SUBMITTED MANUSCRIPT

Cluster Analysis of Trump's Tweets Using Sentiment-Based Methods

Xize Zhang and Chinthaka Kuruwita[a]

[a]Hamilton College, Clinton, New York, USA

ARTICLE HISTORY
Compiled February 19, 2026

ABSTRACT
This paper analyzes Donald Trump's tweets during the first two years of his presidency using a public dataset of approximately 2,000 tweets from 2017—2019. We propose a novel approach for tweet sentiment analysis by using a sentence based deep learning model, SBERT. Our approach provides sentence level representations (embeddings) which are then used to create clusters of tweets for further linguistic analyses.

We successfully identified distinct sentiments and communication styles across clusters using topic modeling via Latent Dirichlet Allocation (LDA) and sentiment score models. These findings show that Trump's communication style varies systematically by topic. Beyond describing these patterns, the study provides a practical template for combining text embeddings, clustering, and sentiment analysis in political communication research.

KEYWORDS
Sentiment analysis, clustering, topic modeling, Donald Trump, computational linguistics

1. Introduction

Disclaimer: This study focuses only on the sentiment analysis and linguistic features of Donald Trump's tweets and **does not express any political stance**.

Donald Trump is the 45th and 47th president of the United States and is also known for his distinctive characteristics in both communication and policymaking. Unlike most other presidents, who are primarily politicians, Trump is a businessman, which means his behavior is arguably less predictable than that of other presidents. In addition, Trump actively uses X (hereafter referred to as Twitter), through which he posts contents covering a wide range of topics, including politics. He is probably most widely known for his catchphrase "Make America Great Again," which reflects his patriotism and communication style. This slogan is often the first association people make with Trump.

One of the most distinctive characteristics is that Trump extensively uses Twitter. While the platform has relatively fewer monthly active users than that of Facebook—586 million compared to 3,070 million in February 2025 (Dixon 2025), it plays a vital role in public events and political communication. Trump uses Twitter as a political media. He frequently uses Twitter to communicate his political stance, announce presidential affairs, and comment on others' opinions. His tweets usually show strong sentiments and topics about both inside and outside the US.

Trump has a significant influence on both politics and economics, impacting not only the US but also other countries internationally. This highlights the importance of analyses of Trump's online discourse, which can provide valuable insights into different aspects, including the events Trump discusses online, the general sentiment revealed from the characteristics of his tweets, and the attitude toward specific topics (e.g., the US economy and the China–United States trade war).

Prior studies have focused both on qualitative and quantitative approaches. For instance, Clarke and Grieve (2019) focused on the changes in Trump's communication style over time. Building on multivariable analysis and Gimpel Twitter Tagger, which can identify elements essential to online communication, such as hashtags, mentions, and URLs, they identified four stylistic patterns in four dimensions, respectively: "conversational," "campaigning," "engaged," and "advisory." For each dimension, they investigated the parts of speech (nouns, verbs, adjectives, etc.), modality, and other text style and found out that, for each of the stages where he employed the respective style, Trump deliberately crafted his tweets to communicate influentially with the public.

By contrast, Hilman, Suganda, and Damayanti (2023) mainly used a qualitative method to analyze the insults in Trump's tweets. According to their introduction, Trump, unlike prior presidents, employs considerable insults in public, including both offline settings, such as campaigns, and online settings on Twitter. They therefore adopted Searle's classification of speech acts, which categorizes speech acts into representatives, directives, commissives, expressives, and declarations (Searle 1976), to classify insult contents and specifically identify the classification of insult words as "obscenity," "abusive swearing," and "swearing." The researchers concluded that Trump used insults to express emotions, attack opponents, and declare his personal identity.

Similarly, Ross and Rivers (2018) researched Trump's "fake news" content. They used comparative keyword analysis and compared Trump's tweets with that of other politicians to find out words that are usually used by Trump but not by other politicians. They then conducted a qualitative analysis and investigated the keyword occurrence within sentences to understand contextually. They concluded that Trump accused the "fake news" in order to mask the truth that he was spreading misinformation.

Some other studies also focused on Twitter as a platform for political communication. Alexandre, Jai-sung Yoo, and Murthy (2022), for instance, studied the tweets about Trump during the first days of Trump's presidency as the 45th president of the US and identified the opinion leaders using social network analysis. They then used the package MALLET (McCallum 2002) for topic modeling to identify topics in text. The researchers also evaluated the degree of polarization of tweets about Trump. Their research discovered the top opinion leaders and the main topics they discussed during the time frame.

However, previous studies did not use sentiment clustering to group text and conduct analyses in each cluster. We focused on Trump's tweets from 2017 to 2019 to analyze the sentiments of the tweets, the connections between his tweets and certain events, and his communication styles. During this time frame, there were several ongoing international crises, including the China–United State trade war, the Russia investigation, the North Korea nuclear issue, and the Mexico–United States border issue. Our research focuses on the following two questions:

(1) Are there recognizable sentiment clusters in Trump's tweets?
(2) If so, what are the topics in each cluster and does Trump employ different communication styles in different clusters?

To address these questions, we used a public dataset from a quantitative analysis of Trump's tweets on Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan (Coe et al. 2020). We first encoded the text into embeddings while capturing the semantic information. Then, we reduced the dimension of embeddings and used the result for sentiment-based clustering. Based on the clustering result, we analyzed the related topics and sentiment in specific clusters. The remainder of this paper is organized as follows: Section 2 describes the dataset used in our study, Section 3 introduces our methodology in detail, Section 4 shows the results of clustering, Section 5 examines the characteristics of each cluster, Section 6 discusses the significance of our study, and Section 7 concludes this paper.

2. Data Source

In contrast to earlier research on Trump's tweets, which primarily focuses on analyzing the communication style and discussed topics using words as units (Clarke and Grieve 2019; Hilman, Suganda, and Damayanti 2023; Ross and Rivers 2018), our study addresses this lack of sentiment-based analysis by combining two processes:

(1) We first encoded the entire sentence in each tweet to an embedding vector and performed clustering based on these numerical representations, in contrast to prior word-based analyses.
(2) We then examined each cluster using word-based approaches, including VADER sentiment scores, topic modeling, and word frequencies.

By doing so, our study not only provides understanding about the topics and communication style of Trump's tweets but also offers application value for using sentiment-based approaches to analyze online social discourse.

Our research relies on the public dataset in the quantitative analysis of Trump's tweets from 2017 to 2019 on ICPSR (Coe et al. 2020). This dataset includes text and basic information, such as likes and retweet counts drawn from Trump Twitter Archive[1]. It also provides 13 additional human codings: for example, "CRIT" stands for whether it "criticizes another person/idea," while "MEDI" stands for whether it is a "derogatory/condescending statements about news media." Table 1 lists six of the human codings and their corresponding meanings defined by the researchers. The content of the description is excerpted directly from the dataset codebook provided by Coe et al. (2020).

Table 1. Coding description (excerpted from dataset codebook).

Coding	Description
CRIT	Criticizes another person/idea (not his own)?
MEDI	Derogatory statements about news media?
INDV	References an individual?
INTN	References other countries/leaders?
PRTY	Reference to partisan/ideological labels?
IMMG	References immigration?

[1] https://www.thetrumparchive.com/

Initially, we considered DEC (deep embedded clustering) (Xie, Girshick, and Farhadi 2016), which is a deep neural network (DNN) approach that can run clustering and learn text features at the same time. This approach can be separated into two general steps. Firstly, it trains an initial autoencoder, and then it drops the decoder layer and uses the output of the encoder to perform k-means clustering. Next, it uses Student's t-distribution and KL divergence to improve the clustering result, which is subsequently used in backpropagation to optimize the DNN (Xie, Girshick, and Farhadi 2016). However, we chose not to use DEC due to our small dataset size—roughly 2,000 rows were not enough to use a DNN approach as it may result in unnecessary complexity and an overfitting problem. Still, we were inspired by the idea of encoding the text, reducing the dimension (motivated by the bottleneck layer in autoencoders), and performing clustering methods.

The dataset includes 2,082 Trump's tweets. After excluding retweets and missing values, 1,805 tweets remained for analysis. For detailed data preprocessing steps, see Appendix A.

3. Methodology

Figure 1 illustrates the components of our research framework. This paper is organized in the way the flowchart shows. After cleaning the tweets, we first encoded the tweets into computer-readable numerical vectors and reduced the dimensionality of the vectors. Three clustering methods were then performed in sequence to determine the optimal result for cluster analysis. Finally, we analyzed the sentiment and topics within specific clusters.

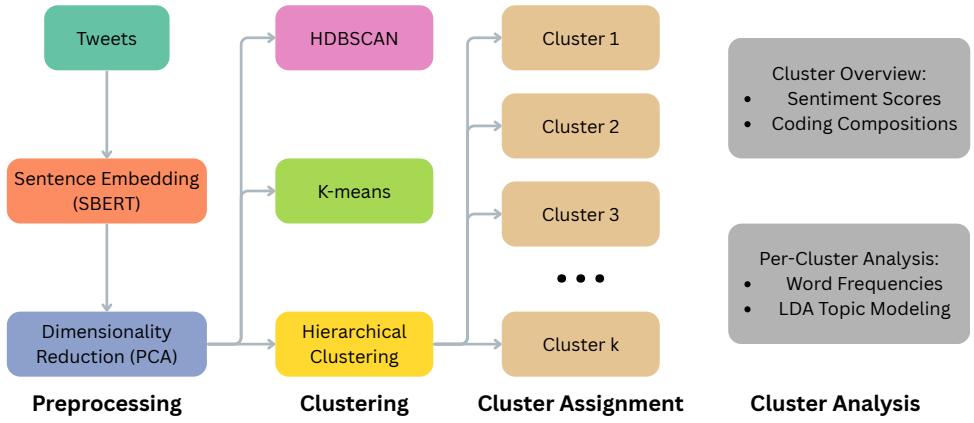

Figure 1. Flowchart of our research.

3.1. *Sentence Embedding*

One of the most crucial components in our study is embedding, which is a method used to represent high-dimensional, unstructured data (e.g., words, sentences, and images) as low-dimensional vectors or matrices while capturing the object's semantic or structural attributes. This makes measuring the similarity between different objects through their distance in vector space plausible (Bengio et al. 2003).

However, word2vec and GloVe are word embedding methods. Although calculating

the average of word vectors as the sentence embedding can carry some semantic information, the effectiveness of this approach is limited. Nowadays, one of the most widely used methods for generating sentence embedding that can capture semantic relationships is Sentence-BERT (SBERT) (Reimers and Gurevych 2019). The method is an extension of BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2019), a deep learning model that is based on transformer encoder architecture and produces vector representations for words with contextual information (i.e., the surrounding words in a sentence) that static word embedding lacks, and is designed to generate embeddings for sentences.

We used pre-trained SBERT model `all-MiniLM-L6-v2`[2] (Reimers and Gurevych 2019; Wang et al. 2020) to encode text into embeddings while keeping its semantic information. We chose SBERT because our analysis on tweets relied on contextual information within text. In addition, `all-MiniLM-L6-v2` is trained on more than 1 billion training pairs and has proved its competence in many datasets. Moreover, the model is efficient and lightweight in terms of dimensionality, generating embedding vectors with 384 dimensions—a relatively small number in the context of embeddings.

Following the encoding of the tweets, we applied principal component analysis (PCA) to reduce the dimensionality of the embedding vectors. By experimenting with various numbers of principal components and utilizing three different clustering methods, HDBSCAN, K-means, and hierarchical clustering, we identified that 30 principal components was the optimal choice.

3.2. *Topic Modeling*

Topic modeling is a crucial part in the cluster analysis as it offers insight into the topics discussed in each cluster that we identified using hierarchical clustering, allowing us to examine the characteristics within each cluster more comprehensively. Topic modeling is an unsupervised machine learning technique that identifies main topics that spread across the documents (i.e., tweets in our study) within a large corpus (i.e., the collective tweets in a cluster). Inspired by Alexandre, Jai-sung Yoo, and Murthy (2022), we also performed topic modeling using MALLET. MALLET implements topic modeling using Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), the most commonly used model for identifying topics. The method assumes that there are a range of topics in the corpus, that each document includes an identical set of topics but differs in proportion, and that each topic contains several keywords. The model randomly generates distributions for topics per documents and distributions for corresponding keywords in each topic. Subsequently, LDA infers the distributions reversely from the documents in the corpus using probabilistic models (Blei 2012).

3.3. *Sentiment Score*

Apart from the previously mentioned methods, we used VADER (Valence Aware Dictionary and sEntiment Reasoner) (Hutto and Gilbert 2014), a rule-based sentiment analysis tool that can identify sentiment in sentences, to evaluate the sentiment distribution in each cluster.

For each text, the VADER compound sentiment score is computed in the following way described by Hutto and Gilbert (2014) and implemented in the vaderSentiment

[2] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

library[3] we use. Since VADER is a lexicon-based method, it sums the sentiment intensity adjusted by linguistic rule (e.g., punctuation and capitalization) of each word. Denote the sum as $S = \sum_i s_i$, where s_i is the intensity for each word, ranging from -4 to 4. Then, the compound sentiment score is derived by normalizing S:

$$\frac{S}{\sqrt{S^2 + 15}}$$

This tool assigns each sentence with a sentiment score, ranging from -1 to 1. Sentences with the score less than -0.05 are classified as negative; those with the score between -0.05 and 0.05 are classified as neutral; and those with the score greater than 0.05 are classified as positive. Table 2 shows three examples with the sentiment classified as positive, negative, and neutral, respectively. We chose VADER because it is "specifically attuned to sentiments expressed in social media" (Hutto and Gilbert 2014), which aligns with our objective to analyze tweets.

Table 2. Examples for VADER sentiment score.

Sentence	Score	Sentiment
Today is much more excellent.	0.6115	Positive
Yesterday was so terrible.	-0.6113	Negative
Today is much more excellent, and yesterday was so terrible.	0.0005	Neutral

4. Results

We used PCA to reduce the dimensions of the embedding vectors to 20, 30, 50, and 80, and then applied the three clustering methods. We evaluated the effectiveness of clustering based on the numbers of tweets that were related to China, Russia, and North Korea, respectively, within each cluster. This approach was chosen because these countries correspond with certain international issues, which were frequently communicated through Twitter. We identify these tweets by checking if a tweet contains certain keywords. For example, tweets that contain either "China" or "Xi" (the president of China) are classified as related to China. Table 3 shows all keywords we used to identify the corresponding tweets. We compiled the keyword lists by adding common keywords (e.g., country names and presidents) and manually checking the exceptions in the dataset.

Table 3. Keywords for countries.

Keywords	Country
China; Beijing; Xi	China
North Korea; Kim Jong; Kim Prepared; Chairman Kim	North Korea
Russia; Putin	Russia

[3]https://github.com/cjhutto/vaderSentiment/blob/master/vaderSentiment/vaderSentiment.py

For the first attempt, we ran HDBSCAN for clustering on the result of PCA because it can automatically identify the number of clusters, which is a key function for the initial stage to explore the data. The number of clusters given by this function may also be used as the hyperparameter K in K-means. Unfortunately, HDBSCAN did not yield valuable results for our dataset.

Table 4 shows the cluster assignments of HDBSCAN for each combination of `n_components` (the number of principal components in PCA) and `min_cluster_size` (the minimal number to form a cluster in HDBSCAN). We tried 10 and 20 for `min_cluster_size` as a lower bound and an upper bound, since any clusters with sizes less than 10 are problematic to form stable semantic relationships, while capturing only clusters larger than 20 may overlook certain semantic groups. Respectively, we set `n_components` to 20, 30, 50, and 80 because the initial dimension of sentence embeddings is 384, and we wanted to find out the minimal number of principal components, or dimensions, that could group text meaningfully.

Table 4. HDBSCAN cluster assignments sorted by cluster size (top 5 largest clusters).

`n_components`	20		30		50		80	
`min_cluster_size`	10	20	10	20	10	20	10	20
Outliers	1,308	1,362	1,283	1,321	1,428	1,221	1,425	1,462
Cluster 1	234	191	213	223	133	485	128	132
Cluster 2	178	165	193	178	125	50	59	113
Cluster 3	20	87	42	52	53	49	57	43
Cluster 4	16	—	19	31	18	—	50	34
Cluster 5	14	—	16	—	14	—	18	21

Note. "—" indicates that no cluster of this rank for the combination of parameters.

For each combination of `n_components` and `min_cluster_size`, the method identified over 1,200 tweets as outliers, where an outlier means a tweet whose embedding lies in a low-density regions of the space. This indicates that the majority of the tweets lie together in a sparse region of the space without distinctly dense areas. In addition to outliers, the remaining cluster assignments were also unstable across each trial, and no meaningful patterns were recognized. This may be attributed to that the embedding vectors were generated from the tweets, which were usually relatively short and led to increased spread in embedding space. This relatively even distribution of embedding vectors in high-dimensional space made density-based HDBSCAN difficult to distinguish the clusters.

The failure of the HDBSCAN model led us to use K-means method for cluster identification. The hyperparameter K was specified to 10 because a lesser number would make the topics in each cluster too broad, while too many clusters would be difficult to analyze. Experimenting with the same set of numbers of principal components, we found out 20 to be the optimal number for the principal components. We assessed the cluster assignments using the same method—counting the number of tweets that relate to the three countries. The numbers of the tweets that relate to either North Korea and Russia in each cluster remained roughly the same across different numbers of principal components. However, the number of the tweets that relate to China varied.

Table 5. Numbers of tweets related to China using K-means (top 3 related clusters).

Cluster	$n_{PC} = 20$	$n_{PC} = 30$	$n_{PC} = 50$	$n_{PC} = 80$
Cluster 1	31	18	18	18
Cluster 2	5	17	17	16
Cluster 3	3	3	3	5

Table 5 shows these numbers in each cluster for different numbers of principal components, where n_{PC} denotes the number of principal components. When n_{PC} was 30, 50, and 80, respectively, the majority of the China-related tweets were grouped into two different clusters, but these tweets converged to one cluster when n_{PC} was 20. Therefore, we identified 30 as an acceptable and relatively small number of principal components that could retain the general semantic information in the tweets to continue our subsequent clustering and analysis. The rest of the paper uses 30 principal component dimensions as the basis for all subsequent analyses.

Having determined the number of principal components as 30, we applied hierarchical clustering on the same sentence embeddings with reduced dimensions. While K-means generated effective clusters, hierarchical clustering has its own advantages. First, unlike K-means grouping tweets into a fixed number of clusters, the assignment of hierarchical clustering is not determined, which allows us to explore the appropriate number of clusters readily. Second, K-means is a space-partitioning method, which evaluates the similarity between the tweets based on the distance in the embedding space, whereas hierarchical clustering offers a different perspective from the partial order of the relationships between tweets. For example, consider that A, B, and C are three of numerous tweets in a two-dimensional embedding space. Suppose that the three points are on a same line. K-means may group A, B, and C into three clusters, while hierarchical clustering may identify the similarity between A and B and the similarity between B and C, so it can recognize the relationship between A and C using B as a bridge.

Following the dendrogram of hierarchical clustering and using the numbers of country-related tweets as the metric, we experimented with different thresholds to cut the dendrogram into clusters. As a result, eight appeared to be the most suitable number of clusters because it yielded a result similar to that of K-means while the clusters are of moderate sizes.

Table 6 lists the numbers of the tweets that are related to China, North Korea, and Russia, respectively, in the three major clusters. The column "Aggregate" counts the tweets related to either one of the three countries. In Cluster 2 and Cluster 8, the tweets are predominantly related to Russia and China, respectively, suggesting country-specific events. By comparison, the tweets in Cluster 7 involve the three countries. Subtracting the summation of the three numbers by the aggregate number, we observed that several tweets need to be related to the two or the three countries simultaneously. This indicates international events that involve multiple countries.

Figure 2 shows the dendrogram generated by hierarchical clustering. The branches are colored based on the 8-cluster assignments we determined, and the number under each subtree indicates the size of the corresponding cluster. Notably, Cluster 3 contains only 27 tweets, which include solely one URL in the text. This specific cluster, not identified using K-means, demonstrates the capability of hierarchical clustering to

Table 6. Numbers of tweets related to the three countries (hierarchical clustering).

Cluster	China	North Korea	Russia	Aggregate
Cluster 2	2	3	67	70
Cluster 7	10	28	19	47
Cluster 8	21	1	0	21

separate clusters with unique traits.

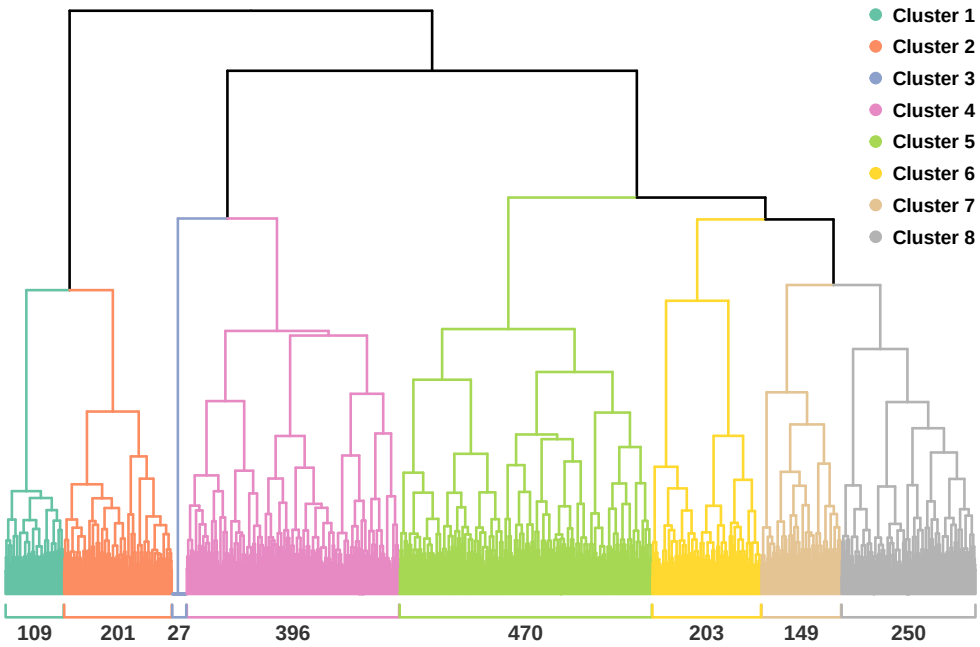

Figure 2. Hierarchical clustering dendrogram with cluster sizes (number of tweets) displayed at the bottom.

5. Sentiment Analysis of Clusters

In this section, we perform a detailed sentiment analysis of the clusters we identified in Section 4. This section is organized as follows: Subsection 5.1 summarizes the eight clusters using codings from the quantitative analysis by Coe et al. (2020) and VADER sentiment score; Subsection 5.2.2, **??**, 5.2, and **??** analyze Cluster 2, Cluster 7, Cluster 8, and Cluster 6, respectively, in detail using LDA topic modeling and word frequencies. We chose Clusters 2, 7, and 8 because they were identified using the keywords related to the three countries in Table 3 and therefore are relevant to international relationships.

5.1. *Overview*

Since Cluster 3 includes only the tweets that have a URL in the text exclusively, we excluded the cluster in the following analysis. Based on the quantitative analy-

sis (Coe et al. 2020), we calculated the percentages of each tweet category, such as CRIT (whether it criticizes another person or idea) and MEDI (whether it contains derogatory statements about news media), for each cluster. See Table 1 for the full description of the categories.

Table 7. Percentages of codings in clusters (see Table 1 for coding description).

	CRIT	MEDI	INDV	INTN	PRTY (Republican)	PRTY (Democrat)	IMMG
Cluster 1	89.9	92.7	27.5	11.0	2.8	4.6	0.0
Cluster 2	88.1	12.4	74.1	34.3	2.0	18.9	0.5
Cluster 4	3.5	0.3	37.9	25.3	0.3	0.0	1.3
Cluster 5	50.4	7.0	62.3	8.9	4.0	13.8	10.0
Cluster 6	70.4	2.5	21.7	8.9	9.4	32.0	62.6
Cluster 7	45.0	5.4	24.2	59.1	0.0	2.0	2.7
Cluster 8	26.0	2.8	8.8	22.4	2.4	0.8	4.0

Table 7 lists the percentages of the codings per cluster. We can observe several unique characteristics of different clusters. In Cluster 1, the percentage of "MEDI" reaches the highest at 92.7%. Indeed, according to the result of topic modeling, the most prominent topic in Cluster 1 significantly outweighs the others. The topic includes the words "fake" and "news," suggesting Trump's common phrase "fake news." In Cluster 4, the percentage of "CRIT" is the lowest among all clusters, and almost no condescending attitude toward medias is expressed as the percentage of "MEDI" shows. Using the result of topic modeling, we identify the topic of this cluster to be about presidential events, including White House meetings, public appearances, and rallies in the United States. Cluster 6 is characterized by high mention of immigration indicated by the percentage of "IMMG." We will analyze this cluster in **??** in detail as it, similar to the other three clusters we will analyze, also features political issues.

We then computed the sentiment score of each tweet using VADER and classified the sentiment as negative, neutral, or positive according to the official criterion. Figure 3 visualizes the distribution of the sentiment scores in each cluster and shows the percentage of each sentiment in the clusters. The black dots and their respective text above them in the left panel of the figure demonstrates the average value of the sentiment scores in the corresponding cluster.

We observe the largest proportion of positive sentiments in Cluster 4. According to the topics of presidential affairs identified in the cluster, the most frequently used words in the most important topics are "great" and "honor," suggesting Trump's positive attitude toward presidential affairs. Clusters 1 and 2 are the two clusters with the most negative sentiments, which corroborates the observation in Table 7 that the two clusters are characterized by the highest percentages of "CRIT." Below are sample tweets from Clusters 4 and 1:

5.2. *Understanding Policy Through Tweets*

5.2.1. Cluster 8

Cluster 8 is selected for its high concentration of China-related tweets. We generated the word clouds filtered using the keywords for China listed in Table 3. Figure 4 shows

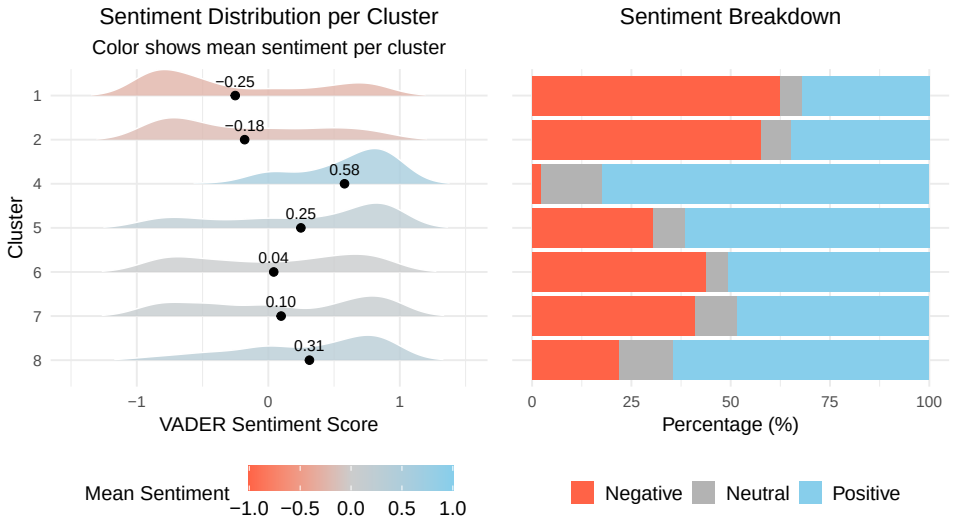

Figure 3. VADER sentiment score distribution and sentiment composition per cluster.

Table 8. Example tweets from Clusters 1 and 4.

Cluster 1 (Negative Sentiment)	Cluster 4 (Positive Sentiment)
While on FAKE NEWS @CNN Bernie Sanders was cut off for using the term fake news to describe the network. They said technical difficulties!	Interview with David Muir of @ABC News in 10 minutes. Enjoy!
The failing @nytimes has disgraced the media world. Gotten me wrong for two solid years. Change libel laws?	A great and important day at the United Nations. Met with leaders of many nations who agree with much (or all) of what I stated in my speech!

the filtered and unfiltered clouds, respectively.

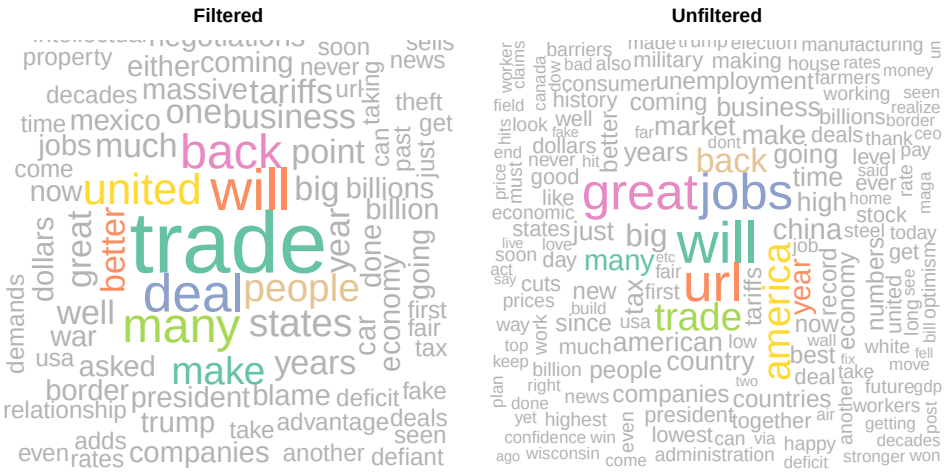

Figure 4. Cluster 8: word clouds with and without regex filtering.

The filtered and unfiltered clouds for Cluster 8 show distinct major topics. The size of the word "trade" in the filtered cloud is observably larger than the other frequent words, suggesting a primary focus on the China–United States trade war. Although the word "trade" is also among the most frequent words in the unfiltered cloud, words such as "great" and "job" are more dominant. This observation suggests a broader theme in the cluster.

To gain a more comprehensive understanding of the topics discussed in each cluster, we applied LDA topic modeling using the MALLET package. Since the clusters we analyze typically contain 150–250 tweets, we selected five as the number of topics, a hyperparameter for LDA topic modeling. This number provides sufficient information for analysis of clusters of this size while being easy to interpret. For each topic, the model generated two output files: one containing the Dirichlet parameter and keywords for each topic, and another listing the composition of topics for each tweet. The Dirichlet parameter of each topic can be interpreted as the importance of the topic in the corresponding cluster. We normalized the Dirichlet parameters so that, within each cluster, they sum to one, enabling interpretation of the share of each topic in the cluster. When visualizing the topics, we display only the top eight keywords to maintain conciseness.

Figure 5 shows the result of topic modeling. The top three topics account for the majority of the content in Cluster 8. Since Topics 1 and 2 share multiple keywords, we expect them to have distinct focus within the same broader topic. Below are typical tweets for Topics 1 and 2.

(1) July is just the ninth month since 1970 that unemployment has fallen below 4%. Our economy has added 3.7 million jobs since I won the Election. 4.1 GDP. More than 4 million people have received a pay raise due to tax reform. $400 Billion brought back from overseas. @FoxNews (Topic 1)

(2) Stock Market had another good day but now that the Tax Cut Bill has passed we have tremendous upward potential. Dow just short of 25000 a number that few thought would be possible this soon into my administration. Also unemployment went down to 4.1%. Only getting better! (Topic 1)

(3) AMERICA will once again be a NATION that thinks big dreams bigger and

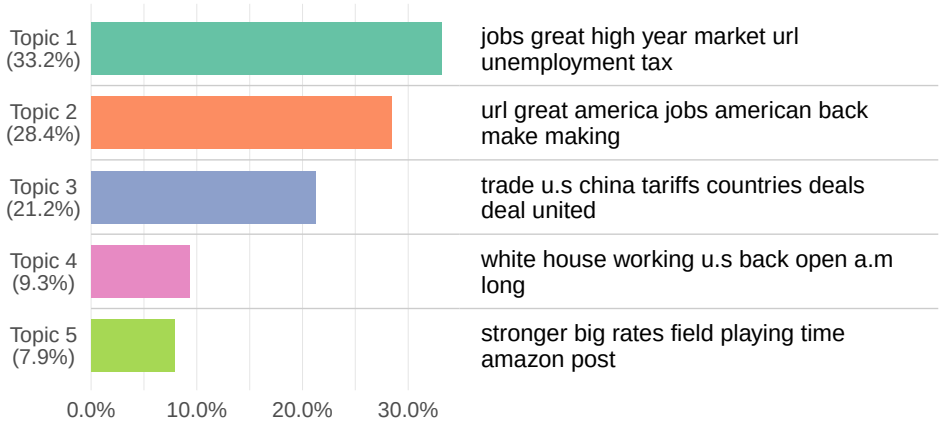

Figure 5. Cluster 8: topic importance (left panel) and corresponding topic keywords (right panel).

always reaches for the stars. YOU are the ones who will shape Americas destiny. YOU are the ones who will restore our prosperity. And YOU are the ones who are MAKING AMERICA GREAT AGAIN! #MAGA (Topic 2)

(4) We are bringing back our factories we are bringing back our jobs and we are bringing back those four beautiful words: MADE IN THE USA! (Topic 2)

The distinction between the two topics is similar to the difference in focus of Topics 1 and 4 in Cluster 7. Topic 1 focuses on the concrete economic outcome, mentioning stock market records, low unemployment rates, and other indices such as the GDP, whereas Topic 2 emphasizes promises ("we are bringing back our jobs") and Trump's famous slogan "MAKE AMERICA GREAT AGAIN." The keywords of Topic 2 include "great," "america," and "make" at the same time, indicating frequent references to the slogan. These two topics together account for approximately 60% of the content, signaling a primary focus in this cluster.

The top keywords of Topic 3—"trade," "u.s," "china," and "tariffs"—strongly suggest discussion centered on the China–US trade war, aligning with the high concentration of China-related tweets illustrated in Table 6. Topics 4 and 5, in contrast, do not have an identifiable focus based on the keywords and constitute a smaller portion of the discussion. Examination of the tweets of these two topics reveals a broad theme on economics. However, compared to the other three identified topics, Topics 4 and 5 appear to be exceptions.

In addition, we calculated the correlations between the topics based on the output file that provides the percentages of the topics in each tweet using Pearson correlation coefficient (PCC), a measure of linear correlation between two variables that ranges from -1 to 1. Before calculation, we first applied Centered Log-Ratio (CLR) transformation because computing the correlation coefficient directly from the topic composition without preprocessing is misleading. Since all topic proportions of a tweet must sum to 1, an increase in the proportion of one topic typically means decreases in the others. CLR addresses this issue by transforming the compositional data from the constrained space (e.g., values summing to 1) into real space. The CLR transformation is described by the following formula:

$$\mathrm{CLR}(x_i) = \ln \frac{x_i}{g(\mathbf{x})},$$

where $g(\mathbf{x})$ is the geometric mean of all components:

$$g(\mathbf{x}) = \left(\prod_{i=1}^{n} x_i \right)^{1/n}.$$

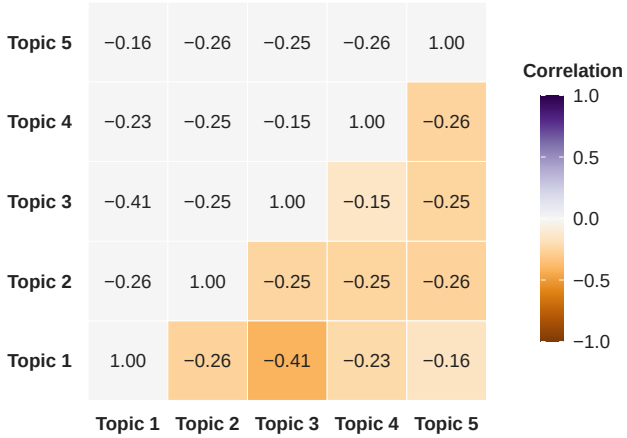

Figure 6. Cluster 8: topic correlations.

Figure 6 shows the correlations between topics in Cluster 8. The exclusivity between Topics 1 and 3 (-0.41) is the greatest among the four clusters we evaluated. Given that Topic 1 focuses on US economic well-being and Topic 3 on the trade war, it is reasonable that Topics 1 and 3 are less likely to coincide.

For the other clusters, we also calculated the topic correlations. However, only the figure for Cluster 8 is shown for the greatest observed exclusivity. The correlation figures for other clusters are attached in Appendix C.

Overall, the topic model effectively identified the theme of economics in this cluster. Although Topics 4 and 5 appear to be less fitting, this does not necessarily disprove the competence of either the cluster assignment or the LDA topic modeling. Considering that the tweets are divided into only eight clusters, the identified topics may be an optimal outcome.

5.2.2. Cluster 2

According to the clustering result shown in Table 6, Cluster 2 is characterized by the high concentration of Russia-related tweets—a third of the tweets in the cluster contain Russia-related keywords. We generated two word clouds: one filtered using the same keywords in Table 3 that we used to evaluate the cluster assignments, and one unfiltered. To make the word clouds more effective in conveying information, we removed the filtering keywords from the word cloud and positioned the most frequent words at the center of the word clouds. The most frequent words are then highlighted for readability.

Figure 7 shows the filtered and unfiltered word clouds of Cluster 2. Following the 2016 U.S. presidential election, U.S. intelligence agencies and Congress investigated whether Russia had interfered in the election. Robert Mueller was appointed to examine the Russian interference, whether Trump's campaign had colluded with Russia, and whether Trump obstructed justice. In the political conversation at that time,

"collusion" refers to the allegation that Trump's campaign may have coordinated with Russia, which was not found in the Mueller report (Mazzetti and Benner 2019). The phrase "witch hunt" was frequently used by Trump to refer to the investigation to suggest his innocence and the unfairness of the investigation (Young 2022). The most frequent words in both word clouds are largely similar, including "collusion," "witch," and "hunt." High usage of these words collectively suggests a prominent focus on this international political issue. Other words and phrases in the word clouds such as "hoax" and "fake news" illustrate Trump's rebuttal to the investigation and his repeated declarations of innocence.

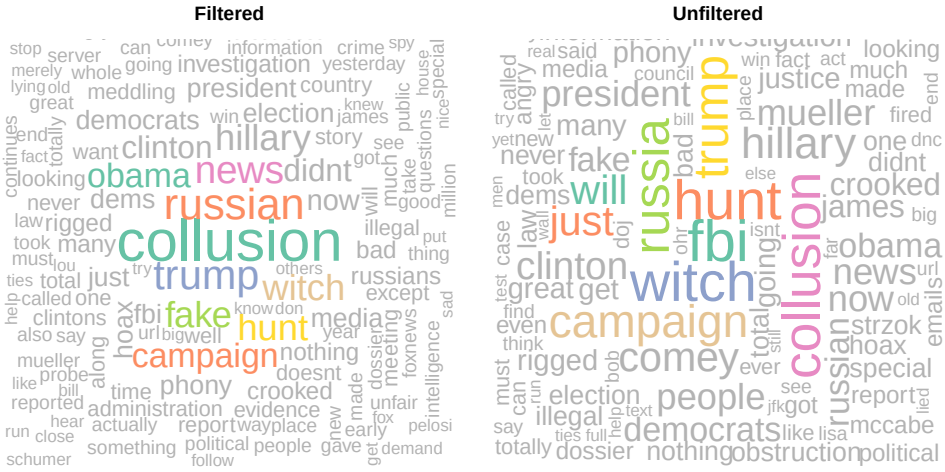

Figure 7. Cluster 2: word clouds with and without regex filtering.

Figure 8 shows the importance of each topic and the corresponding keywords. The left panel of the figure visualizes the importance, or share, of each topic with the values in parentheses indicating the exact share.

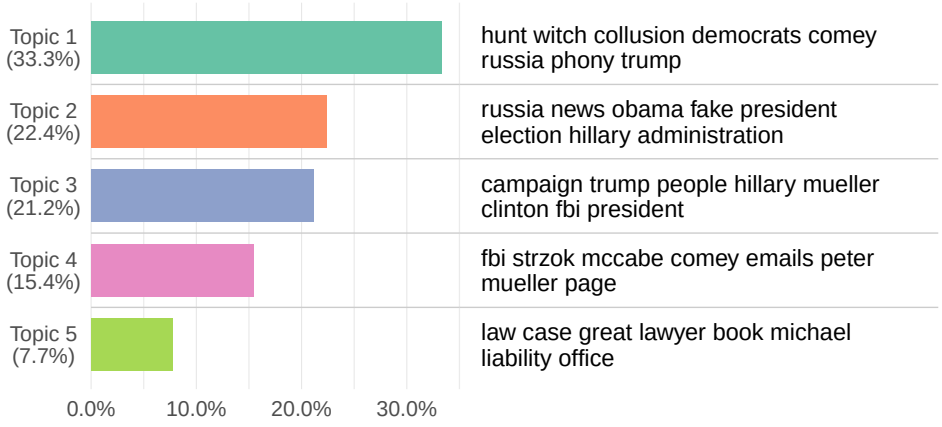

Figure 8. Cluster 2: topic importance (left panel) and corresponding topic keywords (right panel).

The top three identified topics reflect the same theme on the Russia investigation but with different focuses. Topic 1, including "collusion" and "witch hunt," signal a focus on Trump's allegation that the Russia investigation was a politically motivated attack. The keywords of Topic 2 alone do not clearly reveal its focus. Based on the output file that provides the topic composition for each tweet, we selected several

typical tweets with a high percentage of Topic 2, as listed below.

(1) Just out: The Obama Administration knew far in advance of November 8th about election meddling by Russia. Did nothing about it. WHY?
(2) . . . have it. Fake News said 17 intel agencies when actually 4 (had to apologize). Why did Obama do NOTHING when he had info before election?
(3) Funny how the Fake News Media doesn't want to say that the Russian group was formed in 2014 long before my run for President. Maybe they knew I was going to run even though I didn't know!
(4) When you hear the Fake News talking negatively about my meeting with President Putin and all that I gave up remember I gave up NOTHING we merely talked about future benefits for both countries. Also we got along very well which is a good thing except for the Corrupt Media!

These tweets and the topic keywords collectively reveal the focus of Topic 2 on alleged inaction by Obama (illustrated in the first and the second examples), the positive framing of the meeting with Putin (shown in the fourth example), and criticism of medias (expressed in the third and the fourth examples). The keywords of Topic 3, including "campaign" and "mueller," suggest a focus on the Mueller report, the official investigation report. In contrast, Topic 4 focuses on specific subjects: the FBI and the major figures involved (Strzok, McCabe, and Comey). On the other hand, Topic 5 does not relate to the other four topics and appears to be an outlier in Cluster 2. This misclassification may be attributed to limitations of our model, particularly the lack of a robust embedding model, which is crucial for extracting semantic information.

Figure 9 illustrates the correlations between topics. All correlations, except the diagonal correlations, are negative, suggesting a general exclusivity of among the topics. This is indeed expected, given that tweets are generally short. Together with the effectiveness of LDA topic modeling, one tweet is typically dominated by one topic. Nonetheless, we can quantify the relative strengths of exclusivity among the topics. For example, Topics 1 and 3 are more mutually exclusive (-0.32) than Topics 2 and 5 (-0.14).

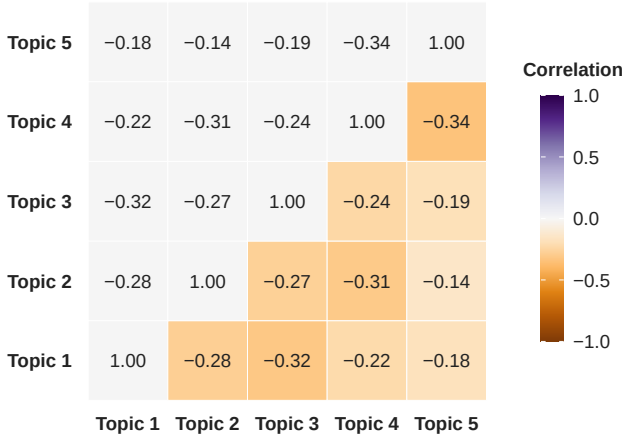

Figure 9. Cluster 2: topic correlations.

6. Discussion

6.1. *Key Findings*

In our study, we experimented with three different clustering methods—HDBSCAN, K-means, and hierarchical clustering—and then determined an eight-cluster assignment based on the dendrogram generated by hierarchical clustering. We then performed an overview that summarizes selected codings provided in the dataset (Coe et al. 2020) and evaluates sentiment distributions for each cluster. Finally, word frequency analysis and topic modeling were applied to discover the topics within chosen clusters (Clusters 8 and 2).

Evaluating the identified clusters, we discovered distinguishable sentiments and topics across them. This observation affirms our first research question on the existence of recognizable sentiment clusters. In the overview, several clusters exhibited distinct characteristics. For example, the proportions of both "CRIT" and "MEDI" in Cluster 1 are the highest among all clusters, reaching approximately 90%. The mean sentiment score in the cluster is also the lowest (-0.25) compared to the others. Examination of the topics in the cluster reveals a dominant topic with frequent "fake news" references. On the other hand, Cluster 4 registers a significantly higher mean sentiment score than those of other clusters (0.58). The fact that over 75% of the tweets express positive sentiment further suggests predominantly favorable discussions within this cluster.

Following the overview, detailed evaluations of selected clusters directly address our second research question by revealing Trump's distinct communication styles regarding different topics. Cluster 8 centers on the broad theme of economics, with a focus on both US economic well-being and the China–US trade war. According to the coding overview shown in Table 7, this cluster exhibits the highest proportion of "MAGA" coding, which is then corroborated by the identified topics. The top keywords of Topic 2—including "make," "america," and "great"—suggest frequent references to his famous campaigning slogan. In addition, the prominent use of "will," illustrated by Figure 4, indicates Trump's promises and commitment over the US economy.

Additionally, the topics identified Cluster 2, concerned with the Russia investigation, indicate Trump's defensive and accusatory tone. For instance, he frequently used the phrase "witch hunt" to frame the investigation as a politically motivated attack in order to establish his innocence, as shown in Topic 1 in Figure 8. In addition, Topic 2 reveals Trump's accusations against his political opponents for their alleged inaction.

These detailed cluster analyses together uncover how Trump expresses different attitudes and employs distinct communication styles for each topic, thereby addressing our second research question.

6.2. *Implications*

The first part of this study clusters tweets by sequentially applying SBERT for vector representations, PCA for dimensionality reduction, and hierarchical clustering. Thus, our approach generates cluster assignments based on the semantic information in the tweets by leveraging embedding vectors, in contrast with most prior studies. As Hanny and Resch (2024) mention, natural language processing techniques have been advanced significantly in recent years, yet few studies have relied on these methods, such as SBERT. In this way, our study not only addresses this gap but also provides application value for analysis of online social discourse.

In addition, by performing an overview for each cluster and conducting thorough

analyses on certain clusters, our study reveals the topics in Trump's discussion during 2017–2019 as well as his attitudes and communication styles regarding different issues. Therefore, the methods and findings in this study are also applicable for societal and political studies.

6.3. *Limitations and Future Directions*

6.3.1. Dataset Limitations

Our research relies on the dataset from a quantitative analysis, which contains solely 1,805 tweets during 2017–2019 after data cleansing. According to Trump Twitter Archive, Trump sent over 6,000 tweets during that time frame, whereas the dataset includes only one-third of them. Nonetheless, given that the dataset is randomly sampled from the 6,000 tweets and that our research does not involve model training, this still dataset provides a sufficient number of samples to produce reliable results. Admittedly, further research may rely on the full corpus to test the robustness of our methods.

6.3.2. Embedding Model Limitations

In this study, we used the model `all-MiniLM-L6-v2` to encode tweets into embedding vectors. Although it is one of the best-performed open models, an improved or specifically fine-tuned model can capture more nuances of the semantic information. Since this was the only embedding model we applied, further research may focus on evaluating the performance of different embedding models.

6.3.3. Potential of HDBSCAN

Among the three clustering methods we applied, HDBSCAN was the only method that did not yield usable cluster assignments. Nevertheless, some studies have successfully employed HDBSCAN on embedding vectors (Hanny and Resch 2024; Yang et al. 2023; Zhang et al. 2024). For instance, Hanny and Resch (2024) reported that HDBSCAN outperformed other clustering models in certain cases. This suggests room for further research on how to utilize HDBSCAN efficiently.

6.3.4. Potential of UMAP

Our study leverages PCA for dimensionality reduction. However, PCA is a linear transformation where each principal component is a linear combination of the original variables. Hanny and Resch (2024) compared the performance of PCA and UMAP, a non-linear dimensionality reduction technique, and found that UMAP was better supported than PCA. Their findings suggest the potential of UMAP in future sentiment-based cluster analyses.

7. Conclusion

Trump is partly known for his distinctive communication styles on Twitter. Our study aims to identify sentiment clusters in his tweets, explore the topics within each cluster, and examine whether he has employed distinguishable communication styles for

different issues. Through performing clustering methods—including HDBSCAN, K-means, and hierarchical clustering—and the subsequent cluster analyses, we achieved an eight-cluster assignment with well-separated sentiments. For instance, Clusters 1 and 2, which focus on the "fake news" accusation and the Russia investigation, respectively, feature predominantly negative tweets, resulting in the two lowest sentiment scores. In contrast, Clusters 4 and 8, centered on presidential affairs and primarily on US economic well-being, achieve the two highest sentiment scores, revealing Trump's positive attitude toward these topics. Other clusters, although not analyzed in detail in our study, also reveal Trump's distinct communication styles.

As the area of artificial intelligence develops rapidly, technological breakthroughs can be revolutionary. The advent of SBERT and other embedding models has advanced the field of natural language processing and computational linguistics significantly. Our study takes a step in performing clustering of tweets based on embeddings generated by SBERT, a method that has been underutilized. Therefore, our research not only provides insight into Trump's online discussion and communication styles but also, at the same time, offers practical value for analyses of online social discourse.

References

Alexandre, Ilo, Joseph Jai-sung Yoo, and Dhiraj Murthy. 2022. "Make Tweets Great Again: Who Are Opinion Leaders, and What Did They Tweet About Donald Trump?" *Social Science Computer Review* 40 (6): 1456–1477. Publisher: SAGE Publications Inc, Accessed 2025-08-22. https://doi.org/10.1177/08944393211008859.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. "A neural probabilistic language model." *J. Mach. Learn. Res.* 3: 1137–1155.

Blei, David M. 2012. "Probabilistic topic models." *Commun. ACM* 55 (4): 77–84. Accessed 2025-08-26. https://dl.acm.org/doi/10.1145/2133806.2133826.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent dirichlet allocation." *J. Mach. Learn. Res.* 3: 993–1022.

Clarke, Isobelle, and Jack Grieve. 2019. "Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018." *PLOS ONE* 14 (9): e0222062. Publisher: Public Library of Science, Accessed 2025-08-20. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0222062.

Coe, Kevin, Julia Berger, Allison Blumling, Katelyn Brooks, Elizabeth Giorgi, Jennifer Jackson, Melanie Lee, et al. 2020. "Quantitative Content Analysis of Donald Trump's Twitter, 2017-2019." Apr. Accessed 2025-08-20. https://www.openicpsr.org/openicpsr/project/118603/version/V1/view;jsessionid=53D7CC774A48311666E73E5CC8DBB15F.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." May. ArXiv:1810.04805 [cs], Accessed 2025-08-27. http://arxiv.org/abs/1810.04805.

Dixon, Stacy Jo. 2025. "Biggest social media platforms by users 2025." Mar. Accessed 2025-08-21. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

Hanny, David, and Bernd Resch. 2024. "Clustering-Based Joint Topic-Sentiment Modeling of Social Media Data: A Neural Networks Approach." *Information* 15 (4): 200. Publisher: Multidisciplinary Digital Publishing Institute, Accessed 2025-09-11. https://www.mdpi.com/2078-2489/15/4/200.

Hilman, Evert Haryanto, Dadang Suganda, and Nani Damayanti. 2023. "Insults on Donald Trump's Twitter: A Study of Pragmatics." *Theory and Practice in Language Studies* 13 (4): 873–880. Publisher: Academy Publication Co., LTD, Accessed 2025-08-20. https://go.gale.com/ps/i.do?p=LitRC&sw=w&issn=17992591&v=2.1&it=r&id=

GALE%7CA746558850&sid=googleScholar&linkaccess=abs.

Hutto, C., and Eric Gilbert. 2014. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." *Proceedings of the International AAAI Conference on Web and Social Media* 8 (1): 216–225. Accessed 2025-08-27. https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

Mazzetti, Mark, and Katie Benner. 2019. "Mueller Finds No Trump-Russia Conspiracy, but Stops Short of Exonerating President on Obstruction." *The New York Times* Accessed 2025-09-06. https://www.nytimes.com/2019/03/24/us/politics/mueller-report-summary.html.

McCallum, Andrew Kachites. 2002. "MALLET:A Machine Learning for Language Toolkit." *http://mallet.cs.umass.edu* Accessed 2025-08-24. https://cir.nii.ac.jp/crid/1570572699312665856.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." Aug. ArXiv:1908.10084 [cs], Accessed 2025-08-20. http://arxiv.org/abs/1908.10084.

Ross, Andrew S., and Damian J. Rivers. 2018. "Discursive Deflection: Accusation of "Fake News" and the Spread of Mis- and Disinformation in the Tweets of President Trump." *Social Media + Society* 4 (2): 2056305118776010. Publisher: SAGE Publications Ltd, Accessed 2025-08-23. https://doi.org/10.1177/2056305118776010.

Schofield, Alexandra, and David Mimno. 2016. "Comparing Apples to Apple: The Effects of Stemmers on Topic Models." *Transactions of the Association for Computational Linguistics* 4: 287–300. Accessed 2025-08-26. https://doi.org/10.1162/tacl_a_00099.

Searle, John R. 1976. "A Classification of Illocutionary Acts." *Language in Society* 5 (1): 1–23. Publisher: Cambridge University Press, Accessed 2025-08-23. https://www.jstor.org/stable/4166848.

Wang, Wenhui, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers." Apr. ArXiv:2002.10957 [cs], Accessed 2025-08-27. http://arxiv.org/abs/2002.10957.

Xie, Junyuan, Ross Girshick, and Ali Farhadi. 2016. "Unsupervised Deep Embedding for Clustering Analysis." May. ArXiv:1511.06335 [cs], Accessed 2025-08-25. http://arxiv.org/abs/1511.06335.

Yang, Yuxuan, Yingmei Wei, Min Gao, Zanxi Ran, and Qi Wang. 2023. "Sentiment Analysis of Social Network Text Based on HDBSCAN and SO-PMI." *Journal of Physics: Conference Series* 2504 (1): 012055. Publisher: IOP Publishing, Accessed 2025-09-12. https://dx.doi.org/10.1088/1742-6596/2504/1/012055.

Young, Cathy. 2022. "No, 'Russiagate' Wasn't the Hoax That Team Trump Claims It Was." *Cato Institute* Accessed 2025-09-06. https://www.cato.org/commentary/no-russiagate-wasnt-hoax-team-trump-claims-it-was.

Zhang, Lutao, Xuesong Su, Yifei Wang, Mei Wang, Xinxin Yang, and Zixuan Xu. 2024. "HDBSCAN-based semantic clustering model in classifying incidents on security and environmental conservation management." In *Ninth International Symposium on Advances in Electrical, Electronics, and Computer Engineering (ISAEECE 2024)*, Vol. 13291, Oct., 1410–1416. SPIE. Accessed 2025-09-12. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/13291/132915N/HDBSCAN-based-semantic-clustering-model-in-classifying-incidents-on-security/10.1117/12.3033910.full.

Appendix A. Implementation Details

A.1. *Data Preprocessing*

The dataset includes 2,082 tweets from Trump's Twitter account, with 273 of these tweets retweeted from other accounts. We excluded all retweeted tweets out of two reasons. Firstly, our research focuses on Trump's tweets, and the retweeted contents are not written by Trump. Secondly, favorite counts for those retweeted tweets are 0 in the dataset, while other non-retweeted tweets generally have more than tens of thousands favorites. This discrepancy may result in inaccurate clustering results. Then, we checked for missing values and found four rows having one missing value in different columns. We removed the four rows from our dataset and retained 1,805 usable rows for our analysis. Before further analysis, we replaced links in text with an identical `<URL>` placeholder to avoid interference.

A.2. *LDA topic Modeling*

When performing LDA topic modeling, we removed stop words—unimportant words such as "the," "is," and "at"—from the tweets since these words appear pervasively across all tweets and do not belong to a single topic. In addition, we did not lemmatize the words in the tweets and kept the forms of the words as they are. To lemmatize means to transform a word to its base form. For example, the result of lemmatization of "using" is "use." According to Schofield and Mimno (2016), lemmatization does not improve the outcome of topic modeling and may even have an adverse effect on the result.

Appendix B. Topics of Other Clusters

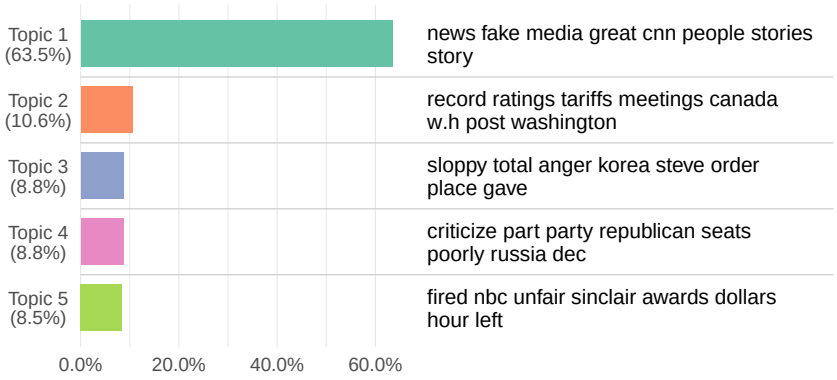

Figure B1. Cluster 1: topic importance (left panel) and corresponding topic keywords (right panel).

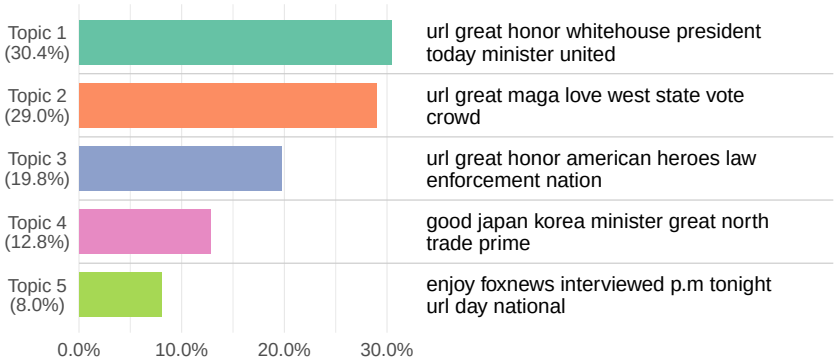

Figure B2. Cluster 4: topic importance (left panel) and corresponding topic keywords (right panel).

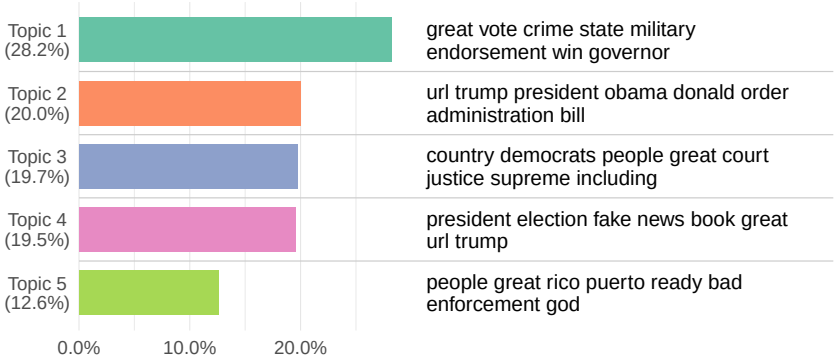

Figure B3. Cluster 5: topic importance (left panel) and corresponding topic keywords (right panel).

Appendix C. Supplemental Correlation Figures

Figure C1. Cluster 6: topic correlations.

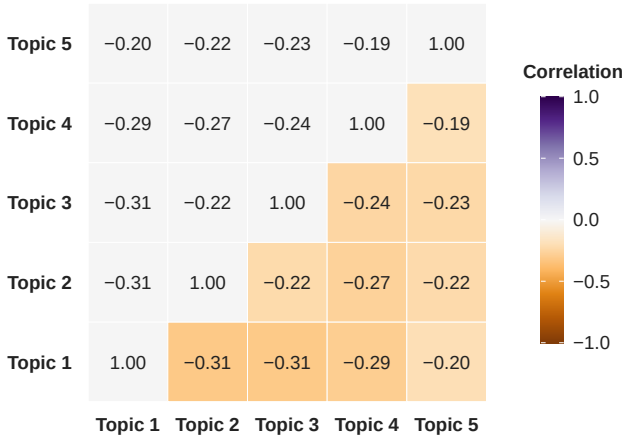

Figure C2. Cluster 7: topic correlations.