
- Justin McConney - @JustinMcConney
- Jason Miller - @JasonMillerinDC
- Katrina Pierson - @KatrinaPierson

- Roger Stone - @RogerJStoneJr
- Sam Clovis - @Clovis2014
- Sarah Huckabee Sanders - @SarahHuckabee
- Michael Biundo - @MichaelBiundo
- Kellyanne Conway - @KellyannePolls
- Omarosa Manigault - @OMAROSA
- Joseph E. Schmitz - @josepheschmitz
- Walid Phares - @WalidPhares
- Ben Carson - @RealBenCarson
- Chris Christie - @ChrisChristie
- Michael Cohen - @MichaelCohen212

totaling 44.459 tweets, 3000 of which belong to Donald Trump. Token distribution for the Trump and his campaign team's tweets after filtering out stop words:

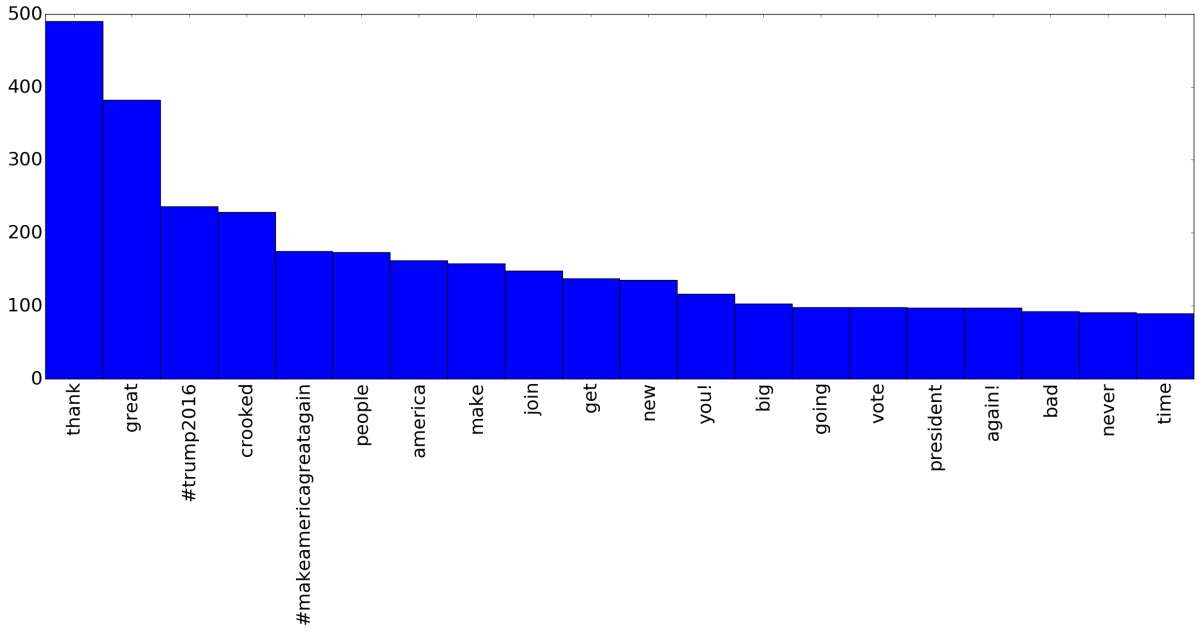

Figure 23. Trump Account Token Distribution

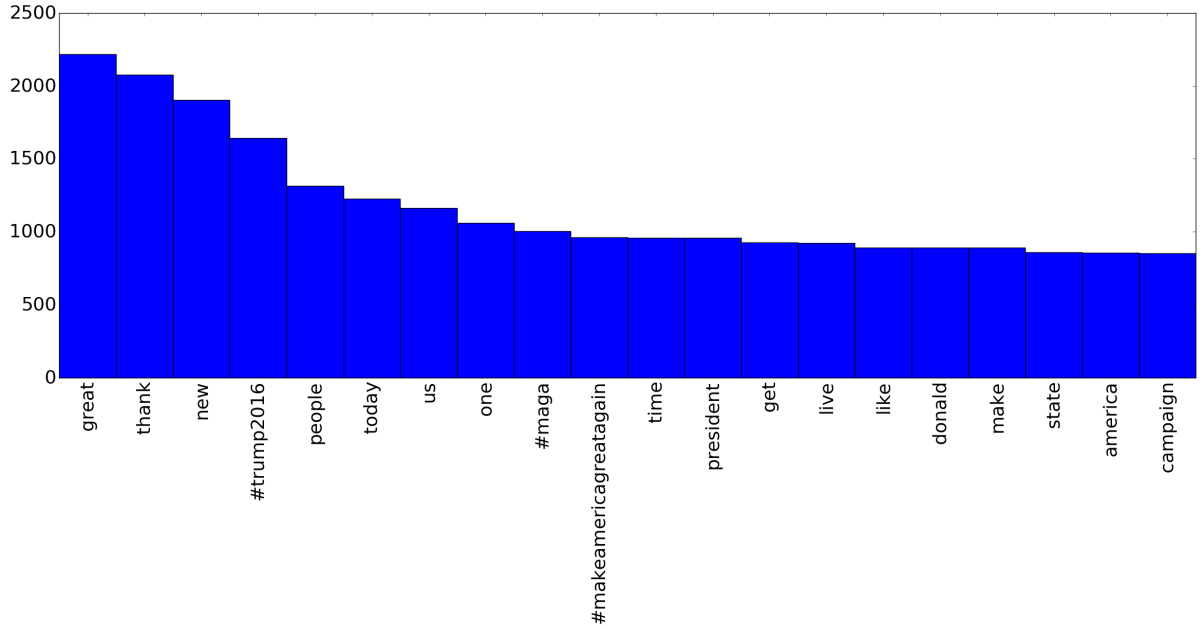

Figure 24. Trump Campaign Team Token Distribution

5.2 Issues

To analyze how Trump's campaign team focused on foreign policy issues during the campaign, several issues highly discussed in the campaign period are selected to analyze, Russia (and President Vladimir Putin), China, Syria – Iraq – Terrorism – ISIS and Iran to find out around which point they revolve around.

5.2.1 Russia-Vladimir Putin

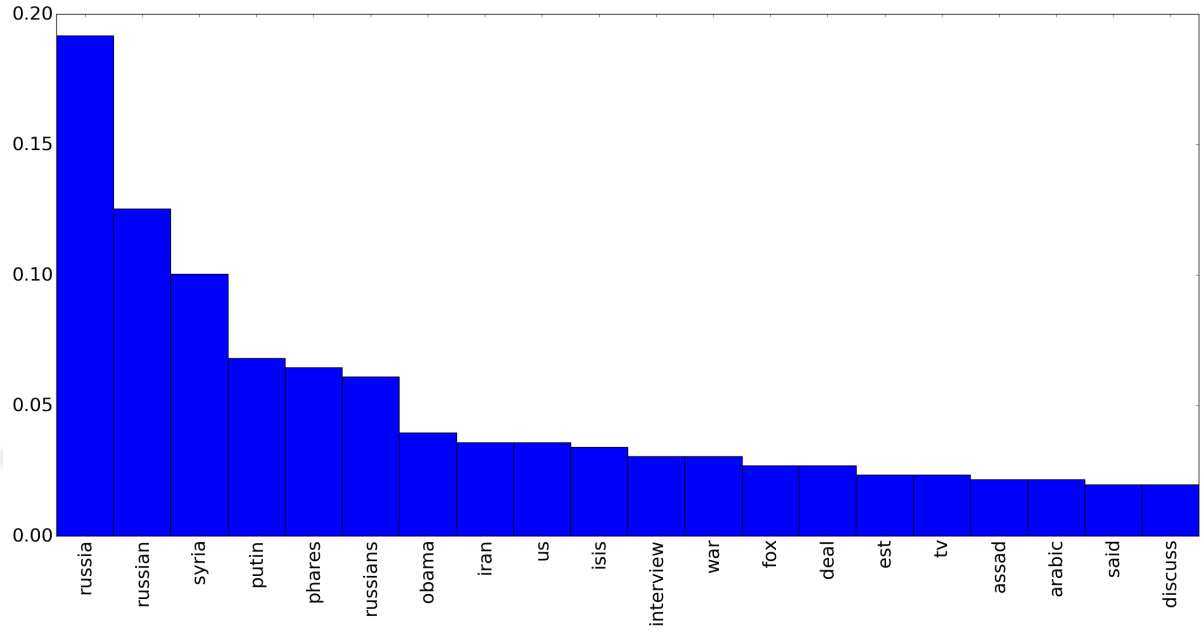

Figure 25. Russia - Vladimir Putin Tweets Token Distribution

As it seen in the above token distribution bar chart, Trump's campaign team's tweets about Russia and its leader, President Vladimir Putin are oriented generally towards the Syrian issue. With the US and Russian involvement in Syrian Civil War for the Trump presidency period, it is highly possible that US-Russian relationships will revolve around the Syria in the near term. Also, it can be expected that because the discussion is highly correlated with Obama we can see a different policy compared to Obama's handling of the Syrian issue with Russia.

On the Iranian issue, it is expected that Trump will try to work with Russia and President Vladimir Putin cooperatively because he sees two issues (Russia and Iran) correlated. However, because Trump stresses his tough character and US hard

power, although he will first try to cooperate with Russia, at the end there can be disputes between USA and Russia.

5.2.2 China

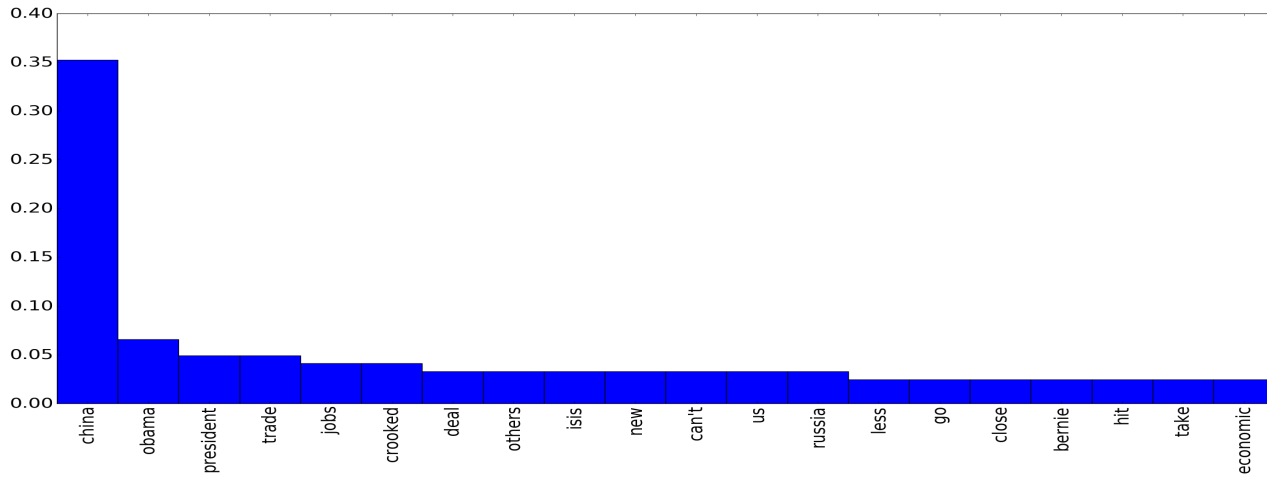

Figure 26. China Tweets Token Distribution

China related tweets reflect the discussion mainly about Obama's foreign policy towards China in his two presidential terms and Trump's emphasis on domestic production of American goods instead of Chinese import, an issue Trump had stressed many times in his campaign speeches. Therefore, we can expect pressures on US firms about moving their productions and factories from China to US. This also reflects Trump's emphasis on American employment and economic isolation against the critics who supports liberal capitalist policies. There can be quotas on Chinese products, too.

On foreign policy issues, however, we can deduce that Trump and his campaign team do not see China as a partner much other than very small correlation

with ISIS in the tweet dataset. However, in North Korean nuclear problem which again Trump emphasizes his toughness there can be cooperation or dispute between US and China, a possibility that tweet data set does not capture.

5.2.3 Syria – Iraq - Terrorism – ISIS

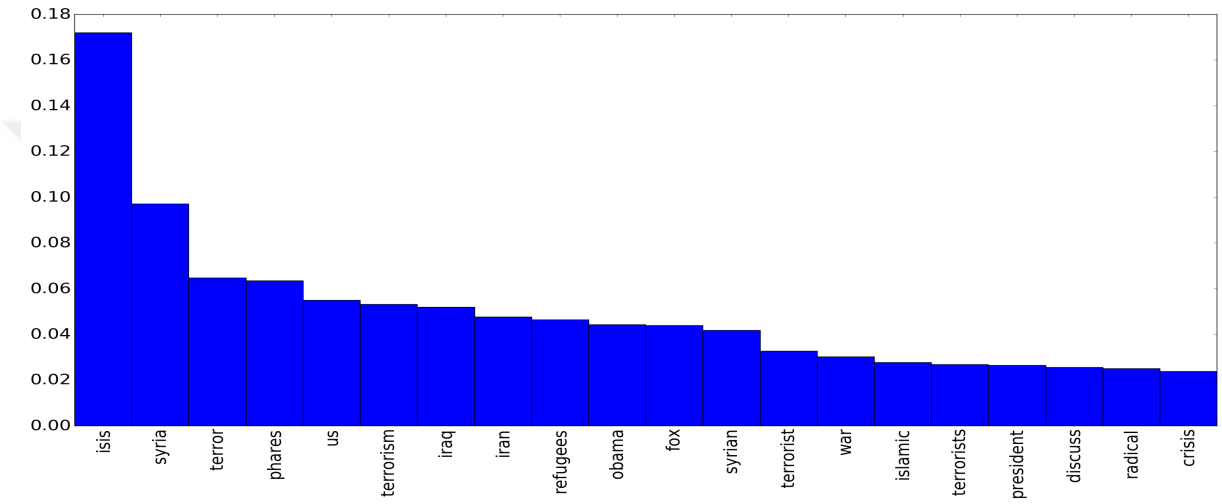

Figure 27. Syria-Iraq-Terrorism-ISIS Tweets Token Distribution

Because of the issues' relatedness, these 4 subjects are analyzed together. As it can be seen from the token distribution, these issues were discussed during the campaign in the context of refugee crisis in Iraq and especially in Syria and constant critic of Obama's foreign policy towards to subjects.

Trump stresses his aim of generally isolating US from the international disputes but in his and his campaign tweets, we see that actually he talk much about Syria and ISIS issues. We may see some actions internationally and domestically like policies about immigration of refugees to US. However, on refugee crisis, because

Trump generally talks about the relationship between Muslims and terrorism, we can see policies which prevent refugee flow to US.

5.2.4 Iran

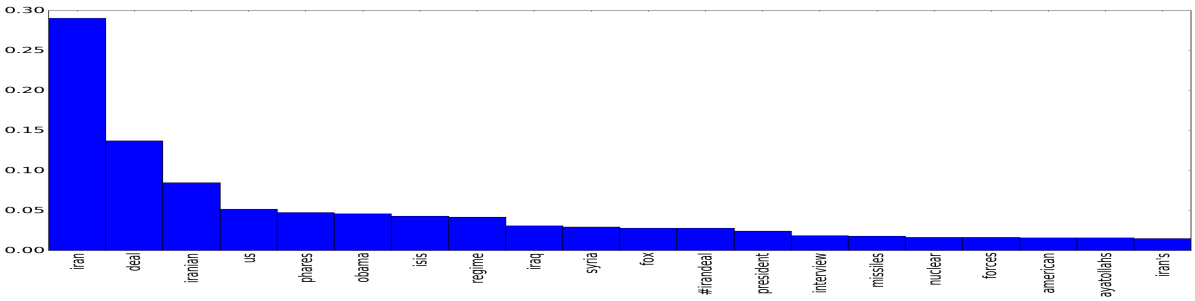

Figure 28. Iran Tweets Token Distribution

Campaign team's discussion about Iran generally revolves around Iranian nuclear deal and Obama's handling of it. What we can expect is however is not definitive from the tweet data. On the one side, Trump criticizes Obama's policies on Iran nuclear and missile issues but he also favors isolationism in foreign policy. Therefore, the best prediction could be Trump will not focus much on Iranian problem but when an extra development occurs in nuclear and missile progress on the Iranian side, he will try to show his toughness and act on it. Here, as it is discussed in the Russia section, Trump first may try to cooperate with Russia but at the end we can see conflicting approaches on US and Russian side.

5.3 Conclusion

Main Foreign Policy issues discussed during the campaign period of US Presidential Election 2016 and related tweets of Trump Campaign Team, almost directly mirror each other in terms of relatedness of the issues among themselves. Token distributions of the related tweets are representative of general public discussion of the issues and their framing by the decision makers.

CHAPTER 6:

CONCLUSION

This final and concluding chapter focuses on several tasks: what was the motivation of the work, revisiting the results so that conclusions of chapters can be combined to get a combined understanding of the study, direct brief responses to this work's main research questions, theoretical and policy relevant implications of the results, limitations and possible future avenues for research.

6.1 Motivation behind the work

The main motivation behind this work can be summarized in: (1) utilizing the growing body of social media data and computer science methodologies to measure public opinion more precisely for better and improved election forecasting, (2) measuring foreign policy orientation of policy makers through their social media data and (3) contributing to Turkish and English NLP literature by increasing the accuracy of semantic analysis. This study achieves these aims by employing cutting edge machine learning and natural language processing algorithms on vast body of publicly available twitter data.

6.2 Revisiting the results

To forecast election results with twitter data first requires an accurate understanding of meanings of political tweets. For this aim, 1.2 million Turkish and 82 million English tweets are preprocessed and fed into machine learning algorithms after careful future engineering using statistical tests and NLP methods. In Turkish and English sentiment analysis, the most accurate algorithm – feature set is Bernoulli Naïve Bayes with statistically selected unigram features. In Turkish, this combination gives 0.92 10-Fold cross validation accuracy in test set and 0.95 in training set. The accuracy values are very close which shows that the algorithm does not fall into overfit trap. Also, with cross validation standard deviation value of 0.01, this combination is very stable across different subsets of test data set. In English tweets, a higher accuracy of 0.96 is achieved. These results are very promising in twitter data set which have 140-character limit per tweet.

After ensuring that the model have an accurate understanding of tweets, this study achieved the successful application of the model to election forecasting in Turkey General Election November 2016 and US primary election 2017. Different approaches are used for these elections: in Turkey, prediction was for the final election result and in US it was for daily increases or decreases for both of the major candidates. The best prediction for Turkey General Election is achieved by measuring the overall tendency of Turkish Twitter data volume without considering each individual as a single voter. With R-Squared value of 0.86 between the predictions

and real election results, the model achieved its objective. In US, daily fluctuations are predicted with 0.958 accuracy.

Against the conclusions of some relevant papers discussed above in the literature review section, tweet volume does not predict election results well. Volume data is should be just a parameter in state-of-the-art machine learning and natural language processing algorithms. And R-Squared values can be considered as somewhat high, predictions cannot capture the relationship between MHP and HDP vote percentages. In all predictions, HDP gets higher vote than MHP, but in the real election MHP gets more vote than HDP. This problem can be considered as the biggest deficiency of the predictive model. Therefore, another important insight for election prediction literature is a successful approach should prove itself on several elections to be statistically important. A novel approach can be carried on a single election and can give good predictive results, but to be able to consider that approach as successful it should forecast several elections accurately.

When we look parties' performances in Twitter, AkParty achieves its degree of popular support it gets in elections almost same in Twitter opinion. It can be concluded that AkParty mobilizes its supporters in Twitter via successful Twitter campaigns. However, Twitter users' perception of CHP is very low compared to CHP's actual election performance. In an era where Twitter is a very important medium to express political opinions and millions of dollars poured into Twitter campaigns, it is worth noting that CHP is unsuccessful to mobilize its young supporters in this medium. Opposite of this true for HDP, HDP has for more support

in Twitter than its election performance, even it has higher predicted percentage in Twitter than MHP. One of the explanations for this is HDP's popularity among young secular people which is also one of the reasons why HDP surpassed 10% election threshold in November 1, 2015 elections.

In the latest aim of this study, then president elect Donald Trump's major foreign policy focuses in campaign period are found to be represented in his and his campaign team's tweets. In Russia and Vladimir Putin related tweets, data is generally oriented around the Syrian issue. The US and Russia's involvement in the Syrian Civil War during the Trump presidency is likely to affect the US – Russia relationships in the near term. Moreover, since the debate is very much related to Obama, it can be expected that we will see a different policy than the Obama's term.

On the Iranian side, Trump may try to cooperate with President Vladimir Putin, and that two issues (Russia and Iran) are correlated in the dataset. Nevertheless, although Trump tries to cooperate with Russia at first, because it emphasizes its difficult character and the toughest strength of the United States, there may eventually be disagreements between the United States and Russia. However, what we can expect is not certain from the tweet data. On the one side, Trump criticizes Obama's policies on Iran nuclear and missile issues but he also favors isolationism in foreign policy. Therefore, the best prediction could be Trump will not focus much on Iranian problem but when an extra development occurs in nuclear and missile progress on the Iranian side, he will try to show his toughness and act on it.

China tweets are about mainly about Obama's economic policy towards China in his presidential terms and Trump's emphasis on the importance of domestic production of American goods instead of Chinese imports, an issue Trump had stressed heavily in his campaign. Hence, we can expect pressures on US firms about moving their operations from China to USA. This also reflects Trump's handling of American employment and stress of economic isolation. There can be quotas on Chinese products, too.

On the other hand, on foreign policy issues we can deduce that Trump and his campaign team do not see China as a partner much other than very small correlation with ISIS in the tweet dataset but in North Korean nuclear problem which again Trump emphasizes his toughness there can be cooperation or dispute between US and China, a possibility that tweet data set does not capture.

Trump stresses his aim of generally isolating US from the international disputes in his campaign talks, but we see that actually he tweets much about Syria and ISIS issues. Therefore, we may see some actions internationally and domestically like policies about immigration of refugees to US from Syria. However, on refugee crisis, because Trump generally talks and tweets on the relationship between Muslims and terrorism, we can expect policies which hardens refugee flow to US.

6.3 Theoretical and Policy Relevant Implications

This interdisciplinary study has several theoretical and practical results for both International Relations and Computer Science fields. In International Relations, this

study shows that foreign policy orientation of key decision makers can be captured with not just their speeches and interviews as in the case of OpCode and LTA but also with twitter data. This information is also very useful to forecast potential future leaders' policy orientations especially when we consider almost all politicians have their official Twitter accounts. With network analysis and natural language processing these politicians' and their close circle can be analyzed by analyzing their tens of thousands of past tweets on specific policy issues.

Furthermore, elections can be forecasted through measuring public opinion in Twitter. These results show the growing importance of social media networks and algorithms on foreign and domestics politics with the possible results of real time opinion polls and predicting in advance the potential foreign policy choices of foreign governments by other governments. By analyzing past elections and past predictions, social media based election forecasts can supplement or can be an alternative against classical public opinion polls. With its real-time nature and ability to analyze millions of public users in an inexpensive way, social media based polls and forecasts have the potential to become the main tool to gauge public opinion on not just political matters but on every issue which classical polls try to measure.

In Computer Science side, even on single tweets with 140-character limit, sentiment can be captured accurately in English and Turkish with state of the art ML, NLP algorithms and statistical analysis.

6.4 Research Questions and Answers

Here, as the last summary, research questions of this study and direct brief responses are given:

1) Can we create a quantified model which can capture the voting intentions of electorate and over performs current methodologies used in the research community? How reliable are these current applied methodologies?

--For Turkish General Election 2015 November:

With R-Squared value of 0.86 we can capture the public opinion to some degree. However, this analysis is just for one election and statistically not enough.

--For US Primaries for Presidential Election:

Tweet sentiment polarity based model predicts fluctuations (increase or decrease) with 0.95833 accuracy.

2) Can we create a model capable of explaining a decision maker's policy orientation on specific issues during the campaign through their own twitter posts and network structure?

-Yes, we can explain the main theme of a foreign policy issue through the campaign team's tweets.

To be able to answer above questions we should be able to capture a tweet's meaning in terms of sentimental polarity (positive vs. negative) for a specific policy issue accurately in a probabilistic model. Therefore, our third research question is:

3) In how much accuracy a tweet's sentimental polarity (positive vs. negative) can be captured?

-In Turkish language: 92%

-In English language: 96%

6.5 Limitations, Future Work

There are several limitations, assumptions and possible future avenues for further research in this study.

(1) People change tendencies they expressed in online world and go to ballot box with different opinions. Some normalization and reweighing may be required to accommodate changes in voting behavior. This can be achieved by analyzing and modeling several elections across the world to get a clear understanding of how people change their voting behavior before and during the election.

(2) Political salience changes especially in times of political instability. Importance of factors before elections may change and affect the outcome or black swan events may occur and shift voting behavior of electorates dramatically. For the first possible danger, it is important to track public opinion closely through the days prior to the election day. And for the second, because every probabilistic model is exposed to the unpredictable black swan events, nothing can be done anything do overwrite and cancel these effects. Every predictive model makes assumptions about the real world and ignores very low probability issues to have a good accuracy on a set of events.

(3) Twitter users are not representative of the voter population and even it is unrepresentative sample of an unrepresentative sample. People eligible to vote are not represented proportionally in Twitter user community hence demographically and geographically twitter users are biased. Also, not all twitter users post their political

opinions and some of them such political partisans, tweets much more heavily than others which creates selection bias in the dataset. Another reason for bias in twitter data is different groups react to different topics which makes dataset heterogeneous in terms of user cluster – political salience pairs. However, it can be stipulated that because the dataset consists of hundreds of thousands of tweets and is large enough statistically and computationally, it may offer unbiased results supported by the law of large numbers and Glienko – Contelli Theorem:

Law of Large Numbers: mean of a sample converges to the mean of the population.

Glienko- Contelli Theorem: the unknown distribution of random variable x of an attribute in a population can be approximated with observed distribution x.
In other words, although the sample is not representative, it may give probabilistically correlative insights even though it may not capture the casual streams among the variables (Wisdom of crowds).

(4) Tweets contain spam, name ambiguity, humor, sarcasm, name ambiguity other types of not clear data. This work tries to eliminate this kind of biases partially by employing a rule based spam filtering mechanism although the model gets more accurate when it does not employ his spam filtering mechanism. However, a machine learning model to identify spams and other types of noises will be much better in terms of predictive accuracy.

(5) In Turkish General Election modeling, there is only one election and its result. Therefore, although this work's model capture the variance of the predicted results with R-Squared value of 0.86, it is just one election and is not enough, statistically. To

capture the voting intentions of Turkish constituents, several elections should be modeled. In US Presidential Primaries, to solve the deficiency of election forecasting modeling in Turkish case, grand truth is chosen as the opinion polls over a long period, represented as time series.

(6) In foreign policy orientation case, Trump campaign team's foreign policy orientation during the campaign period, analysis can be done not just descriptionally but also predictively for the Trump's Presidential term 2017-2021.

REFERENCES

Abramowitz, A. (2012). Forecasting in a polarized era: The time for change model and the 2012 presidential election. PS: *Political Science & Politics*, 45(4), 618-619.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In Proceedings of the workshop on languages in social media (pp. 30-38). Association for Computational Linguistics.

Aisopos, F., Papadakis, G., Tserpes, K., & Varvarigou, T. (2012, June). Content vs. context for sentiment analysis: a comparative analysis over microblogs. In Proceedings of the 23rd ACM conference on Hypertext and social media (pp. 187-196). ACM.

Albrecht, S., Lübcke, M., & Hartig-Perschke, R. (2007). Weblog campaigning in the German Bundestag election 2005. *Social Science Computer Review*, 25(4), 504-520.

Bafumi, J., Erikson, R. S., & Wlezien, C. (2014). National Polls, District Information, and House Seats: Forecasting the 2014 Midterm Election. PS: *Political Science & Politics*, 47(4), 775-778.

Barbosa, L., & Feng, J. (2010, August). Robust sentiment detection on twitter from biased and noisy data. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 36-44). Association for Computational Linguistics.

Bautin, M., Vijayarenu, L., & Skiena, S. (2008, April). International Sentiment Analysis for News and Blogs. In ICWSM.

Bifet, A., & Frank, E. (2010, October). Sentiment knowledge discovery in twitter streaming data. In International Conference on Discovery Science (pp. 1-15). Springer Berlin Heidelberg.

Blumenthal, M. (2014). Polls, forecasts, and aggregators. PS: *Political Science & Politics*, 47(2), 297-300.

Campbell, J. E. (2014). The 2014 midterm election forecasts. PS: *Political Science & Politics*, 47(4), 769-771.

Caruana, R., & Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (pp. 161-168). ACM.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.

Daumé III, H. (2004). Notes on CG and LM-BFGS optimization of logistic regression. Paper available at http://pub. hal3. name# daume04cg-bfgs, implementation available at http://hal3. name/megam, 198, 282.

Davidov, D., Tsur, O., & Rappoport, A. (2010, August). Enhanced sentiment learning using twitter hashtags and smileys. In Proceedings of the 23rd international conference on computational linguistics: posters (pp. 241-249). Association for Computational Linguistics.

Franch, F. (2013). (Wisdom of the Crowds) 2: 2010 UK election prediction with social media. *Journal of Information Technology & Politics*, 10(1), 57-71.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780), 1612.

Gayo Avello, D., Metaxas, P. T., & Mustafaraj, E. (2011). Limits of electoral predictions using twitter. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. Association for the Advancement of Artificial Intelligence.

George, A. L. (1969). The" operational code": A neglected approach to the study of political leaders and decision-making. *International Studies Quarterly*, 13(2), 190-222.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(12).

Goldberg, Y., & Levy, O. (2014). word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.

Goldstein, P., & Rainey, J. (2010). The 2010 elections: Twitter isn't a very reliable prediction tool. Retrieved January, 10, 2012.

Hermann, M. G., & Milburn, T. W. (1977). *A Psychological Examination Of Political Leaders.* Free Press.

Hernández, S., & Sallis, P. (2011, November). Sentiment-preserving reduction for social media analysis. In Iberoamerican Congress on Pattern Recognition (pp. 409-416). Springer Berlin Heidelberg.

Huberty, M. (2015). Can we vote with our tweet? On the perennial difficulty of election forecasting with social media. *International Journal of Forecasting*, 31(3), 992-1007.

Jansen, H. J., & Koop, R. (2006). Pundits, ideologues, and the ranters: The British Columbia election online. *Canadian Journal of Communication*, 30(4).

Jungherr, A., Jürgens, P., & Schoen, H. (2012). Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im "predicting elections with twitter: What 140 characters reveal about political sentiment". *Social Science Computer Review*, 30(2), 229-234.

Kim, E., Gilbert, S., Edwards, M. J., & Graeff, E. (2009). Detecting sadness in 140 characters: Sentiment analysis of mourning michael jackson on twitter. *Web Ecology*, 3, 1-15.

Khan, A. Z., Atique, M., & Thakare, V. M. (2015). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE), 89.

Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg!. *Icwsm*, 11(538-541), 164.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3, 211-225.

Lewis-Beck, M. S., & Stegmaier, M. (2014). US Presidential Election Forecasting. *PS: Political Science & Politics*, 47(2), 284-288.

Lewis-Beck, M. S., & Tien, C. (2014). Congressional election forecasting: Structure-X models for 2014. *PS: Political Science & Politics*, 47(4), 782-785.

Lindsay, R. (2008). Predicting polls with Lexicon. Available at: languagewrong. tumblr. com/post/55722687/predicting-polls-with-lexicon.

Linzer, D. A. (2014). The future of election forecasting: More data, better technology. *PS: Political Science & Politics*, 47(02), 326-328.

Martínez-Cámara, E., Martín-Valdivia, M. T., Urena-López, L. A., & Montejo-Ráez, A. R. (2014). Sentiment analysis in Twitter. Natural Language Engineering, 20(01), 1-28.

Mitchell, T. M. (1997). *Machine Learning*. 1997. Burr Ridge, IL: McGraw Hill, 45(37), 870-877.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Mnih, A., & Hinton, G. E. (2009). A scalable hierarchical distributed language model. In Advances in neural information processing systems (pp. 1081-1088).

Nigam, K., Lafferty, J., & McCallum, A. (1999, August). Using maximum entropy for text classification. In IJCAI-99 workshop on machine learning for information filtering (Vol. 1, pp. 61-67).

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.

Pak, A., & Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In LREc (Vol. 10, No. 2010).

Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). 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). Association for Computational Linguistics.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct), 2825-2830.

Sanders, E., & Van Den Bosch, A. (2013). Relating Political Party Mentions on Twitter with Polls and Election Results. In DIR (pp. 68-71).

Sang, E. T. K., & Bos, J. (2012, April). Predicting the 2011 dutch senate election results with twitter. In Proceedings of the workshop on semantic analysis in social media (pp. 53-60). Association for Computational Linguistics.

Schafer, M., & Walker, S. G. (2006). Operational code analysis at a distance: The verbs in context system of content analysis. In Beliefs and leadership in world politics (pp. 25-51). Palgrave Macmillan US.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, 62(2), 406-418.

Traugott, M. W. (2014). Public opinion polls and election forecasting. *PS: Political Science & Politics*, 47(02), 342-344.

Tumasjan, Andranik, et al. "Election forecasts with Twitter: How 140 characters reflect the political landscape." *Social Science Computer Review* 29.4 (2011): 402-418.

Walker, S. G., Schafer, M., & Young, M. D. (1998). Systematic procedures for operational code analysis: Measuring and modeling Jimmy Carter's operational code. *International Studies Quarterly*, 42(1), 175-189.

Williams, C., & Gulati, G. (2008). What is a social network worth? Facebook and vote share in the 2008 presidential primaries. American Political Science Association.

Yenigun, G. E. (2013). Social Networks and Collective Action Outcomes: Do Mobilization and Alliance Structures Matter? (Doctoral Dissertation, University of Pittsburgh).