Rhetoric Analysis: Tweets vs Speeches

Pranav Shah
School of Computing
University of Utah
Utah, USA
u1266563@utah.edu

Mike Swenson
School of Computing
University of Utah
Utah, USA
u0585863@utah.edu

Sara Nurollahian
School of Computing
University of Utah
Utah, USA
u1217653@utah.edu

Matthew Timpson
School of Computing
University of Utah
Utah, USA
u1241011@utah.edu

ABSTRACT
With the development of a variety of social media platforms, humans tend to interact differently on different platforms. At times, our social media personality is different from who we are in real life. Browsing twitter is a considerable time drain for many demographics. Considering how easy it is to obsess over social media, influencers and people who hold a position of power make a significant impact on the world population. In this project, we will be analyzing the ex-president of the United States –President TRUMP. We will be comparing his social media interaction to his speeches. This will help us understand the rhetoric and behavior differences found in formal speeches vs. social media. Additionally, we will provide sentimental analysis of the words used on both platforms. This will justify the power of social media in politics.

Keywords
Twitter Data, Trump, President, Political Analysis, Data Mining, Sentimental Analysis, Polarity, Natural Language Processing, Similarity Matrices, Data Preprocessing, Twitter Dataset, Topic Modeling, LDA, BTM, Cosine Similarity, Jaccard Similarity, personality matching.

1. INTRODUCTION
On social media, you can edit what people see [1]. You may take hours to decide on what you want to post and calmly reply to the comments without the spontaneity and body language. If we consider the COVID-19 pandemic, people spend more time on social media than ever before. Social media has become a powerful tool for some and a weapon for others. It is important to know that we are looking at a filtered version of a person on social media. Hence, people may have different reactions to comments when they are in-person. When we are talking about electing a person to be the President, we need to be careful about their social media personality too. The importance of this analysis is found in the differences that arise when comparing social text and public oration. This helps us understand the true nature of a person. Considering the recent chaos about the elections in the United States of America, there are a lot of questions which go unanswered such as: Is the President acting differently online and in-person? How does the President act on social media during special events? What is the people's reaction to the speeches and tweets of the President? Does the polarity of tweets influence the public's attention? And many more questions. To answer all of these, it is important to analyze the data.

2. PROBLEM STATEMENT
It is challenging to understand the character of Politicians. The consequences of words spoken and written by political figures is great and far reaching. Thus, it is important to understand the impact of their words. Considering this, the need to build a tool to identify the rhetoric and behavioral differences found in formal speeches vs. social media is apparent. To build our tool we will be studying the ex-President of the United States of America – President TRUMP.

3. DATASET
We used two different datasets to study ex-President Trump. The two datasets include Twitter tweets and the speeches given by the ex-President Trump.

3.1 Twitter Dataset [2]
This dataset is from Kaggle that includes all of Donald Trump tweets from 2009 - 2020.

3.2 Donald Trump Transcripts [3]
This dataset is from Kaggle that includes all of Donald Trump's transcripts for his speeches.

4. METHODOLOGY
The workflow of the project includes data preprocessing followed by applying data mining techniques such as K-Grams, Similarity and Semantic Matrices, Sentiment Analysis and Topic Modeling approaches such as LDA and BTM.

4.1 Clean the Data with NLP Techniques:
In order to be able to analyze the text data, preprocessing is the first step to be taken. Since raw data is unstructured and includes redundant information. For this goal we used the well-known NLTK (Natural Language Toolkit).

- **Tokenizing:** The first step completed was tokenizing the transcript dataset. This was as simple as opening the file and splitting the words based on spaces.

  The Twitter data set had a similar approach, but the data contained more information than just the content of a message that must be removed. This data was filtered for hashtags, URLS, and mentions.
- **Stop Word and Special Characters Removal:** The tokens were then filtered resulting in the removal of the stop words and special characters.
- **Lemmatization:** The words were then lemmatized, reverted to their base meaning, and converted to lower-case for more uniform comparisons. For instance, "rocks" become "rock" or "better" becomes "good."
- **POS Tagging:** Marking and identifying each word with its part of speech

4.2 Generation of Word Clouds using Unigrams and Unigrams in combination of frequent Bigrams:

We used words as our n-gram items. This provides a more semantically meaningful analysis than the choice of characters. Our approaches were:

- **Unigram and Bigrams:** The creation of unigrams was trivial because of the previous steps of tokenization and filtering. Also Unigrams in combination with Bigrams were produced using NLTK and word cloud collocations.
- **Word Clouds:** Word clouds were created using the unigrams and unigrams-bigrams as input.

4.3 Similarity Analysis:

The initial analysis we conducted is the measure of Jaccardian and Cosine similarities.

- **Jaccard Similarity:** This returned a similarity coefficient calculated as the ratio between the intersection and union of the twitter set and transcript set. Formally defined below:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

- **Cosine Similarity:** This returned the cosine of the angle between two vectors. This is a more accurate measure of similarity when the dimensionality of the data is high, such is the case with our data.
  Formally defined below:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

- **Semantic Similarity:** This measure calculates semantic similarity scores between two documents using definitions and synonyms from wordnet synsets.

4.4 Sentiment Analysis:

While communicating with others, besides sharing the information, people also express their feelings and emotions such as fear, anxiety, grief and gratitude. To analyse the sentiment conveyed in the Ex. president tweets on social media vs on public transcripts, we decided to use the well-known LIWC (Linguistic Query and Word Count), which is a dictionary-based lexicon approach that is widely used in many areas especially in social and psychological computing research [5].

Following are the definitions of the LiWC parameters that we have used to calculate sentiment scores:

- **Clout**: Refers to the relative social status, confidence, or leadership that people display through their writing or talking.
- **Authenticity**: When people reveal themselves in an authentic or honest way, they are more personal, humble, and vulnerable.
- **Social**: These are words that make reference to other people (e.g., they, she, us, talk, friends). Generally, people who use a high level of social words are more outgoing and more socially connected with others.

The polarity score consists of two categories, positive and negatively charged emotional words. The negative emotions specifically, anger, anxiety, and sadness, are a subgroup of negative emotions that we reviewed and their definitions are self explanatory.

4.5 Topic Modeling:

Topic model is a type of statistical model for text classification and discovering the abstracts (topics) in a collection of documents. Using topic modeling, we can get better insights about topics of transcripts vs tweets [6].

In this study, we used the well known Latent Dirichlet Allocation model (LDA), to find the hidden semantic structures in the transcript. LDA learns topics from document-level word co-occurrences by modelling each document as a mixture of topics [7,10] that has been proved to perform well in many topic modelling of long documents. However, LDA suffers from sparsity of word co-occurrence patterns in the short documents and thus, is not capable of providing precise models on the social media platforms with short texts like twitter[10]. For this reason, we used Biterm Topic Modeling (BTM) on the tweets. BTM learns topics by modeling the generation of word co-occurrence patterns using unordered word pair co-occurring in short context[10].

Using the coherence measure, we found that using 20 as the number of topics, gives us the highest coherence score. Figure 1 shows 3 topic examples from tweets and transcripts.

Top words for transcript topics examples	transcript- examples
right, wall, build, better, forget, catch	And we will build the wall. Don't worry about it, we will.
test, vaccine, go, china, number	I met yesterday with the biggest drug companies, Pfizer, Johnson and Johnson. We are moving at a maximum speed to develop the therapies, not only the vaccines but the therapies. I can call it china virus, …
Isis, deal, thank, Iran, Israel	We obliterated the ISIS caliphate, and we killed the leader of ISIS, Soleimani is dead. I withdrew from the last administration's disastrous Iran Nuclear Deal. I recognized the capital of Israel…
Top words for tweet topics examples	**tweet- examples**
Obama, Iran, attack, Isis, kill, bad, Syria	It is time for Iran to face serious consequences. Obama's plan to have Russia stand up to Iran was a failure. Iran was planning to attack the Israeli and Saudi DC embassies. We respond accordingly
border, wall, country, security, illegal, immigration, Mexico	The fight against ISIS starts at our border. ISIS have been caught crossing Mexico. border. Build a wall, the border is wide open for cartels &amp, terrorists, build a wall, deduct the costs from Mexican foreign aid
china, country, trade, deal, pay, great, tariff job	China is closing a massive oil deal w/ Russia, taking advantage of the Ukraine conflict. China will now pass our economy this year,

Figure 1: 3 Example Topics out of 20 for Transcripts & Tweets

5. RESULTS

5.1 K-Grams (Unigram and Bigram):

Four word clouds, each composed of 200 words, were generated. Figure 2 and figure 3 show the visual representations of the top 200 words that appeared most often in transcripts and tweets respectively. In each figure, the left image represents word clouds using unigrams and the right image consists of Unigram and Bigram together.

Figure 2: Transcripts' word clouds (Unigrams and Unigrams in combination of bigrams)

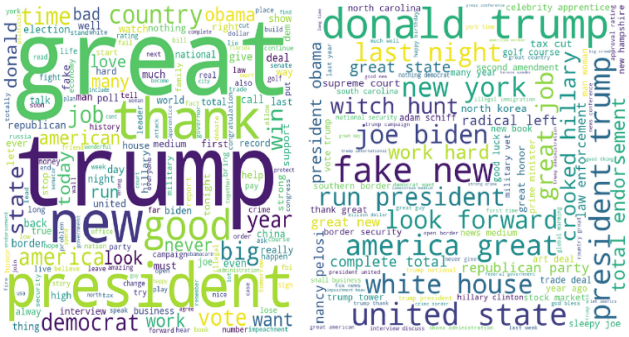

Figure 3: Tweets' word clouds (Unigrams and Unigrams in combination of bigrams)

As it can be seen from the above figures, using unigrams in combination with bigrams leads to more intuitive results in comparison to only using unigrams. For example, "White house", "supreme court", "Joe Biden" are some examples of words that make more sense when coming together and combinations of unigram and bigrams could capture these words together.

5.2 Similarity Measure:

The initial measure of Jaccard similarity between Trump's transcripts and tweets was done using words in each dataset. Since they are relatively small data sets, an intersection magnitude of 5504 and union size of 36360. Thus, Jaccard similarity is~15.1% between tweets and the transcripts. It was then recalculated using the top 200 most common words from each set, generated for the word clouds, and that measured at ~28.2%. The reason for this recalculation is that Jaccard has bias toward the length of the document [9], and we want to make this bias as little as possible.

Then we computed Cosine similarity on all words in tweets vs transcripts as ~36%. We will not try cosine on the 200 most frequent words since, the frequency of words can change the cosine similarity [9].

5.3 Sentiment Analysis: LiWC:

LiWC, pronounced "luke", is a method sentiment analysis that clusters words into containers of similar meaning and provides scoring for each category. These are the definitions of the different containers:

This analysis in figure 4 showed us that the transcripts scored higher on the categories of clout and authenticity.

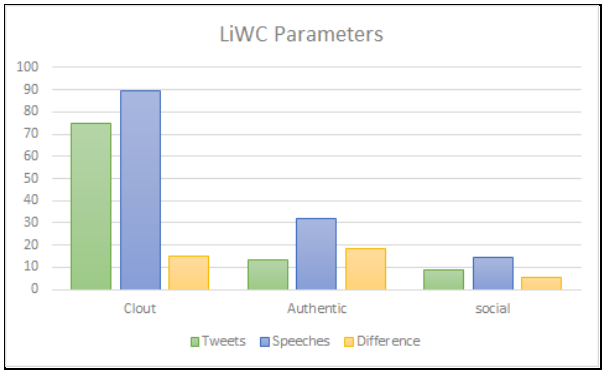

Figure 4: Sentiment Analysis

Figure 5 states that the specific scoring of the tweets and transcripts for polarity was almost the same for positivity. However, the two data sets differed significantly in regards to negativity. The tweets scored approximately 75% higher for negative words than the transcripts. This can be corroborated by the scoring found in anxiety, anger, and sadness.

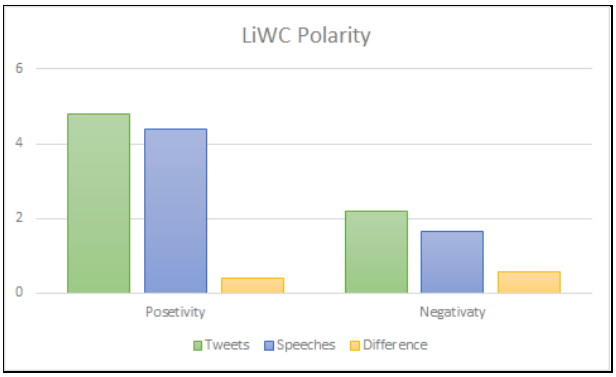

Figure 5: Sentiment Analysis(Polarity)

5.4 Topic Modeling:

Entire transcripts and tweets were clustered into 20 topics each where every topic consisted of 20 top words.

5.4.1 Sentiment Analysis: Polarity and Subjectivity

We have used Spacy to calculate Polarity and Subjectivity for each cluster of words (topics).

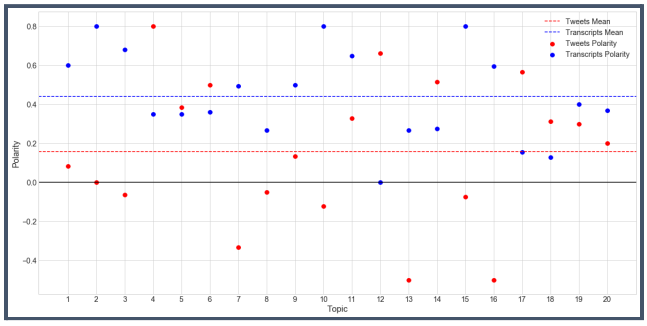

Figure 6: Sentiment Analysis: Polarity for 20 Topics

In the above figure 6, we can see that the blue dots represent the Polarity for Transcript topics and red dots represents Polarity for Tweet topics. The dotted lines represent the average score, respectively. As we can see Transcripts have a higher average polarity score than Tweets. Also, there are about 7 Tweet Topics which have a negative score, which also explains why the Negative emotion is higher for Tweets than Transcripts (Refer Figure: LiWC).

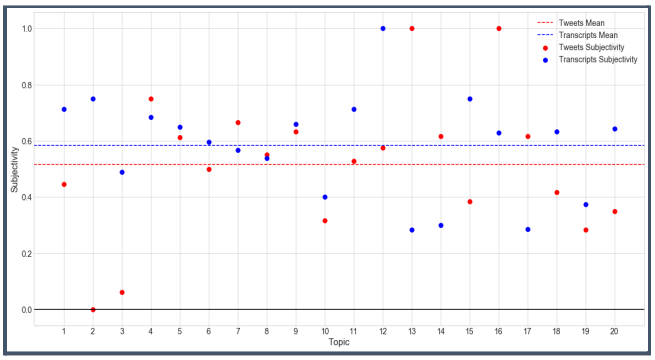

Figure 7: Sentiment Analysis: Subjectivity for 20 Topics

In the above figure 7, we can see that the blue dots represent the Subjectivity for Transcript topics and red dots represents Subjectivity for Tweet topics. The dotted lines represent the average score, respectively. As we can see Transcripts have a higher average subjectivity score than Tweets. Higher subjectivity means that the text content contains more personal feelings and opinions. This is calculated by comparing the text across a predefined set of words. From our LiWC model, we can see Transcripts have a higher Clout, Authenticity and Social score. This explains why our average subjectivity for Transcripts is higher than Tweets from Topic Modeling.

5.4.2 Semantic Similarity Analysis:

Here, we have implemented our own semantic similarity measure methodology. We calculated similarity scores between every pair of Transcript and Tweet Topic clusters. We noticed that the highest score we could achieve was 47% between Topic 12 from Tweets and Topics 18 from Transcripts. This also relates to the fact that we had a low similarity score between the overall data of Tweets and Transcripts (Refer section 5.2). And even after clustering we have a low similarity score. That justifies that our clusters are well formed. And the data from both sources is not completely similar.

6. CONCLUSIONS

We have evaluated the soundness of our plan and believe our results are reasonable and judicious. The word clouds provide great information about most frequent words in transcripts vs tweets. We noticed that there is a significant difference between the top words used on both platforms.

This was confirmed through our initial analysis using Jaccard and Cosine similarity. Both of these metrics scored quite low for similarity between these sets specifically, Jaccard had ~15% and cosine had ~36%.

However, Jaccard and Cosine similarity measures are quite shallow because they do not capture the sentiments or compare semantics between the words. Thus, we implemented other methods like sentiment analysis (LiWC) and topic modeling.

From Topic Modeling, we see that Tweets have 7 topics which have a negative polarity whereas the transcripts have just 1. This can be confirmed from LiWC which states that Trump tends to be more negative with more expressions of anger and sadness on Twitter than he is in his speeches. Our results may not entirely justify him being banned on Twitter but gives a hint in the right direction.

Beyond that, subjectivity is high for topics from the speech transcripts. From the transcripts' higher clout value, derived by LiWC, we see that Ex-President Trump's more subjective nature in public Speeches comes off as more authentic to the general public.

Hence, we conclude that people might act differently on various platforms. So, before following, judging, voting, or being influenced by someone, we need to analyze their different personalities, starting from public speaking to making statements on social media.

7. FUTURE WORK

Per our expectations, we were able to extract insights from ex-President Trump and study the behaviour on social media and during speeches. Some additional work may be to model the ex-President's behaviour on social media before and after his term as the POTUS.

Furthermore, it would be interesting to evaluate how the society reacts to the ex-President on twitter during special events or topics such as impeachments, elections and so on.

8. REFERENCES

[1] https://choma.co.za/articles/628/the-difference-between-social-media-and-real-life .

[2] https://www.kaggle.com/ironicninja/all-of-trumps-tweets-20092020 Tavel, P. 2007. Modeling and Simulation Design. AK Peters Ltd., Natick, MA.

[3] https://www.kaggle.com/arnavsharmaas/all-donald-trump-transcripts Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.

[4] https://english.kyodonews.net/news/2020/11/6b5a0c4fed46-chronology-of-major-events-under-trump-administration.html

[5] https://www.cs.cmu.edu/~ylataus/files/TausczikPennebaker2010.pdf

[6] https://en.wikipedia.org/wiki/Topic_model

[7] https://www.aclweb.org/anthology/W17-4420.pdf

[8] http://billchambers.me/tutorials/2014/12/21/tf-idf-explained-in-python.html

[9] https://towardsdatascience.com/overview-of-text-similarity-metrics-3397c4601f50

[10] https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=677