Comparing Viewer Sentiment on Political Topics

A Youtube Channel Analysis on Trump and Biden

Authors:
Lourenço Santos Moitinho de Almeida (4643607)
Tymo Staal (7277202)
Tom Mayer (4595238)
Sajad Ahmadi Jozdani (5328381)

Seminar Group: 2

Instructor's Name: Prof.dr.Antal van den Bosch

Utrecht University

Text and Media Analytics

Word Count: 3155

1 Introduction

Online social networks provide many people with unlimited access to information and connections. The content generated on these platforms greatly influences personal opinion and has a wide-reaching effect on society (Ferrara et al., 2013). Among the numerous possibilities, analyzing public opinion and sentiment has been a prominent topic in the domain of natural language processing (NLP). Formally, this research design is commonly known as a sentiment analysis or opinion mining which involves extracting behaviors, opinions, and sentiments from texts, chats, and similar content using natural language processing and information retrieval techniques (Neshan & Akbari, 2020).

A sentiment analysis is able to provide indispensable insights on human-computer interaction for professionals, researchers, and individuals in domains such as sociology, marketing, psychology, economics, and political science to name a few (Hutto & Gilbert, 2014). For example, academicians have published extensive research on analyzing sentiment during presidential elections. However, to our knowledge, limited work has been done to examine latent topics within this context. Furthermore, unlike Twitter and Facebook, Youtube and the videos published on the platform seem to be less investigated by researchers as evident in our literature review. As a result, this paper aims to capitalize on this opportunity and address the NLP research gap by applying the synergy of topic modeling and a sentiment analysis on the videos uploaded onto Youtube. Consequently, this led us to formulate the research question: **How does viewer sentiment vary between political topics across YouTube's online platform?**

In order to address this question, Latent Dirichlet Allocation (LDA) was used in order to uncover latent topics. The growing popularity and availability of resources for LDA topic modeling within natural language processing led us to its inclusion into this study (Negara et al., 2019). In a similar fashion, the sentiment analysis was applied through the simplistic yet functional design of the TextBlob library (Kaur & Sharma, 2020). By incorporating this research design with the available data on Youtube, this study aims to uncover invaluable insights for politicians and their strategic online presence.

2 Literature Review

Within the domain of natural language processing, analyzing sentiment across social media platforms has been a prominent theme in studies due to the diverse applications and implications of the obtained results. For instance, policymakers equipped with the knowledge gained from analyzing public opinion and emotion towards real-world events across online platforms could result in improved decision making. In fact, politicians have utilized online platforms such as Facebook and Twitter in the past in order to gauge the public consensus on themselves

(Dorle & Pise, 2018). During political elections these insights could prove to be an invaluable asset. In a similar setting, Dorle and Pise (2018) collected crawled data from Facebook and Twitter that was used to inform politicians on public opinion as opposed to the outdated and tedious door-to-door surveys that were commonly used in the past. Meanwhile, Ilyas et al. (2020) analyzed sentiment beyond a political context by monitoring Twitter discussions using the "Brexit" keyword. Results from their analysis revealed underlying Brexit effects on the stock market and on the exchange rate of the British pound.

Among all the possible use cases of a sentiment analysis, researchers have been particularly drawn to this type of investigation during presidential elections. The sheer magnitude of the results from this analysis has the potential to influence the nation for the following years to come. As an illustration of the possible implications, Oyewola et al. (2023) analyzed public sentiment during the 2023 prudential elections in Nigeria by classifying tweets related to the presidential candidates. Their findings offered valuable insights into the public's general consensus on each of the candidates and the effectiveness of their campaigns. Similarly, researchers (Budiharto & Meiliana, 2018) have analyzed sentiment in order to predict outcomes in the Indonesian presidential election. Between March and July of 2018, this study examined tweets from two election candidates by pinpointing key hashtags related to the election. By developing a model trained on this data researchers were able to uncover insights on sentiment polarity across the platform. As a result, the study was able to accurately predict Jokowi's election victory.

Undeniably, sentiment studies have proven to be imperative whether in a political or alternative context yet researches have utilized alternative approaches such as topic modeling as well. This method is an unsupervised machine learning technique that examines and clusters a corpus of documents through text similarity. As a result, topic modeling serves as an effective method for rapidly categorizing and understanding the core themes within a large body of (unlabeled) text (Naseem et al., 2020). To demonstrate, Anwar et al. (2021) conducted an analysis on tweets related to the US 2020 election containing the term "QAnon". Contrary to the widespread assumption that QAnon discussions on social media were predominantly present in states leaning towards the republican party, their findings indicated otherwise. Their research uncovered that a significant majority of Twitter users that were publishing QAnon related tweets were in fact supporters of Donald Trump and tended to have greater conservative and nationalistic views. While in the report of Golino et al. (2022), the researchers utilized topic modeling in order to investigate underlying topics within the 2016 US elections. In particular, dynamic exploratory graph analysis (DynEGA) was incorporated into their study which uncovered the intentions and influence of Russian state-sponsored Twitter accounts on the elections.

Despite the recurring themes of investigating sentiment and topic modeling across online platforms, limited research has been done that has incorporated the two methods together especially within a political context. In any case, Wisnu et al. (2020) attempted to address this dilemma by combining topic modeling and a sentiment analysis towards the 2018 Central Java Gubernatorial Election on Twitter. Their study aimed to connect patterns and topics uncovered through Latent Dirichlet Allocation (LDA) with the results of sentiment gathered through Twitter. Their findings revealed two dominant topics and the corresponding sentiment for each of the candidates which was later used to predict the outcome of the election. Evidently, this study demonstrated the effectiveness of synergizing topic modeling with a sentiment analysis.

3 Data & Methodology

3.1 Dataset Overview & Supplementation

In order to examine viewer sentiment regarding political topics across Youtube's online platform, the focal point of our study will begin with two datasets collected through YouTube's API. The software was then applied to videos published in 2023 from the official Trump and Biden Youtube channels. Hence, each dataset corresponds with one of the channels. In total, the data consisted of 43 and 41 unique videos from Biden and Trump respectively and each contained 13 variables. Overall the collected data encompassed a wide variety ranging from insightful video comments to more obscure statistics such as the url profile images of the authors of each comment. With that said, only two out of the 13 variables will be utilized from each of the original datasets. A brief description of the variables of interest can be seen below:

(i) Video ID: A unique video identifier found at the end of video urls.

(ii) Video Comments: A maximum of 500 comments ranked by relevance.

Furthermore, due to the nature of our research design (explained below), we supplemented the datasets with additional metrics. Through the reuse of the YouTube API, data on the amount of video likes, comments, and views was collected. Next, the Python package named 'youtube-transcript-api' along with the original video ids allowed the corresponding video transcript to be collected as well (Depoix, 2023). Moreover, it's important to address the potentially misleading name of the package for which a clear distinction needs to be made between the Python library and the official Youtube API. Lastly, due to the possible ethical concerns regarding web scraped data, only the required data for this study was collected.

3.2 Methodology Overview

In this paper, a multi-step analysis was conducted that aimed to compare viewer sentiment and topic engagement on the official YouTube channels of Donald

Trump and Joe Biden. We begin by merging datasets collected from YouTube's API, followed by extracting video transcripts for detailed content analysis. We assess the polarity and subjectivity of each video and identify and categorize prevalent topics within the video content by performing topic modeling on the transcript of the videos. Finally, conducting a sentiment analysis on the comment sections within the specific topics to understand how viewer sentiment varies between topics comparing Trump and Biden. An overview of the framework used can be seen in Figure 1.

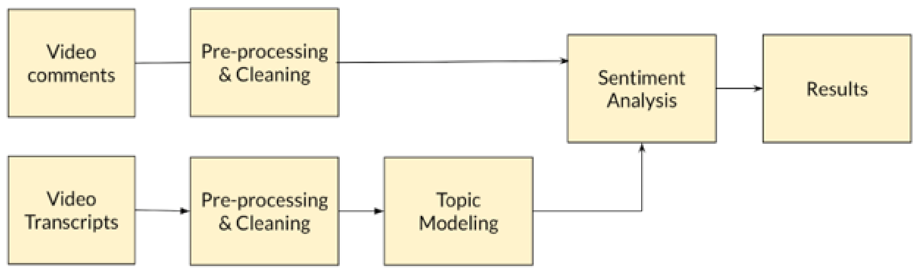

Figure 1: Methodology Framework Overview

3.3 Pre-processing & Cleaning

The differences in format and content between the video transcripts and comments required differing pre-processing steps.

3.3.1 Comments

Pre-processing comments started with checking if the input text was missing and returning an empty string if true. In the comment text a lot of emojis were included which led to the the Emoji library to be used. This process converted emojis into descriptive text in order to retain their emotional or contextual significance in a form suitable for the sentiment analysis. Following this, punctuation was removed and all characters were converted to lowercase to standardize the text, making it uniform for further processing or analysis by eliminating variations due to capitalization or punctuation marks.

3.3.2 Transcripts

The pre-processing of the transcripts started by retrieving the data from the API. Unfortunately, 4 out of the 84 unique videos had either disabled or unavailable transcripts. We decided not to take these videos into account, which was not a big problem since the data was well-balanced. Additionally, the data was not limited to only videos but included Youtube shorts as well. Further investigation revealed that these videos were significantly shorter (as the name suggests) which in turn reduced the length of the collected transcript.

The true pre-processing started by removing stopwords (common words that add little value to analysis) from the input text, including a custom list of additional stopwords relevant to the context (e.g. names and transcript-specific words such as [applause] or [music]). Afterwords, the text was then converted to lowercase and split into tokens (words). Lastly, these tokens (excluding stopword and punctuation tokens) were then lemmatized into their base form.

3.4 Topic modeling

In order to find meaningful topics within the transcripts numerous approaches were tested ranging from optimizing the amount of LDA topic clusters with respect to maximize coherence scores to more complex models such as BERTopic. Among the approaches tested, the best topic modeling results originated from adjustments to the LDA model. The first preliminary step of the process required the processed transcripts to be turned into a corpus that was compatible with the LDA model. Next, a model with five distinct clusters was run to obtain the top 30 words within each of the topics. Each transcript was then assigned to one of the topic clusters based on the probabilistic results of the model. To interpret these topic modeling results and understand the themes within each cluster, Bard was utilized, providing insights and interpretations of the clusters.

3.5 Sentiment analysis

Through the lexicon-based TextBlob library, a sentiment analysis focusing on polarity and/or subjectivity was conducted on the video comments and transcripts. Unlike the pre-processing steps above, the sentiment analysis between the two were more comparable. Results from the polarity analyses ranged from -1 to 1. Transcript polarity results retained this raw value while the polarity of the comments was further categorized into negative, positive, and neutral sentiment based on a threshold of less than -0.1, greater than 0.1, and between -0.1 and 0.1 respectively. This additional step was included to improve the interpretability of our results. On the other hand, the exploration of sentiment subjectivity was limited to only the transcript data. Unlike the sentiment results of polarity which will be later compared and evaluated, subjectivity results mainly served to aid in the division of latent topics during the LDA topic modeling process.

4 Results

4.1 Topic Distribution Analysis

As displayed in Figure 2, the topic distribution varied significantly between Trump and Biden. However, the dynamics of the clusters were comparable. There was a certain concentration around the two biggest topics: “American

Aspirations and Unity" and "America's Global Leadership and Energy Security". This overlap in topic focus could be attributed to several factors. For example, in the political context of the US where themes such as unity and shared aspirations often emerge and are used as a powerful tool to connect with a broad spectrum of voters. On the "Drug Crisis and Public Health" we see a modest contribution compared to the rest of the topics and with a significant difference between Trump and Biden. By examining the content of these topics, it's interesting to see that drugs and public health were within the same cluster. Furthermore, the content of Biden tendded to focus on topics such as affordable healthcare whereas Trump was more focused on crimes and drugs. The distribution of the Economic topic was very similar and for the "Social Justice and Individual rights" we see a big difference yet again. Trump almost has double the percentage of Biden and again within this topic we see a distinction between the two. Trump has more of a focus on freedom and rights associated with this while Biden aimed to address more on abortion and the rights of minorities.

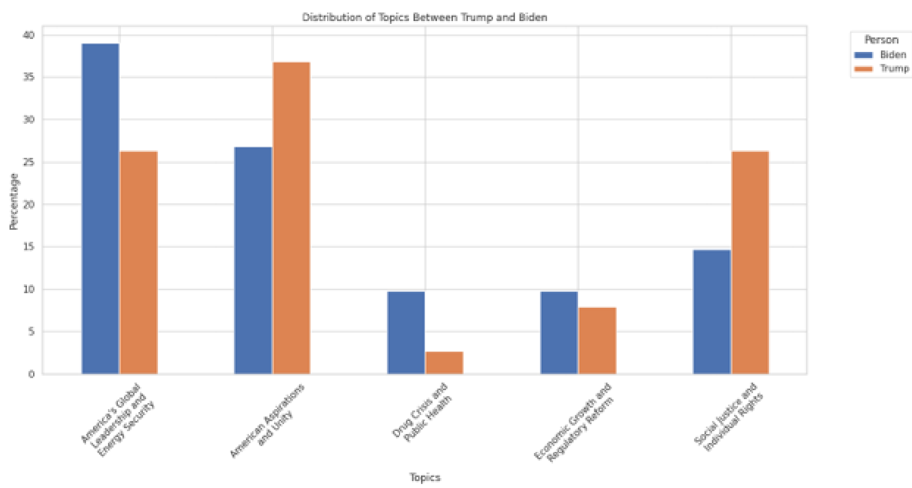

Figure 2: Topic Distribution between Trump and Biden

4.2 Comment Polarity Comparison

Figure 3 shows a comparison of the polarity of comments about Trump and Biden. Polarity is a measure of sentiment, and can range from -1 (negative) to 1 (positive). The chart shows that the average sentiment of comments about Trump was more positive than the average sentiment of comments about Biden. However, it is important to note that the chart also shows a wider range of sentiment for Trump in contrast to Biden. This suggests that while there are more positive comments about Trump on average, there are also more negative comments about him which consequently make him the more polarizing political figure among the two.

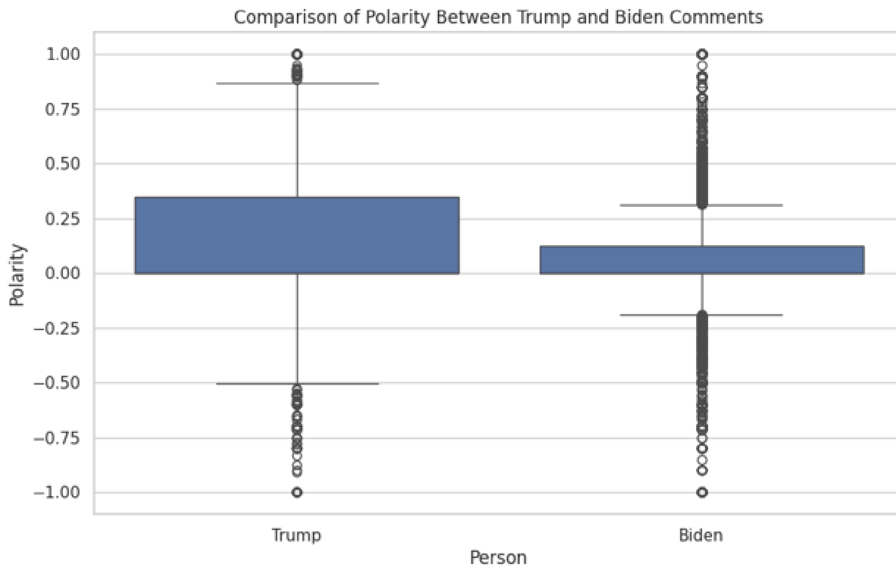

Figure 3: Comparison of Polarity between Trump and Biden Comments

4.3 Comment Polarity within Topics

To further understand the sentiment landscape within the topics discussed on the Trump and Biden YouTube channels, an analysis of comment polarity within the topic clusters was conducted. The polarity of comments grouped in this manner offers insight into how viewers react and engage with content related to different topics. Initially, a general sentiment analysis was carried out to establish a baseline for both Trump and Biden. The overall proportion of both positive and negative comments was calculated for Trump and Biden separately. As described in the method section to determine if a polarity score was positive or negative we used a threshold value. Quite a lot of comments were neutral which allowed an ideal way of finding polarizing topics and comparing the sentiment within the topics. As mentioned before Trump had an higher polarity average which meant that the proportion of positive comments was also higher. The higher positivity could be indicative of the type of audience Trump attracts or the nature of his communication style, which may be more aligned with the sentiments of his viewers.

In the results, "America's Global Leadership and Energy Security" emerged as a relatively highly polarized topic for both Donald Trump and Joe Biden, drawing an above-average volume of both positive and negative comments which indicated the nature of the discussions around America's foreign policy and energy strategy. For Joe Biden, the "Social Justice and Individual Rights" topic showed a notably lower proportion of negative comments, reflecting a more favorable view of his policies in these areas among his audience. Conversely, Donald Trump's coverage of "Economic Growth and Regulatory Policies" attracted the highest proportion of negative feedback, with relatively fewer positive responses, highlighting critical views on his economic and regulatory strategies.

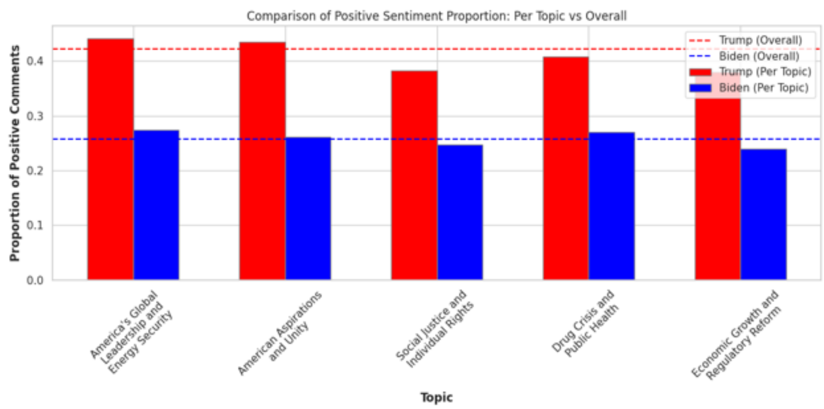

Figure 4: Comparison of Positive Sentiment Proportion: Per Topic vs Overall

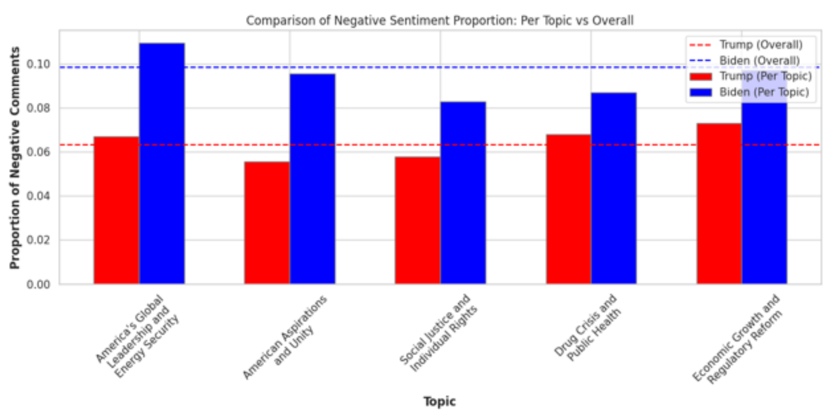

Figure 5: Comparison of Negative Sentiment Proportion: Per Topic vs Overall

5 Discussion and Conclusion

In this study, we investigated viewer sentiment on political topics across YouTube's online platform, specifically focusing on the channels of Donald Trump and Joe

Biden. Using a multimodal approach, we investigated latent topics and performed sentiment analyses using data from transcripts and video comments. Our results offer insightful information about the shifting landscape of viewer sentiment and how it differed between the two political figures.

With 43 and 41 videos from Biden and Trump, respectively, the dataset enabled us to use YouTube's API in addition to other metrics for a thorough examination. The methodology comprised distinct analyses of the transcripts and video comments, which were then integrated to present a comprehensive picture. The difficulties in text pre-processing and transcript topic modelling highlight the difficulty of analysing alternative types of content.

The findings showed that Trump and Biden had different topic distributions, with concentrations around general concepts like "American Aspirations and Unity" and "America's Global Leadership and Energy Security." Notably, the study showed that the two prominent individuals' sentiment polarities differed, with Trump drawing, on average, more positive remarks but also a wider range of sentiments, indicating his polarising influence.

Moreover, the polarity of comments on particular subjects revealed subtle trends in the responses of viewers. While Biden's talks on subjects like "Social Justice and Individual Rights" showed clear differences, Trump's higher average positivity suggested a distinctive audience engagement or communication style.

Future research is necessary to address issues arising from small dataset sizes, explore alternative models, and refine methodologies due to the challenges encountered, particularly in transcript topic modelling. In order to obtain a thorough understanding of viewer sentiment patterns, the study also highlights the possibility of integrating sentiment analysis across a variety of social media platforms.

TextBlob offers a useful sentiment analysis tool, but little is known about its internal operations. To gain a deeper understanding of the sentiment classification process, future research could examine the nuances of sentiment analysis tools and how they affect outcomes.

As previously mentioned, the difficulties with topic modeling, particularly the imprecise outcomes from transcript data, make it difficult to determine whether the existing methods are appropriate for small dataset sizes. Further research endeavors may investigate substitute models or employ a supervised methodology to enhance the precision of topic clustering.

Including Twitter and other social media sites in the analysis in addition to YouTube might yield a more thorough understanding of viewer sentiment. Identifying whether sentiment patterns are consistent across different online channels or platform-specific would be possible with cross-platform analysis.

Expanding the dataset to include more videos or a wider range of political figures may improve the findings' generalizability. Furthermore, delving into aspects other than sentiment polarity, like detecting sentiment in comments, may offer a more thorough comprehension of viewer interaction, including instances of support or possible hate speech.

6 References

Anwar, A., Ilyas, H., Yaqub, U., & Zaman, S. (2021). Analyzing Qanon on twitter in context of us elections 2020: Analysis of user messages and profiles using vader and bert topic modeling. DG. O2021: The 22nd Annual International Conference on Digital Government Research

Budiharto, W., & Meiliana, M. (2018). Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis. *Journal of Big data*, *5*(1), 1-10.

Depoix, J. (2023) Youtube-transcript-api (Version 0.6.2) [Computer Software].Python Software Foundation.https://pypi.org/project/youtube-transcript-api

Dorle, S., & Pise, N. (2018). Political sentiment analysis through social media. 2018 second international conference on computing methodologies and communication (ICCMC)

Ferrara, E., Varol, O., Menczer, F., & Flammini, A. (2013). Traveling trends: Social butterflies or frequent fliers? Proceedings of the first ACM conference on Online social networks

Golino, H., Christensen, A. P., Moulder, R., Kim, S., & Boker, S. M. (2022). Modeling latent topics in social media using Dynamic Exploratory Graph Analysis: The case of the right-wing and left-wing trolls in the 2016 US elections. *Psychometrika*, 1-32.

Hutto, C., & Gilbert, E. (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

Ilyas, S. H. W., Soomro, Z. T., Anwar, A., Shahzad, H., & Yaqub, U. (2020). Analyzing Brexit's impact using sentiment analysis and topic modeling on Twitter discussion. The 21st Annual International Conference on Digital Government Research

Kaur, C., & Sharma, A. (2020). Twitter sentiment analysis on coronavirus using textblob. *EasyChair2516-2314*.

Naseem, U., Razzak, I., Musial, K., & Imran, M. (2020). Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems*, *113*, 58-69.

Negara, E. S., Triadi, D., & Andryani, R. (2019). Topic modelling twitter data with latent dirichlet allocation method. 2019 International Conference on Electrical Engineering and Computer Science (ICECOS),

Neshan, S. A. S., & Akbari, R. (2020). A combination of machine learning and lexicon based techniques for sentiment analysis. 2020 6th international conference on web research (ICWR)

Oyewola, D. O., Oladimeji, L. A., Julius, S. O., Kachalla, L. B., & Dada, E. G. (2023). Optimizing sentiment analysis of Nigerian 2023 presidential election using two-stage residual long short term memory. *Heliyon*, *9*(4).

Wisnu, G. R. G., Muttaqi, A. R., Santoso, A. B., Putra, P. K., & Budi, I. (2020). Sentiment analysis and topic modelling of 2018 Central Java gubernatorial election using Twitter data. 2020 International Workshop on Big Data and Information Security (IWBIS)