Sentiment Analysis on the Potential Ban of TikTok

Background

On March 23rd at the 32nd congressional hearing regarding the privacy and use of data within big tech companies, TikTok's CEO, Shou Chew, was grilled for four-and-a-half hours by a US committee, where he attempted to protect his company from a potential ban or forced sale in the country. The potential ban of TikTok in the US raises several social issues, including data privacy, national security, and freedom of speech. Data privacy and national security have been continual concerns expressed by the US government. Big technology companies such as Meta have already been under fire by Congress for the alleged misuse of user data. TikTok has been accused of collecting user data and sharing it with the Chinese government, which has raised concerns about the safety and security of user information. Chew defended TikTok, arguing that it is no different than other social media giants and that the company has put in place even stronger safeguards than its competitors. However, the purpose of this paper is not to conclude who is right or wrong in this complex situation. The purpose is to perform sentiment analysis on a reddit post, posted in r/technology, with the title of "There's a 90% chance TikTok will be banned in the US unless it goes through with an IPO or gets bought out by mega-cap tech, Wedbush says." This title refers to an article posted by Business Insider, with the title specifically belonging to a quote by a Wedbush Securities analyst.

Data Preprocessing

The post itself was posted on March 27th, and since then has garnered over 49,000 upvotes, 5,800 comments and has gained significant traction. To properly clean the comments for sentiment analysis, spaCy and NLTK were used. They are both natural language processing (NLP) libraries in Python for processing text data and various functions for text analysis. I used the Natural Language Toolkit (nltk) package to download stopwords. Once the stopwords were removed from the text, I processed the remaining words using the spaCy model. Using spaCy, I removed words with part-of-speech tags such as 'PRON', 'ADV', 'SCONJ', 'DET', 'AUX'. Stopwords and these tags do not meaning to the text. I created a 'reddit_clean()' function defined to perform various text cleaning techniques such as converting text to lowercase, removing apostrophes, hashtags, and URLs. In the final steps of preprocessing, I tokenized each comment by splitting the text into individual words and stemmed the words in each comment. By tokenizing and

stemming, I reduced each individual word to its base form. By breaking down the text into individual tokens and reducing them to their root forms, these preprocessing techniques help to clean and standardize the text data, making it easier to analyze and model. More can be read within the comments of the code.

Sentiment Analysis

It is critical to understand that the purpose of analyzing the sentiment is not to conclude whether the actions done by the committee or TikTok CEO Shou Chew are objectively right or wrong. The analysis of the post is to give a perspective on what the Reddit users are expressing and expand into their intuition. The overall sentiment of the Reddit post was largely negative with users commenting on the US congressional committee failing to ask the correct questions regarding actual data privacy. From the word cloud (figure 1), it is apparent that there are many words associated with negative connotation, expressing the sentiment of the users. Words such as ‘threat’, ‘propaganda, ‘stupid’, ‘wrong’,’ worse’ seem to describe the overall approach used by the congressional committee in efforts to attack Shou Chew. In addition, words such as ‘companies’, ‘meta’, ‘facebook’ ‘twitter’, ‘instagram’ seem to connote that users are actively comparing this hearing to past hearings of social media companies, as well as comparing the actions or interests of these companies to TikTok. Finally, words such as ‘data’, ‘security’, ‘privacy’, and ‘freedom’ express the users’ acknowledgement of the core of this social issue, how TikTok, an international company, may use data in ways that might compromise its users.

Using VADER’s ‘SentimentIntensityAnalyzer()’, the comments were split into positive, neutral or negative, as well as the overall polarity of the comments were produced. In addition, the use of Textblob determined the subjectivity of each comment. within the context of the top 10 neutral words (figure 2) within the comments, I specifically analyzed the use of the words within context. For example, the top 10 words used within a comment classified under negative sentiment (figure 3) was extracted. After, the process was repeated with positive (figure 4) and neutral sentiment (figure 5). We can see that the top 10 words for each sentiment is not all the same. For example, the words ‘tiktok’ and ‘people’ was used much more in a negative sentiment context as opposed to a positive context. The word ‘banned’ only appears within the top 10 words included in negative sentiment. Reddit commenters seem to be reflecting in a negative manner on the committee attacking Chew's company's data security and content moderation

policies, and the impact of the app on its US users. Figure 6 shows descriptive statistics for both polarity and subjectivity, and figure 7 displays a table depicting counts for sentiment and subjectivity. Figure 6 shows numerical evidence that the mean polarity of the comments was negative, with a polarity average of -0.026. From figure 7, it can be determined that there were more negative comments as opposed to neutral or positive. Because the text is from a Reddit post, where commenters have no obligation for objectivity, it makes sense that neutral comments saw a much higher objectivity rate of 89% as opposed to 62% and 66% objectivity rates from positive and negative comments, where Reddit users may be more inclined to express subjective opinions. Finally, the sentiment distribution (figure 8) and subjectivity distribution (figure 9), further strengthens our findings that most of the comments were only slightly positive or negative, and most comments were objective.

Conclusion

In conclusion, the analysis of the sentiment of the Reddit post regarding the US congressional committee's questioning of TikTok CEO Shou Chew reveals that the overall sentiment was negative. Reddit users expressed their dissatisfaction with the committee's line of questioning and highlighted concerns regarding TikTok's data privacy and content moderation policies. The use of VADER's SentimentIntensityAnalyzer and Textblob provided numerical evidence supporting these findings. The results suggest that while most comments were objective, negative sentiment comments outweighed positive sentiment comments. This analysis highlights the importance of understanding public sentiment and opinion when addressing social issues, especially in the context of social media platforms regarding data privacy and security.

Limitations

Most of the sentiment analysis was done using spaCy. I believed it would give more accurate results because spaCy is generally better with social media sentiment analysis, but the results may have been different with a full analysis using Textblob. In addition, perhaps I did not remove all parts-of-speech needed to give the most accurate sentiment and polarity scores. Although the overall polarity was negative, I believe it would be even more negative with better data preprocessing. A fuller understanding of these NLP techniques in the future would yield better results.

Appendix

Figure 1

Figure 2

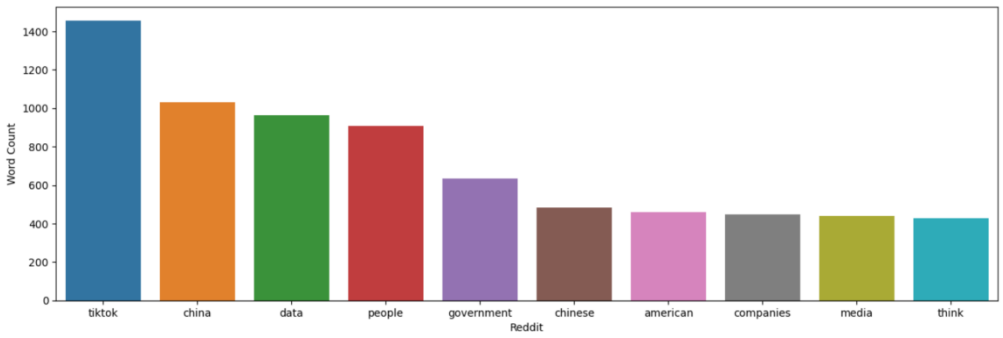

Figure 3

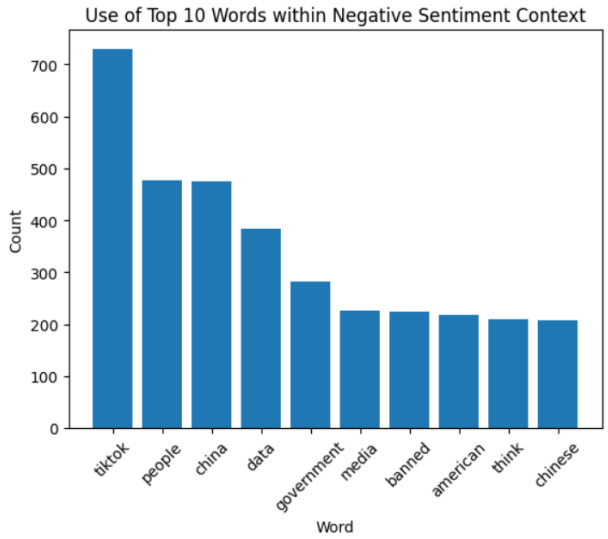

Figure 4

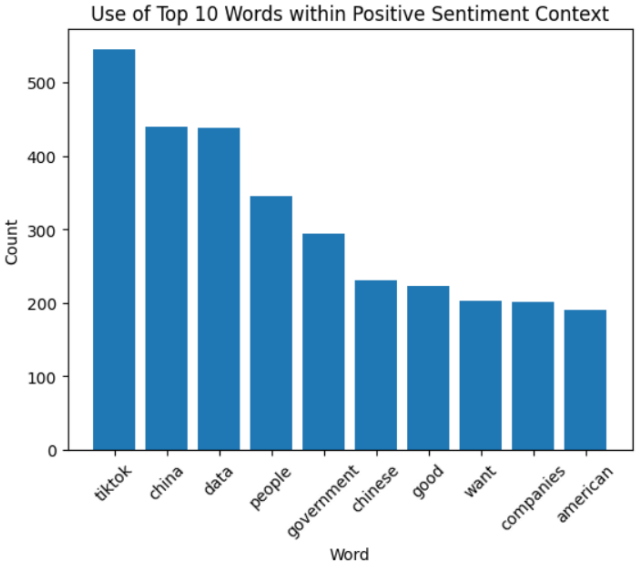

Figure 5

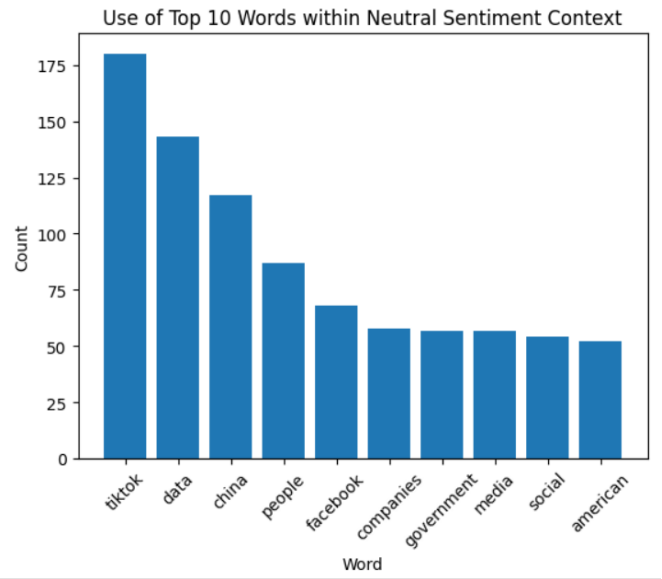

Figure 6

	polarity	subjectivity
count	4998.000000	4998.000000
mean	-0.026755	0.338540
std	0.476118	0.300228
min	-0.995600	0.000000
25%	-0.401900	0.000000
50%	0.000000	0.344097
75%	0.340000	0.555952
max	0.994200	1.000000

Figure 7

subjectivity2 sentiment	Objective	Subjective
Negative Sentiment	1231	691
Neutral Sentiment	1163	138
Positive Sentiment	1101	674

Figure 8

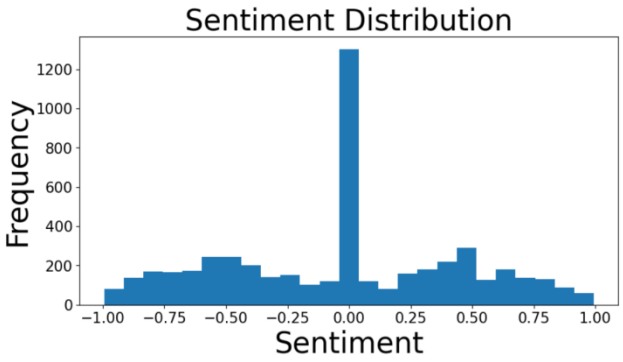

Figure 9