A Fine-Grained Analysis of Public Opinion toward Chinese Technology Companies on Reddit

Enting Zhou*, Yurong Liu*, Hanjia Lyu, Jiebo Luo

Department of Computer Science
University of Rochester
Rochester, USA
{ezhou12, yliu217}@u.rochester.edu, hlyu5@ur.rochester.edu, jluo@cs.rochester.edu

Abstract—In the face of the growing global influence and prevalence of Chinese technology companies, governments worldwide have expressed concern and mistrust toward these companies. There is a scarcity of research that specifically examines the widespread public response to this phenomenon on a large scale. This study aims to fill in the gap in understanding online public opinion toward Chinese technology companies using Reddit data, a popular news-oriented social media platform. We employ the state-of-the-art transformer model to build a reliable sentiment classifier. We then use LDA to extract the topics associated with positive and negative comments. We also conduct content analysis by studying the changes in the semantic meaning of the companies’ names over time. Our main findings include the following: 1) Notable difference exists in the proportions of positive comments (8.42%) and negative comments (14.12%); 2) Positive comments are mostly associated with the companies’ consumer products, such as smartphones, laptops, and wearable electronics. Negative comments have a more diverse topic distribution (notable topics include criticism toward the platform, dissatisfaction with the companies’ smartphone products, companies’ ties to the Chinese government, data security concerns, 5G construction, and general political discussions); and 3) Characterization of each technology company is usually centered around a particular predominant theme related to the company, while real-world political events may trigger drastic changes in users’ characterization.

Index Terms—public opinion, Chinese technology company, social media, text analysis, transformer-based language model

I. INTRODUCTION

In the past few years, China has been quickly building up its global influence, especially in the realm of technology. The internationalization of Chinese technology companies has also contributed to the growing global influence. The ubiquitous presence of Chinese technology companies has unnerved governments worldwide, especially in the West. Many governments have put forward legislation to tightly regulate Chinese technology companies’ operations in their countries. In some extreme cases, governments have explicitly forbidden certain Chinese technology companies from providing service in their countries [6], [11].

Previous studies investigated governments’ reactions to Chinese technology companies and sought to explain the governments’ mistrust toward these companies [4], [8]. Few previous studies have focused on the large-scale public reaction to the growing influence and presence of Chinese technology companies. Therefore in this study, we attempt to study public opinion toward Chinese technology companies, especially ones with great influence outside China, such as Huawei, Tencent, ByteDance (Tiktok), and Alibaba. We aim to answer the following research questions:

• RQ1: What sentiment do Reddit users generally express in discussions of Chinese technology companies?
• RQ2: What are the topics associated with the positive and negative sentiments?
• RQ3: What are the dominant features in the discussions of a particular company? Do they change over time?

We approach these questions using a series of computational methods. We adopt a human-guided machine learning framework based on a transformer model to classify comments’ sentiments. We model the latent topics of both positive and negative comments. We investigate the users’ characterization of each particular company by training a set of word2vec models to generate word embeddings and calculating the most semantically similar words to the company names. To summarize, we find:

• The public often show more negative attitude than positive attitude toward Chinese technology companies.
• Positive comments are mainly about companies’ products, while the topics are more diverse in the negative comments.
• The results of the word embeddings show predominant themes in the discussions regarding Chinese technology companies. In addition, we have identified various real-world events that have a significant impact on how the general public perceives these companies.

Compared to previous studies focusing on the relations between Chinese technology companies and the Western governments [4], [8], we contribute to a better understanding of large-scale public sentiment toward Chinese technology companies. To our best knowledge, this is the first study that investigates public opinion toward Chinese technology companies. Additionally, we make our dataset publicly available to the research community to facilitate future work.

II. RELATED WORK

Many studies have been conducted to monitor and analyze the opinion on Reddit. Some studies collected posts under a
TABLE I: The labeling scheme.

| Class              | Description                                                                 |
|--------------------|-----------------------------------------------------------------------------|
| Positive           | i. Expressing positive sentiment toward the Chinese tech companies themselves |
|                    | ii. Expressing positive sentiments toward the tech companies’ products       |
|                    | (i.e., social platforms, smartphones, laptops, etc.)                       |
| Negative           | i. Expressing negative sentiment toward the Chinese tech companies themselves |
|                    | ii. Expressing negative sentiment toward Chinese tech companies’ products   |
|                    | (i.e., social platforms, smartphones, laptops, etc.)                       |
|                    | iii. Promoting/arguing in favor of conspiracy theories about the             |
|                    | Chinese tech companies                                                      |
|                    | iv. Advocating for a ban or sanction toward the Chinese tech                |
|                    | companies                                                                  |
| Neutral or         | i. News regarding Chinese tech companies with no written opinion            |
| Irrelevant         | from the commenters                                                         |
|                    | ii. Including Chinese tech companies and the commenters’ opinions,          |
|                    | but the focus is something else (i.e., politics, economics, etc.)           |
|                    | iii. Comments with neutral opinion toward Chinese technology companies      |
|                    | iv. Comments/questions on Chinese tech companies entities but with           |
|                    | unclear meanings                                                           |

small number of selected related subreddits. For instance, Shen and Rudzicz [12] investigated anxiety on Reddit by comparing the posts under a group of anxiety-related subreddits to the posts under a control group of subreddits that are unrelated to anxiety. Farrell et al. [3] analyzed the phenomenon of misogyny by collecting posts under a set of carefully selected subreddits around the topics of men’s rights and difficulty in relationships. Other studies have used off-the-shelf datasets that contain all the posts on Reddit in a given period. For instance, Soliman, Hafer, and Lemmerich [13] studied the characterization of the political community using a readily available dataset that includes all submissions and comments on Reddit from 2005 to 2018. Such a data collection method provides the scale that we require for our study, but it introduces enormous difficulty to distinguish between relevant and irrelevant content to our study subjects. Wu, Lyu, and Luo [15] studied public sentiment toward COVID-19 vaccines. In this paper, we adopt their methodology to perform keyword research on a sitewide basis, which provides a breadth of data while keeping data noise at a minimum.

Computational methods have been employed to study public sentiment using social media data. To study public sentiment toward COVID-19 vaccines, Lyu et al. [7] applied the state-of-the-art transformer model - XLNet to mining opinions. Tahmasbi et al. [14] aimed at revealing online Sinophobic behaviors during the COVID-19 pandemic using word embeddings. By modeling Chinese-related terms on Twitter and 4chan forums as word vectors, they found that Chinese-related terms are associated with racial slurs on both Twitter and 4chan, thus revealing the rise of Sinophobic behaviors in a cross-platform manner. By analyzing the word embeddings' temporal change, they discovered new racial slurs related to China and the tendency to blame China and Chinese people as the pandemic escalates. In our study, we apply a similar method to understanding public opinion about Chinese technology companies.

A. Data Collection

To collect users’ comments, we employ a Python Reddit API Wrapper (PRAW). We perform keyword searches with names of the companies and the names of their CEOs (e.g., “Tencent”, “Huawei”, “Ma Huateng”, “Ren Zhengfei”) on a sitewide basis. We pull comments over two years, from November 1, 2019 to November 1, 2021. Duplicate, non-English, and automatic moderator comments are pruned from the dataset. Our final dataset contains 294,610 comments from 172,453 distinct authors.

B. Data Preprocessing

To prepare the data for further sentiment classification and language modeling, we perform a text cleaning process. We convert all words to lowercase and remove all uniform resource locators (URLs) and numbers from the text. We use a dictionary of English stop words provided by the Natural Language Toolkit (NLTK) to remove all stop words from the text. Additionally, we perform a text lemmatization using NLTK.

IV. Text Sentiment Classification

A. Method

XLNet [16] is a generalized autoregressive pretraining method that can capture left and right contexts jointly in sentences. We employ XLNet Base for our task and limit each comment_body to the first 512 tokens. Next, we use the Adam optimizer to fine-tune our XLNet model for three epochs. It predicts a probability for each of the three possible categories (i.e., positive, negative, neutral/irrelevant) for comment_body. We have also experimented with VADER [5], a lexicon and rule-based sentiment analysis tool which is specifically designed for social media posts.

B. Data Labeling

To study the sentiment of each comment toward a Chinese technology company, we classify each comment into three categories: (1) positive, (2) negative, and (3) neutral/irrelevant according to the labeling scheme in Table I. Our initial collected data presents a challenging class imbalance problem among these three classes where the majority of the data is irrelevant to the opinion toward these companies. To address this problem, we adopt the approach by Lyu et al. [7] to employ a human-guided machine learning framework based on the state-of-the-art transformer model.

More specifically, we build our initial training dataset by randomly sampling 1,600 comments from the entire corpus of 294,610 comments. For each comment, two researchers independently read and label the comment with one of the three categories presented in Table I. If the two labels given by the two researchers are different, then a third researcher would discuss with the group to determine the consensus label of the comment. It is worth noting that three researchers’ decisions reach a Cohen’s Kappa score [9] of 0.77, which indicates a good agreement among the annotated labels.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
TABLE II: Performance of the XLNet model and VADER.

| Model | Class          | Precision | Recall | F1-score |
|-------|----------------|-----------|--------|----------|
| VADER | Neutral/Irrelevant | 0.83      | 0.17   | 0.28     |
|       | Negative        | 0.36      | 0.59   | 0.45     |
|       | Positive        | 0.29      | 0.90   | 0.44     |
|       | Overall (Weighted) | 0.65      | 0.38   | 0.34     |
| XLNet | Neutral/Irrelevant | 0.86      | 0.80   | 0.83     |
|       | Negative        | 0.62      | 0.69   | 0.65     |
|       | Positive        | 0.74      | 0.82   | 0.78     |
|       | Overall (Weighted) | 0.79      | 0.78   | 0.78     |

We train an XLNet model $H$ using these 1,600 labeled comments as the initial training corpus ($n = 1,600$). However, due to the severe imbalance issue, $H$ performs poorly at predicting positive comments. We then use $H$ to construct a new batch of training corpus of 1,000 comments. Out of these 1,000 comments, 45% are the comments that $H$ predicts to be most likely positive, 45% are the comments that $H$ predicts to be most likely negative, and 10% are randomly sampled to increase diversity. The new batch of 1,000 comments is labeled by researchers and added to our original training corpus. Next, we train $H$ using this new training corpus ($n = 2,600$). This entire process is considered one iteration. We repeat two iterations before obtaining our final training corpus, with balanced data between positive and negative categories. This framework actively searches for the most possible negative and positive comments to increase the size of our training data and strike toward a more balanced class distribution.

C. Evaluation

Table II summarizes our final model’s performance. We use the “weighted” F1-score to evaluate the overall performance on a test set that is also annotated in the same way. Note that this test set is not used during model training. An overall F1-score of 0.78 is obtained by the XLNet model, as well as similar F1-scores for positive and negative classes, which are sufficiently reliable for our further analysis. The XLNet model outperforms the lexicon and rule-based sentiment analysis tool VADER. This further supports the necessity of our human-guided machine learning framework.

Based on the classification results by our fine-tuned XLNet model on the entire dataset, we find, while our study samples comprise primarily (77%) of neutral or irrelevant content, there are notable differences in the proportions of positive comments (8.42%) and negative comments (14.12%), where there are 68% more negative comments than positive comments. This finding gives insight into the general sentiment the Reddit community holds toward Chinese technology companies. Our subsequent topic analysis and content analysis will delve into the underlying reasons for these variations in sentiment, with meticulous attention to detail.

V. TOPIC MODELING

A. LDA

To characterize different topics associated with positive and negative comments, we apply Latent Dirichlet allocation (LDA) [2] separately for negative and positive comments. To better differentiate possible topics, we remove the names of the companies we are studying from the processed text corpus. To determine the optimal number of topics, we train a set of LDA models, compute their coherence scores, and read the representative keywords. For negative comments, the number of topics is set to 8. The model has a coherence score of 0.54. The number of topics for positive comments is set to 2 which gives rise to a coherence score of 0.56.

B. Result

Fig. 1 visualizes the keywords in each negative comment topic in the form of a word cloud and displays the percentage of comments belonging to each topic. We adopt a common approach in topic modeling by manually assigning topic labels according to keywords in the topics generated from the model. We also categorize each comment to its dominant topic. Interestingly, we can observe two major groups of topics. The first, second, third, and seventh topics could be grouped together, comprising 2/3 of all negative comments, because they are mostly concentrated on the “technology” aspect of these companies, while the fourth, fifth, and sixth topics could be grouped together, comprising 1/3 of all negative comments, because they are concentrated on the “Chinese part” of Chinese technology companies.

We have also analyzed the topic distributions from a cross-sectional perspective (Fig. 2). The topic distributions correspond to the subreddit’s theme well. Subreddits that are centered around games, such as r/PUBGMobile and r/CallOfDutyMobile, have the largest proportions of negative comments of the gaming industry topic, whereas subreddits that are centered around smartphones, such as r/Xiaomi and r/Android, have the largest proportions of negative comments of phone complaints and criticism topic. More importantly, subreddits that have a more political theme, such as r/worldnews, r/Canada, and r/China, have apparently more discussions surrounding Chinese technology companies’ ties to the Chinese government, global politics, and 5G constructions. The coherence between topic distributions and subreddits’ themes shows the good performance and robustness of our topic model.

Fig. 3 shows the two topics that are associated with positive comments. Compared to negative comments, positive comments are more monotonic. They are largely associated with the companies’ physical products. The positive comments are approximately equally divided between these two topics. The first one focuses on the general aspect of physical products, which constitutes 55.19% of the total comments, and the second one is more specifically related to the companies’ smartphone products, which accounts for 44.81%.

VI. TEMPORAL & CONTENT ANALYSIS

To further understand the context of the words and the changes of public opinion on Chinese technology companies over time, we employ the skip-gram algorithm [10]. Specifically, we train a set of word2vec models on specific
For positive comments and negative comments, we train one word2vec model separately for each month. After training, we calculate the cosine similarity of the word vectors obtained from each model to study the similarity among words. We conduct an analysis of the top four mentioned companies, Huawei, Tencent, ByteDance, and Xiaomi. However, we discover the discourse around Xiaomi is mainly about its consumer electronics, which coincides with our finding in the previous section using LDA and offers no new insight, therefore we omit the discussion of Xiaomi in this section and focus solely on the other three companies.
TABLE III: Top 25 most semantically similar words to “Huawei” obtained from the monthly word2vec models of negative comments from November 1, 2019 to November 1, 2021. Words related to telecommunication technology, political entity, and network security are labeled red, and words related to consumer electronic products are labeled blue, respectively.

A. Case Study 1: Huawei

First, we look into the overall use of words in negative comments about Huawei. Table III illustrates the top 25 most similar words to “Huawei” each month from November 1, 2019 to November 1, 2021. We notice that there are three dominant groups of frequent words in negative comments:

- Words closely related to telecommunication technology: “5G”, “network”, “equipment”, “infrastructure”, “tech”, “technology”, “hardware”.
- Words related to political entity: “countries”, “UK”, “US”, “Canada”.
- Words that express concern on network security: “security”, “spy”, “backdoor”, “concern”, “risk”.

The frequent occurrence of these three topics as the top most similar words indicate that these words have the most similar context to “Huawei” the majority of the time.

While Huawei’s telecommunication infrastructure and the alleged spying activity is the dominant theme in the discussion of Huawei throughout our study period, we observe a group of smartphone-related words, such as “flagship”, “pro” (Huawei’s smartphone’s premium production line), "camera", “screen” and “rom” (Read-Only-Memory), that overtakes telecommunication and spying topics as the top similar words of “Huawei”. These comments often complained about certain aspects of smartphones or criticized the company Huawei as a smartphone manufacturer.

We analyze the use of words among positive comments using the same methodology. The top similar words are mostly related to only one topic: Huawei’s consumer electronics.

Overall, we have found that among negative comments,
Huawei’s involvement in the telecommunication industry and 5G network construction have the closest connection to Huawei in the course of the discussion, during which Huawei is denounced as a Chinese spyware company and a security threat. We find that positive comments, on the other hand, focus solely on Huawei’s role as a smartphone manufacturer. This finding shows a meaningful difference in how different opinion groups characterize Huawei and provides insights into the reasons behind different sentiments toward Huawei.

B. Case Study 2: Tencent

Although in the previous study case of Huawei, we conduct and outline the result for positive and negative sentiment separately, we only include the analysis of negative sentiment for Tencent here, for the number of Tencent-related positive comments is small, accounting for less than 0.4% of the entire corpus. Table IV illustrates the top 25 most similar words to “Tencent” each month from November 1, 2019 to November 1, 2021.

There are several interesting observations here. First, “money” and “invest” are the underlying theme in the discussion of Tencent during the entire study period, as “money” occurs in every month as the top similar words, and “invest” occurs 50% of the time. By examining comments with these keywords, we have found they are mostly related to two topics. The first topic focuses on commenters’ discontent toward Tencent’s economic investment, because of Tencent’s Chinese root and Tencent’s alleged close relations with the Chinese government. The second topic concerning “money” and “investment” represents commenters’ discontent toward Tencent’s and its invested companies’ monetization behavior in the gaming industry.

Another interesting term that persists through the entire study period, occurring as the top similar words to “Tencent” 80% of the time, is “Reddit”. We have found that there exists large dissatisfaction within the Reddit community toward Tencent’s investment in Reddit itself in February 2019.

Our last observation is through monitoring the change of top similar words to “Tencent”. As we have described above, the discussion around Tencent’s economic investment and Tencent’s monetization behavior are the underlying themes in the entire discussion, but we can also observe that in certain months, the top similar words composition deviates from these underlying themes. For instance, in April 2020, the top similar words to “Tencent” are a group of words that explicitly refer to China, such as “Chinese”, “CCP”, and “government”, which do not exist in the preceding five months. We have found that a particular news event has triggered a backlash in the Reddit community: Tencent’s fully-owned video game company Riot has released a video game called Valorant, which has been found to have a built-in invasive anti-cheat system, that runs continuously even when the game is not booted. Many comments express concern and anger toward this invasive anti-cheat system, drawing links to the Riot’s parent company Tencent and accusing Tencent of deliberately planting the anti-cheat system as spyware to collect user information for the Chinese government.

C. Case Study 3: TikTok

Using the same methodology, we study the Reddit users’ use of words related to the company ByteDance. Note here although we are interested in studying the company ByteDance itself, we use the word “Tiktok”, as a majority of people refer to the company as “Tiktok” rather than “ByteDance”. We also omit the analysis of Tiktok-related positive comments, for the quantity of Tiktok-related positive comments is small, with less than 200 comments in total, to conduct any meaningful analysis. Therefore we only outline the top 25 words that are semantically similar to “Tiktok” among negative comments in Table V.

Our attention is first caught by a set of words including “hate”, “videos”, “content”, “cancer”, “platform”, “bad”, “stupid”, “shit”, etc, which frequently occur during the entire study period. Given that TikTok is a video-focused social networking platform that hosts a variety of short-form user videos,1 a connection can be drawn between Reddit users’ negative attitude and the popular videos on TikTok as well as the Tiktok user community. We find that many commenters expressed harsh criticism toward the contents of Tiktok and the whole Tiktok community, referring to both as “toxic”.

From the above, we can see that the discussion of Tiktok is closely surrounding its role as a social media platform, its user community, and the user-generated content on the platform. This is very different from the case of Huawei and Tencent, as the users’ characterization of Tiktok has less emphasis on Tiktok’s Chinese origin and its ties to the Chinese government. Another evidence of this fact is that Tiktok is discussed in a similar context with other non-Chinese social media platforms, such as “Twitter”, “Facebook”, “Reddit”, and “Instagram” rather than its Chinese counterpart, such as “Huawei” and “Tencent”. We can observe from Table V that these non-Chinese social media platforms appeared as Tiktok’s most similar words most of the time. We also find evidence of this in many comments, where users put Tiktok and these companies in juxtaposition.

While Tiktok is discussed as a social media platform and not characterized closely with China most time during our study periods, we do observe certain exceptional time periods, where Tiktok became closely related to China and the Chinese government. One such period is from July 2020 to September 2020. Not only does “ban” become the most related word with “Tiktok”, other words including “spy”, “ccp”, “china”, and “collect” occur. We believe this change is closely related to Donald Trump’s announcement in July 2020 to divest China’s ByteDance’s ownership of Tiktok, and that the US government was considering banning TikTok [1]. We have found that many Reddit users resonate with this political event. Many users show support for the Tiktok ban. The most dominant concern is unsurprisingly the same as the concerns for Tencent and

1https://en.wikipedia.org/w/index.php?title=TikTok&oldid=1059570012
TABLE IV: Top 25 most semantically similar words to “Tencent” obtained from the monthly word2vec models of negative comments from November 1, 2019 to November 1, 2021. Words related to Tencent’s economic investment are labeled blue. Words related to Tencent’s involvement in the gaming industry are labeled green. Words related to Tencent’s tie to the Chinese government are labeled red.

| 2019 | 2020 |
|------|------|
| Nov  | Dec  | Jan  | Feb  | Mar  | Apr  | May  | Jun  | Jul  | Aug  | Sep  | Oct  |
| money | money | money | epic | studios | riot | epic | game | epic | epic | game | epic | game | epic | game | epic | game |
| give | give | give | mobile | riot | money | game | money | game | money | game | money | game | money | game | money | game |
| care | reddit | care | mobile | game | mobile | trust | game | money | riot | play | money | play | money | play | money | play |
| game | spend | legal | money | trust | game | law | mobile | video | spend | spend | spend | spend | spend | spend | spend | spend |
| riot | make | pirate | game | play | give | money | mobile | video | spend | spend | spend | spend | spend | spend | spend | spend |
| Blizzard | people | make | skin | blizzard | play | game | money | game | money | game | money | game | money | game | money | game |
| make | fuck | want | game | control | fuck | give | make | play | stop | fuck | mobile | company | company | company | company | company |
| free | want | make | spend | spend | spend | spend | spend | spend | spend | spend | spend | spend | spend | spend | spend | spend |
| way | free | way | free | activation | give | pc | control | spend | care | codm | big | pay | big | pay | big | pay |
| epic | stop | people | partially | pc | free | players | ccp | shit | cheaters | shit | spend | stake | reddit | reddit | reddit | reddit |
|uck | shit | play | invest | play | invest | invest | invest | invest | invest | invest | invest | invest | invest | invest | invest | invest |
| every | stake | every | stake | every | stake | every | stake | every | stake | every | stake | every | stake | every | stake | every |
| mean | fund | reason | reason | reason | reason | reason | reason | reason | reason | reason | reason | reason | reason | reason | reason | reason |
| ggg | keep | pay | pay | pay | pay | pay | pay | pay | pay | pay | pay | pay | pay | pay | pay | pay |
| right | mobile | right | mobile | right | mobile | right | mobile | right | mobile | right | mobile | right | mobile | right | mobile | right |
| want | Blizzard | want | Blizzard | want | Blizzard | want | Blizzard | want | Blizzard | want | Blizzard | want | Blizzard | want | Blizzard | want |
| invest | pubg | invest | pubg | invest | pubg | invest | pubg | invest | pubg | invest | pubg | invest | pubg | invest | pubg | invest |
| players | investment | say | players | video | make | report | rootkit | people | people | people | people | people | people | people | people | people |
| boycott | large | try | number | try | number | try | number | try | number | try | number | try | number | try | number | try | number |

Huawei: data security issues and the alleged possibility that Tiktok may be utilized by the Chinese government to spy on its users.

VII. DISCUSSIONS AND CONCLUSIONS

In this paper, we present a fine-grained study of public opinion toward Chinese technology companies on Reddit using computational methods. We employ a state-of-the-art transformer model to classify our data into three categories: positive, negative, and irrelevant. We find that the online public generally shows more negative attitude than positive attitude toward Chinese technology companies. Next, we use LDA to model the latent topics among negative comments and positive comments, respectively. Positive comments are usually associated with the companies’ consumer products, such as smartphones and laptops. We also found that negative comments have a more diverse topic distribution. Notable topics include criticism of the platform, dissatisfaction with companies’ products, concerns over companies’ ties to the Chinese government, data security and 5G construction, and...
CONCERNING THE PUBLIC'S characterization of Tiktok, we have found that negative views are focused toward the threat of spooky activities. Positive comments, on the other hand, focus on its role as a smartphone manufacturer. The most frequent similar words to “Huawei” are the names of its smartphone products and common compliment words for smartphones.

For Tencent, we have found that negative comments are usually associated with Tencent’s economic investment, predatory monetization behavior, and its tie to the Chinese government. These themes among the negative comments have dominated the online discussion and are consistent during our study period.

Concerning the public’s characterization of Tiktok, we have found that negative views are focused toward the threat of spooky activities. Positive comments, on the other hand, focus on its role as a smartphone manufacturer. The most frequent similar words to “Huawei” are the names of its smartphone products and common compliment words for smartphones.

For Tencent, we have found that negative comments are usually associated with Tencent’s economic investment, predatory monetization behavior, and its tie to the Chinese government. These themes among the negative comments have dominated the online discussion and are consistent during our study period.

Concerning the public’s characterization of Tiktok, we have found that negative views are focused toward the threat of spooky activities. Positive comments, on the other hand, focus on its role as a smartphone manufacturer. The most frequent similar words to “Huawei” are the names of its smartphone products and common compliment words for smartphones.

For Tencent, we have found that negative comments are usually associated with Tencent’s economic investment, predatory monetization behavior, and its tie to the Chinese government. These themes among the negative comments have dominated the online discussion and are consistent during our study period.

Concerning the public’s characterization of Tiktok, we have found that negative views are focused toward the threat of spooky activities. Positive comments, on the other hand, focus on its role as a smartphone manufacturer. The most frequent similar words to “Huawei” are the names of its smartphone products and common compliment words for smartphones.

For Tencent, we have found that negative comments are usually associated with Tencent’s economic investment, predatory monetization behavior, and its tie to the Chinese government. These themes among the negative comments have dominated the online discussion and are consistent during our study period.

Concerning the public’s characterization of Tiktok, we have found that negative views are focused toward the threat of spooky activities. Positive comments, on the other hand, focus on its role as a smartphone manufacturer. The most frequent similar words to “Huawei” are the names of its smartphone products and common compliment words for smartphones.

For Tencent, we have found that negative comments are usually associated with Tencent’s economic investment, predatory monetization behavior, and its tie to the Chinese government. These themes among the negative comments have dominated the online discussion and are consistent during our study period.

Concerning the public’s characterization of Tiktok, we have found that negative views are focused toward the threat of spooky activities. Positive comments, on the other hand, focus on its role as a smartphone manufacturer. The most frequent similar words to “Huawei” are the names of its smartphone products and common compliment words for smartphones.

For Tencent, we have found that negative comments are usually associated with Tencent’s economic investment, predatory monetization behavior, and its tie to the Chinese government. These themes among the negative comments have dominated the online discussion and are consistent during our study period.
this social media platform, the platform community and the content. Furthermore, our analysis reveals that Tiktok can be distinguished by its relatively weaker association with China and the Chinese government compared to other social media platforms overall, with only certain exceptional time periods affected by political events, such as when Trump announced his plan to ban Tiktok in the United States in July 2020. Tiktok became tightly connected to China, the Chinese government, and spying allegations as its fellow Chinese technology companies in the following two months.

Overall, we have found that Reddit users hold more negative sentiment toward Chinese technology companies. We have characterized users’ concerns and discontent, and explored the reasons behind them. These findings serve to increase our understanding of the wide public opinion toward Chinese technology companies both in the political sense and commercial sense. We believe the findings of our study can be of interest to those who attempt to understand public sentiment toward Chinese technology companies and could provide useful insights into the reasons behind the public sentiment.

Our study has its limitations. For instance, although our XLNet classifier achieves acceptable results in classifying all three categories, we believe the performance of the classifier can still be improved especially in predicting negative and positive comments. In the future, we intend to address this issue by obtaining a larger and more balanced training dataset. Apart from that, another concern is the potential data collection biases that are introduced by the ranking algorithm of Reddit.\(^2\) Reddit’s algorithm (1) considers early submissions more important, and (2) favors topics that have an overwhelming opinion (i.e., The topic is less controversial where most users hold similar opinions). To verify if the sentiment distributions vary across different data collection periods, we conduct additional analysis on the sentiment values in different periods - pre-COVID (October to November 2019) and during COVID (2020 to 2021). To account for seasonality, we choose the same two months in 2020 and 2021. We find that the sentiment compositions of Reddit comments are consistent across the three periods. There are approximately 1.5 times more negative comments than positive comments. We also intend to conduct our analysis using Twitter data. Although the discussion quality is higher on Reddit, Twitter, as a more popular platform, can allow us to access broader-scale opinions toward our study subjects.

As the first work to characterize public opinion on Chinese technology companies, our study has the following broader implications. First of all, unlike the previous studies that focus on the relations between Chinese technology companies and the Western governments [4], [8], we shed light on the reasons behind the public sentiment, especially behind the public’s discontent toward these companies in a commercial sense, such as the discussions around the companies’ consumer products and economic investment. The finding of our study can provide useful information to not only these companies and other general Chinese companies that seek to expand their overseas market, to assess their companies’ image in areas outside China, but also companies that do business with Chinese companies. Second, our study also reveals a wide mistrust toward Chinese technology companies out of political and ideological concern, and identifies several dominant aspects of such mistrust, such as the data privacy concern and companies’ close ties to the Chinese government. The finding of our studies can be informative to scholars who study China’s foreign relations and China’s emerging technology and economic influence in the global sphere.

\[^{2}\text{https://github.com/reddit-archive/reddit}\]

**References**

[1] Riya Bhattacharjee, Amanda Macias, and Jordan Novet. Trump says he will ban tiktok through an executive action. CNBC, 2020.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. the Journal of machine Learning research, 3:993–1022, 2003.

[3] Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. Exploring misogyny across the manosphere in reddit. In Proceedings of the 10th ACM Conference on Web Science, WebSci ’19, page 87–96, New York, NY, USA, 2019. Association for Computing Machinery.

[4] Jesse Friedlander. How the arrest of huawei’s meng wanzhou may mark the start of the great china-us technology decoupling. 2018.

[5] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media, volume 8, pages 216–225, 2014.

[6] Don Lee. Trump bans huawei in u.s. markets, saying chinese firm poses security threat. Los Angeles Times, 2019.

[7] Hanjia Lyu, Junda Wang, Wei Wu, Viet Duong, Xiyang Zhang, Timothy D Dye, and Jiebo Luo. Social media study of public opinions on potential covid-19 vaccines: informing dissent, disparities, and disinfection. Intelligent medicine, (2021):1–12, 2022.

[8] Bruno Mascalitelli and Mona Chung. Hue and cry over huawei: Cold war tensions, security threats or anti-competitive behaviour? Research in Globalization, 1:100002, 2019.

[9] Mary L McHugh. Interrater reliability: the kappa statistic. Biochemia medica, 22(3):276–282, 2012.

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119, 2013.

[11] Olivier Noyan. Eu countries keep different approaches to huawei on 5g rollout. EURACTIV, 2021.

[12] Judy Hanwen Shen and Frank Rudzicz. Detecting anxiety through Reddit. In Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality, pages 58–65, Vancouver, BC, Aug. 2017. Association for Computational Linguistics.

[13] Ahmed Soliman, Jan Hafer, and Florian Lemmerich. A characterization of political communities on reddit. In Proceedings of the 30th ACM Conference on Hypertext and Social Media, HT ’19, page 259–263, New York, NY, USA, 2019. Association for Computing Machinery.

[14] Fatemeh Tahmashi, Leonard Schild, Chen Ling, Jeremy Blackburn, Gianluca Strimplini, Yang Zhang, and Savvas Zannettou. “go eat a bat, chung!”: On the emergence of sinophobic behavior on web communities in the face of covid-19. In Proceedings of the Web Conference 2021, WWW ’21, page 1122–1133, New York, NY, USA, 2021. Association for Computing Machinery.

[15] Wei Wu, Hanjia Lyu, and Jiebo Luo. Characterizing discourse about covid-19 vaccines: A reddit version of the pandemic story. Health Data Science, 2021:9837856, Aug 2021.

[16] Zhulin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019.