COMPUTATIONAL ANALYSIS OF THE CHINESE GOVERNMENT'S USE OF ONLINE SOCIAL NETWORKS

by

Zhuo Cheng

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2023

Approved by:

Dr. Samira Shaikh

Dr. Wlodek Zadrozny

Dr. Razvan Bunescu

Dr. Tiffany Gallicano

Dr. Min Jiang

ABSTRACT

ZHUO CHENG. Computational Analysis of the Chinese Government's Use of Online Social Networks. (Under the direction of DR. SAMIRA SHAIKH)

Governments worldwide are increasingly engaging with both citizens and non-citizens via online social networks. This trend reflects the fact that these platforms have become primary sources of news and crucial forums for public discussion. China is no exception to this phenomenon.

Existing research into the Chinese government's use of online social networks and the public's reaction to it primarily offers qualitative insights rather than quantitative evidence. There is a noticeable absence of publicly available models and datasets that could enrich the research community's understanding of this topic. This dissertation aims to fill this void by proposing novel frameworks to examine the Chinese government's use of online social networks and the public's response.

In the first section, this dissertation proposes a fresh framework for identifying persuasive techniques in textual posts. The proposed framework employs a divide-and-conquer strategy to isolate and detect each persuasive technique individually. This approach is found to be effective in extracting features specific to each technique. Furthermore, the framework leverages GPT-3.5 to generate additional training samples at a much lower cost than other methods reliant on human-annotated external data. The model derived from our framework surpasses the performance of the previous state-of-the-art model.

In the second section, the dissertation investigates the similarities and differences in the Chinese government's posts across various platforms, each targeting distinct audiences. It is found that the government's focus varies from platform to platform, presumably due to the diverse audiences each network attracts.

The final section of this dissertation provides a quantitative analysis of users' opinions towards the content produced by the Chinese government on online social networks. This analysis assesses whether posts by government-affiliated accounts reach their intended audiences and explores the stance characteristics of the responses.

Overall, this dissertation aims to contribute to our understanding of Chinese government's engagement on social media platforms.

ACKNOWLEDGEMENTS

First and foremost, I want to thank my advisor, Dr. Samira Shaikh. I could not have completed this dissertation without her continuous support and guidance. Words cannot truly express how grateful I am. I would also like to extend my gratitude to my esteemed committee members, Dr. Razvan Bunescu, Dr. Tiffany Gallicano, Dr. Min Jiang, and Dr. Wlodek Zadrozny, for their invaluable support and feedback.

Furthermore, I want to express my gratitude to my friends and lab mates. Their support and encouragement made this journey much more manageable.

Last but certainly not least, I want to thank my parents and sister. Nothing would have been possible without their unwavering support and love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI4SG	Artificial Intelligence for Social Good
NLP	Natural Language Processing
NLP4SG	Natural Language Processing for Social Good
OSN	Online Social Networks

CHAPTER 1: INTRODUCTION

1.1 Background

Driven by the fact that people use social media for news access and political discourse, governments around the world attempt to engage with the domestic audience and the international audience through online social networks. The prevalence of online social networks has made it possible for both state and non-state actors to reach a large audience at a smaller expense. Almost all members of parliament have Twitter accounts in 32 European countries of the European Union, the European Free Trade Association, and the United Kingdom [4]. In the United States, all Senators and almost all Representatives adopt Twitter accounts as well [5].

China is no exception. Although Twitter is blocked in China, the number of China's diplomatic Twitter accounts increased to 80 as of January 17, 2020, from 13 as of October 20, 2018 [6]. Notably, most of the Chinese diplomats with Twitter accounts do not have Weibo accounts. Within China, as of December 2021, there are 1,032 million Internet users in China [7], accounting for 73% of the whole population. Weibo has become an important online social network where public opinion is exchanged since 2009 [8]. As of December 2019, there are 138,854 government Weibo accounts [9], indicating the government is active on this online social network.

This dissertation aims to understand the Chinese government's use of online social networks from three perspectives.

- First, we propose a new framework that detects how posts are expressed (persuasive techniques). Existing research only analyzes the topics of the posts by the Chinese government without detailing how those posts are expressed. The persuasive technique provides a great perspective to gain more knowledge of the government's behavior. This dissertation proposes a new framework to detect persuasive techniques at a lower cost compared with previous models.

- Secondly, we compare the similarities and differences between the Chinese government's activities on domestic and international online social networks. Given that multiple major international online social networks are blocked in China, the government's posts on different platforms are targeting different audiences. This dissertation aims to analyze whether the Chinese government approaches different audiences in different ways.

- At last, we conduct a detailed analysis of the public opinion received by the Chinese government accounts. Even though the Chinese government aims to tell China's story well, however, the effect hasn't been studied in detail. Previous research only provides qualitative studies without providing quantitative evidence. This dissertation provides such analysis to reveal the public opinion towards the posts by the Chinese government-affiliated accounts.

1.2 Motivation

We summarize the motivations and novelty of our research as follows.

- Existing research analyzing the Chinese government's use of online social networks focuses only on what without detailing how the content is expressed. The persuasive technique could provide such a lens to improve our understanding of how the posts are expressed by the Chinese government-affiliated account.

- Existing models to detect persuasive techniques have a very complex structure and uses external data source annotated by humans. This raises three questions. 1) Could we use simpler models? 2)Is the external dataset necessary for the detection of all persuasive techniques? 3) Could we have a less expensive way to generate additional training data? We design a new framework to tackle those three questions.

- Existing comparison study provides only qualitative observation of the Chinese government's activities on domestic and international online social networks. Huang and Wang [6] find that China's diplomatic posts on both Twitter and Weibo use information sources from major Chinese state-owned news outlets, exhibiting a hierarchical structure. Posts on Weibo tend to manifest a harder attitude on diplomatic issues than on Twitter. However, the evidence is based on observation of certain specific tweets without showing quantitative evidence. With the help of the proposed two-dimensional analysis framework, we will compare the similarities and differences in the Chinese government's use of online social networks as to both what the government is posting and how the government is posting. Our analysis will be based on a larger dataset so that the result could be generalized and robust.

- When analyzing the engagement with the Chinese government's online social network accounts, existing research focuses on the authenticity of the engagement. Only a case study [10] reveals that the replies the Chinese Zhao Lijian receives are charged with repulsion and hatred. What's the people's stance towards the Chinese government's posts in general and the characteristics of different groups remain to be studied.

1.3 Organization

Chapter 2 reviews existing research in two major fields. One field is the analysis of the Chinese government's use of social media. The other field is the detection of persuasive techniques, which provides an additional dimension to understanding the posts of the Chinese government-affiliated accounts. Chapter 3 proposes a new framework that detects persuasive techniques. Chapter 4 introduces the work that analyzes the similarity and differences of the behavior when the Chinese government accounts post on different platforms. Chapter 5 provides a detailed analysis that reveals the stance of the users towards the Chinese government's posts and examines the similarities and differences across groups using different languages. Chapter 6 summarizes the main findings of this dissertation.

CHAPTER 2: Related Work

This Chapter reviews the literature regarding two major fields. The first part reviews literature concerning the Chinese Government's use of domestic and international online social networks. The second part reviews the literature that compiles the persuasive techniques and designs models to detect them. We want to emphasize that the literature in the second part categorizes those techniques as propaganda techniques. However, there is no widely accepted and operational process to distinguish what is propaganda and what is not. Therefore, we use the term persuasive techniques instead of propaganda techniques to avoid predisposition and bias.

2.1 The Chinese Government's Use of Online Social Networks

We divide this part into two, the Chinese government's use of domestic online social networks and its use of the international online social networks.

Domestic

As of December 2021, there are 1,032 million Internet users in China [7], accounting for 73% of the whole population. Weibo has become an important online social network where public opinion is exchanged since 2009 [8]. As of December 2019, there are 138,854 government Weibo accounts [9], indicating the government is active on the online social network.

Research in this field studies the government's use of domestic online concerns two aspects — the adoption of domestic online social networks, and content analysis. In terms of the adoption of the use of online social networks, previous research analyzes the drivers, challenges, and capabilities of the government, as well as factors associated with this innovation diffusion among local governments [11, 12]. In terms of content analysis, previous research examines the pattern of interaction, information dissemination, and the government's method to influence public opinion on domestic online social networks. Frequent topics of the content are examined through manual coding of a small sample [8, 13, 14].

Zheng [12] conducts brainstorming activities among 78 civil servants and interviewed 12 managers of influential government Weibo accounts to explore the drivers, challenges, and capability of the government's adoption of the online social network. The external drivers of the government's adoption of the online social network include the pervasiveness of IT devices, increasing citizen participation, and international trends. The challenges or concerns are the digital division - online social network users may not be very representative of the whole population, people's low trust in government, hackers and astroturfing, and sustainability of the popularity of online social networks. The strengths of the government include adequate information and communications technology (ICT) infrastructure, and rich information resource. The weaknesses of the government's capabilities are lack of attention and support from leadership, inexperience in managing the online social network, no designated positions, shortage of funds, complicated content review process and centralized power, risk-averse culture, little collaboration, and information security.

From the perspective of organization innovation diffusion, Ma [11] applies multivariate regression and finds that government size, internet penetration rate, regional competition and learning, and upper-tier pressure are positively associated with the government's early adoption of Weibo. Fiscal revenue, economic development, and economic openness - proxies for financial resources and economic conditions - are not significantly associated with the government's adoption of the online social network. Those factors are only important in cases where vast investment is needed to adopt the innovation, while the online social network is free for use and requires little fiscal resources.

As to content analysis, several studies find that Chinese governments use Weibo mainly for information dissemination rather than interacting with the public [8, 14]. However, from 2011 to 2012, the Chinese government made significant progress on online social media by posting more service-related messages, using less formal language, providing more timely information, and being more interactive through more frequent forward and push to better serve the public. However, both forward and push are indirect interactions. The Chinese government seldom interacts with the public directly on online social media. A more recent case study of the Beijing Police Department (BPD) shows that BPD provides information not only related to crime and enforcing the law, but also useful information to the city's newcomers. The Chinese government posts information not only through its official social media but also conducts secretive operations by requiring government employees to post pro-government content [15]. This strategy adopted by the Chinese government is to cheerlead the government's policy and distract the public rather than engage in direct argument.

International

Driven by the fact that people use social media for news access and political discourse, governments around the world attempt to engage with the international audience through online social networks [16] and China is no exception. The prevalence of social media has made it possible for both state and non-state actors to reach a larger audience at a smaller expense. Almost all members of parliament have Twitter accounts in 32 European countries of the European Union, the European Free Trade Association, and the United Kingdom [4]. The American diplomacy community is actively researching how to make the most of digital technologies [17, 18]. In the United States, all Senators and almost all Representatives adopt Twitter accounts [5]. At least 56 foreign policy officials of the United States have Twitter accounts [17]. Although Twitter is blocked in China, the number of China's diplomatic Twitter accounts increased to 80 as of January 17, 2020, from 13 as of October 20, 2018 [6]. Between June 2020 and February 2021, PRC diplomats and state-backed media post much more frequently on Twitter than on Facebook [19].

One thread of existing research regarding China's public diplomacy on international online social networks captures its characteristics of network structure and content [20, 6, 10, 21]. Despite the fact that some Chinese missions communicate and interact with foreign audiences and counterparts while some do not, they share a common primary information source - the state-owned media outlets, indicating a hierarchical structure of the communication network within the Chinese diplomatic active Twitter accounts. The top three topics most frequently tweeted about include China-foreign cooperation in economic and social aspects, political relationship with foreign countries, and promotion of Chinese culture and society [20], most of which are moderate and trying a create a friendly international environment for China. During the US-China Trade War, even though China's related posts on Weibo generally show a tough attitude, China's Twitter posts show a relatively softer gesture [6]. However, there is one exception. Zhao Lijian, the current spokesperson of the Chinese Ministry of Foreign Affairs and previous diplomat affiliated with the Chinese embassy in Pakistan, actively tweets content with characteristics of polarization and strong emotion [10]. Zhao Lijian tweets in a manner not aligned with the traditional purposes of public diplomacy, which generally tend towards boosting mutual understanding, promoting nation branding, and creating a friendly international environment for foreign policies [22]. Zhao Lijian's proactive tweeting style is claimed to be integral to Chinese diplomacy's shift from forbearance, and softness to proactivity and assertiveness [21].

Another thread of research aims to detect the Chinese government's usage of bots on international online social networks. Despite media reports of bot activity backed by the Chinese government on Twitter, there is no solid academic evidence supporting this kind of claim. There is some academic evidence showing that automation is used to spread anti-government information [23]. The two major groups using automation to spread anti-government information are *the 1989 bot group* and the *pan-Asia* group. In later research in 2021, the same research group find on Twitter, the top 1% super-spreader account for 50% of the retweets of the PRC accounts [19]. However, instead of claiming this disproportion as suspicious automation, the authors suggest this may be genuine support for the PRC. A large portion of the accounts contributing substantially to the share of the engagement with PRC accounts are later suspended by Twitter. However, there is no way to assess which specific rule those accounts violate on Twitter. This also reveals an obstacle to all the articles tackling the authenticity of the engagement with the Chinese government accounts — it is difficult to find ground truth.

Overall, existing research about the Chinese government's use of international online social networks shows that the Chinese government is active on international online social networks such as Twitter and Facebook. The network of those accounts exhibits a hierarchical structure where most accounts' information source is the state-owned news outlets. Though the main topics within the posts of the Chinese government's accounts are promoting a positive image of its own and cooperation with other nations, there is an outlier whose behavior is inconsistent with the traditional goal of public diplomacy. Zhao Lijian, the spokesperson of the Chinese Ministry of Foreign Affairs, tweets aggressively and proactively. Despite media claims that China is using automation to disseminate information, fake popularity, and amplify the engagement, no solid academic evidence has been found either in the Chinese government's activity on international online social networks or in the engagement with the Chinese government's accounts. On the contrary, automation has been found in spreading anti-government information. However, the researchers also acknowledge that the ground truth is difficult to get in their study.

Summary

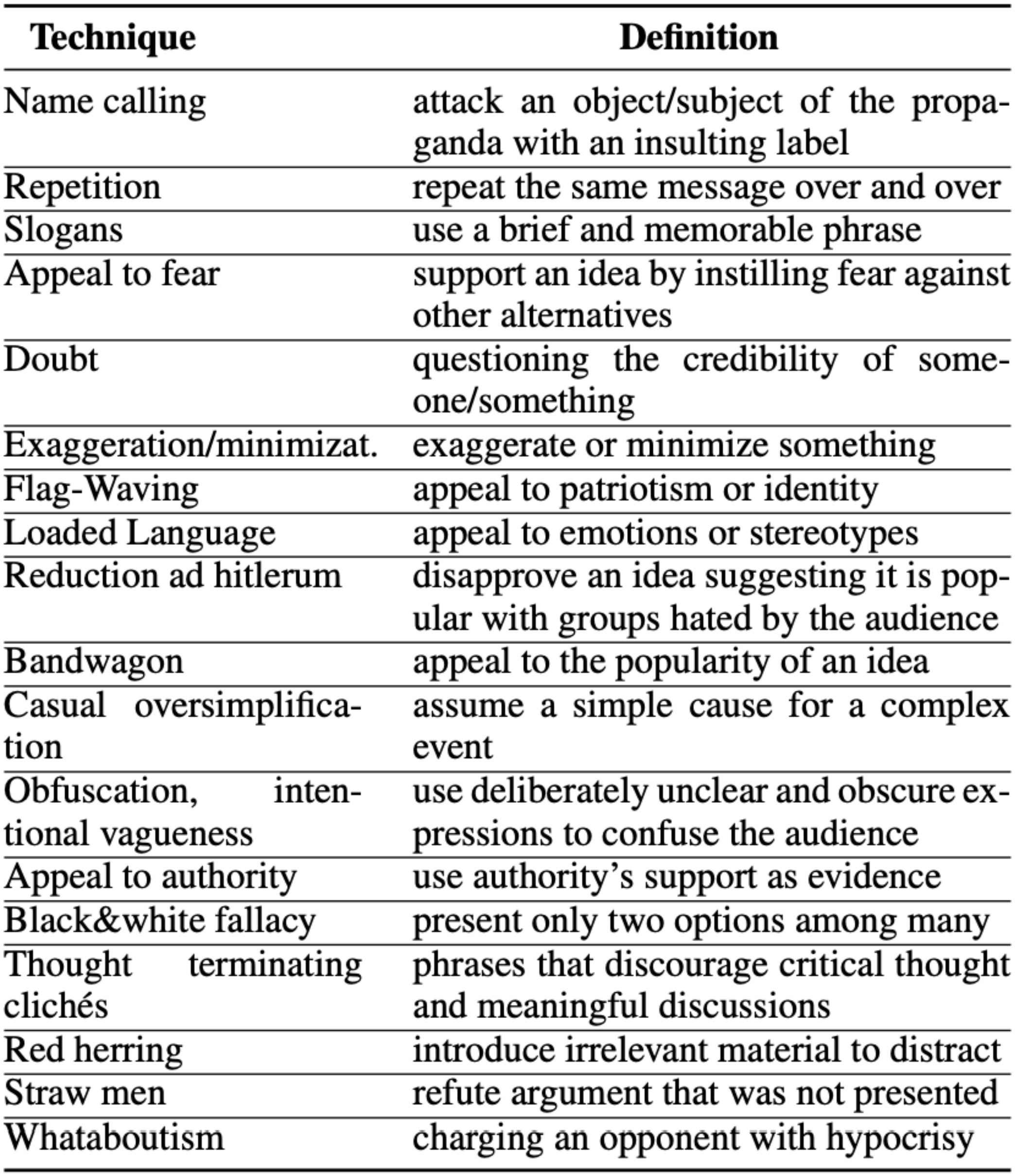
The existing research regarding the Chinese government's use of online social networks doesn't explore the details of the engagement with the Chinese government's accounts. Though Guo [10] shows that on Twitter, replies to Zhao Lijian's posts are characteristic of repulsion and hatred, it provides little quantitative information regarding the proportion of replies of this kind and omits the replies of other kinds. What's more, the case study limits the generalization of its conclusions. To get a better grasp of public opinion, we need a larger dataset to be representative of the whole population, a new annotation schema that characterizes different kinds of replies to the Chinese government's posts, and a new model which can be used to extrapolate the labels to the large dataset. As for the domestic online social networks, since there is widespread censorship on Weibo, it's understandable that few research explores the public opinion on the government's posts.

Previous research on the content of the Chinese government account posts only focuses on what the government is posting and the conclusions are drawn through manual coding on a small sampled data. This small size of the dataset and the focus on only one dimension of the content limits both the generalization of the conclusions and the breadth of our understanding of the Chinese government's behavior on online social networks. Therefore, a multi-dimension analysis framework that covers not only what the government is posting but also how the content is posted, is needed to improve our understanding of the Chinese government's use of online social networks. What's more, a public dataset is needed to benefit the research community.

The difference between the Chinese government's use of domestic online social networks and international online social networks is underexplored. Existing research observes that the Chinese diplomatic accounts have a more moderate tone on Twitter in certain issues [6]. The conclusions are drawn from very specific observations without providing quantitative evidence.

2.2 Persuasive Techniques and the Detection

In the related literature of propaganda, 18 persuasive techniques are compiled and a task is created to detect those techniques [1]. Those techniques include psychological and rhetorical techniques that are used to persuade people and influence their opinion. Fig. 2.1 lists those techniques and their definitions.



Technique	Definition
Name calling	attack an object/subject of the propaganda with an insulting label
Repetition	repeat the same message over and over
Slogans	use a brief and memorable phrase
Appeal to fear	support an idea by instilling fear against other alternatives
Doubt	questioning the credibility of someone/something
Exaggeration/minimizat.	exaggerate or minimize something
Flag-Waving	appeal to patriotism or identity
Loaded Language	appeal to emotions or stereotypes
Reduction ad hitlerum	disapprove an idea suggesting it is popular with groups hated by the audience
Bandwagon	appeal to the popularity of an idea
Casual oversimplification	assume a simple cause for a complex event
Obfuscation, intentional vagueness	use deliberately unclear and obscure expressions to confuse the audience
Appeal to authority	use authority's support as evidence
Black&white fallacy	present only two options among many
Thought terminating clichés	phrases that discourage critical thought and meaningful discussions
Red herring	introduce irrelevant material to distract
Straw men	refute argument that was not presented
Whataboutism	charging an opponent with hypocrisy

Figure 2.1: List of the 18 propaganda techniques and their definitions [1]

This list of techniques is not necessarily exclusively being used in the context of propaganda. Black and white fallacy, and red herring are both logical fallacies that people could make out of pure mistakes. Though the techniques' definitions can guide people to discern them in the text, existing literature doesn't provide an actionable process to distinguish propaganda. It would give them people the impression that we are claiming public diplomacy and everything in the Chinese government's posts as propaganda if we adopt the term propaganda techniques. We have no such intention. In our proposed framework, we refer to this list of techniques as persuasive techniques since this list of techniques is nonetheless providing an additional dimension as to how the Chinese government is posting the content. The aspect hasn't been explored in the context of the posts of the Chinese government's accounts, a gap our proposed two-dimensional framework aims to fill.

Given that one dimension our framework aims at is to detect the persuasive techniques, we present the development of the models to detect the persuasive techniques. Previous models are designed to tackle the detection of propaganda on different levels — the document level, the sentence level, and the token level.

Document level detection. Some researchers detect propaganda on the article level to see whether an article has propaganda content or not, which is a document-level binary classification task. Rashkin et al. [24] collected a news corpus with varying reliability. The types of news articles are trusted, satire, hoax, and propaganda. They use an LSTM model with GLOVE word embeddings concatenated with LIWC measurements to predict the category of the news articles and find that lexicon features may not help improve the prediction accuracy but it does help understand the differences between reliable and unreliable news sources. Barron-Cedeno et al. (2019) created a real-time system called Proppy to detect propaganda existing in news articles. They use the DBSCAN clustering algorithm to identify events and assign a propaganda score to each article using a maximum entropy classifier with features of n-grams, LIWC measurements, style, vocabulary richness, readability, and NELA features which include sentiment, bias, and complexity. Habernal et al. [25] created a corpus with cases of ad hominem arguments annotated and they found that CNN outperformed Bi-LSTM with word2vec embeddings [26] as input features for both models. They label these arguments of ad hominem arguments with more specific labels like illiteracy insult and condescension in the corpus to see what makes an argument ad hominem. Habernal et al. [27] created a game named Argotario to educate players to understand and create fallacies including ad hominem, red herring, and irrelevant authority.

textbfSentence level and token level detection Some researchers detect propaganda in finer granularity — at the sentence level and token level. They tend to train the model jointly to achieve better performance. Da san et al. [1] labels spans in news articles with 18 propaganda techniques which can be identified by annotators without referring to outside information other than the article itself. Their classifier is built on BERT and they released an online system Prta to support users to analyze propaganda fragments in the text they input [28]. Want et al. [2] later incorporated textual knowledge and first-order logical information to further improve the performance of Transformer-based model. Recently, in the context of detecting propaganda techniques in Tweets rather than news, the context, the previous Tweet of the target in a thread, is used as additional information to detect the propaganda techniques [3].

Transformer-based models The state-of-the-art models to detect propaganda techniques are using the Transformer architecture. Here we will elaborate on the architectures of these models.

Multi-Granularity Network [1] created a Multi-Granularity Network where the higher-granularity task uses the output of the lower-granularity task output as input information. In their paper's case, the higher-granularity task is fragment-level classification (FLC) of the 18 propaganda techniques and the lower-granularity task is binary sentence-level classification (SLC) which detects whether the sentence contains at least one propaganda technique.

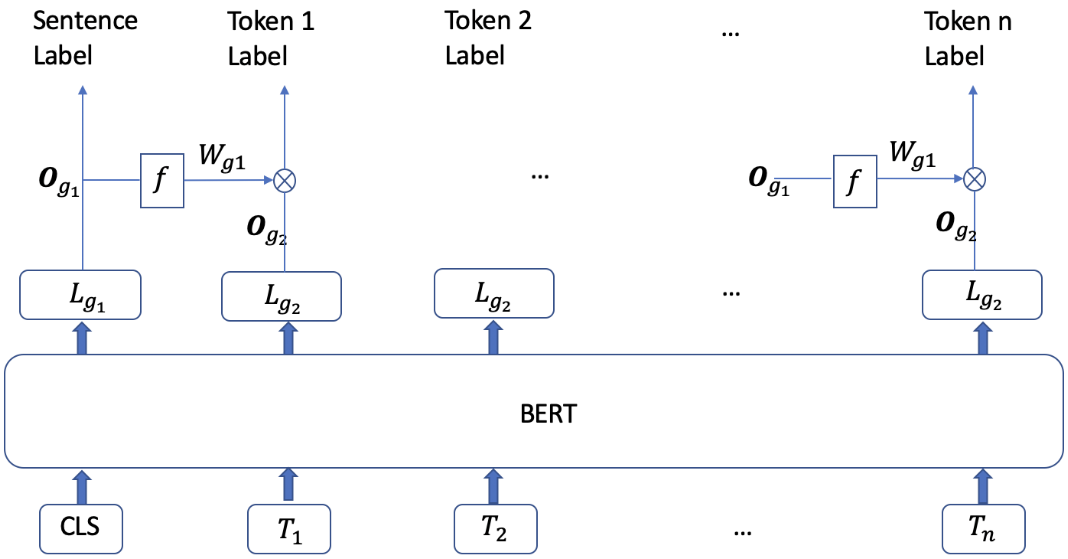


Figure 2.2: Structure of Multi-Granularity Network [1]

As shown in Figure 2.2, in MGN, the output of the lower-granularity task will be fed to the higher-granularity task. Each task g_k has its own classification layer L_{g_k} and its output is $\boldsymbol{o}_{g_k}$. The dimension of $\boldsymbol{o}_{g_k}$ is the number of labels of task g_k. Suppose the there are d_k labels of task g_k. g_{k+1} is the task of next granularity of task g_k. The output $\boldsymbol{o}_{g_k}$, generated by L_{g_k} will be used to generate a weight w_{g_k} for the task g_{k+1} using a trainable function f where

$$w_{g_k} = f(\boldsymbol{o}_{g_k}). \tag{2.1}$$

f projects $\boldsymbol{o}_{g_k}$ of d_k dimensions to a one dimension number w_{g_k}. The higher granularity classification layer's output $\boldsymbol{o}_{g_{k+1}}$ is updated by multiplying its element with w_{g_k} and becomes $\boldsymbol{o}'_{g_{k+1}}$:

$$\boldsymbol{o}'_{g_{k+1}} = w_{g_k} * \boldsymbol{o}_{g_{k+1}}. \tag{2.2}$$

In this way, the information from the lower-granularity task can be used by the higher-granularity task. For instance, assume g_k is the binary sentence classification of propaganda techniques, and g_{k+1} is the span detection of propaganda techniques. If the sentence classifier finds that the sentence contains no propaganda technique, then w_{g_k} would be 0 and $\boldsymbol{o}'_{g_{k+1}}$ would also be 0.

LatexPRO: Logical and Textual Knowledge for Propaganda Detection. MGN cannot guarantee the consistency between sentence-level prediction and fragment-level detection. To solve the problem of inconsistency and also introduce human knowledge into the model, Wang et al. [2] proposed ***LatexPRO*** that leverages **l**ogical **a**nd **tex**tual knowledge for **pro**paganda detection.

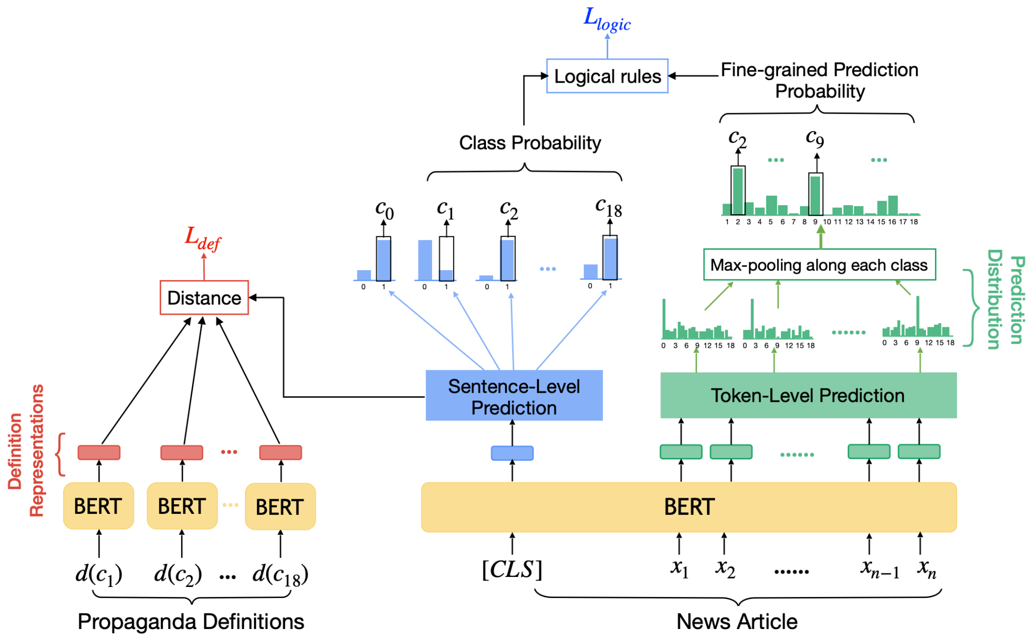


Figure 2.3: LatexPRO Structure [2]

There is logical rule between the sentence-level and fragment-level classification tasks. If the sentence has no propaganda then there should not be propagandistic fragment. A new loss function that hard-coded this rule was designed to take full advantage of this association.

Let $f_c(x)$ be the probability that the sentence that contains x belongs to class c and $g_c(x)$ be the probability that the fragment classifier detects the input which contains x belongs to class c. $g_c(x)$ is derived by max-pooling over all the probabilities of class c produced by the classifier for every token. Then equation 2.3 can be rewritten as

$$P(F) = P(f_c(x))(P(g_c(x)) - 1) + 1 \tag{2.3}$$

The objective here is to maximize $P(F)$, which is equal to minimizing $L_{logic} = -log(P(F))$.

To use textual knowledge of the definition, the distance $dist$ between the BERT representation of the technique c_i's definition, $D(c_i)$ and the representation of the predicted label $W(c_i)$ should be minimized. The overall loss L_{def} of all techniques is

$$L_{def} = \sum_{i=1}^{18} dist(W(c_i), D(c_i)) \tag{2.4}$$

A weighted sum L_j derived by summing up the token-level loss L_{tok}, the fine-grained sentence-level loss L_{sen}, the textual definition loss L_{def} and the logical loss L_{logic} is used to train the whole model. α, β, λ and γ are predefined hyperparameters.

$$L_j = \alpha * L_{tok} + \beta * (L_{sen} + L_{def} * \lambda) + \gamma * L_{logic} \tag{2.5}$$

The most recent model, a transformer-based multi-view propaganda detection model (MV-PROP), expands the existing dataset by creating a new corpus with annotated Tweets [3]. As shown in Figure 2.4, this model uses several encoders for the same pair of a Tweet and its context, an action based on the assumption that different encoders could grasp different features of different propaganda techniques. The output of different encoders is merged in a similar fashion as the attention structure.

The whole model looks like a transformer of transformers.

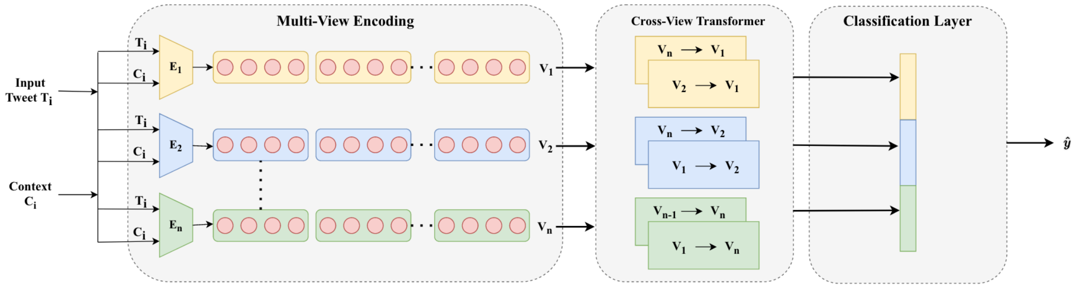


Figure 2.4: Illustration of the MV-PROP model [3]

To summarize, the complexity of the model design is increasing. The model evolves from one transformer to a transformer of transformers. The authors try to incorporate information from various sources, such as textual information of the label, the association of the label of the sentence with the label of the fragment, and the context information.

CHAPTER 3: Detection of Persuasive Techniques

3.1 Introduction

When analyzing the content of the Chinese government's posts, existing research normally focuses on what the government is posting without detailing how [14, 10, 20]. It's noteworthy that when analyzing the Chinese government's use of Twitter, which belongs to public diplomacy, some research uses the word propaganda to refer to the activities of the Chinese government [29, 6] and it is not used to refer to democratic countries in the same thread of research [30]. However, there is no clear explanation why this word is reserved only for the Chinese government. There is also debate as to whether public diplomacy should be regarded as propaganda [31]. Therefore, it would be inappropriate to refer to the Chinese government's use of online social networks and related activities as propaganda.

Nevertheless, the list of persuasive techniques compiled in the related literature of computational propaganda detection offered a new perspective to examine how the content is expressed [1]. Those techniques include psychological and rhetorical techniques that are used to persuade people and influence their opinion. Fig. 2.1 lists those techniques and their definitions.

This list of techniques is not necessarily exclusively being used in the context of propaganda. The black-and-white fallacy, and red herring are both logical fallacies that people could make out of pure mistakes. While it would be controversial to claim the Chinese government's use of online social networks as related activities, this list of techniques is nonetheless contributing to our understanding of how the government is expressing the content. This set of techniques can not only be used to analyze the Chinese government but also governments all over the world.

To analyze the persuasive techniques used in the textual content posted by the Chinese government-affiliated accounts, this Chapter aims to develop a better model at a lower cost to detect those techniques. First, inspired by ChatGPT's success across various NLP tasks, we test its ability to detect persuasive techniques with several popular prompting strategies. Secondly, we present a new framework that adopts a divide-and-conquer approach to detect the persuasive technique in textual content. This framework utilizes additional samples generated by ChatGPT. Even though ChatGPT cannot beat fine-tuned BERT-style models on the persuasive technique detection task, it is useful for generating additional training data to improve the model's performance. Compared with models in existing work [32] that uses external data annotated by humans, the cost of deriving additional training data in our framework is much lower. The model developed with this framework beat the state-of-the-art in the dataset of SemEval 2021 task 6.

3.2 Existing Datasets

There are two public textual datasets of persuasive techniques. The first is from Da San Martino et al. [1]. This dataset contains 18 persuasive techniques. The second dataset is from the SemEval 2021 task 6 [33] and it contains 20 persuasive techniques. The additional two techniques in the second dataset are *Smears* and *Glittering Generalities(Virtue)*.

There are three major papers with experiments run on the first dataset [1, 2, 3]. However, only the paper [1] that introduced the dataset provides public access to the model's source code. The other two papers do not provide open-source code. What's more, Want et al. [2] didn't use the test dataset of Da San Martino et al. [1] for comparison citing that back then the test dataset was not available. At last, those papers fail to provide the test result of each technique. On the other hand, the SemEval 2021 dataset is used for competition, which is publicly and the workshop papers list the metric for each technique. The metric is also trustworthy since all results are public accessible and recorded. The SemEval 2021 task 6 dataset also includes two additional persuasive techniques. Therefore, we use the SemEval 2021 task 6 dataset for the experiments in this chapter.

3.3 Existing Models

The dominant paradigm of existing research to tackle the detection of persuasive techniques in sentences is fine-tuning BERT-based models. Different modifications of the model architecture, additional features, and various engineering techniques have been explored to improve detection accuracy. The model of [1] trains two tasks jointly — the first task is to detect whether there is at least one persuasive technique used in the sentence and the second task is to detect the fragment and the specific technique in the sentence. The overall loss of the model is a linear combination of these two sub-tasks. [2] creates a new loss function that explicitly uses the association between the two tasks, that is if a sentence uses at least one persuasive technique, then there must be fragments with specific techniques in the sentence. TWEETSPIN [3] builds a large model which uses multiple encoders to extract features, under the intuition that each group of encoders could specialize in extracting features relevant to different persuasive techniques. The models also exploit a Tweet's previous Tweet as context and feed the context to the model as an additional feature. However, the latter two papers haven't made their models or data public, nor have they provided the details of the model's performance for each task.

In the SemEval 2021 task 6 competition subtask 1 — identity which of the 20 techniques are used in the given text, the best team MinD [32] mainly adopts three techniques to achieve outstanding performance. First, they use transfer learning — they initially train the model on the external human-annotated dataset [34], which consists of more than 20,000 sentences for 18 persuasive techniques. Secondly, an ensemble model with five pre-trained language models is trained and the final probability of the system is the average of the five components. Thirdly, they use post-processing to detect the technique of *Repetition*, in which they assign the technique of *Repetition* to the text if there exists a bigram that appears more than 3 times.

Considering the complexities of the existing models and the costly annotation of external training data sources, it's natural to ask the following questions.

1. Could we use simpler models?
2. Is the external dataset necessary for the detection of all persuasive techniques?
3. Could we have a less expensive way to generate additional training data?

To tackle those questions, we first explore ChatGPT's performance in detecting the persuasive technique with different prompting strategies.

The introduction of the GPT series, especially ChatGPT, has drawn the attention of researchers in the field and NLP and AI in general. GPT 1 was proposed in 2017 [35]. Different from BERT which uses bi-directional self-attention components, it's an autoregressive generative model with only decoder stacks that use uni-directional attention. GPT-2 was released in 2019, which utilized the same architecture as its predecessor but was trained on a larger dataset and had a larger model size. The larger model size and training data make it able to solve many new tasks without the need for supervised learning. GPT-3 uses more parameters and is trained on an even larger dataset.

However, the increase in scale doesn't necessarily strengthen the model's capability to generate satisfying responses from a human's perspective. Therefore, the InstructGPT model is proposed to solve this issue. InstructGPT uses Reinforcement Learning from Human Feedback (RLHF). Given the input, humans not only provide some demonstrations of the desired model behavior but also rank the outputs. GPT-3 is fine-tuned on that dataset. As a result, ChatGPT, which uses the technique of InstrctGPT and more human feedback, shows excellent quality in generating content aligning with human expectations.

Existing research compared ChatGPT with State-of-the-Art (SOTA) on various

NLP tasks [36, 37, 38]. ChatGPT doesn't beat previous SOTA on every task but achieves better or comparable performance on some. Kocoń [36] find that on 25 diverse analytical NLP tasks, the average loss of performance of ChatGPT compared with SOTA is 25%. The more difficult the task is, the larger the loss is. For inference tasks and the stance detection task, ChatGPT outperforms BERT-style models [37, 38]. However, the performance of ChatGPT on the persuasive technique detection task hasn't been carefully investigated. This chapter aims to reveal the performance of ChatGPT on this task and find that it cannot beat the fine-tuned BERT-style models on this task.

We then introduce a new framework that utilizes simple binary classifiers without future modification of the BERT-style models. This approach is under the assumption that a divide-and-conquer approach that deals with each technique individually could achieve good performance because each binary classifier would be good at extracting features pertinent to the persuasive technique targeted. ChatGPT (GPT-3.5) is used to generate additional training datasets, which cost much less than the human annotation.

The experiment of this chapter includes two major parts. In the first part, the performance of ChatGPT on the persuasive detection task is explored. The result shows that ChatGPT falls short in this task despite various prompt strategies utilized. In the second part, we demonstrate that a new framework that adopts a divide-and-conquer approach with additional training samples generated by ChatGPT could outperform the previous BERT-style SOTA model.

3.4 Dataset

The dataset we use is from the SemEval 2021 task 6. It includes training and test samples for the 20 persuasive techniques listed in Table 3.1. The specifics of the number of instances for each technique are shown in Table 3.1. The dataset is quite imbalanced, reflecting the difference of frequencies that people use different persuasive techniques. *Loaded Language*, *Name calling* and *Smears* have at least 200 training instances while *Obfuscation*,*Red Herring*,*Reductio ad hitlerum*,*Repetition* and *Bandwagon* have less then 10 training instances.

Table 3.1: Number of Positive Training Samples and Test Samples for Each Persuasive Technique

Label	**# of Training**	**# of Testing**
Loaded Language	358	100
Name calling/Labeling	218	53
Smears	200	45
Doubt	48	28
Slogans	44	19
Exaggeration/Minimisation	52	19
Glittering generalities (Virtue)	32	11
Whataboutism	40	10
Appeal to fear/prejudice	43	10
Black-and-white Fallacy/Dictatorship	18	7
Appeal to authority	13	7
Flag-waving	27	6
Thought-terminating cliche	20	6
Presenting Irrelevant Data (Red Herring)	1	4
Reductio ad hitlerum	9	3
Causal Oversimplification	27	3
Obfuscation	4	1
Repetition	8	1
Bandwagon	2	1
Straw Man	20	1

3.5 ChatGPT's Performance with Different Prompting Strategies

We adopt several prompt strategies to test its capability on this task of persuasive technique detection. The GPT series have shown great potential in various NLP tasks [38, 36, 37]. However, the task of persuasive technique detection in a given text piece is a relatively new task and hasn't been explored in detail. Therefore, we conduct experiments to see whether ChatGPT could beat the previous fine-tuned BERT-style models on this task.

3.5.1 Prompt Strategies

A prompt is a set of instructions provided to the Large Language Models (LLMs) [39]. Prompting offers a natural and interactive way for people to engage with LLMs [40]. Different prompts may have a great influence on the LLMs' performance on various tasks [41, 39]. For the field of text classification, different prompting engineering strategies have been explored extensively [42, 43, 44, 45]. Some popular ones are chain-of-thought (CoT) prompting [46], few-shot prompting (in-context learning) [47]. We use four different types of prompts.

The first prompting strategy we use is a **simple zero-shot prompt** without additional demonstrations or instructions. We directly ask the LLM to detect whether the text uses a specific persuasive technique. Though the GPT series is known for its ability to provide an explanation for its response [36, 48], We device the prompt in a way that it gives the answer without explanation, easy for post-processing. Below is the simple template we use, where *technique* stands for a specific persuasive technique and *text* stands for textual content for detection.

Does the text use the persuasive technique {technique}?

Only answer in yes or no.

[text]

The second prompting strategy is similar to **Chain-of-Thought (CoT)** [46]. CoT prompting is inspired by human's multi-step problem-solving method. CoT prompts decompose the problem in multiple steps and provide a series of instructions for each step before reaching the final answer. In our case, we insert the definitions and explanations to identify the persuasive techniques into the simple prompt, expecting the LLM to follow the definition and explanation to increase its accuracy in detecting the persuasive techniques. Below are two prompts for the persuasive technique of *Loaded Language* and *Whataboutism.*

Loaded Language Template

Detect the whether loaded language, a persuasive technique, is used in a given piece of text. Only respond in yes or no.

Loaded language refers to the intentional use of words or phrases that carry strong emotional or biased connotations, aiming to influence the reader's opinion or perception of a particular topic.

[text]

Whataboutism Template

Detect the whether Whataboutism, a persuasive technique, is used in a given piece of text delimited by three backticks. Only respond in yes or no.

Whataboutism or whataboutery (as in "what about...") denotes in a pejorative sense a procedure in which a critical question or argument is not answered or discussed, but retorted with a critical counter-question which expresses a counter-accusation. Whataboutism may involve responding to an accusation or concern with a counter-question like "What about...?" or by pointing out perceived hypocrisy or double standards.

[text]

Few-shot prompt with random samples. Few-shot prompting or in-context learning provides some examples in the prompt as demonstrations for LLMs to achieve better performance [47]. On top of the CoT prompt, we design a prompt template where random samples were added to the template of CoT.

Few-shot prompt with selective samples. To check whether the choice of examples for demonstrations could influence the LLM's performance, in design a new template where the examples used for demonstrations are chosen from the samples where the LLM made mistakes with CoT template.

For both few-shot templates with samples, those examples from chosen from the training dataset. When using examples from the test dataset, we find that the LLM would remember those demonstrations and return the correct response for those examples.

3.5.2 Experiment Setting

We use the API provided by OpenAI [1] to test the LLM's performance on the task of persuasive technique detection. This API has two important parameters. The first parameter is the version of the model. In this experiment, we use the stable ChatGPT version *gpt-3.5-turbo*. The second parameter is the temperature. The value of temperature could be set between 0 and 2. The higher the temperature, the more random the outcomes would be. Since our task type is text classification, we set the temperature to zero so that the response could be consistent and deterministic.

3.5.3 Result

Even with different prompts adopted, ChatGPT cannot beat the previous SOTA model MinD [32] on the task of persuasive technique detection. As shown in Table 3.2, the micro F1 scores for ChatGPT are 0.2706, 0.2746, 0.2751, and 0.2970 respectively when using the zero-shot prompt, CoT prompt, few-shot prompt with random samples and few-shot prompt with selective samples. The micro F1 score of the previous SOTA MinD is 0.5933.

The different prompt strategies here don't exhibit a great impact on ChatGPT's performance on this task. Only a few-shot prompt with selective samples where ChatGPT makes mistakes using the CoT prompt. The highest micro F1 score from ChatGPT is 0.2970, far below the SOTA model MinD's score of 0.5933. Compared with ChatGPT's impressive performance on sentiment analysis, stance detection, and inference tasks, ChatGPT cannot handle this niche task with the same sophistication. Persuasive technique detection is not as widely studied and applied in the industry compared with those tasks where ChatGPT prevails. Previous research also shows that when the data is more likely to be used during the pre-training of ChatGPT,

[1] https://openai.com/

Table 3.2: ChatGPT's Performance with Different Prompting Strategiess

Technique	**Plain**	**CoT**	**Few-shot 1**	**Few-shot 2**	**MinD**
Loaded Language	0.5787	0.6791	0.6952	0.7232	0.8190
Name calling	0.3750	0.4854	0.5362	0.5123	0.6666
Smears	0.3721	0.3894	0.4068	0.4098	0.5113
Doubt	0.2647	0.3846	0.3448	0.3571	0.4000
Slogans	0.2800	0.2482	0.2626	0.2242	0.1538
Exaggeration	0.2529	0.2645	0.2424	0.2207	0.5500
Virtue	0.0800	0.1538	0.2069	0.2143	0.2857
Whataboutism	0.4138	0.3226	0.2667	0.1772	0.3750
Appeal to fear/prejudice	0.2500	0.2308	0.2105	0.2338	0.5217
Black-and-white	0.1053	0.1316	0.1136	0.1905	0.4000
Appeal to authority	0.2857	0.2308	0.2941	0.2857	0.0000
Flag-waving	0.1333	0.1538	0.2273	0.2500	0.6154
Cliche	0.1250	0.0952	0.0698	0.0714	0.0000
Red Herring	0.0594	0.0571	0.0952	0.0606	0.0000
Reductio ad hitlerum	0.2500	0.1176	0.0769	0.1176	0.0000
Oversimplification	0.0645	0.0506	0.0412	0.0182	0.5000
Obfuscation	0.0000	0.0000	0.0164	0.0233	0.0000
Repetition	0.0260	0.0247	0.0000	0.0000	0.0000
Bandwagon	0.0000	0.0000	0.2222	0.0000	0.0000
Straw Man	0.0541	0.0714	0.0299	0.0377	0.0000
MICRO F1	0.2706	0.2746	0.2751	0.2970	0.5933
MACRO F1	0.1985	0.2046	0.2179	0.2073	0.2899

ChatGPT is more likely to perform well on those tasks. Therefore, the small size of the persuasive technique detection dataset and its narrow application may contribute to ChatGPT's moderate performance on this task.

3.6 A Divide-and-Conquer Framework

Could we use simpler models? Is the external dataset necessary for the detection of all persuasive techniques? Could we have a less expensive way to generate additional training data? To solve those questions, we propose a new framework to develop a model to detect persuasive techniques in textual content.

Previous SOTA MinD uses a multi-label ensemble model comprised of several BERT-styles models with external data resources labeled by humans. Since there are 20 techniques and each technique relies on different features to be identified, we assume that a divide-and-conquer approach that deals with the persuasive techniques individually could yield good performance since each binary classifier could extract features pertinent to that technique. MinD uses the external data for training purposes. However, the distribution of samples of each persuasive technique is quite imbalanced. Our framework initially trains the classifiers without additional datasets to examine whether an external dataset is needed from every technique. What's more, even though ChatGPT is not comparable with fine-tuned BERT-style models on the task of persuasive technique detection directly, it could be used to generate samples for the training purpose of those binary classifiers, which costs much less than human annotation. Based on those assumptions, we propose a new framework to develop a new model that could achieve better performance than the previous SOTA MinD.

3.6.1 Framework

Our framework is illustrated in Figure 3.1. In the first step, we want to use simple binary classifiers for each persuasive technique instead of the ensemble model in MinD [32]. To examine whether the external dataset used by MinD [32] is needed to achieve comparable performance, we only use the original training dataset from SemEval 2021 task 6 without external resources. In our experiment, the binary classifier uses RoBERTa as its backbone and we use the validation data to select the best model. The maximum number of training epochs is 15. We do not necessarily choose the model derived after all epochs. Instead, we choose the best model from the epoch where the model gets the highest F1 score. After the first step, we would get 20 binary classifiers for each persuasive technique.

Then we compare the 20 binary classifiers with the previous SOTA MinD. For those binary classifiers that have better or comparable performance, we would use them as the final classifiers. However, for those binary classifiers outperformed by MinD, we need to retrain them. In the retraining process, we need additional training datasets for data augmentation. In this step, we differ from MinD, which uses an external dataset annotated by humans by using ChatGPT to generate additional samples. This approach of data augmentation is less expensive in terms of both money and time, compared with human annotation. To generate samples as diverse as possible, we devise the prompt with specific instructions so that the generated samples could contain textual content across various topics. Specifically, in our experiment, ChatGPT is used to generate 50 samples of a target persuasive technique for the retraining purpose.

Finally, 20 binary classifiers would consist of the binary classifiers retained in the first step and the new classifiers from the retraining process. This framework is able to develop a model that outperforms the previous SOTA model.

The techniques where the binary models are outperformed by MinD are *Doubt*, *Exaggeration/Minimisation*, *Appeal to fear/prejudice*, *Black-and-white Fallacy/Dictatorship* and *Flag-waving*. For those techniques, we use ChatGPT to generate 50 additional training samples for each technique and retrain the binary classifiers for those techniques with those additional generated training samples. Finally, our framework could

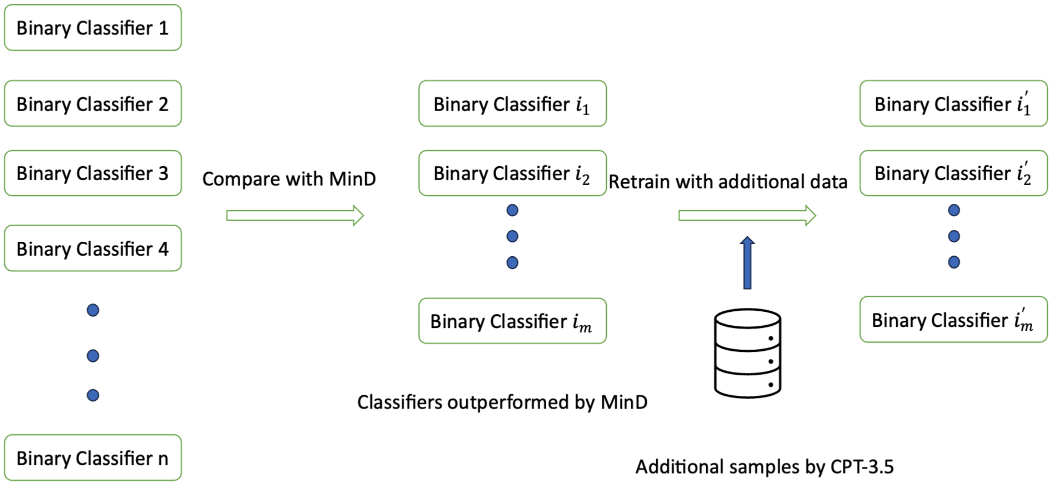


Figure 3.1: A Divide-and-Conquer Framework to Detect Persuasive Techniques

yield a model that beats MinD on performance.

3.6.2 Result and Analysis

We find that simple binary classifiers without additional could already beat or be comparable with the multi-label classification ensemble model of MinD on many techniques, showing that those binary models are good at extracting useful features to identify the specific technique, as shown in Table 3.3. What's more, those techniques include all three techniques with more than 100 training examples, indicating the data size needed to train a competent classification for those techniques. Those techniques are *Loaded Language*, *Name calling/Labeling*, and *Smears*. More interestingly, we find that for the technique of *Applal to authority*, the F1 score of the binary model without additional training examples is 0.615 while MinD only yields an F1 score of 0. This could potentially be caused by the diversion of the annotation of the external data source with regard to the persuasive technique of *Appeal to authority*. More training datasets may not necessarily produce greater performance if there is a data shift [49] between the original training dataset and the additional training dataset.

Table 3.3: Model Performance Comparison in Step 1

Technique	**Ours (Step 1)**	**MinD**
Loaded Language	0.8272	0.8190
Name calling/Labeling	0.6604	0.6666
Smears	0.5091	0.5113
Doubt	0.3913	**0.4000**
Slogans	0.2963	0.1538
Exaggeration/Minimisation	0.4138	**0.5500**
Glittering generalities	0.4211	0.2857
Whataboutism	0.5263	0.3750
Appeal to fear/prejudice	0.2609	**0.5217**
Black-and-white Fallacy	0.2500	**0.4000**
Appeal to authority	0.6154	0.0000
Flag-waving	0.4615	**0.6154**
Thought-terminating cliche	0.0000	0.0000
Red Herring	0.0000	0.0000
Reductio ad hitlerum	0.0000	0.0000
Causal Oversimplification	0.5455	0.5000
Obfuscation	0.0000	0.0000
Repetition	0.0000	0.0000
Bandwagon	0.0000	0.0000
Straw Man	0.0000	0.0000
MICRO F1	0.5804	0.5933
MACRO F1	0.3089	0.2899

Samples generated by ChatGPT could be quite simplistic and similar if not given the proper prompt. If the prompt is just asking ChatGPT to generate a certain number of persuasive technique, it will only generate simple sentences mostly of one topic and the topic tend to be commercial. What's more, if the prompt instructs ChatGPT to generate sentences, those sentences tend to be short, and not consistent with the samples shown in the original dataset. The prompt has to be designed in a clear way demanding ChatGPT to generate diverse short paragraphs across various topics so that ChatGPT could generate more diverse training examples beyond the commercial category. We still use the API provided by OpenAI and choose the stable ChatGPT version *gpt-3.5-turbo*. The temperature is set at 0.8 so that the generated samples could be diverse. Here we list two prompts and some samples generated with them by ChatGPT to show the observations we have.

Prompt without instruction to generate samples across various topics.

This is the definition of the persuasive technique of Exaggeration or Minimisation: Either representing something in an excessive manner, e.g., making things larger, better, worse ("the best of the best", "quality guaranteed"), or making something seem less important or smaller than it really is, e.g., saying that an insult was just a joke. Generate 20 examples of sentences that uses the persuasive technique exaggeration of minimisation. Those examples are created to train a classification model. So those examples should be as diverse as possible.

Examples of generated samples.

- This product will change your life forever.
- This car is the safest vehicle on the planet.
- Our cleaning service guarantees a spotless home every time.
- You won't believe the incredible taste of our chocolate cake.
- This app is the ultimate solution to all your problems.

- Our team of experts will transform your business overnight.
- These shoes will make you run faster than the speed of light.
- Our customer service is second to none.
- This diet plan will make you lose weight effortlessly.
- Our shampoo will give you hair like a superstar.

Prompt with instruction to generate samples across various topics. *This is the definition of the persuasive technique of Exaggeration or Minimisation: Either representing something in an excessive manner, e.g., making things larger, better, worse ("the best of the best", "quality guaranteed"), or making something seem less important or smaller than it really is, e.g., saying that an insult was just a joke. Generate 10 examples of short paragraphs that uses the persuasive technique Exaggeration or Minimisation. Those topics of those examples can be news, politics, comments, commercial, etc. Those examples are created to train a classification model. So those examples should be as diverse and distinct from each other as possible.*

Examples of generated samples.

- The economic downturn is being portrayed as a minor hiccup, but the truth is, it's a seismic shock that threatens the livelihoods of millions. Let's not downplay the gravity of the situation and work towards comprehensive solutions.
- The political scandal is being brushed off as a simple misunderstanding, but it's a blatant betrayal of public trust and a stain on democratic values. Let's not trivialize the importance of accountability and transparency in governance.
- The environmental crisis is being undermined as an overblown concern, but it's a looming catastrophe that demands immediate action. Let's not underestimate the irreversible damage that our planet is facing and take responsibility for sustainable practices.

- The controversial policy is being downplayed as a minor adjustment, but it jeopardizes fundamental rights and freedoms. Let's not diminish the impact it has on marginalized communities and stand up for justice and equality.

- The hate speech is being dismissed as harmless words, but it perpetuates discrimination and fosters a climate of fear and division. Let's not minimize the harm caused by such rhetoric and strive for a more inclusive society.

- The healthcare crisis is being portrayed as manageable, but it's an urgent matter that affects the well-being of countless individuals. Let's not overlook the struggles faced by those without access to quality care and advocate for comprehensive reforms.

- The corruption scandal is being framed as isolated incidents, but they expose a deep-rooted culture of unethical behavior. Let's not ignore the systemic issues that allow corruption to thrive and demand transparency and accountability.

- The human rights violation is being treated as minor infractions, but they are gross violations of dignity and freedom. Let's not undermine the suffering endured by victims and work towards a world where human rights are protected for all.

- The income inequality is being diminished as a natural outcome of a meritocratic system, but it perpetuates social disparities and hinders social mobility. Let's not turn a blind eye to the growing wealth gap and advocate for equitable opportunities.

- The global refugee crisis is being minimized as a regional issue, but it's a humanitarian emergency of staggering proportions. Let's not forget the plight of displaced individuals and strive for compassionate responses and sustainable solutions.

For the techniques where the ensemble model with external dataset beats the initial binary classifiers, a new binary classifier trained with additional training samples generated by ChatGPT could achieve comparable or better performance as the previous SOTA which uses external data source that requires expensive human annotation. As shown in Table 3.4, for the techniques *Doubt*, *Exaggeration/Minimisation*, *Appeal to authority*, *Black-and-white Fallacy/Dictatorship* and *Flag-waving*, if we retrain new binary models for them with the additional diverse training examples generated by ChatGPT, we could yield better or comparable F1 scores as MinD.

Table 3.4: Model Performance Comparison in the Retrain Step

Technique	**Ours (Retrain)**	**MinD**
Doubt	**0.4000**	**0.4000**
Exaggeration/Minimisation	**0.5500**	**0.5500**
Appeal to fear/prejudice	**0.5455**	0.5217
Black-and-white Fallacy/Dictatorship	**0.4285**	0.4000
Flag-waving	**0.6666**	0.6154

Overall, our framework could develop a model that outperforms the previous SOTA MinD at a lower cost.

3.7 Conclusion

In this Chapter, we aim to develop a better tool to identify persuasive techniques in the textual content posted by government-affiliated accounts.

We first show that ChatGPT could not beat fine-tuned BERT-style models on the task of persuasive techniques detection even with different prompt strategies explored. The performance only varies slightly when different prompts are provided to ChatGPT for the detection task, its performance is still far worse than the fine-tuned models.

We also propose a new framework to answer the three questions.

1. Could we use simpler models?

2. Is the external dataset necessary for the detection of all persuasive techniques?
3. Could we have a less expensive way to generate additional training data?

We first demonstrate the binary models that deal with each technique prove to be good at identifying a single technique even without extra training samples, showing their capacity at extracting features pertinent to the specific technique. The external dataset is not needed for each persuasive technique when developing the model to achieve the comparable performance of the previous SOTA. We also noticed that for certain persuasive techniques, the utilization of an external dataset annotated dataset doesn't necessarily improve the performance. This could be caused by the data shift among different datasets for the specific persuasive technique.

What's more, though ChatGPT could not beat the fine-tuned BERT-style models, it is a useful tool in terms of generating additional training examples for those techniques where the initial binary models fall behind MinD. We find that without giving specific instructions, ChatGPT tends to generate samples from the commercial category. To generate diverse samples from various topics, specific instructions have to be given in the prompt.

Overall, the new framework could develop a better model for identifying persuasive techniques at a lower cost.

CHAPTER 4: A Comparative Study of the Chinese Government's Use of Domestic and International Online Social Networks

4.1 Introduction

Driven by the fact that people use social media for news access and political discourse, governments around the world attempt to engage with domestic audience and international audience through online social media. The prevalence of social media has made it possible for both state and non-state actors to reach a large audience at smaller expense. Study has found that almost all members of parliaments have Twitter accounts in 32 European countries of European Union, the European Free Trade Association and the United Kingdom [4]. In the United States, all Senators and almost all Representatives adopt Twitter accounts as well [5]. Although Twitter is blocked in China, the number of China's diplomatic Twitter accounts increased to 80 as of January 17, 2020 from 13 as of October 20, 2018 [6]. Notably, most of the Chinese diplomats with Twitter accounts do not have Weibo accounts.

Chinese government has presence on both domestic online social networks and international social networks. China's diplomats employ Twitter as an arena for public diplomacy. The presence of China's diplomatic Twitter accounts align with the goal to "tell China's story well" [50], set by China's president Xi Jinping back in 2013 [6]. Public diplomacy is a term first coined in 1965 by Edmund Gullion [51], it means a direct communication initiated by one government to non-state actors to influence their government. A government can use public diplomacy to build relationship with non-state actors and their government, promote a positive image and perception of itself and enhance understanding from the outside, thus creating a favorable environment to implement its diplomatic policies [6, 10].The prevalence of social media provided chance for diplomats to engage with more non-state actors from outside than they could in the traditional channels like in-person activities. Social media, with a core of interactivity [52], not only provide an additional channel for public diplomacy, but also makes once a predominantly one-way communication more interactive [20]. Domestically, China's government uses Weibo to for various purposes, such as information provision, disseminating educative information to a city's newcomers, dealing with crime and enforcing law [14], and promoting the government's image by requiring government employees to post pro-government content [15].

This dissertation aims to study the Chinese government's use of online social networks, both within and without China.

Although Twitter is blocked along with other international social networks and news outlets within China, China's diplomats are active on Twitter. A better understanding f the government's behavior could lead to people making better decisions [53]. This paper contributes to the understanding the similarities and difference in the Chinese government's use of domestic and international online social networks, with a focus on Zhao Lijian, who is prolific on both Twitter and Weibo. We aim to answer the following research questions:

- RQ1: What topics characterize Zhao's posts on Twitter and Weibo?
- RQ2: What sentiments characterize Zhao's posts on Twitter and Weibo?
- RQ3: What are the factors affecting people's engagement with Zhao's posts on Twitter and Weibo?

4.2 Related Work

Study regarding China's public diplomacy on Twitter captures its characteristics of network structure and content dynamics both in general and in specific events. Despite the fact that some Chinese missions communicate and interact with foreign audiences and counterparts while some do not, they share a common primary information source - the China's state-owned media outlets, indicating a hierachical structure of the communication network within the China's diplomatic active Twitter accounts. The top three topics most frequently tweeted about are the China-foreign cooperation in economic and social aspects, political relationship with foreign countries and promotion of Chinese culture and society [20], most of which are moderate and trying a create a friendly international environment for China. Study also shows that during the U.S.-China Trade War, even though China's related posts on Weibo are generally showing a tough attitude, China's Twitter posts are relatively showing a softer gesture [6]. However, there is one exception. Lijian Zhao, the current spokesperson of the Chinese Ministry of Foreign Affairs and previous diplomat affiliated with the Chinese embassy in Pakistan, are actively tweeting contents with characteristics of polarization and strong emotion [10]. Lijian Zhao tweets in a way not aligned with the traditional purposes of public diplomacy, which generally are trying to boost mutual understanding, promote nation branding and create friendly international environment for home country's foreign policies [22]. Lijian Zhao's proactive tweeting style is integral to China's diplomacy's shift from forbearance, softness to proactivity and assertiveness [21].

Regarding Weibo, scholars reveals specific strategies and methods government use to influence public opinion. Evidence has been found that China's government asks government employees to post pro-government content on the platform, contrary to previous claims that China recruits people with money to post pro-government content [15]. One case study finds the Beijing Police Department (BPD) uses Weibo to disseminate *positive energy*, a phrase referring to healthy, active emotions and attitudes, which is consistent with Xi administration's ideological agenda[14]. On Weibo, scholars have found little evidence that China's government uses bots to disseminate information or to fake popularity [29, 14].

In terms of comparison study, Huang and Wang [6] find that China's diplomatic posts on both Twitter and Weibo use information source from major Chinese state owned news outlets, exhibit hierarchical structure and that posts on Weibo tends to manifest a harder attitude on diplomatic issues than on Twitter. Due to the fact many diplomats do not have Weibo accounts, Huang and Wang [6] collects data from 80 Twitter accounts and only 2 Weibo accounts. Such comparison study helps people better understand the government's behavior, especially considering that Twitter is blocked along with other social networks in China.

By examining the difference/similarity of the content and engagement of the posts of the spokesperson for China's Ministry of Foreign Affairs with the help of NLP tools, our analysis provides more information to better infer the government's agenda behind the posts delivered or not delivered to the China's citizens. We also provide additonal quantitative evidence to previous qualitative findings in this area.

4.3 Dataset

We collected 515 posts on Weibo and 3233 posts on Twitter from February 20, 2021 to September 17, 2021, 2310 posts on Weibo and 3239 posts on Twitter from November 2, 2021 to May 22, 2022. We conduct the comparison analysis in two time periods to see whether the findings hold across time. The numbers are summarized in Table 4.1. Weibo allows users to comments when they repost other people's posts. On the other hand, Twitter doesn't allow long post. Twitter's API allows collecting a user's Tweets, Retweets, replies and Quote Tweets up to around 3200. Actually, more than 70% of Lijian Zhao's timeline posts are just Retweets.

We mainly focus on analyzing the original text by Lijian Zhao. On Weibo, it includes the comments when he repost and the his original post. To analyze Zhao Lijian's Weibo content, we use his comments when he reposts and his original Weibo posts. On Twitter, we use the text of his Tweets, replies and Quote Tweets. In the following of this paper, unless explicitly stated, the posts refer to those with Zhao

Lijian's original text.

In the rest of the paper, we use time period 1 and time period 2 to refer to February 20, 2021 to September 17, 2021 and November 2, 2021 to May 22, 2022 for convenience.

Table 4.1: Data Summary

Platform	Time Period 1	Time Period 2
Weibo	515	2308
Twitter	3233	3239
Weibo (w/ original text)	515	2308
Twitter (w/ original text)	863	922

4.4 Research Questions

RQ1: What topics characterize Zhao's posts on Twitter and Weibo? Do they exhibit difference across the two platforms?

Methods. To understand the topics, we use Named Entity Recognition to see what objects Zhao Lijian is frequently talking about on social media. We use pre-trained models from PaddleNLP [54] and Flair [55] to process Chinese text and English text respectively. We also look into the way of wording when Lijian Zhao described certain entities by visualizing the frequent co-occurrence words with those entities.

Results. We list below the top 10 entities mentioned by Lijian Zhao Below.

Top 10 entities mentioned by Zhao Lijian February 20, 2021 to September 17, 2021,

- **Weibo**: *(1) China, (2) the United States, (3) The Ministry of Foreign Affairs, (4) Zhao Lijian, (5) Xinjiang, (6) Covid, (7) Wang Yi , (8) Japan, (9) Afghanistan, (10) the World Heath Organization.*

- **Twitter**: *(1) China, (2) the United States, (3) Xinjiang, (4) Pakistan, (5) Afghanistan, (6) Wang Yi, (7) Japan, (8) Covid, (9) UK, (10) Taiwan.*

Top 10 entities mentioned by Zhao Lijian from November 2, 2021 to May 22, 2022,

- **Weibo**: *(1) China, (2) the United States, (3) The Ministry of Foreign Affairs, (4) Wang Yi, (5) Zhao Lijian, (6) Covid, (7) Ukraine , (8) Beijing, (9) Xinjiang, (10) the Winter Olympic Games.*

- **Twitter**: *(1) China, (2) the United States, (3) Xinjiang, (4) Beijing, (5) the Winter Olympic Games, (6) Ukraine, (7) Afghanistan, (8) Lithuania, (9) Pakistan, (10) Wang Yi.*

Zhao Lijian have gives different priorities of certain topics on Weibo and Twitter. While the United States is the most frequent mentioned entity on both platforms by Zhao Lijian,the Xinjiang issue is given higher priority on Twitter than on Weibo. In both time periods of observation, Xinjiang remains the third most frequent entity mentioned by Zhao Lijian, while on Weibo, it is the fifth and ninth most frequently mentioned entity in the two time periods. In the second time period, during which the 2022 Beijing Winter Olympic Games took place and the Ukraine War broke out, on Weibo, Xinjiang is mentioned less than both of these two events while on Twitter, it keeps the third place.

Lijian Zhao acts more like a broadcaster on Weibo. On Weibo, his comments composes of the headline of the reposted content along with his additional words. He also reposts a lot from the Spokesperson's Office account, whose posts usually start with "*Zhao Lijian responded to ...*" or "*Zhao Lijian said ...*". That's why Zhao Lijian himself is a frequent mentioned entity of his own text. On the other hand, Zhao Lijian comment more directly on Twitter rather than just repeating the headline of the quoted Tweets and replied Tweets.

RQ2: What sentiments characterize Zhao's posts on Twitter and Weibo? Do they exhibit difference across the two platforms?

Methods. We use pre-trained models which have been fine-tuned on sentiment tasks provided by Cardiff NLP [56] and PaddleNLP [54] to classify the sentiment on English text and Chinese text respectively. We manually sampled and annotated the results and found both models yield accuracy above 0.8.

Results. Zhao Lijian post significant portion of negative text on both platforms. In time period 1, there are over half of the posts are negative on both platforms (62.52% and 50.52% on Weibo and Twitter respectively). In time period 2, the ratios of negative posts drops to 40.42% and 30.8%. The numbers are shown in Table 4.2. Such a high ratio of negative posts on Twitter resonates with previous findings that Zhao Lijian posts in an aggressive way inconsistent with the traditional goal of public diplomacy to create a favorable international environment for diplomatic policies.

The ratio of negative posts on Weibo is higher on Weibo than on Twitter in both time periods. Zhao Lijian is even more aggressive on Weibo, cricitizing Western Countries especially the United States. There is claim that China's citizens welcome the so-called "Wolf Warrior" diplomacy, so Zhao may need to behave more aggressively to align with the public opinion to consolidate the government's legitimacy in the domestic platform [57].

Zhao Lijian has distinct ways of posting about *friends* and *enemies*. With keywords search, we looked into posts regarding the United States and Pakistan and found extremely high ratios of negative sentiment and positive sentiment respectively. The numbers of time period 2 are shown in Table 4.3. Results of time period 1 are similar. Zhao Lijian constantly criticizes the United States for its intervention with the China's internal affairs while show friendship and mutual support between China and Pakistan. This kind of finding pattern is also found in previous research by Guo [10]. Our sentiment analysis provides more evidence to this claim and show that this pattern is consistent on both platforms.

RQ3: What are the factors affecting people's engagement with Zhao's posts on Twitter and Weibo? Do they exhibit difference across the two platforms?

Table 4.2: Ratio of Negative and Non-negative Posts on Weibo and Twitter

Platform	**Time Period 1**		**Time Period 2**	
	Non-negative	**Negative**	**Non-negative**	**Negative**
Weibo	37.48%	62.52%	59.58%	40.42%
Twitter	49.48%	50.52%	69.2%	30.8%

Table 4.3: Ratio of Negative and Non-negative Pakistan and U.S. Related Posts on Weibo and Twitter (time period 2)

Platform		**Count**	**Non-negative**	**Negative**
Weibo	Pakistan	85	78.82%	21.18%
	U.S.	631	20.5%	79.46%
Twitter	Pakistan	20	85%	15%
	U.S.	243	34.98%	65.02%

Considering that Zhao Lijian posts frequently about the United States and posts a lot in negative tone. We explore whether the posts charged with negative emotions and posts related to the United States invokes more engagement.

Methods. On Weibo, we calculated the average number of favorite, forward and comment of negative and non-negative posts. On Twitter, we calculated the the average number of Retweet and favorite. Twitter API doesn't provide information of the number of comments to ordinary users.

Results. On Weibo, we find users engage more with non-U.S. content and non-negative content. The metrics on Weibo are shown in Table 4.4 and Table 4.5. In time period 1, 7 out of 10 posts with most favorites are positive. in time period 2, 9 out of 10 posts with most favorites are positive. However, there is no consistent pattern across the two time periods on Twitter. Notably, on Weibo, in both time periods, the top 1 posts with most favorites are positive and contain beautiful natural scenes. While on Twitter, one criticizes the West and the other one criticizes Japan's handling of the nuclear water.

Table 4.4: Engagement on Weibo in Time Period 1

Metric	**Non-negative**	**Negative**	**Non-U.S.**	**U.S.**
favorite	9053	5152	7558	5174
forward	524	315	427	343
comment	509	313	444	298

Table 4.5: Engagement on Weibo in Time Period 2

Metric	**Non-negative**	**Negative**	**Non-U.S.**	**U.S.**
favorite	4146	2600	3944	2397
forward	322	183	297	183
comment	229	198	227	189

4.5 Conclusion

In this Chapter, we examined Zhao Lijian's use of Weibo and Twitter with a focus on his own text, with the help of pre-trained NLP models. We find both similarities and difference of the content posted by Zhao Lijian on Weibo and Twitter. Xinjiang is given higher priority on Twitter than on Weibo while the United States is the biggest target on both platforms. We also find that Zhao Lijian post a larger ratio of negative content on Weibo than on Twitter. However, on both platforms, he constantly post negatively about the United States and positively about Pakistan. At last, we find that Weibo users engage more with non-U.S. content and postive content while there is no clear pattern of engagement on Twitter. This Chapter has been published in ASONAM 2022 [58][1].

[1] ©2022 IEEE. Reprinted, with permission from [Z. Cheng and S. Shaikh, "A comparative study of china's foreign ministry spokesperson's use of weibo and twitter," in 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 552 - 555, IEEE, 2022.]

CHAPTER 5: Towards Understanding the Public Stance towards the Chinese Government's Use of Online Social Networks

5.1 Introduction

Engagement is the responses of users to posts from other users on online social networks [19]. Engagement with the Chinese government's accounts can, on the one hand, can expand the reach of the government's information dissemination. On the other hand, it also reveals whether the government achieves its agenda on online social networks, such as promoting its image and creating a friendly environment for its foreign policies. Existing research focuses on studying the authenticity of the engagement with the Chinese government's accounts.

Twitter has been taking down state-linked accounts of the PRC since 2018[1] and suspending accounts that are suspected of being used to amplify the Chinese government's information on Twitter [59, 60]. Those suspended accounts make up for around 10% of the PRC diplomat retweets [19].

Despite all the claims in the existing research that there is evidence the Chinese government is using inauthentic approaches to promote its online content, the evidence supporting those claims is not bulletproof. It's challenging to distinguish accounts genuinely posting patriotic content from accounts that are controlled by the Chinese government. The ground truth of authenticity is difficult to get and current research relies on the suspended accounts disclosed by Twitter to find inauthentic accounts [19].

Existing research hasn't examined the content of engagement with Chinese online

[1] https://blog.twitter.com/en_us/topics/company/2021/disclosing-state-linked-information-operations-we-ve-removed

social networks in detail. The academic focus is on the inauthentic engagement by the Chinese government. But they still cannot come across the obstacle of getting the ground truth. Guo [10] finds that the comments Zhao Lijian receives on Twitter are mostly resistance and hate speech, charged with hatred and sarcasm through manual coding on the comments Zhao Lijian received. However, there is no public dataset or specific number attached to this paper. The evidence from a case study of Zhao Lijian also limits the generalization of the finding. Another stream of research investigates how people's opinion expression is influenced [61] under the widespread and sophisticated censorship within China [62, 63, 64, 65]. However, the opinion outside China hasn't been explored in detail. To mitigate this gap, we propose to create a new corpus that investigates the characteristics of the comments the Chinese government received on Twitter. We provide a quantitative analysis of public opinion toward the Chinese government's posts from different perspectives. We aim to answer those research questions.

- Q1: are the replies generated by a large number of users or mainly by a small number of users?
- Q2: what languages are those replies written in?
- Q3: what stance characterizes the replies?

Q1: are the replies generated by a large number of users or mainly by a small number of users? Public diplomacy aims to build good relationships with people from other countries. The first step is to reach as many audiences as possible. To see whether the engagement involves a large number of users, we have to investigate the distribution of the number of replies by the users. If the replies are mainly generated by a small number, then the number of audiences those posts have reached is not as large as it seems. However, we want to make it clear here that even though some research uses the number of replies as a key feature to identify bot behavior, it does not necessarily indicate that the users generating a large number of replies are bots. They may just be enthusiastic users engaging with those government-affiliated accounts. Even if there are some accounts generating large numbers of replies, it cannot be used as evidence that the government uses bots to fake popularity.

Q2: which languages are the replies in? Since Twitter is blocked in China and the posts are mostly written in languages other than Chinese, it's obvious that those Tweets are not targeting Chinese citizens. Instead, those posts aim people outside China, especially citizens of other countries. This is also consistent with the goal of public diplomacy, which seeks to establish a good relationship with people from other countries. Therefore, this raises a question, do the posts by the Chinese government-affiliated accounts really reach the target audience as expected?

Q3: what sentiments are characteristic of the posts? Even if the posts by the Chinese government-affiliated accounts may reach the target audience, they may not receive desirable engagement. Public diplomacy tries to create a friendly international environment so that foreign policies could be supported. What kind of posts do the Chinese government-affiliated accounts receive? This research question seeks to understand public opinion towards those posts by identifying the sentiments of the replies to those posts.

5.2 Dataset

We use the Twitter API V2 to collect replies to the original posts by the affiliated Chinese government accounts. Those accounts are provided by Alliance for Securing Democracy[2].

The original list we retrieve contains 264 accounts affiliated with the Chinese government, however, we use to Twitter API V2 to check whether those accounts are still available. After filtering out suspended or deleted accounts, we get 253 valid accounts remaining. Those accounts are divided into two categories, Diplomatic/Government

[2] https://securingdemocracy.gmfus.org

and Media, by Alliance for Securing Democracy. The accounts not longer valid are @chinaembturkey, @indurban1, @generalkonsuldu, @dr_zhaoqinghua, @cgchina_cpt, @wdonghua, @drzhaoyanbo, @cctvenespanol, @cntvfrancais, @fullframecgtn, @wangguanbeijing. Some examples of valid accounts are shown in Table 5.1. Typical Diplomatic/Government accounts are accounts of diplomats while typical Media accounts are those of the government news outlets and editors from those news outlets.

Table 5.1: Examples of Diplomatic/Government Accounts and Media Accounts Affiliated with the Chinese Government

User Account	**Category**
@ambcina	Diplomatic/Government
@chineambdjibout	Diplomatic/Government
@ambchineburundi	Diplomatic/Government
@ambchinecmr	Diplomatic/Government
@chineambassade	Diplomatic/Government
@pdchinalife	Media
@beijingreview	Media
@cctv	Media
@cctvasiapacific	Media
@cctvenespanol	Media

For those valid accounts, we collect the most recent original English out of the 3200 recent Tweets allowed by Twitter API. The data was retrieved on May 15, 2023. To identify the original Tweets, we make sure that *in_reply_to_user_id* and *referenced_tweets* are not keys of the Tweet object. What's more, we only collect Tweets where the value of the field *lang* is English. In the end, we successfully collect 209903 Tweets from those accounts, where the latest post date is May 15, 2023. Then we calculate the number of replies from those original posts and choose the top 10 accounts with the most replies. The accounts and the number of replies of those accounts are listed in Table 5.2. Noteworthy, the distribution of the number of replies is imbalanced among accounts, where huxijin_gt received more than half a million replies to his original Tweets while xinhuatravel receives less than 20,000 replies to its 3176 original Tweets.

We collect replies to the original posts from the top 10 accounts. Since the top 10 accounts receive more than 77% of the replies of all replies from all 253 valid accounts, representative of the overall population, we choose them as the study object and collect the replies from them. We use the ID of the original post as the Conversation ID in the search query to get replies to the original posts. We also make sure those replies are directly replying to the original post. The query doesn't necessarily return all replies implicated by the number of replies in Table 5.2. Finally, we collect 1,025,639 replies to the original English posts by the top 10 Chinese government-affiliated accounts. The statistics of the number of replies for each account are shown in Table 5.3.

Table 5.2: Top 10 Accounts with Most Replies

Username	**# of Original Posts**	**# of Replies**	**Average per Post**
huxijin_gt	2237	592944	265.0621368
spokespersonchn	2048	302998	147.9482422
mfa_china	1280	128437	100.3414063
zlj517	861	117129	136.0383275
liuxininbeijing	2482	35225	14.19218372
ambliuxiaoming	769	20596	26.78283485
globaltimesnews	2781	19345	6.956130888
chineseembinus	861	19261	22.37049942
xinhuatravel	3176	16405	5.165302267
shen_shiwei	1831	15904	8.685963954

5.3 Research Questions

After getting the dataset, we conduct a variety of analyses to answer the research questions we proposed. In this section, we will illustrate the experiments we have done and the conclusions drawn from the results.

Table 5.3: The Number of Replies Collected for the Top 10 Accounts

Username	**# of Replies Collected**
huxijin_gt	429041
spokespersonchn	276482
mfa_china	115967
zlj517	97803
liuxininbeijing	27347
ambliuxiaoming	22682
globaltimesnews	17697
chineseembinus	13436
xinhuatravel	10851
shen_shiwei	14333

5.3.1 RQ1: Are the Replies Mainly Created by a Small Number of Users?

Public diplomacy aims to reach a large number of audience to engage with so that a friendly international environment could be cultivated to promote the country's foreign policy. In our dataset, although the top 10 accounts receive more than one million replies, it does not equal to that they engage with one million people. The power law in the context of digital work estimate that 20% users generate 80% of the content [66]. Therefore, it's worthwhile to investigate whether those replies are generated by a disproportionately small number of users.

Methods. We first check how many users are involved in our collected replies. Then the Gini index, an index indicating the dispersion of the distribution of the replies, is derived. We then generate the Lorenz curve to show the distribution of the replies among users to check whether the replies are created by only a small number of users.

Result. The distribution of replies from users is quite similar to what's predicted by the power law. There are 290,569 authors of the 1,025,639 replies we collected. The index of the distribution of replies among users is 0.64, indicating a highly uneven distribution. In Figure 5.1, the green line indicates the scenario where each user contributes the same amount of work of replies. The black line is the Lorenz curve of the distribution of replies to the top 10 Chinese government-affiliated accounts. The black line is well below the even distribution line. In fact, the bottom 80% users contribute only 28.05% of the replies while the top 20% users contribute 71.95% of the replies. The top 5% users contribute almost 51% of the replies. Even though the top 10 Chinese government-affiliated accounts receive more than 1 million replies, nearly half of the replies are generated by the top 5% active users.

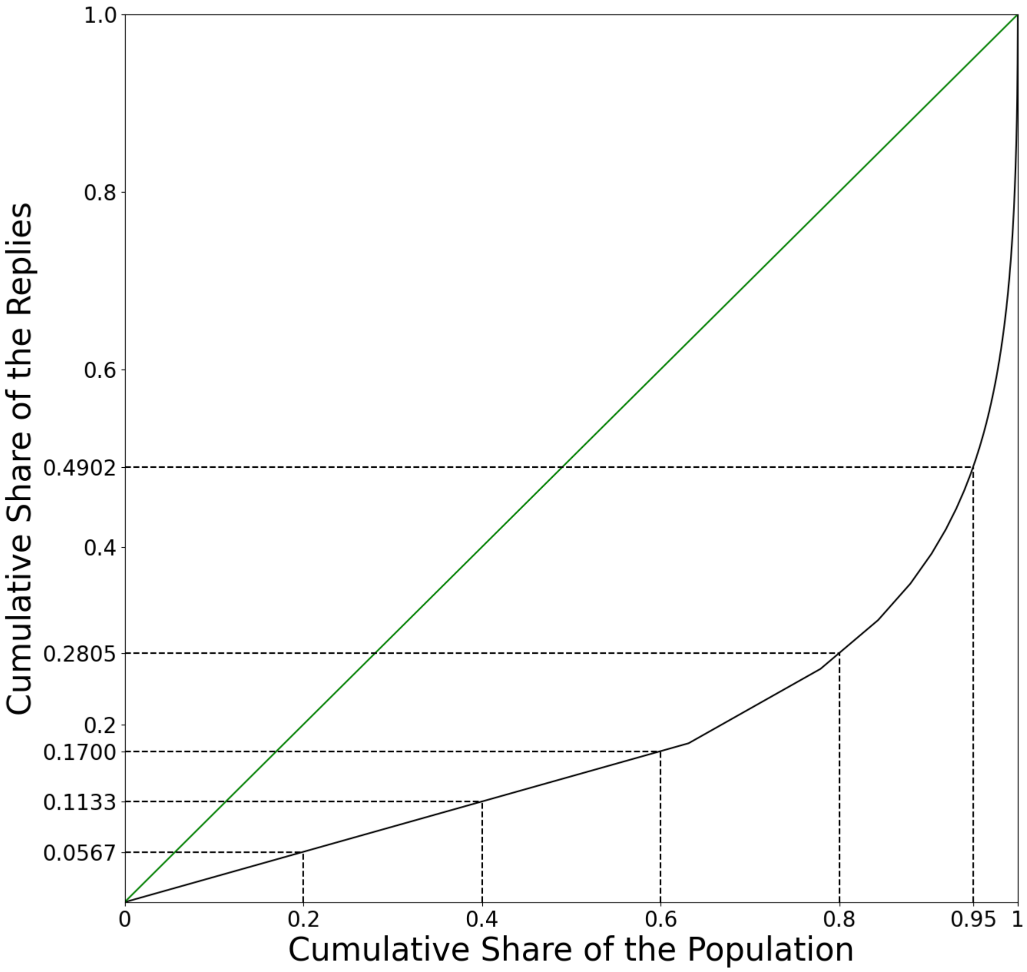


Figure 5.1: Lorenz Curve Depicting the Replies Distribution

5.3.2 RQ2: Which Languages Are the Replies in?

Public diplomacy aims to engage with citizens of other countries. This is why those government-affiliated accounts post the majority of their posts in various languages other than Chinese. What's more, Twitter is blocked in China. Contents posted by the Chinese government on it are therefore not targeting Chinese citizens. This raises the question of whether the government's posts successfully reach the target audience.

Methods. To answer this question, we examine which languages are those replies written in and the number of users for each language. We then generate the Lorenz curves for the major groups to see whether the distribution of replies is more uneven in some groups than others.

Result. Firstly, the majority of replies are in English and Chinese. There are 581,414 English replies and 234,573 Chinese replies. Then replies of links, Japanese, Spanish, Hindi, Indonesian, Tagalog, replies of mentions, and French. The number of replies in the top 10 languages/categories is shown in Table 5.4. The number of English and Chinese replies is significantly larger than any other language.

Table 5.4: The Number of Replies in Top 10 Languages/Categories

Lang	**Note**	**Count**
en	English	581414
zh	Chinese	234573
qme	Link	136741
ja	Japanese	28194
es	Spanish	4832
hi	Hindi	4729
in	Indonesian	3753
tl	Tagalog	3554
qam	Mention	2600
fr	French	1867

The average English replies and Chinese replies per author are 3.27 and 2.77 respectively, much higher than that of other languages. The numbers of authors of the replies in each language are shown in Table 5.5.

Table 5.5: Number of Authors of Replies in Top 8 Languages

Language	**Number of Replies**	**Number of Users**	**Average**
English	581414	177768	3.27
Chinese	234573	84625	2.77
Japanese	28194	18304	1.54
Spanish	4832	3561	1.35
Hindi	4729	3644	1.29
Indonesian	3753	3231	1.16
Tagalog	3554	3149	1.13
French	1867	1362	1.37

People who post Chinese replies also tend to post English replies. Since there are way more English replies and Chinese replies than in other languages, we focus on those two languages. We find that there are 14,454 users and 17.08% of the users that reply in Chinese also reply in English. What's more, those users generate 83675 Chinese and 114,141 English replies. To make sure the English replies are generated by citizens from countries other than China, we remove those English replies by users that reply in both Chinese and English. After removing those users and their replies in the English post, we have 467,273 English posts left and 163,314 users.

The distributions of work of replies are uneven for both Chinese and English replies. The distribution of the English replies is slightly more uneven than the Chinese replies. The index of the Chinese replies is 0.56 while the index of the English replies is 0.58. The top 5% users generate 45.75% of the English replies while the top 5% users generate 52.98% Chinese replies. The Lorenz curves of both Chinese and English replies are shown in Figure 5.2. They both fall under the curve of even distribution.

The Chinese replies are dominated by replies in simplified Chinese. As shown in Table 5.6. 89% of the Chinese replies are in simplified Chinese, much higher than the proportion of replies written in traditional Chinese, which accounts for 11% of the

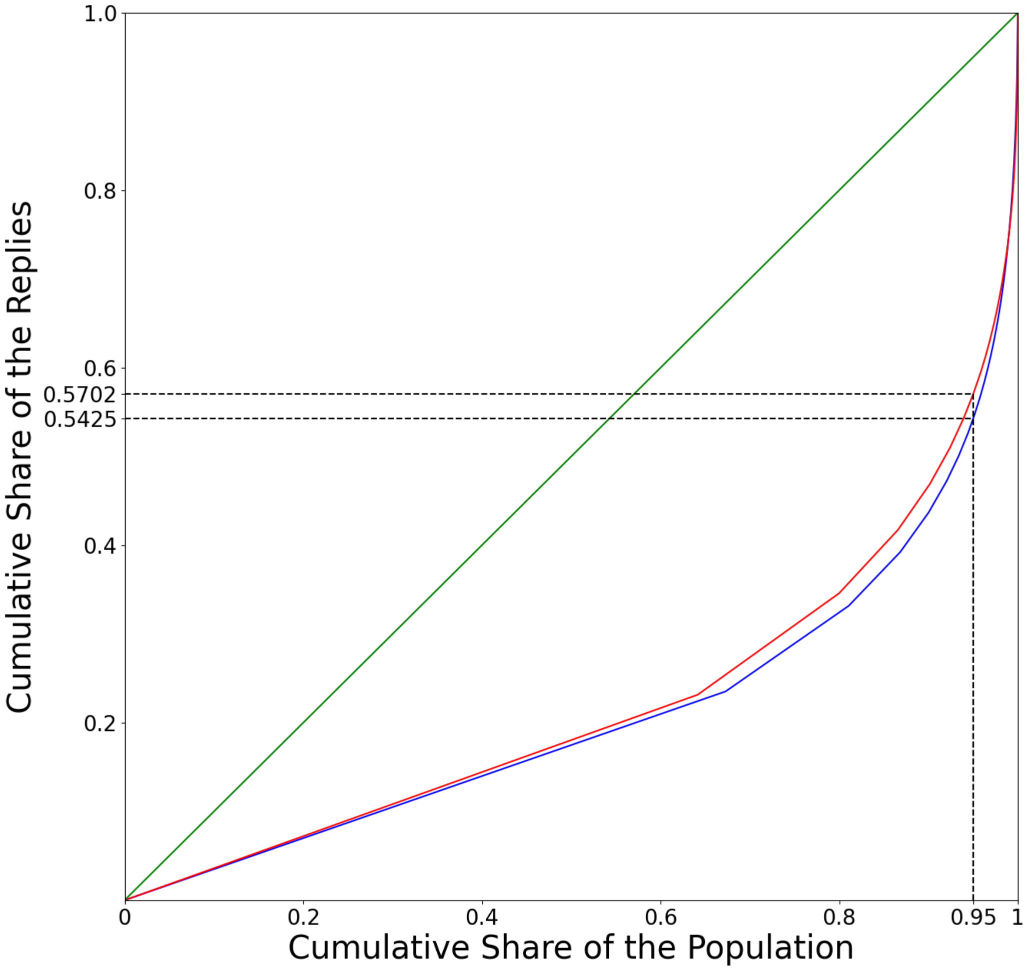


Figure 5.2: Lorenz Curve Depicting the English and Chinese Replies

Chinese replies.

Table 5.6: Proportion of Simplified-Chinese Replies and Traditional-Chinese Replies

Type of Chinese	**Number of Replies**	**Proportion of Chinese Replies**
Simplified Chinese	208634	89%
Traditional Chinese	25939	11%

5.3.3 RQ3: Which Stance Characterizes the Replies?

Public diplomacy aims to create a friendly international environment to promote the government's foreign policies. The Chinese government also aims to "tell China's story well". This research question attempts to understand how people perceive posts by Chinese government-affiliated accounts. We investigate the stance of replies in Chinese and English to see stances vary among people using different languages.

Methods. ChatGPT is used to identify the stance of the replies to the original English Tweets by the Chinese government-affiliated accounts. Existing study has shown that ChatGPT's performance in stance detection is comparable to or even outperforming previous models which are mainly fine-tuned on specific datasets [37]. We first randomly select the 1% of the replies in both English and Chinese replies. Then we create a prompt to let ChatGPT detect whether the reply is favoring or against the referenced Tweet. We feed ChatGPT with both the pairs of replies and Tweets instead of just asking ChatGPT to only detect the sentiment of replies because replies with negative sentiment could be also favoring the original Tweet. To make sure that our conclusion is robust, we repeat these process for another 4 times and we get similar conclusions.

Result.

As shown in Figure 5.3, in both Chinese and English replies, the proportions of replies with a stance against the Chinese government's affiliated accounts are dominant. Of the English replies, 88% of them are against the Chinese government-affiliated accounts and only 12% of the replies are favoring the posts. For the Chinese replies, 80% of the replies are against the Chinese government-affiliated accounts while 20% of the replies are favoring the Chinese government's posts.

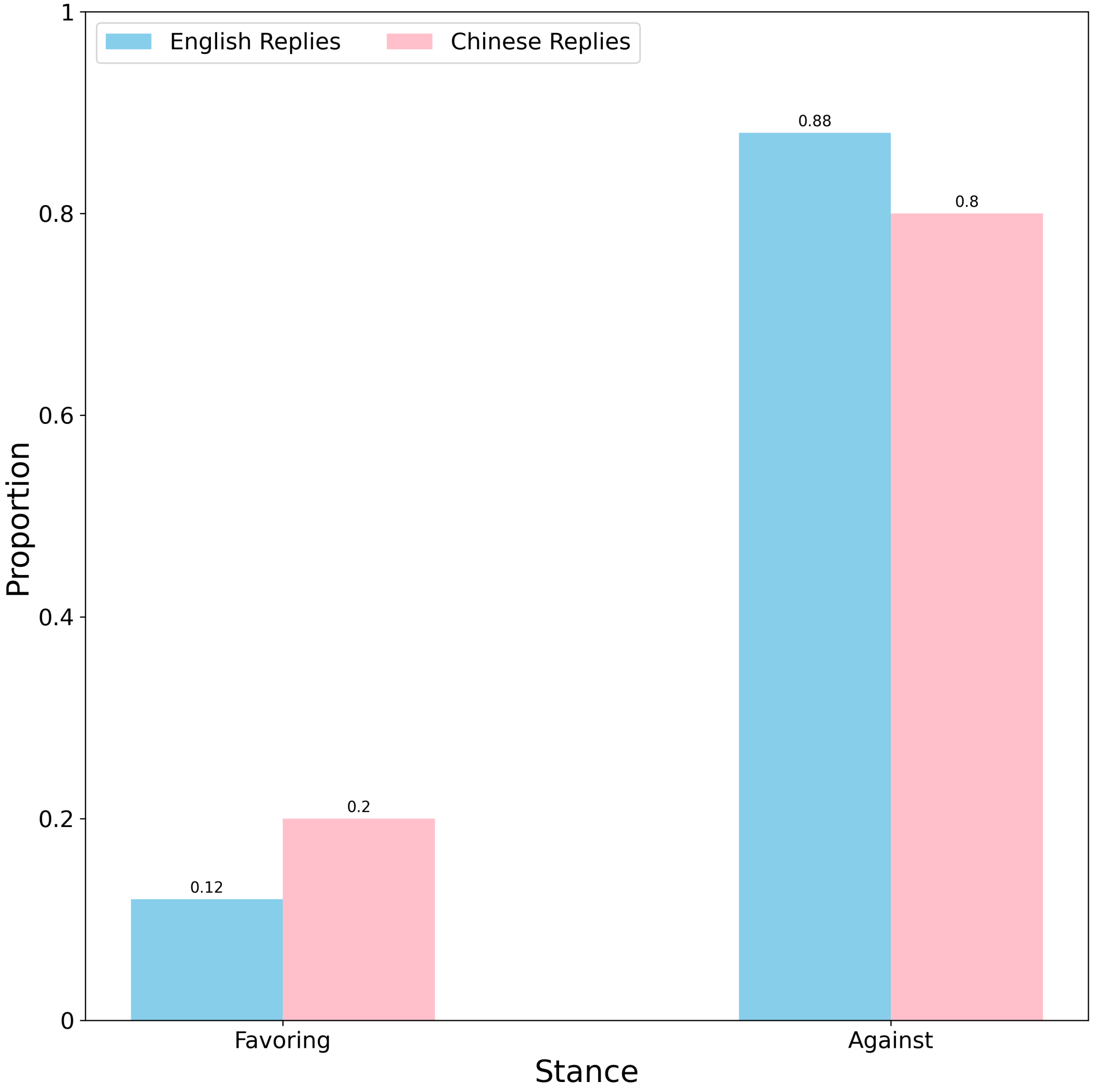


Figure 5.3: Stance Distribution of Chinese and English Replies

It's also noteworthy that even though the Chinese replies are also dominated by the stance against the Chinese government's posts, the proportion of replies favoring the government's posts is higher than the English reply. VPN has been known by people within China to access online social networks blocked by the Chinese government. Previous research has also shown that actually, the popularity of the government's policies is actually high among Chinese citizens [67]. This may account partially for the higher proportion of stances favoring the government's posts in the Chinese replies.

Within the Chinese replies, the replies in simplified Chinese replies have a higher proportion of replies favoring the Chinese government-affiliated accounts' posts. As shown in Table 5.7, there are 21% of replies in simplified Chinese favoring the government posts while only 14% of replies in traditional Chinese are favoring the Chinese government posts.

Table 5.7: Stance of Simplified-Chinese Replies and Traditional-Chinese Replies

Type of Chinese	**Proportion (Favoring)**	**Proportion (Against)**
Simplified Chinese	21	79%
Traditional Chinese	14	86%

5.4 Conclusion

This Chapter examines the stance of the replies to the posts by the Chinese government-affiliated accounts.

First, we find that the distribution of the work of replies among users is highly uneven, with the top 5% users generating almost half of the replies.

Regarding the languages those replies are written in, the dominant languages of the replies are English and Chinese, showing that a large portion of the audience of the posts are still people using Chinese even though those posts are written in English. Other languages mainly used in those replies include Japanese, Spanish, Hindi, Indonesian, Tagalog, and French, with only the number of Japanese replies surpassing 10,000, showing that Japanese users have a higher interest in Chinese-related topics compared with other language users. Within the Chinese replies, there are much more simplified-Chinese replies than traditional-Chinese replies.

Users that reply in Chinese also tend to reply in English sometimes. There are 14,454 users and 17.08% of the users that reply in Chinese also reply in English. What's more, those users generate 83675 Chinese and 114,141 English replies.

We compared the index and the Lorenz curves of the Chinese replies and English replies, finding that the distribution of the Chinese replies is more uneven compared with the English replies.

The English replies and Chinese replies are both dominated by the stance against the Chinese government's posts. However, the proportion of replies favoring the Chinese government's posts is higher in Chinese replies compared with English replies. Within the Chinese replies, the proportion of replies favoring the Chinese government's posts is higher in the simplified-Chinese replies than the traditional-Chinese replies.

CHAPTER 6: Conclusion

6.1 Conclusion

This dissertation helps improve people's understanding of the behavior of the Chinese government's use of online social networks by the following steps.

First, since existing research only provides qualitative analysis of topics of the Chinese government posts on online social networks, we propose that how those posts are written also provides new perspectives to understand the government's behavior. The list of persuasive techniques compile in former research provides exactly such a lens through which we can understand how the posts are written. Therefore, this dissertation presents a new framework to develop a better model to identify persuasive techniques. Compared with previous models, our approach uses simple binary classifiers instead of complex ensemble models. Our experiment shows that those binary classifiers are good at extracting features pertinent to the specific technique they target. It also shows that not every technique needs an external data source to train a good classifier. For some techniques, due to the data shift among datasets, the external data source may not help improve the performance at all. In terms of data augmentation, ChatGPT is used in this framework instead of external data sources annotated by humans. Compared with human annotation, ChatGPT excels in both monetary cost and time cost.

Secondly, inspired by the fact that the Chinese government is active on many online social networks which are blocked in China, we infer that the Chinese government may behave differently on different platforms because they are targeting different audiences. We, therefore, examined Zhao Lijian's use of Weibo and Twitter with a focus on his own text, with the help of pre-trained NLP models. We find both similarities and differences in the content posted by Zhao Lijian on Weibo and Twitter. Xinjiang is given higher priority on Twitter than on Weibo while the United States is the biggest target on both platforms. We also find that Zhao Lijian posts a larger ratio of negative content on Weibo than on Twitter. However, on both platforms, he constantly posts negatively about the United States and positively about Pakistan. At last, we find that Weibo users engage more with non-U.S. content and positive content while there is no clear pattern of engagement on Twitter.

At last, even though the government aims to tell China's story well, which is consistent with the goal of public diplomacy to create a friendly international environment to promote a country's diplomatic policies, the opinion on those online social networks hasn't been studied with quantitative methods in detail. This last section of the dissertation examines the stance of the replies to the posts by the Chinese government-affiliated accounts. First of all, we find that the top 10 popular government accounts receive most of the replies. Secondly, those replies are distributed highly unevenly, with the top 5% user generating almost half of the replies. Thirdly, we find that even though the replies are dominated by the stance against the government's posts, there is a higher proportion of replies favoring the government's posts in the Chinese replies than in the English replies. there is also a higher proportion of replies favoring the government's posts in the Chinese replies than in the English replies.

REFERENCES

[1] G. Da San Martino, S. Yu, A. Barrón-Cedeno, R. Petrov, and P. Nakov, "Fine-grained analysis of propaganda in news article," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5640–5650, 2019.

[2] R. Wang, D. Tang, N. Duan, W. Zhong, Z. Wei, X.-J. Huang, D. Jiang, and M. Zhou, "Leveraging declarative knowledge in text and first-order logic for fine-grained propaganda detection," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3895–3903, 2020.

[3] P. Vijayaraghavan and S. Vosoughi, "Tweetspin: Fine-grained propaganda detection in social media using multi-view representations," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3433–3448, 2022.

[4] M. Haman and M. Školník, "Politicians on social media. the online database of members of national parliaments on twitter," *Profesional de la información*, vol. 30, no. 2, 2021.

[5] J. M. Cook, "Twitter adoption and activity in u.s. legislatures: A 50-state study," vol. 61, no. 7, pp. 724–740. Publisher: SAGE Publications Inc.

[6] Z. A. Huang and R. Wang, "Exploring chinaâs digitalization of public diplomacy on weibo and twitter: A case study of the u.s.âchina trade war," vol. 15, no. 0, p. 28. Number: 0.

[7] C. I. N. I. Center, "The 49th statistical report on chinaâs internet development," 2022.

[8] L. Zheng and T. Zheng, "Innovation through social media in the public sector: Information and interactions," *Government information quarterly*, vol. 31, pp. S106–S117, 2014.

[9] C. I. N. I. Center, "The 45th statistical report on chinaâs internet development," 2020.

[10] J. Guo, "Crossing the 'great fire wall': A study with grounded theory examining how china uses twitter as a new battlefield for public diplomacy," *Journal of Public Diplomacy*, vol. 1, no. 2, pp. 49–74, 2021.

[11] L. Ma, "The diffusion of government microblogging: Evidence from chinese municipal police bureaus," *Public Management Review*, vol. 15, no. 2, pp. 288–309, 2013.

[12] L. Zheng, “Social media in chinese government: Drivers, challenges and capabilities,” *Government Information Quarterly*, vol. 30, no. 4, pp. 369–376, 2013.

[13] G. King, J. Pan, and M. E. Roberts, “How the chinese government fabricates social media posts for strategic distraction, not engaged argument,” *American political science review*, vol. 111, no. 3, pp. 484–501, 2017.

[14] Y. Luo and T. M. Harrison, “Examining daily activities and propaganda on government social media in china: A case study of weibo use by beijing police department,” in *DG.O2021: The 22nd Annual International Conference on Digital Government Research*, DG.O’21, pp. 89–103, Association for Computing Machinery.

[15] G. King, J. Pan, and M. E. Roberts, “How the chinese government fabricates social media posts for strategic distraction, not engaged argument,” vol. 111, no. 3, pp. 484–501. Publisher: Cambridge University Press.

[16] M. D. Dodd and S. J. Collins, “Public relations message strategies and public diplomacy 2.0: An empirical analysis using central-eastern european and western embassy twitter accounts,” *Public relations review*, vol. 43, no. 2, pp. 417–425, 2017.

[17] S. D. Collins, J. R. DeWitt, and R. K. LeFebvre, “Hashtag diplomacy: twitter as a tool for engaging in public diplomacy and promoting us foreign policy,” *Place branding and public diplomacy*, vol. 15, no. 2, pp. 78–96, 2019.

[18] N. Strauß, S. Kruikemeier, H. van der Meulen, and G. van Noort, “Digital diplomacy in gcc countries: Strategic communication of western embassies on twitter,” *Government Information Quarterly*, vol. 32, no. 4, pp. 369–379, 2015.

[19] M. Schliebs, H. Bailey, J. Bright, and P. N. Howard, “China’s public diplomacy operations: understanding engagement and inauthentic amplifications of PRC diplomats on facebook and twitter,” Publisher: Programme on Democracy and Technology, Oxford University.

[20] Z. A. Huang and R. Wang, “Building a network to ‘tell china stories well’: Chinese diplomatic communication strategies on twitter,” vol. 13, no. 0, p. 24. Number: 0.

[21] Z. A. Huang, “Wolf warrior and china’s digital public diplomacy during the covid-19 crisis,” *Place Branding and Public Diplomacy*, vol. 18, no. 1, pp. 37–40, 2022.

[22] M. Aronczyk, *Branding the Nation: The Global Business of National Identity*. OUP USA. Google-Books-ID: FQ8oAAAAQBAJ.

[23] G. Bolsover and P. Howard, “Chinese computational propaganda: automation, algorithms and the manipulation of information about chinese politics on twitter and weibo,” *Information, communication & society*, vol. 22, no. 14, pp. 2063–2080, 2019.

[24] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, “Truth of varying shades: Analyzing language in fake news and political fact-checking,” in *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 2931–2937, 2017.

[25] I. Habernal, H. Wachsmuth, I. Gurevych, and B. Stein, “Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 386–396, 2018.

[26] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[27] I. Habernal, R. Hannemann, C. Pollak, C. Klamm, P. Pauli, and I. Gurevych, “Argotario: Computational argumentation meets serious games,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 7–12, 2017.

[28] G. D. S. Martino, S. Shaar, Y. Zhang, S. Yu, A. Barrón-Cedeño, and P. Nakov, “Prta: A system to support the analysis of propaganda techniques in the news,” *arXiv preprint arXiv:2005.05854*, 2020.

[29] G. Bolsover and P. Howard, “Chinese computational propaganda: automation, algorithms and the manipulation of information about chinese politics on twitter and weibo,” vol. 22, no. 14, pp. 2063–2080. Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2018.1476576.

[30] S. C. Woolley and P. Howard, “Computational propaganda worldwide: Executive summary,” 2017.

[31] D. Kruckeberg and M. Vujnovic, “Public relations, not propaganda, for us public diplomacy in a post-9/11 world: Challenges and opportunities,” *Journal of Communication Management*, 2005.

[32] J. Tian, M. Gui, C. Li, M. Yan, and W. Xiao, “Mind at semeval-2021 task 6: Propaganda detection using transfer learning and multimodal fusion,” in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pp. 1082–1087, 2021.

[33] D. Dimitrov, B. B. Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, and G. Da San Martino, “Semeval-2021 task 6: Detection of persuasion techniques in texts and images,” in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pp. 70–98, 2021.

[34] G. Da San Martino, A. Barrón-Cedeno, H. Wachsmuth, R. Petrov, and P. Nakov, “Semeval-2020 task 11: Detection of propaganda techniques in news articles,” in

Proceedings of the 14th International Workshop on Semantic Evaluation, SemEval, 2020.

[35] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training,"

[36] J. Kocoń, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz, *et al.*, "Chatgpt: Jack of all trades, master of none," *Information Fusion*, p. 101861, 2023.

[37] B. Zhang, D. Ding, and L. Jing, "How would stance detection techniques evolve after the launch of chatgpt?," *arXiv preprint arXiv:2212.14548*, 2022.

[38] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, "Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert," *arXiv preprint arXiv:2302.10198*, 2023.

[39] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[40] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers," *arXiv preprint arXiv:2211.01910*, 2022.

[41] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," in *Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021*, pp. 3816–3830, Association for Computational Linguistics (ACL), 2021.

[42] K. Hambardzumyan, H. Khachatrian, and J. May, "Warp: Word-level adversarial reprogramming," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4921–4933, 2021.

[43] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, 2021.

[44] T. Schick, H. Schmid, and H. Schütze, "Automatically identifying words that can serve as labels for few-shot text classification," in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5569–5578, 2020.

[45] T. Schick and H. Schütze, "Exploiting cloze-questions for few-shot text classification and natural language inference," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 255–269, 2021.

[46] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24824–24837, 2022.

[47] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[48] K. I. Roumeliotis and N. D. Tselikas, "Chatgpt and open-ai models: A preliminary review," *Future Internet*, vol. 15, no. 6, p. 192, 2023.

[49] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning.* Mit Press, 2008.

[50] J. Xi, "Xi jinping: Ideological work in an extremely important task of the party." http://www.xinhuanet.com//politics/2013-08/20/c_117021464.html, 2013.

[51] P. Sharp, "Revolutionary states, outlaw regimes and the techniques of public diplomacy," in *The New Public Diplomacy: Soft Power in International Relations* (J. Melissen, ed.), Studies in Diplomacy and International Relations, pp. 106–123, Palgrave Macmillan UK.

[52] E. P. Bucy, "Interactivity in society: Locating an elusive concept," vol. 20, no. 5, pp. 373–383. Publisher: Routledge _eprint: https://doi.org/10.1080/01972240490508063.

[53] D. M. Ivester, "The constitutional right to know note," vol. 4, no. 1, pp. 109–164.

[54] P. Contributors, "Paddlenlp: An easy-to-use and high performance nlp library." https://github.com/PaddlePaddle/PaddleNLP, 2021.

[55] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1638–1649, Association for Computational Linguistics.

[56] D. Loureiro, F. Barbieri, L. Neves, L. Espinosa Anke, and J. Camacho-collados, "TimeLMs: Diachronic language models from twitter," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 251–260, Association for Computational Linguistics.

[57] S. Fang, X. Li, and A. Y. Liu, "Chinese public opinion about USâchina relations from trump to biden," vol. 15, no. 1, pp. 27–46. Publisher: Oxford Academic.

[58] Z. Cheng and S. Shaikh, "A comparative study of china's foreign ministry spokesperson's use of weibo and twitter," in *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 552–555, IEEE, 2022.

[59] R. Zhong, A. Krolik, P. Mozur, R. Bergman, and E. Wong, "Behind china's twitter campaign, a murky supporting chorus," *The New York Times*, vol. 8, 2020.

[60] B. Nimmo, I. Hubert, and Y. Cheng, "Spamouflage breakout: Chinese spam network finally starts to gain some traction," 2021.

[61] Y. Zhu and K.-w. Fu, "Speaking up or staying silent? examining the influences of censorship and behavioral contagion on opinion (non-) expression in china," *New Media & Society*, vol. 23, no. 12, pp. 3634–3655, 2021.

[62] R. Creemers, "Cyber china: Upgrading propaganda, public opinion work and social management for the twenty-first century," *Journal of contemporary China*, vol. 26, no. 103, pp. 85–100, 2017.

[63] M. Jiang, "Managing the micro-self: The governmentality of real name registration policy in chinese microblogosphere," *Information, Communication & Society*, vol. 19, no. 2, pp. 203–220, 2016.

[64] M. Jiang and K.-W. Fu, "Chinese social media and big data: big data, big brother, big profit?," 2018.

[65] G. Yang, "The co-evolution of the internet and civil society in china," *Asian Survey*, vol. 43, no. 3, pp. 405–422, 2003.

[66] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[67] E. Cunningham, T. Saich, and J. Turiel, "Understanding ccp resilience: Surveying chinese public opinion through time," *Ash Center for Democratic Governance and Innovation*, vol. 18, 2020.