Dataset of Propaganda Techniques of the State-Sponsored Information Operation of the People’s Republic of China

Rong-Ching Chang
g08350003@thu.edu.tw
Tunghai University
Taiwan

Kai-Lai Chang
g08350009@thu.edu.tw
Tunghai University
Taiwan

Chun-Ming Lai
cmlai@thu.edu.tw
Tunghai University
Taiwan

Chu-Hsing Lin
chlin@thu.edu.tw
Tunghai University
Taiwan

ABSTRACT

The digital media, identified as computational propaganda provides a pathway for propaganda to expand its reach without limit. State-backed propaganda aims to shape the audiences’ cognition toward entities in favor of a certain political party or authority. Furthermore, it has become part of modern information warfare used in order to gain an advantage over opponents.

Most of the current studies focus on using machine learning, quantitative, and qualitative methods to distinguish if a certain piece of information on social media is propaganda. Mainly conducted on English content, but very little research addresses Chinese Mandarin content. From propaganda detection, we want to go one step further to provide more fine-grained information on propaganda techniques that are applied.

In this research, we aim to bridge the information gap by providing a multi-labeled propaganda techniques dataset in Mandarin in detecting state-backed information operation dataset provided by Twitter. We labeled 9,950 tweets in total with 21 propaganda techniques. In addition to presenting the dataset, we apply a multi-label text classification using fine-tuned BERT. We have observed consistency in the promoted message and used techniques by state-backed propaganda operations toward certain entities or topics. Viewing country and political party as an entity, we could view state-backed propaganda detection on different topics as stance detection tasks. Potentially our research could help future research in detecting state-backed propaganda online, especially in a cross-lingual context and cross-platform manner.

CCS CONCEPTS

• Social and professional topics → Political speech.

KEYWORDS

propaganda, information operation, social media

1 INTRODUCTION

Propaganda has the purpose of framing and influencing opinions. With the rise of the internet and social media, propaganda has adopted a powerful tool for its unlimited reach, as well as multiple forms of content that can further drive engagement online and offline without disclosing the writers’ identity. Computational propaganda is defined as propaganda being created or distributed using computational or technical means [5]. Exploiting social media is considered as one of the low-cost and high-impact techniques in information warfare, driving and manipulating human behavior with various psychological manipulations [1]. How information is conveyed is by using propaganda techniques. Propaganda techniques are not only used for political content, but also for marketing, and religious content for persuasion purpose. Propaganda techniques, commonly used in disinformation and misinformation, are the way that propaganda is conveyed [10], such detection requires for more fine-grained analysis and detection, not only distinguishing if it is propaganda, but characterizing where it might come from. The propaganda activity launched by foreign adversaries could be particularly concerning to a country as the usual goal may include steering discord, spreading fear, influencing beliefs and behaviors, diminishing trust, and threatening the stability of a country [1]. Various state-backed official and unofficial departments, organizations, and agencies were established to address information warfare include the Internet Research Agency of Russia [12], 50 Cent Party [15] [14] of Chinese Communist Party (CCP) and the Public Opinion Brigades of the Communist Party of Vietnam [6].

Most of the recent work has been focused on propaganda detection, in another word, identifying if the information is propaganda or not. This has been done using various methods such as qualitative analysis, quantitative analysis [4], and machine learning [20] [7]. The main features for this detection task could be divided into two parts, content-driven, and network-driven. Some of the current propaganda text corpora open data sets on document levels include Rashkin et al. [18] which labeled texts into trusted, satire, hoax, and
propaganda on news. Barrón-Cedeno et al. [3] further increased the corpus [18] with more propaganda articles and metadata of the articles. The currently available fine-grained propaganda technique dataset is the one presented by Da San Martino et al. [10]. From news articles, they labeled 18 propaganda techniques on a word-by-word sentence level, so that the position of where the propaganda technique was applied from start to end was being documented. All of the mentioned data sets are in English. Baisa et al. [2] released a propaganda technique dataset for Czech based on newspapers. Another open data source is Twitter, a popular social media platform, the dataset discloses state-linked information operations that took place on their platform. However, the Twitter dataset is not labeled with propaganda techniques but the Twitter account metadata and media information only. The previously labeled propaganda technique in news article texts could be quite different linguistically compared to texts on social media. Tweets, messages posted on Twitter, tend to be more casual with slang and emojis. They are also shorter as the platform has a text length limit. Da San Martino et al. [9] conducted a survey of computational propaganda, mentioned that there is limited propaganda detection research based on text features due to the lack of annotated data sets. Yet we think text content is an important feature for performing cross-platform detection, in user-identity linking, and in information origin tracing. Since the network feature may differ from platform to platform, text content is more consistent in that regard. To our knowledge, there is no existing propaganda technique dataset for Mandarin Chinese.

To address such a gap, we present our dataset\(^1\) that focuses on propaganda techniques in Mandarin based on state-linked information operations dataset from the PRC released by Twitter in July 2019. The dataset consists of multi-label propaganda techniques of the sampled tweets. Additionally, we employed a fine-tuned BERT model for the multi-label classification task.

2 PROPAGANDA TECHNIQUES

Below we explained a list of selected propaganda techniques we have considered based on various studies [10] [2] [21]. Using the same assumption as [10], we labeled our data based on the linguistic and language use that can be judged directly without retrieving extra information. The propaganda techniques we considered are as follows:

1. Presenting Irrelevant Data
   Also called Red Herring. Introducing irrelevant information or issues to an argument.

2. Misrepresentation of Someone’s Position (Straw Man)
   Substituting one’s opinion with a distorted version rather than the original one.

3. Whataboutism
   Defaming the opponents with hypocrisy.

4. Oversimplification
   Overly generalizing information or the complexity of certain issues to favor a party.

5. Obfuscation, intentional vagueness, confusion
   Purposefully being vague with the intention for the audience to develop false recognition toward the subject.

6. Appeal to authority
   Supporting the opinion or claim unconditionally as long as it comes from the government or an expert.

7. Black-and-white Fallacy
   Presenting only two opposite possibilities, one favoring a certain party and one presented by the opponent.

8. Stereotyping, name-calling, labeling
   Labeling the target with the intention of arousing prejudices or making an association with stereotypes.

9. Loaded Language
   Using emotional words to influence audience opinions.

10. Exaggeration or Minimisation
    Overly amplifying or reducing the importance of something.

11. Flag-waving
    Justifying or presenting as a nation or group or idea. In our case, we also consider Flag-waving when one person is presented as their opinion represents the entire nation or group.

12. Doubt
    Questioning or steering uncertainty or trust toward something, an entity, or a group.

13. Appeal to fear or prejudice
    Spreading a sense of anxiety, fear, or panic toward the audience or entity.

14. Slogans
    A brief sentence that includes labeling, stereotyping, or certain cognitive belief.

15. Thought-terminating cliché
    Using simple and generic sentences to discourage detail in discussions.

16. Bandwagon
    Persuading the audience to align with the bigger crowd who appear to have an advantage or better situation, or implying a certain entity will lose or have a worse situation.

17. Guilt by association or Reductio ad Hitlerum
    Associating an opponent or target with the usually disliked object or group.

18. Repetition
    Repeating the same message or idea several times.

Additional to the usual propaganda techniques, we also introduce the following that has been seen in the dataset:

1. Neutral Political
   This includes the international political news that’s being written objectively.

2. Non-Political
   This includes the non-political related content, which could be written with a neutral or angry, or happy tone.

3. Meme humor
   This is the content that used sarcastic humor toward an entity.

3 DATA

Twitter disclosed 936 accounts with identified state-backed information operations from the People’s Republic of China (PRC) government departments, agencies, and party-supported organizations.
The dataset was divided into two batches that consist of 744 accounts and 196 accounts separately on Twitter’s information operations disclosure. In our study, we sampled tweets from a batch of 744 accounts. The available data disclosed containing account metadata (account created time, a user profile description, user-reported location, etc), tweet metadata (tweet created time, tweet text, tweet language, tweet hashtags, etc), and shared media content (images and videos). In our study, we only focus on the tweet metadata.

The total number of tweets sent by the 744 accounts is 1,898,108. We first filter it by language, and duplicates were dropped. The total number of tweets in Chinese contained in the dataset is 74,277, we randomly selected 9,950 tweets out of that number for labeling.

Uren et al. [19] conducted a detailed quantitative and qualitative analysis on these accounts and suggested that this cluster could be re-purposed spam accounts as they altered the used language through different periods of time. These findings are aligned with ours. Figure 2 shows the top 15 tweet language usage out of 50 total used languages. The top 5 languages used in this cluster of accounts are Indonesian (in), English (en), Portuguese (pt), Chinese (zh), and Tagalog (tl).

Figure 1: Language usage of more than 10,000 times each year

In Figure 1 we plot the language used more than 10,000 times each year and we can see that Chinese was only used by this cluster of accounts after 2017. The primary used language appears to have been clear cut in different years, which indicate that this cluster could be spam accounts that were created and used by entities with different backgrounds and purposes at different time period.

Manual annotation was done by two annotators separately on different portions of the dataset: the first 4,950 tweets and the rest 5,000 tweets respectively. Both annotators are native mandarin speakers in graduate degree programs. This was designed intentionally to insure the alliance of opinions in the dataset [13]. This design will increase the annotator consistency, reduce noise and have better model performance. In the annotation process, we iterate the process of reviewing, data labeling, and documenting political entities being named in the sentence. After the first 1,000 labeling, we observed a consistent propaganda techniques alignment on the tweets toward certain topics and entities. Based on such observation, we built pre-defined labels according to the political entities mentioned. Human annotators iterate the action of updating the labels from the rule-based labels when necessary, adding more entity keywords for rule update, and updating the further unseen data labels based on the updated rules. The category of keywords for the pre-defined labeling rules are shown in Table 1, they can be divided into four categories: exiled or anti-government Chinese, Hong Kong protest, Taiwan independence, International Geo-Political related topics. The usage of keywords is not to be exact but assisting the human annotators.

Table 1: Aggregation of keyword category for the pre-defined labeling rules

<table>
<thead>
<tr>
<th>Keyword Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exiled or anti-government Chinese</td>
<td>5,406</td>
</tr>
<tr>
<td>Hong Kong protest</td>
<td>209</td>
</tr>
<tr>
<td>International Geo-Political</td>
<td>1,718</td>
</tr>
<tr>
<td>Taiwan independence</td>
<td>2</td>
</tr>
</tbody>
</table>

The propaganda techniques are only labeled on the political-related content, there could be non-political content using propaganda techniques but this is not labeled as it was not our focus. Such content will be labeled as a non-political class.

In total, we have 21 different propaganda techniques, we showed a label statistic in Table 2. This is an imbalanced dataset, as the most frequently used label is the non-political content that was used for 6,117 times. Loaded Language was used the most at 2,609 times, followed by Whataboutism 2,509 and Name-Calling 2,313 are the most used propaganda techniques on political-related content. A few techniques occurred rarely, especially the Thought-terminating cliché was not used. We suspect that this is due to the nature of spam accounts. That is, building a relationship with other accounts was
not their primary goal. Thought-terminating clichés might be used more in circumstances where building a relationship with other accounts is one of the target goals. An example of how the dataset was formatted can be seen in Table 3, the tweets were translated for the purpose of display.

Table 2: Data set label statistics

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Propaganda Techniques</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Presenting Irrelevant Data</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>Straw Man</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>Whataboutism</td>
<td>2,509</td>
</tr>
<tr>
<td>4</td>
<td>Oversimplification</td>
<td>37</td>
</tr>
<tr>
<td>5</td>
<td>Obfuscation</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>Appeal to authority</td>
<td>50</td>
</tr>
<tr>
<td>7</td>
<td>Black-and-white</td>
<td>265</td>
</tr>
<tr>
<td>8</td>
<td>Name Calling</td>
<td>2,313</td>
</tr>
<tr>
<td>9</td>
<td>Loaded Language</td>
<td>2,609</td>
</tr>
<tr>
<td>10</td>
<td>Exaggeration or Minimisation</td>
<td>114</td>
</tr>
<tr>
<td>11</td>
<td>Flag-waving</td>
<td>81</td>
</tr>
<tr>
<td>12</td>
<td>Doubt</td>
<td>147</td>
</tr>
<tr>
<td>13</td>
<td>Appeal to fear or prejudice</td>
<td>141</td>
</tr>
<tr>
<td>14</td>
<td>Slogans</td>
<td>37</td>
</tr>
<tr>
<td>15</td>
<td>Thought-terminating cliché</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>Bandwagon</td>
<td>64</td>
</tr>
<tr>
<td>17</td>
<td>Reductio ad Hitlerum</td>
<td>83</td>
</tr>
<tr>
<td>18</td>
<td>Repetition</td>
<td>60</td>
</tr>
<tr>
<td>19</td>
<td>Neutral Political</td>
<td>915</td>
</tr>
<tr>
<td>20</td>
<td>Non-Political</td>
<td>6,117</td>
</tr>
<tr>
<td>21</td>
<td>Meme humor</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 3: Data set sample display

<table>
<thead>
<tr>
<th>Tweetid</th>
<th>Translated Tweet</th>
<th>Propaganda Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>990189929</td>
<td>The truth and hypocrisy under the false democratic face of Guo Wengui, the clown jumping beam, is now undoubtedly exposed!</td>
<td>3,8,9</td>
</tr>
<tr>
<td>836699648</td>
<td></td>
<td></td>
</tr>
<tr>
<td>114879827</td>
<td>We must severely punish the rioters and return Hong Kong to peaceful</td>
<td>8,9,13,14</td>
</tr>
<tr>
<td>6281364480</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5 EVALUATION

The training and testing size was set to 80% and 20% respectively. The results are shown in the Table 4. We only trained it for 2 epochs yet we saw the loss decreased drastically from 0.71102 to 0.05953. In the experiment, we trained for more than 2 epochs; however, the accuracy did not improve. Thus 2 epochs appear to be optimal in our experiment. The evaluation metrics used were accuracy, micro-averaged F1-score, and macro-averaged F1-score. Micro-averaged F1-score aggregate all the samples to compute the average metric of the True Positives out of the Predicted Positives. Macro-averaged F1-score aggregated each class and compute the metrics based on each class. In our case, our accuracy is 0.80352 with micro-averaged F1-score of 0.85431 and macro-averaged F1-score of 0.20803. This indicates that our model performed well in predicting overall samples, however, the performance on each label varied a lot. This is expected as our dataset is skewed, some labels have many data while a few labels have very little data labeled in the dataset.

Table 4: Classification results

<table>
<thead>
<tr>
<th>Measurement Name</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loss : Epoch 0</td>
<td>0.71102</td>
</tr>
<tr>
<td>Loss : Epoch 1</td>
<td>0.05953</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.80352</td>
</tr>
<tr>
<td>F1 Score (Micro)</td>
<td>0.85431</td>
</tr>
<tr>
<td>F1 Score (Macro)</td>
<td>0.20803</td>
</tr>
</tbody>
</table>

Two main activity directions of the dataset were to target opponents of the CPC, such as exiled Chinese, human rights lawyers, relevant personnel and to vilify the protesters against the national security law in Hong Kong. This finding was aligned with what was found in [19], [5], where the spam accounts flooded content in Mandarin with the purpose of dominating search results on Twitter.
We presented a new dataset on propaganda techniques from the People’s Republic of China. When it comes to certain topics, propaganda operates wanted the search results to be skewed toward a perspective that favored the CCP and eschewed the certain community. By doing so the propaganda operators wanted the search results to be skewed toward a perspective that favored the CCP and eschewed the certain community.

6 DISCUSSION
In this paper, we presented the first propaganda technique dataset of state-backed information operation accounts from PRC in Mandarin based on a dataset released by Twitter. We applied 21 propaganda techniques and we annotated a total of 9,950 sentences under a multi-label setting. Machine learning models driven by propaganda research can be particularly benefited by our data set. As we labeled political content with propaganda techniques while labeling it, this data set’s linguistic feature or the propaganda techniques may not apply to all.

Upon the organization structure of PRC, different departments and agencies may lunch online operations targeting the same or different groups of audiences, with different linguistic characteristics. Thus, this data set’s linguistic feature or the propaganda techniques may not apply to all.

7 CONCLUSION
We presented a new dataset on propaganda techniques from the state-backed information operation accounts from PRC in Mandarin. We trained a fine-tuned BERT model to perform multi-label classification on our dataset. In the times where information on social media is part of information warfare strategies. Our dataset could be beneficial in propaganda, political research, and beyond.

By considering the country, political party, or authority as an entity, we could initially view state-backed propaganda on different topics as a stance detection of texts from such an entity. And propaganda techniques could be viewed as a writing style feature. This could help future research in clustering and identifying how likely it is that the information is coming from the same entity or agency.

One state could launch several propaganda texts that have a similar stance or opinion in different countries with different languages. Thus we hope to see our dataset inspire or provide useful information on multilingual or cross-platform state-back propaganda research, using the propaganda techniques as the universal features across languages.

REFERENCES