

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

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

University of Rochester

{ezhou12,yliu217}@u.rochester.edu, hlyu5@ur.rochester.edu, jluo@cs.rochester.edu

* these authors contributed equally

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 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.

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.^{1 2}

Previous studies investigated governments' reactions to Chinese technology companies and sought to explain the governments' mistrust toward these companies. Friedlander (2018) argued that the distrust and restriction imposed on Chinese technology companies stem from the fact that technology has increasingly strategic importance in today's world, especially in the realms of national defense, commerce, and social organization. Mascitelli and Chung (2019) found that the mistrust toward Chinese technology firms is not only ideological but also commercial. The deliberate mistrust is built in efforts to suppress China's growing economic presence in the global market with technology firms, such as Huawei, as the face of such rising economic power.

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. This topic is itself very sensitive to daily real-world events. We choose Reddit as our source of data. It is one of the largest news-oriented social media platforms, serving as a good data source for capturing and analyzing public opinion on such a news-sensitive topic. Reddit has 530 million monthly active users, and a diverse user base across the world, providing us with a copious amount of data.³ Reddit also has a more flexible comment length constraint - up to 40,000 characters per post compared to merely 280 characters on Twitter. As a result, Reddit stands out as a superior platform for retrieving high-quality conversational data and discussions. Another special advantage of Reddit is that unlike Twitter or other major social media platforms, Reddit does not require users to fill in their personal information, therefore creating an anonymous environment on Reddit, which is crucial in their effort to safeguard users' freedom of speech (Centivany and Glushko 2016). Reddit has been adopted by many researchers as a reliable resource for

¹<https://www.euractiv.com/section/digital/news/eu-countries-keep-different-approaches-to-huawei-on-5g-rollout/> [Accessed Jan. 8, 2022]

²<https://www.latimes.com/business/la-na-pol-trump-huawei-ban-20190515-story.html> [Accessed Jan. 8, 2022]

³<https://backlinko.com/reddit-users> [Accessed Jan. 8, 2022]

large-scale, relatively high-quality posts on a topic of broad interest (e.g., COVID vaccine, rape, e-cigarette) (Wu, Lyu, and Luo 2021; Kamarudin et al. 2018; Chen et al. 2020). In this paper, we collect data from Reddit to examine public sentiment in the discussions of Chinese technology companies. 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 (Friedlander 2018; Mascitelli and Chung 2019), 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.

Related Work

Traditional methods of analyzing public opinion rely on soliciting responses to a poll, which have several disadvantages. They are expensive and time-consuming. Their quality may easily suffer from poor sampling and a low response rate (Cody et al. 2016). With growing influence, Social media platforms have become an alternate source for researchers to obtain massive text-based datasets. They provide easy access to public opinion in real time (Glenski and Weninger 2017), and have enabled researchers to shed new light on human behavior and opinion related to major issues of global importance (Cody et al. 2016). There has been a significant amount of research in public opinion that analyzed social media data. For instance, Cody et al. (2015)

used Twitter data to study public sentiment toward the climate change problem. They have uncovered that on Twitter, the responses to climate change news are predominately from climate change activists rather than climate change deniers, indicating that Twitter is a valuable resource for the spread of climate change awareness. More recently, Lyu et al. (2022) focused on public opinion on vaccines in the face of the COVID-19 pandemic. Empowered by the rich information provided on social media platforms, they found correlations between user demographics and their attitudes toward vaccine uptake.

Many studies have been conducted to monitor and analyze the opinion on Reddit. Some studies collected posts under a small number of selected related subreddits. For instance, Shen and Rudzicz (2017) 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. (2019) 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 (2019) 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 (2021) 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. (2022) applied the state-of-the-art transformer model - XLNet to mining opinions. Tahmasbi et al. (2021) 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.

Datasets

A common problem faced by all approaches that attempt to measure public opinion is the problem of selecting appropriate samples that will generalize to the public population. Inappropriate selection of the sample will expose the study to self-selection bias that limits the generalizability outside the study sample. Social media study might be more prone

to this problem, as people’s motivation to participate in social media conversations regarding a particular topic may be diverse and hard to be distinguished as opposed to conventional survey research, which is able to assert study subjects’ motivation directly (Chen and Tomblin 2021). Furthermore, Reddit itself underwent significant political polarization since the 2016 US election, resulting not from increasing individual polarization but from system-level shifts driven by the arrival of new users (Vitak, Shilton, and Ashktorab 2016). Therefore, while we acknowledge that the study population of this study may have polarization issues and inherent bias, we think likewise as Chen and Tomblin (2021) that this built-in bias in the study population will facilitate us to identify and better understand issues that may be previously unknown, unintentionally ignored or marginalized in conventional studies.

Data Collection

Reddit consists of individual micro-communities that are based on certain topics (i.e., *r/worldnews*, *r/technology*), which are referred to as subreddits. Under each subreddit, users can make posts, known as submissions in Reddit. Users can also make comments under submissions. To obtain the data for our study, we collect publicly available content from Reddit using a list of keywords, which contain the names of Chinese technology companies and the names of chief executive officers (CEOs). Note that we choose not to collect submissions from Reddit, because compared to comments, most submissions are opinion-neutral content, such as news and questions, and usually do not indicate users’ opinions toward our study subjects. Therefore, we choose *users’ comments* as our data source.

To collect users’ comments, we employ a Reddit API Wrapper called PRAW.⁴ 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.

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)⁵ to remove all stop words from the text. Additionally, we perform a text lemmatization using NLTK.

Text Sentiment Classification

Method

XLNet (Yang et al. 2019) is a generalized autoregressive pretraining method that can capture left and right contexts

⁴<https://praw.readthedocs.io/en/latest/> [Accessed Jan. 8, 2022]

⁵<https://www.nltk.org/> [Accessed Jan. 8, 2022]

jointly in sentences. It was claimed that XLNet outperforms BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2019), a pretrained language model with transformer architecture that is designed to perform downstream NLP tasks after fine-tuned, on 20 tasks such as question answering, natural language inference, sentiment analysis, and document ranking. Previous studies have used XLNet models fine-tuned for emotion classification, sentiment analysis (Alshahrani et al. 2020), and classification of censored tweets (Ahmed and Kumar M. 2021), *etc.* 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 (Hutto and Gilbert 2014), a lexicon and rule-based sentiment analysis tool which is specifically designed for social media posts.

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 1. 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. (2022) 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 1. 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 Fleiss’ Kappa score of 0.77, which indicates a good agreement among the annotated labels.

We train an XLNet model H using these 1,600 labeled comments as the initial training corpus ($n = 1600$). 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 ($n = 2600$). 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. The distributions of three categories of the initial and final training corpus are displayed in Fig. 1. 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.

Table 1: 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, <i>etc.</i>)
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, <i>etc.</i>) 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/Irrelevant	i. News regarding Chinese tech companies with no written opinion from the commenters ii. Including Chinese tech companies and the commenters' opinions, but the focus is something else (i.e., politics, economics, <i>etc.</i>) iii. Comments with <i>neutral</i> opinion toward Chinese technology companies iv. Comments/questions on Chinese tech companies entities but with unclear meanings

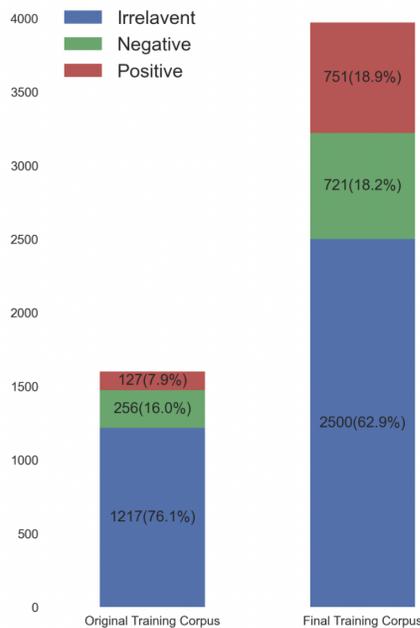


Figure 1: Distributions of three categories of the initial training corpus and the final training corpus.

Evaluation

Table 2 summarizes our final model’s performance. We use the “weighted” F1-score to evaluate the overall performance. 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, which may be because (1) the pre-built sentiment analysis tool cannot handle the complex semantic meanings, and (2) sentiment is not opinion (i.e., a comment can contain many negative words but express positive opinion toward the Chinese technology companies). This further supports the necessity of our human-guided machine learning framework.

Table 2: Performance of the XLNet model and VADER.

Model	Class	Precision	Recall	F1-score
VADER	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	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

Based on the classification results by our fine-tuned XLNet model on the entire dataset, we plot a side-by-side bar chart (Fig. 2) to compare the difference between the number of positive comments and the number of negative comments in the overall discussion and specific discussion on the top four most mentioned technology companies. 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. One interesting observation can be made from the company-wise comparison: Two companies that have the most positive comments, Xiaomi and Huawei, are all major producers of consumer electronics such as smartphones and laptops, while the two companies that have predominantly negative comments, Tencent and Tiktok, are companies that offer online services, such as social media platforms and video games.

Table 3: Eight topics of negative Reddits generated by LDA and their associated example comments.

ID	Topics	% of comments	Example Comment
1	Social/Media Platform	23.01%	"i am tired of tiktok with their insensitive trends. the community is terrible, i see a ton of racist, homophobic and abliest people on that app. i dont mind tiktok but i hate the community, it reminds me of school becoss of how similar they are."
2	Smartphone Products	18.86%	"I would NEVER ever buy huawei now. I have Mate series phone, although I get upgrades still. One of the upgrade has taken by ability to change my launcher from default to anything I want..."
3	Gaming Industry	17.40%	"Tencent really only cares about the skins and them making money and trying to make people spend money with this blackpink whatever this is all they really need to do is start banning this cheaters the devs need to know if the game has like no hackers to a little people would actually consider spending money in the game."
4	Politics/5G construction	12.31%	"China will just steal the ip of those exports once they can and produce it themselves huawei is based off nortel a canadian company that fell because of industrial espionage they didnt even change the stolen source code. China forcing nations on 5g just goes to prove it will be used for spying"
5	Ties to Chinese Government	11.79%	"Chinese propaganda goes a long way. all this acting to make us believe all is well and people are happy instead of showing the massacre of Uighurs. xi jinping and tencent can go to hell..."
6	Data Privacy/Security	8.78%	"Every major company in China is required to have a party board member and every Chinese company by law in China has to share it's data with the government. Also Tencent has significant ties to CCP including having built part of the Chinese government state surveillance platform."
7	Online Shopping	7.30%	"Me too. I have used Alibaba and wish, all about the same wait times unless you pay gobs. The shitty part is that the refund process can take way longer."
8	Slur/Curse	0.50%	"Yup. Absolutely. F**k TikTok."

Topic Modeling

LDA

Latent Dirichlet allocation (LDA) by Blei, Ng, and Jordan (2003) is a topic modeling method that has been used extensively in studies on social networks and microblogging environments (Jelodar et al. 2019). Albalawi, Yeap, and Benyoucef (2020) found that LDA can achieve better performance in a short text context compared to other topic modeling methods, such as Latent Semantic Analysis (LSA) (Landauer, Foltz, and Laham 1998), Principal Component Analysis (PCA) (Pearson 1901), and random projection (RP) (Bingham and Mannila 2001). It also produces higher quality topics and more coherent topics than other topic modeling methods. Therefore, we choose LDA as our method for topic modeling.

To characterize different topics associated with positive and negative comments, we apply LDA separately for neg-

ative 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.

Result

Table 3 itemizes the eight topics that are associated with negative sentiment and their example comments. Fig. 3 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

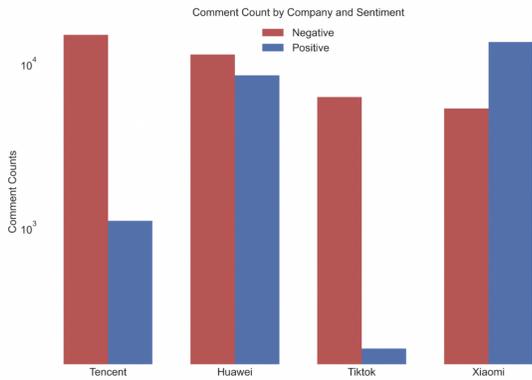


Figure 2: Comment counts in logarithmic scale by sentiment.

model. We also categorize each comment to its dominant topic.

Topic 1 indicates that the most frequent topic, accounting for 23% of the overall negative comments, is criticism toward social media platforms. The most heavily criticized companies are TikTok and Tencent which account for up to 80% of the total comments on this topic. Many commenters express sharp negative sentiment toward these two companies' communities and online platforms with swear words such as "hate", "stupid", "suck", *etc.* While TikTok's platform content is frequently negatively commented on as stupid and cringe, Tencent is blamed for its poor control of cheaters in its gaming communities where the anti-cheat system is not efficient enough. Another interesting observation is that "Reddit" appears as a keyword in the topic. We find that the Reddit community itself is commonly mentioned in this topic, and the relationships between Reddit and these two companies are very different. Since Tencent has invested in Reddit, certain Reddit users argue that the involvement of Tencent has introduced low-quality content and heavy censorship on the Reddit platform. On the other hand, Reddit users compare their platform with TikTok and imply that Reddit is a superior social media platform.

Topic 2 illustrates the second most frequent topic - criticism toward smartphone products of certain companies such as Huawei and Xiaomi. The third most frequent topic corresponds to negative sentiment toward the overarching influence of Chinese technology companies, Tencent in particular, on the gaming industry. A prevalent criticism on this front is that Tencent's monopoly and significant profits from investing in major gaming companies have resulted in little to no positive impact on the gaming industry, and in some cases, even a negative impact.

The next three topics focus on the negative sentiment toward the "Chinese" part of the companies. The fourth most frequent topic is the discussion about politics between China and foreign governments, the trade war with the US, and some recent events, such as Meng Wanzhou's arrest in Canada.⁶ We can observe how these political events' influ-

ence may help shape the public's view on Huawei from some example comments. For instance, a few comments express mistrust toward Huawei in response to two Canadian citizens being detained by the Chinese government. Another important subtopic is Chinese technology companies' involvement in 5G construction. Huawei, as the leading company of the world's telecommunication providers, becomes the target of these criticisms.

The fifth most frequent topic is the link between these companies to the Chinese government. Many representative comments express negative sentiments toward these companies because of their alleged close relationship with the Chinese government. Some comments hold these companies responsible for assisting the Chinese government's alleged "mass surveillance" or "genocide of the Uighurs". Topic 6 talks about how the comments perceived these companies as a risk or danger to users' data privacy. Many associated comments refer to these companies as the "state surveillance platform", "national threat concerns" or "spyware created by the CCP (Chinese Communist Party)".

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. 4 shows the distributions of different topics by the top 10 subreddits that have the most negative comments. 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. 5 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%. After further investigation, we find that the comments on both topics are mainly associated with Huawei and Xiaomi,

⁶<https://www.cnn.com/2018/12/05/tech/huawei-cfo-arrested->

<canada/index.html> [Accessed Jan. 12, 2022]



Figure 3: Word cloud of the negative comment topics.

the two consumer electronic manufacturers in the top four most mentioned Chinese technology companies. This also explains the large number of positive comments for Huawei and Xiaomi displayed in Fig. 2.

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 (Mikolov et al. 2013). Specifically, we train a set of word2vec models on specific groups of comments, which reflect the most closely associated terms to each company during each month of our study period. 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.

Case Study 1: Huawei

First, we look into the overall use of words in negative comments about Huawei. Table 4 illustrates the top 25 most similar words to “Huawei” each month from November 1, 2019 to November 1, 2021. Table 5 shows the top 20 most frequent words in Table 4. 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 words indicates that these words have the most similar context to “Huawei” the majority of the time. This leads us to suspect the discussion of Huawei in negative comments may be closely knitted with Huawei’s espionage controversy over the allegations of using its telecommunication device to spy on its users for the Chinese government.⁷ After reviewing comments with these terms as keywords, we find evidence of this suspicion. Many commenters expressed their belief that Huawei is culpable in the spying allegations:

- “Duh, Huawei IS spying on users, its not rumors, its fact, they have confirmed it, they have whole departments connected to the Chinese military. All Chinese companies are directly or indirectly connected to the Chinese government, telecom ones are of huge interest”
- “aNd tHe cHiNeSe hUaWeI aRe sPyInG oN uS !!!”

We also observe the word “ban” sometimes co-occurs with the three categories above in several months. In examining closer, we discovered that many comments also echo many western governments’ ban on Huawei.⁸ These comments often characterize Huawei as “national threat”, “security risk”, and “Chinese spyware”, or believed that Huawei has planted “backdoors” in their devices. These negative characterizations of Huawei are reflected as the top most similar words of “Huawei” in Table 4. These comments often showed support for a specific country’s existing ban on

⁷<https://www.cfr.org/backgrounder/huawei-chinas-controversial-tech-giant> [Accessed Jan. 8, 2021]

⁸<https://www.cnet.com/tech/services-and-software/huawei-ban-timeline-detained-cfo-makes-deal-with-us-justice-department/> [Accessed Jan. 8, 2021]

Table 4: 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.

2019		2020									
Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
5g	5g	network	network	google	build	5g	5g	5g	5g	issue	apple
use	network	5g	5g	software	5g	us	us	network	india	oneplus	use
security	build	equipment	spy	apple	samsung	security	security	countries	allow	lg	google
network	tech	national	us	devices	use	network	google	security	build	pro	android
products	brand	backdoors	infrastructure	service	network	google	countries	spy	risk	flagship	oneplus
equipment	samsung	concern	uk	products	google	equipment	issue	us	backdoors	camera	device
tech	germany	us	tech	network	oneplus	devices	apple	national	military	forget	equipment
find	oneplus	allow	equipment	use	apple	apple	tech	uk	usa	mi	samsung
trust	find	ban	military	support	pro	work	apps	equipment	use	screen	5g
apple	apple	use	security	brand	mi	hardware	devices	infrastructure	countries	5g	great
mate	devices	country	intelligence	5g	brand	years	equipment	technology	western	devices	apps
steal	products	spy	state	security	redmi	canada	spy	send	issue	rom	hardware
risk	ban	countries	ban	pixel	note	samsung	canada	tech	threaten	iphone	service
national	equipment	canada	countries	oneplus	buy	use	allow	threat	hardware	worse	camera
devices	security	devices	government	technology	devices	build	backdoors	apps	technology	problem	screen
build	canada	security	foreign	build	equipment	spy	steal	ban	sanction	oppo	instal
canada	technology	spyware	national	screen	release	trump	sanction	steal	cisco	great	pixel
access	evidence	tech	technology	tech	service	chip	pixel	risk	trade	range	price
usa	steal	threat	issue	hardware	first	uk	build	arrest	threat	terrible	infrastructure
root	redmi	surveillance	world	cheap	stock	brand	infrastructure	users	country	device	compare
infrastructure	uk	evidence	allow	intelligence	infrastructure	mi	work	user	network	p20	miui
concern	lenovo	infrastructure	threat	allow	oppo	oneplus	national	concern	international	hardware	battery
google	issue	risk	trade	years	device	device	p40	collect	political	moto	os
us	least	build	telecom	avoid	issue	countries	uk	india	destroy	chip	pro
iphone	ericsson	trump	concern	equipment	uk	arrest	citizens	sanction	service	pixel	annoy
2020		2021									
Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
network	5g	5g	plus	device	australia	zte	countries	use	security	devices	ban
equipment	network	hardware	steam	massive	telecom	os	5g	work	tablet	use	5g
google	spy	block	os	back	install	devices	equipment	apps	normal	apple	security
build	technology	network	day	support	network	hardware	security	case	disappoint	iphone	us
canada	tech	devices	code	phone	home	pixel	example	smartphone	crash	service	products
tech	equipment	uk	replace	devices	eavesdrop	iphone	major	avoid	average	pixel	big
sony	intelligence	device	poor	usa	lock	eu	wifi	honor	anymore	camera	equipment
infrastructure	backdoor	samsung	intellectual	important	apps	work	network	sanction	hardware	mi	use
spy	agencies	apple	5g	info	infrastructure	ios	illegal	p30	past	security	network
security	service	google	life	total	connect	lite	allow	equipment	iphone	great	steal
interest	national	watch	assume	america	fraud	tower	mate	switch	days	poco	place
os	devices	europa	work	concern	compare	fingerprint	military	5g	5g	device	google
zte	request	consider	past	news	warranty	screen	connect	glad	disable	oneplus	tech
currently	find	apps	honest	extreme	tv	software	build	countries	instal	brand	national
whether	chip	pro	screen	supply	none	mate	trump	nortel	cloud	suck	chip
5g	chinas	lead	billion	network	attempt	5g	hack	wrong	provide	oppo	data
create	ask	miui	uk	year	5g	try	technology	zte	mate	hardware	build
recommend	claim	wifi	source	canada	apple	p30	australia	military	ads	battery	trust
military	countries	brand	nearly	days	wait	drop	old	non	p30	apps	spy
vivo	back	issue	rip	dollars	devices	push	future	either	totally	end	business
sanction	us	new	lite	remember	story	issue	foreign	personally	components	os	canada
higher	access	iphone	pick	terrible	different	network	bite	support	laptop	experience	product
available	western	budget	question	americans	service	chip	near	due	thank	old	stop
use	americans	ill	face	invest	block	backdoors	mine	apple	wife	products	trade
nortel	infrastructure	os	non	domestic	examples	gb	name	brand	makers	better	ago

the top 20 most frequent top similar words to “Huawei” from November 1, 2019 to November 1, 2021.

Table 6: Top 20 most frequent semantically similar words to “Huawei” among positive comments. “Count” refers to the number of the corresponding word’s occurrence in the monthly top 25 most semantically similar words to “Huawei”.

word	count	word	count
p20	10	p40	7
mate	8	kind	6
series	7	compare	6
model	6	options	5
service	5	change	5
lite	5	problem	5
least	5	believe	5
move	5	nova	5
love	5	cameras	5
buy	5	plus	5

Consistent with our finding in the previous topic modeling section, positive comments are mostly associated with consumer electronics, such as smartphones and laptops. As shown in Table 6, the most frequent similar words to “Huawei” are either the names of their smartphone models, such as “p20” (Huawei’s 2018 flagship model), “mate” (Huawei’s high-end smartphone series), and “lite” (economic version of the flagship models) or the commonly used words for compliments, such as “love” and “buy”. The comments usually applaud certain aspects of a particular model, and possibly are being posted to recommend this particular model and phone brand to other people:

- *“Huawei mate 20 x the best choice, it’s very amazing phone, it combines the fantastic design, the very large battery and the best performance. The rest phones are good, but Huawei is the better than them.”*

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.

Case Study 2: Tencent

To study the temporal pattern of public sentiment toward Tencent, we train one word2vec model for comments on Tencent of each month and outline the most similar words to “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 7 illustrates the top 25 most similar words to “Tencent” each month from November 1, 2019 to November 1, 2021. Table 8 shows the top 20 most frequent words in Table 7.

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. A portion of the commenters advocate for an economic boycott of Tencent and the companies Tencent has invested in:

- *“What I can do is stay up to date with events here. Delete all blizzard games. Stop buying new blizzard games. stop playing games that take in Chinese money. Stop watching movies backed by Tencent and other Chinese companies that invest heavily into American entertainment. And continue to support the Hong Kong people here. That’s all I can do from the confines of my home. If others are able to go to Hong Kong and help, then they should absolutely do so. They need all the help they can get.”*
- *“Well tencent is one of the production companies pumping dirty money into this movie. Maybe china could foot the bill on those tickets instead. YOU WANT TO MAKE A BETTER GLOBAL NEIGHBORHOOD? DONT PAY A CENT TO TENCENT.”*

The second topic concerning “money” and “investment” represents commenters’ discontent toward Tencent’s and its invested companies’ monetization behavior in the gaming industry. A group of game-related terms, such as “game”, “riot” (a subsidiary video game company of Tencent), “epic” (a video game company of which 40% shares are owned by Tencent), and “ggg” (a video game company acquired by Tencent), persist in the top most similar words to “Tencent” during the entire study period, and often co-occur with “money” and “investment”. After examining comments with these terms, we have found that commenters usually show discontent toward Tencent and its invested companies’ predatory monetization schemes and disregard for users’ gameplay experience. The following example comment showcases a typical comment on this topic:

- *“Epic probably wouldn’t get such a bad rap for it though, if it wasn’t for the fact that most of their monetization systems are predatory and their company is held about 50% by Tencent, a company also known for predatory monetization (they also designed a great deal of Epic’s monetization systems today) and close ties with the Chinese government.”*

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.

Table 7: 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	epic	game
give	give	mobile	riot	money	game	riot	epic	game	money	fuck	riot
care	reddit	play	game	mobile	trust	game	riot	riot	play	play	money
game	spend	epic	money	timi	company	reddit	money	money	play	money	ggg
riot	make	game	play	pubg	give	money	reddit	mobile	game	game	epic
blizzard	people	skin	blizzard	play	money	give	play	stop	fuck	mobile	company
make	fuck	want	steam	game	control	fuck	make	play	reddit	riot	league
free	want	make	spend	spend	reddit	pc	fuck	fuck	make	reddit	players
way	free	free	activision	give	pc	control	spend	care	codm	big	pay
epic	stop	riot	mobile	cheat	spend	company	cheat	pubg	activision	hate	make
people	partially	pc	free	players	ccp	shit	cheaters	shit	spend	stake	reddit
fuck	shit	spend	devs	care	valorant	share	people	pay	mobile	give	spend
play	invest	players	care	people	fuck	chinese	hate	free	shit	ggg	free
reddit	care	fuck	players	guess	government	invest	activision	spend	give	make	stop
every	stake	give	store	create	american	anti	mobile	give	trust	share	play
mean	fund	reddit	greedy	hackers	access	ByteDance	players	pubgm	pubg	spend	cheaters
ggg	reason	community	know	doubt	play	mobile	pubg	support	pubgm	pubg	care
keep	pay	pubg	pc	scam	chinese	ccp	free	players	right	cheaters	skin
stop	amount	industry	shit	purpose	chinese	free	invest	buy	players	hackers	hand
right	mobile	activision	profit	skin	epic	care	give	cheaters	part	invest	majority
want	blizzard	way	take	epic	support	cheat	sell	make	league	sell	ruin
invest	pubg	hackers	cheaters	cod	computer	access	part	nothing	skin	cheat	sell
players	investment	invest	try	fuck	data	run	stake	hackers	free	activision	everything
say	players	video	make	stop	blizzard	big	blizzard	reddit	pay	steam	china
boycott	large	try	number	report	rootkit	people	hand	loud	steam	players	blame
2020		2021									
Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
game	epic	reddit	care	give	ccp	riot	epic	riot	company	ccp	money
riot	game	riot	chinese	epic	riot	company	reddit	game	game	epic	fuck
play	reddit	post	reddit	large	game	epic	ccp	fuck	chinese	game	play
reddit	riot	game	fuck	company	company	ccp	people	money	epic	company	want
money	money	fuck	ccp	money	reddit	game	china	epic	china	money	game
ea	play	activision	become	stake	epic	fuck	company	people	ccp	reddit	people
fund	mihoyo	epic	epic	force	control	chinese	riot	make	government	riot	care
mobile	fuck	part	know	product	make	money	control	company	money	hand	keep
corporation	hate	stake	people	share	money	mean	game	ccp	riot	china	give
league	invest	league	game	hard	fuck	reddit	tie	play	control	control	everything
devs	give	everything	china	long	china	china	ByteDance	china	business	video	epic
tie	care	small	arm	riot	people	control	invest	chinese	party	tie	way
cheat	blizzard	take	fair	play	involve	government	give	hand	us	people	know
fuck	steam	believe	developers	local	invest	give	money	give	right	mihoyo	pay
greedy	activision	money	bad	fund	stake	know	fund	say	play	ByteDance	try
players	see	evil	league	game	ggg	take	influence	reddit	state	give	bad
fortnite	part	target	blow	crappy	allow	people	stake	league	communist	anti	go
ccp	people	right	free	invest	large	league	reason	big	mean	world	sell
market	ea	industry	hand	top	care	invest	free	know	media	take	every
partially	shit	leave	money	decent	entire	care	hand	control	people	invest	back
give	make	alibaba	million	make	pc	big	profit	stop	reddit	big	free
right	stop	community	see	old	market	make	investment	fund	everything	influence	stock
involve	china	lol	censorship	group	mobile	hand	corporations	mobile	tech	profit	time
milk	long	greedy	everyone	biggest	give	hate	share	every	fuck	include	market
blizzard	everything	ggg	give	consider	profit	market	data	data	part	hold	shit

Table 8: Top 20 most frequent semantically similar words to “Tencent” among negative comments. “Count” refers to the number of the corresponding word’s occurrence in the monthly top 25 most semantically similar words to “Tencent”.

word	count	word	count
money	24	invest	12
game	23	free	12
epic	22	care	12
riot	20	mobile	12
give	20	players	11
reddit	20	company	10
fuck	19	spend	10
play	16	ccp	10
make	14	china	9
people	14	control	8

Many commenters show concerns and discontent about potential censorship, data security issues, and the emergence of communist propaganda on Reddit with Tencent. They express a negative attitude toward Tencent being a shareholder of Reddit because, in their opinion, Tencent is a subordinate company of the Chinese government. The following example comments showcase the typical comments regarding this issue:

- *“Reddit is owned by the CCP firm Tencent. China didn’t take any vaccines. This is a Propaganda containment site.”*
- *“China partially own reddit through tencent, there’s a squadron of power users who just mindlessly defend the CCP and Chinese interests across the site.”*

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:

- *“... I wasn’t planning to install Valorant or any other Tencent/CCP China spyware, but good to have confirmation that’s the right decision.”*
- *“Giving people freedom? That’s not how Tencent is used to running games because that’s not how Chinese are*

used to running a society. Remember you’re working with an entity ONE HUNDRED PERCENT owned by Tencent, a Chinese state-run mass media corporation.”

Overall, these findings provide a detailed understanding of how and why the Reddit community expresses negative sentiments toward Tencent. The discontent toward Tencent’s economic investment, its predatory monetization behavior, and its tie to the Chinese government often intertwine to a degree that the commenters rarely express their negative view on Tencent in one single aspect. These findings also show that the reasons behind users’ negative views are very consistent throughout our study period. It is because that Tencent’s economic investment, its predatory monetization behavior, and its tie to the Chinese government are intrinsic to the company itself, which are Tencent’s business model and its Chinese origin, and thus unaffected by local events. Therefore, we anticipate the same pattern of public opinion toward Tencent will persist in the near future because Tencent most likely will not change its business model or cut ties with the Chinese government.

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 9 and summarize the top 20 most frequently occurred terms in Table 10.

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,⁹ 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”. For instance, some comments stated:

- *“Wtf is wrong with people on tiktok. I’ve seen some of the most toxic shit on that app”*
- *“Can’t change the Fact that TikTok is Shit and just stupid”.*

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

⁹<https://en.wikipedia.org/w/index.php?title=TikTok&oldid=1059570012> [Accessed Dec. 23, 2021]

Table 9: Top 25 most semantically similar words to “TikTok” obtained from the monthly word2vec models of negative comments from November 1, 2019 to November 1, 2021. Words related to Tiktok’s social media platform, user community, and user content are labeled blue. Words related to China, the Chinese government, and the spying controversy are labeled red.

2019		2020									
Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
Bytedance	see	people	hate	hate	hate	hate	hate	apps	ban	ban	hate
call	post	memes	people	content	post	videos	people	app	american	app	instagram
result	censorship	post	fuck	reddit	bad	people	platform	information	user	tiktoks	reason
share	dumb	funny	reason	post	videos	bad	users	collect	collect	say	social
purchase	hand	content	app	people	fake	content	app	facebook	spy	fact	bc
financial	propaganda	shit	reddit	hat	see	kid	instagram	spyware	information	wechat	facebook
interest	need	videos	bad	videos	stuff	post	content	access	facebook	nation	friends
platform	citizens	cringe	cringe	cringe	please	shit	hat	users	steal	facebook	stupid
content	power	stupid	fortnite	see	app	platform	dance	spy	us	children	toxic
foreign	hate	kid	use	sub	content	instagram	opinion	american	Bytedance	microsoft	attention
mainland	videos	cancer	video	kid	cringe	cringe	info	steal	access	cancel	watch
gain	content	fortnite	shit	cancel	community	dumb	cancel	ban	china	people	say
delete	foreign	popular	awful	social	sub	funny	cringe	beijing	ccp	laws	twitter
regime	share	popular	kid	fuck	shit	stupid	post	tech	citizens	users	dislike
million	full	fair	platform	popular	love	app	stuff	data	microsoft	user	wrong
harvest	essentially	guy	mean	instagram	everyone	mean	censorship	law	users	personal	music
via	everyone	platform	post	censorship	annoy	fortnite	agree	share	force	deal	general
supporters	memes	enjoy	deserve	toxic	instagram	censorship	overlords	us	tech	majority	type
conglomerate	heavily	sub	everything	delete	comment	cancel	machine	ccp	law	applause	ass
youtube	connect	instagram	good	bad	redditors	love	bad	user	trump	funny	hurt
fuck	ly	annoy	simply	app	show	facebook	view	private	laws	clear	girls
cancel	collect	video	ruin	Bytedance	things	concentration	children	send	share	ask	suck
mass	political	close	social	cancel	half	community	kid	social	wechat	threat	although
obviously	site	community	videos	spread	hell	dance	promote	privacy	need	share	sad
collect	send	easy	funny	evil	platform	delete	see	technology	private	involve	child
2020		2021									
Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
shit	facebook	hat	news	months	legal	platform	collect	dumb	fuck	public	invest
post	twitter	process	past	full	guy	hate	harvest	breach	video	disappear	private
site	hat	gamble	kind	group	threat	dislike	intelligence	video	people	platforms	servers
people	post	content	add	term	grow	social	internet	stupid	reddit	hate	facebook
ea	stupid	emote	include	film	yu	cringe	matter	face	take	toxic	american
guy	people	learn	show	massive	friends	laws	govt	short	everything	attention	parent
real	bad	mind	kid	build	history	fee	foreign	exactly	call	propaganda	operate
videos	kind	videos	post	million	chance	videos	military	facebook	pretty	internet	act
hate	platforms	behind	videos	local	abuse	asian	access	wrong	try	political	directly
say	ruin	awful	ui	lead	send	wrong	allow	anyway	make	believe	foreign
person	comment	improve	criticize	hold	political	whenever	large	law	infrastructure	dude	reddit
fuck	seriously	corrupt	sub	anti	suppose	owners	people	depend	influence	monitor	listen
platform	instagram	proof	Bytedance	kind	cringe	song	private	board	hard	sense	clearly
hear	hate	completely	start	hell	entire	god	tie	ones	western	limit	fund
annoy	tik	argue	assume	better	collect	ask	corporations	corporation	start	privilege	piss
codm	account	credit	dead	content	complain	important	citizens	censorship	post	deserve	businesses
million	whole	happy	behind	video	counter	fine	investment	subject	interest	western	include
pretend	incredibly	sake	edit	different	damn	violations	heavily	authoritarian	jack	negative	position
eye	hell	matter	sad	fine	clear	host	request	content	everyone	pressure	prove
face	team	sit	rip	wrong	follow	conspiracy	understand	harvest	stop	comment	news
history	trend	entire	day	pubg	track	society	talk	include	trade	kid	help
subreddit	communism	message	produce	riot	server	media	completely	awful	little	reduction	infiltrate
fortnite	whatever	everyone	lie	deny	collection	article	tell	hide	things	damage	privacy
holy	white	later	guy	videos	prove	reality	american	address	real	call	outside
comment	studio	minority	person	users	tik	regardless	profit	hat	human	twitter	forget

Table 10: Top 20 most frequent semantically similar words to “TikTok” among negative comments. “Count” refers to the number of the corresponding word’s occurrence in the monthly top 25 most semantically similar words to “TikTok”.

word	count	word	count
videos	11	instagram	7
hate	11	collect	6
post	11	bad	6
people	10	share	5
content	10	cancel	5
cringe	8	users	5
platform	8	social	5
facebook	8	stupid	5
app	7	shit	5
kid	7	cancel	5

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 9 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:

- *“Twitter is great, Reddit too of course. IG is ok but TikTok is mostly cringe”*
- *“*[sigh]* I hate Twitter. And Facebook. And Instagram. And Tumblr. And TikTok. Yes, even Reddit sometimes.”*

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.¹⁰ We have found that many Reddit users resonate with this political event. Many users show support for the TikTok ban. For instance, one commenter notes: *“TikTok is cancer and Trump is right to ban it, change my mind.”* The most dominant concern is unsurprisingly the same as the concerns for Tencent and Huawei: data security issues and the alleged possibility that TikTok may be utilized by the Chinese government to spy on its users:

- *“tiktok was not banned for being Chinese, it was banned for being a chinese spy tool”*

¹⁰<https://www.cnbc.com/2020/07/31/trump-says-he-will-ban-tiktok-through-executive-action-as-soon-as-saturday.html> [Accessed Jan. 8, 2022]

- *“People be like but TikTok was my lifetime career you can’t just ban it, I’d rather lose my job than have fucking China spy on all of my shit”*
- *“It’s being banned because it’s spyware that has been providing user data to the CCP while censoring anything critical of them. That’s not a conspiracy, it’s what they’ve said they’re doing. You seriously need to get informed before you speak on this any more. Tiktok is a terrible company.”*

Overall, the negative comments on TikTok mostly surround the social media platform, platform community, and user-generated content, where many commenters express harsh criticism to these three aspects and describe them as toxic and stupid. We also find for most of the time, Reddit users’ characterization of TikTok has a meaningful difference from the way they characterize Huawei and Tencent, where they mostly refer to TikTok as a general social media platform and put less emphasis on TikTok’s Chinese origin and ties with the Chinese government. However, real-world political events may completely change this characterization in certain periods, such as when Trump announced his plan to ban TikTok for national security concerns in July 2020, and on Reddit, TikTok became closely knitted with China, the Chinese government and a term related to spying activities. These findings show interesting insights into Reddit users’ characterization of TikTok (ByteDance). They show how TikTok (ByteDance) is viewed differently than other Chinese technology companies in terms of its closeness to China and the Chinese government, and how this difference could be quickly changed by the influence of governmental actions.

Discussion and Conclusion

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 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 general political discussions.

By investigating the most semantically similar words in the comments to the names of three Chinese technology companies over time, we have discovered the predominant themes, and identified several real-world events that affect the public’s characterization of these companies. We take the top 3 companies that have the largest volume of comments as our study cases: Huawei, Tencent, and TikTok. We study the negative comments and positive comments separately.

For Huawei, we have found that negative comments are heavily focused on its role as a telecommunication device provider throughout our study period. The most dominant concern/discontent about the company rests on its alleged threat of spying 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 usually targeted toward 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.¹¹ 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

¹¹<https://github.com/reddit-archive/reddit> [Accessed Aug. 14, 2022]

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 (Friedlander 2018; Mascitelli and Chung 2019), 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.

References

- Ahmed, S. S.; and Kumar M., A. 2021. Classification of Censored Tweets in Chinese Language using XLNet. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, 136–139. Online: Association for Computational Linguistics. doi:10.18653/v1/2021.nlp4if-1.21. URL <https://aclanthology.org/2021.nlp4if-1.21>.
- Albalawi, R.; Yeap, T. H.; and Benyoucef, M. 2020. Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Frontiers in Artificial Intelligence* 3: 42. ISSN 2624-8212. doi:10.3389/frai.2020.00042. URL <https://www.frontiersin.org/article/10.3389/frai.2020.00042>.
- Alshahrani, A.; Ghaffari, M.; Amirizirtol, K.; and Liu, X. 2020. Identifying Optimism and Pessimism in Twitter Messages Using XLNet and Deep Consensus. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8. doi:10.1109/IJCNN48605.2020.9206948.
- Bingham, E.; and Mannila, H. 2001. Random Projection in Dimensionality Reduction: Applications to Image and Text Data. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’01, 245–250. New York, NY, USA: Association for Computing Machinery. ISBN 158113391X. doi:10.1145/502512.502546. URL <https://doi.org/10.1145/502512.502546>.
- Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirich-

- let allocation. *the Journal of machine Learning research* 3: 993–1022.
- Centivany, A.; and Glushko, B. 2016. "Popcorn Tastes Good": Participatory Policymaking and Reddit's. *CHI '16*, 1126–1137. New York, NY, USA: Association for Computing Machinery. ISBN 9781450333627. doi:10.1145/2858036.2858516. URL <https://doi.org/10.1145/2858036.2858516>.
- Chen, K.; and Tomblin, D. 2021. Using Data from Reddit, Public Deliberation, and Surveys to Measure Public Opinion about Autonomous Vehicles. *Public Opinion Quarterly* 85(S1): 289–322. doi:10.1093/poq/nfab021. URL <https://doi.org/10.1093/poq/nfab021>.
- Chen, L.; Lu, X.; Yuan, J.; Luo, J.; Luo, J.; Xie, Z.; Li, D.; et al. 2020. A social media study on the associations of flavored electronic cigarettes with health symptoms: observational study. *Journal of medical Internet research* 22(6): e17496.
- Cody, E. M.; Reagan, A. J.; Dodds, P. S.; and Danforth, C. M. 2016. Public Opinion Polling with Twitter.
- Cody, E. M.; Reagan, A. J.; Mitchell, L.; Dodds, P. S.; and Danforth, C. M. 2015. Climate Change Sentiment on Twitter: An Unsolicited Public Opinion Poll. *PLOS ONE* 10(8): 1–18. doi:10.1371/journal.pone.0136092. URL <https://doi.org/10.1371/journal.pone.0136092>.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
- Farrell, T.; Fernandez, M.; Novotny, J.; and Alani, H. 2019. Exploring Misogyny across the Manosphere in Reddit. In *Proceedings of the 10th ACM Conference on Web Science, WebSci '19*, 87–96. New York, NY, USA: Association for Computing Machinery. ISBN 9781450362023. doi:10.1145/3292522.3326045. URL <https://doi.org/10.1145/3292522.3326045>.
- Friedlander, J. 2018. How the arrest of Huawei's Meng Wanzhou may mark the start of the great China-US technology decoupling URL <https://www.scmp.com/comment/insightopinion/united-states/article/2179520/how-arrest-huaweis-meng-wanzhou-may-markstart>.
- Glenski, M.; and Weninger, T. 2017. Rating Effects on Social News Posts and Comments. *ACM Trans. Intell. Syst. Technol.* 8(6). ISSN 2157-6904. doi:10.1145/2963104. URL <https://doi.org/10.1145/2963104>.
- Hutto, C.; and Gilbert, E. 2014. 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, 216–225.
- Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X.; Jiang, X.; Li, Y.; and Zhao, L. 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications* 78(11): 15169–15211. ISSN 1573-7721. doi:10.1007/s11042-018-6894-4. URL <https://doi.org/10.1007/s11042-018-6894-4>.
- Kamarudin, N. S.; Rakesh, V.; Beigi, G.; Manikouda, L.; and Liu, H. 2018. A study of reddit-user's response to rape. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 591–592. IEEE.
- Landauer, T. K.; Foltz, P. W.; and Laham, D. 1998. An introduction to latent semantic analysis. *Discourse Processes* 25(2-3): 259–284. doi:10.1080/01638539809545028. URL <https://doi.org/10.1080/01638539809545028>.
- Lyu, H.; Wang, J.; Wu, W.; Duong, V.; Zhang, X.; Dye, T. D.; and Luo, J. 2022. Social media study of public opinions on potential COVID-19 vaccines: informing dissent, disparities, and dissemination. *Intelligent medicine* 2(01): 1–12.
- Mascitelli, B.; and Chung, M. 2019. Hue and cry over Huawei: Cold war tensions, security threats or anti-competitive behaviour? *Research in Globalization* 1: 100002. ISSN 2590-051X. doi:<https://doi.org/10.1016/j.resglo.2019.100002>. URL <https://www.sciencedirect.com/science/article/pii/S2590051X19300024>.
- Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.
- Pearson, K. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2(11): 559–572.
- Shen, J. H.; and Rudzicz, F. 2017. Detecting Anxiety through Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, 58–65. Vancouver, BC: Association for Computational Linguistics. doi:10.18653/v1/W17-3107. URL <https://aclanthology.org/W17-3107>.
- Soliman, A.; Hafer, J.; and Lemmerich, F. 2019. A Characterization of Political Communities on Reddit. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media, HT '19*, 259–263. New York, NY, USA: Association for Computing Machinery. ISBN 9781450368858. doi:10.1145/3342220.3343662. URL <https://doi.org/10.1145/3342220.3343662>.
- Tahmasbi, F.; Schild, L.; Ling, C.; Blackburn, J.; Stringhini, G.; Zhang, Y.; and Zannettou, S. 2021. "Go Eat a Bat, Chang!": On the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19. In *Proceedings of the Web Conference 2021, WWW '21*, 1122–1133. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383127. doi:10.1145/3442381.3450024. URL <https://doi.org/10.1145/3442381.3450024>.
- Vitak, J.; Shilton, K.; and Ashktorab, Z. 2016. Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW '16*, 941–953. New York, NY, USA: Association for Computing Machinery. ISBN 9781450335928. doi:10.1145/

2818048.2820078. URL <https://doi.org/10.1145/2818048.2820078>.

Wu, W.; Lyu, H.; and Luo, J. 2021. Characterizing Discourse about COVID-19 Vaccines: A Reddit Version of the Pandemic Story. *Health Data Science* 2021: 9837856. doi:10.34133/2021/9837856. URL <https://doi.org/10.34133/2021/9837856>.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. URL <https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf>.