Insights of United States Presidential Elections 2016

A project report submitted as a part of Natural Language Processing Course (CS6320)

by

Sai Vivek Kanaparthy (NetID: sxk163030)

Fall 2016

ABSTRACT

US Presidential elections are the most awaited event in the last quarter of 2016. Before the elections there were three debates between the two nominees, Donald J. Trump and Hillary Clinton. Analyzing these presidential debates will give an overview about the propaganda of the candidates. To read deeply inside the debates, text analytics is performed to analyze the word usage of the candidates. Opinion Mining of the candidate responses to the questions raised by the host and other nominee. The overall reaction of the audience to the candidates responses is also evaluated. Analyzing the debate text to get insights about the topics emphasized by both the candidates.

Keywords: Named Entity Recognition, Chunking, Sentiment Analysis, Bigram, Unigram

Contents

Chapter 1

INTRODUCTION

1.1 Overview

In November 2016, the president for United States was elected. Before which three presidential debates were held between the two nominees, Donald J. Trump and Hillary Clinton. These debates are very resourceful in analyzing the agenda of the candidates which can be analyzed from their responses during the debates. The three debates are individually analyzed to get insights of each of them separately. Questions like the following can be inferred from the project:

- What are the topics spoken by each candidate?
- How did the audience react to the candidates during the debate?
- What is the sentiment of the candidates response?
- What is the frequency analysis of the words spoken by the candidates?

The recent debates have drawn the attention of everyone in changing their opinion about the candidate. So, a deep study of the debates to get better insights is done in this project as the debates have impacted the polls of elections. Debates have an impact on the peoples perception about the candidate. Thus, it has an effect on the percentage of the votes voted for a candidate. The way the candidate responds to the questions posed will enlighten people about the candidates nature of reacting to situations and his perceptions in handling them.

1.2 Related Work

Benoit made a study in [1], about the primary debates in the Republican party of 2012 US elections. The results of the primary debates were used to analyze the presidential debates between President Barack Obama and Mitt Romney. The analysis was mainly on general goals and personal character of the candidates and the past activities of the candidates.

Chapter 2

PROPOSED METHODOLOGY

2.1 Dataset Description

The dataset, ”2016 US Presidential Debates, Full transcripts of the face-off between Clinton Trump” is taken from kaggle datasets. The dataset contains a conversational format of the responses of the candidates along with the host and the audience.

Format of the dataset:

- Line: Line Number
- Speaker: Candidates - Trump or Clinton, Host - Holt,Wallace,Cooper,Raddatz and Audienc
- Text: Text spoken by the Speaker
- Date: Date of the debates (2016-09-26, 2016-10-04, 2016-10-09, and 10/19/2016) to distinguish events

To access the data as data frames, Pandas library in Python is used.

2.2 Pre-Processing Techniques

Removal of Stop Words
A list of stopwords is maintained which inlude words which do not contribute to the meaning of the sentence rather they are used only in relating the phrases. Punctuation marks are also processed and removed for better analysis.
Lemmatization
For a better relation between the words, the lemma of the words are being evaluated. Wordnet is used for this purpose.

2.3 Word Usage of the Candidates

The frequency analysis of the words spoken by the candidates are being evaluated to identify the usage of the words. Two analysis are made for words by taking unigrams and bigrams.

A frequency analysis of the words tokenized directly did not lead proper results. So a better measure to evaluate the important words spoken by the candidate, 'TF-IDF' is used. TF-IDF stands for Term-Frequency Inverse Document Frequency. The value for a word increases with more usage in a document but is held back by the frequency count of the corpus.

$$tf - idf_{t,d} = (1 + \log tf_{t,d}) \cdot \log \frac{N}{df_t}$$

t - term; d- document; N- total number of documents in the corpus ;

2.4 Named Entity Recognition

Named-entity recognition (NER) is a technique used to extract and locate the named entities into a set of categories like person, organization, value(monetary), locations, quantities and time.
Example:
I think you have to knock out ISIS. Right now, Syria is fighting ISIS.
Tree('S', [('I', 'PRP'), ('think', 'VBP'), ('you', 'PRP'), ('have', 'VBP'), ('to', 'TO'), ('knock', 'VB'), ('out', 'RP'),
Tree('NE', [('ISIS', 'NNP')]), ('.', '.'), ('Right', 'RB'), ('now', 'RB'), (',', ','),
Tree('NE', [('Syria', 'NNP')]), ('is', 'VBZ'), ('fighting', 'VBG'),
Tree('NE', [('ISIS', 'NNP')]), ('.', '.')])

2.5 Chunking

Chunking is the process to identify entities. Chunking is the process of selecting a set of tokens to form a meaningful sense. Noun-Phrase Chunking is used to select noun phrases in the corpus. In this project, I have used NP - Chunking which contain adjectives. This chunks are analyzed for sentiments.
In order to do NP-Chunking, the grammer is defined as follows:
$grammar = NP : < DT|PP\$ >? < JJ > + < NN >$

2.6 Sentiment Analysis

Opinion Mining is a synonym for Sentiment Analysis. Classifying the polarity of the given document, text or feature based on the opinion in the document, text or feature as positive, negative or maybe neutral. Pang [2] has applied different methods for determining the polarity of movie reviews.

The sentiment of the adjective phrases are evaluated to determine the positve and negative responses of the candidates to the questions raised by the host and the other contestant. I have used bag of words approach from nltk toolkit for evaluating sentiments.

Chapter 3

RESULTS

3.1 Phase I

Unigrams - Frequency Distribution
The Word frequency of Hillary Clinton:
The words that are spoken maximum by Hillary Clinton can be seen in the following table in the respective debates. She speaks more about people, country, women, insurance, job, tax, business etc. 'Donald' is also spelt many times but it is to refer the other candidate, so, it has less significance.

Table 3.1: Clinton Word Usage during debates

Word(10/19/2016)	Frequency	Word(10/9/2016)	Frequency	Word(9/26/16)	Frequency
people	38	donald	33	people	33
country	31	president	25	donald	30
donald	31	country	25	need	23
woman	30	insurance	16	job	21
president	23	american	16	tax	20
american	20	first	16	really	20
job	20	work	15	business	19
kind	18	woman	15	good	18
work	16	child	15	work	17
tax	15	health	14	country	17

The Word frequency of Donald Trump: The words that are spoken maximum by Donald Trump can be seen in the following table in the respective debates. He speaks more about people, country, border,percent,Mosul, tax, Russia etc.

Table 3.2: Trump Word Usage during debates

Word(10/19/2016)	Frequency	Word(10/9/2016)	Frequency	Word(9/26/16)	Frequency
people	53	country	36	country	64
country	39	hillary	27	people	35
many	23	state	24	year	28
border	23	tax	24	many	28
hillary	19	great	22	company	28
great	19	isis	19	secretary	27
year	18	year	19	job	27
bad	18	money	17	deal	22
percent	18	disaster	17	clinton	22
mosul	17	russia	17	tax	21

Table 3.3: Clinton Bigram Usage during debates

Bigram(10/19/2016)	Frequency	Bigram(10/9/2016)	Frequency	Bigram(9/26/16)	Frequency
united state	7	supreme courty	11	new job	7
second amendment	7	donald trump	8	united state	5
president obama	6	health insurance	8	criminal justice	5
donald think	6	affordable care	5	tax return	5
supreme court	5	care act	4	middle class	5
social security	5	never apologized	4	trade deal	4
roe wade	5	united state	4	justice system	4
new job	5	health care	4	secretary state	4
clinton foundation	4	secretary state	4	nuclear weapon	4
planned parenthood	4	insurance company	4	young african	3

Table 3.4: Trump Bigram Usage during debates

Bigram(10/19/2016)	Frequency	Bigram(10/9/2016)	Frequency	Bigram(9/26/16)	Frequency
second amendment	9	united state	13	secretary clinton	22
million people	8	hillary clinton	12	trade deal	8
united state	7	inner city	10	law order	7
percent percent	6	bad judgment	8	sean hannity	7
open border	6	president obama	7	long time	6
hillary clinton	6	bernie sander	6	president obama	6
strong border	5	energy company	5	inner city	6
year ago	5	carried interest	5	middle east	6
trade deal	5	room talk	5	north korea	5
changed law	4	locker room	5	new york	5

Bigrams make better sense when compared to the unigrams which is evident from bigrams like second amendment, President Obama, Clinton Foundation, planned parenthood etc.

3.2 Phase II

Audience Reactions
Interruption for a candidate corresponds to the host interrupting when the candidate is speaking. The host can interrupt either when the candidate is speaking beyond his time limit or the candidate is speaking controversial. Applause represents the number of times audience gave an applause for the candidates response. Similarly the count is for laughter. In Table 3.5, three columns Int. App. and Lau. represent Interruptions, Applause and laughter respectively for three debates.

Table 3.5: Candidate Vs Interruptions, Applause, Laughter

Candidate	Int.(10/19)	App.	Lau.	Int.(10/9)	App.	Lau.	Int.(9/26)	App.	Lau.
Hillary	98	6	5	61	5	5	26	4	4
Trump	219	7	7	137	6	4	62	5	1

3.3 Phase III

Named Entity Recognition
The NER technique is applied on the debate corpus to identify the topics discussed by the candidates. The NER recognizes the entities like people,organization,locations,quantities etc.

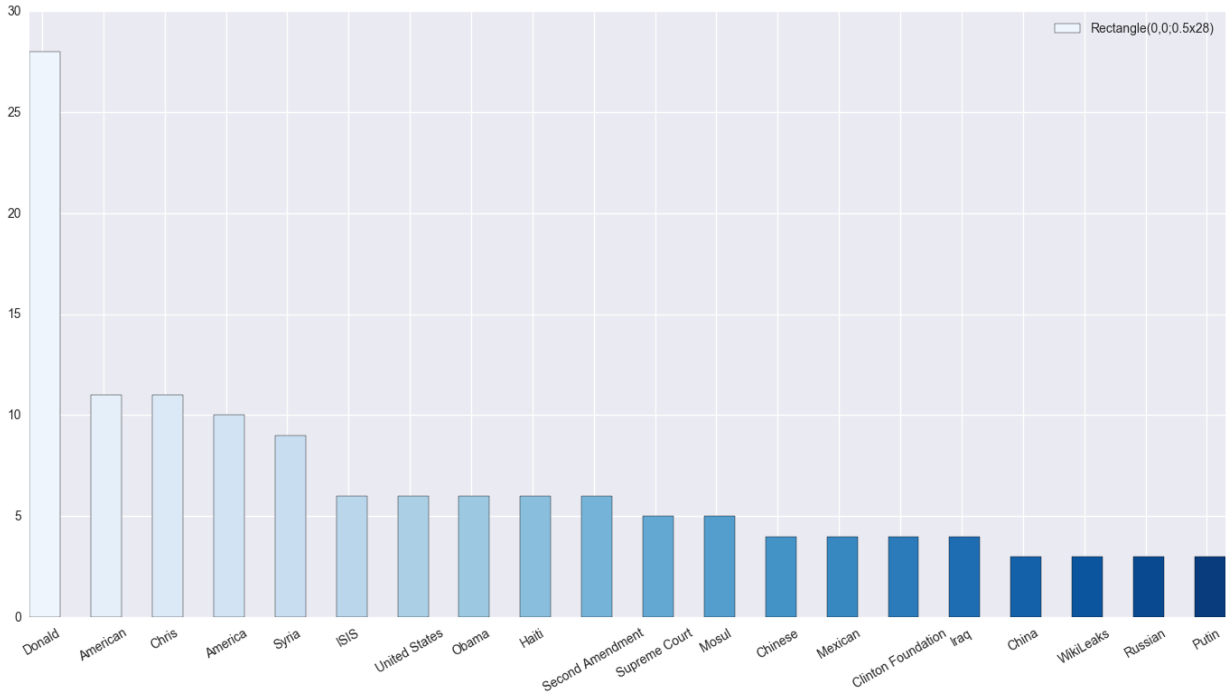

Figure 3.1: Topics spoken by Hillary Clinton on 10/19/16

Thus the important topics revolve around these entities. A frequency distribution is estimated for all the debates separately with respect to the candidate. Sample bar plots for these frequency distribution of topics is shown in Figure 3.1 and 3.2

- The topics spoken by Hillary Clinton are Donald, Chirs, America, Syria, ISIS, United States, Obama, Haiti, Second Amendment, Supreme Court, Mosul,Chinese, Mexican, Clinton Foundation, Iraq, WikiLeaks, Russian,Putin.
- The topics spoken by Donald Trump are Mosul, Putin,Iran, Iraq, ISIS, Obama, Hillary, Russia, Second Amendment, Nobody, Chris, NAFTA, United States, Assad, Hillary Clinton, Wrong, Syria, Obamacare, FBI.

3.4 Phase IV

Sentiments of Chunked Noun Phrases with Adjectives
The noun phrases with adjectives are chunked to estimate the opinion of the candidate to the questions raised by the host. To estimate the opinion, sentiment of the chunked phrases are evaluated. A sample screen shot of Hillary Clinton's debate response is shown in Figure 3.2

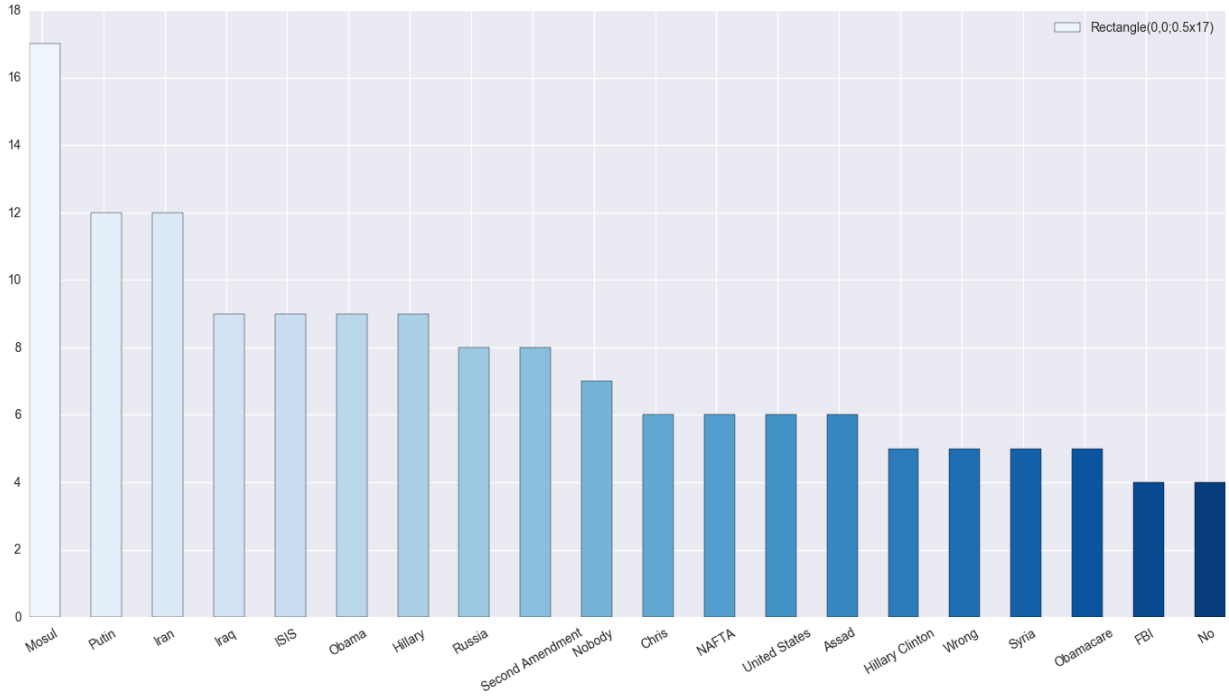

Figure 3.2: Topics spoken by Donald Trump on 10/19/16

```
10/9/16-Clinton
Number of positive sentiments:  26
Number of negative sentiments:  15
Adjective Phrases with positive sentiment:
fair share
a good chance
a comprehensive energy
clean energy
the great privilege
religious freedom
a great job
a great display
carried interest
a deep devotion
Adjective Phrases with negative sentiment:
serious mistake
a terrible mistake
a serious problem
this catastrophic war
financial crisis
violent jihadist
a damaging effect
a terrible time
a big problem
a billion-dollar loss
```

Figure 3.3: Sentiments of a candidate's responses and sample phrases

Chapter 4

Conclusion

Initially the unigram and bigram frequency distribution was analyzed to detect the most spoken words of the candidates which gave information about the topics but was vague. A clear understanding of the propaganda of both the candidates was obtained through the topics discussed by both of them. Named Entity Recognition was used to identify the topics discussed. The topics found to be significant as they dealt with many sensitive topics. It was evident that Donald Trump focused on topics like ISIS, Mosul, Iran and Iraq(top 4 topics) and their relation with United States with a higher frequency where as Clinton was speaking more about Donald, America, Chris and Syria. The third debate was very influential and ended with a positive impact on Trump.
It is found that Trump was interrupted more when compared to Clinton either because of talking beyond the time limit or was controversial But the audience clearly applauded more for Trump compared to Clinton but the difference was small. The noun phrase chunking with adjectives has given more insights on the positive and negative sentences spoken by the candidates. The sentiments of these phrases are evaluated to identify the positive and negative phrases spoken by the candidates. This gave a better overview about the candidates response nature towards the questions which was evaluated by chunking the noun phrases with adjectives and determining their sentiments.
Implemented MALLET(machine learning for language toolkit) but found less accurate results as this data would not be sufficient to fit the model because it works on large multiple documents for topic modelling. So using NER, wordnet, chunking and sentiment analysis from nltk, the objective of the project is achieved.

4.1 References

[1] Benoit, William L. "Political Election Debates." The International Encyclopedia of Political Communication (2013).

[2] Bo Pang; Lillian Lee and Shivakumar Vaithyanathan (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 7986.

[3] http://www.nltk.org/book/ch07.html

[4] https://en.wikipedia.org/wiki/Tfidf

[5] http://nlp.stanford.edu/software/CRF-NER.shtml