Investigating the Impact of Facebook Polarization on the 2021 Taiwanese Referendum

Taiwan AI Labs

December 2021

1 Introduction

Since 2010, many democratic countries worldwide have been dealing with disinformation campaigns manifesting from both within their borders and without. According to a report, one of the most harmful disinformation strategies employed by foreign powers is the use of coordinated messaging through social media to sway public opinion and behavior [Marshall, 2020]. For example, when democratic countries face change in the form of social policy decisions or elections, such coordinated media manipulations can lead to "public opinion polarization."

Public opinion polarization occurs when people are repeatedly exposed to the same ideas, either through their social networks or through propaganda, and they become more set in their viewpoints and less likely to consider opposing views. The level of polarization is therefore an indicator of the potential for a decrease in civil debate. When voters are highly polarized, the ideological middle ground shrinks, and the opinions of the masses on various major issues or national policies can become divided into "us versus them," negatively impacting the governance of the country [Abrams and Fiorina, 2012, Niemi et al., 2001, Fiorina, , Fiorina et al., 2008]

Taiwan has four important upcoming referendums on December 18, 2021. "Pork Imports" will decide whether or not to import pork containing the chemical ractopamine from the United States. "Referendum Dates" will decide whether to place voting times for important referendums on the same days as elections. "Nuclear Plant" will decide whether to activate the fourth nuclear power plant in New Taipei City. "Algae Reef Protection" will decide whether to build a receiving terminal to produce natural gas in Taoyuan's Datan Algal Reef. According to previous research, disinformation and manipulation of the media could influence public opinion on topics like these referendums.

This study will examine how Facebook has affected the polarization of public opinion on these topics over the previous eleven months. The phenomenon of public opinion polarization and the factors affecting policy support and political attitudes will be analyzed using artificial intelligence technology. The polarization impact of pages that demonstrated evidence of potential media manipulation through coordinated behavior were considered. Within these pages, those that were shared most often are referred to as "amplifiers." This research aims to study the effect of these amplifiers and on Facebook users and overall online public opinions regarding the referendum issues. The expected research value is to establish a link between cross-strait social issues and political polarization at a theoretical level, and to provide a reference for subsequent research on classifying different public groups.

2 Methodology

To analyze the data in this study, a three-step process was employed. First, posts were collected, then the information flow graph between users, groups, and pages was built. Second, two sets of target results were established to determine amplifiers in the graph. Third, amplifier behavior was analyzed, as was its impact on the polarization of normal Facebook users.

Posts between January 1, 2021, and November 30, 2021, were crawled from Facebook. Only posts containing one of the four following keywords

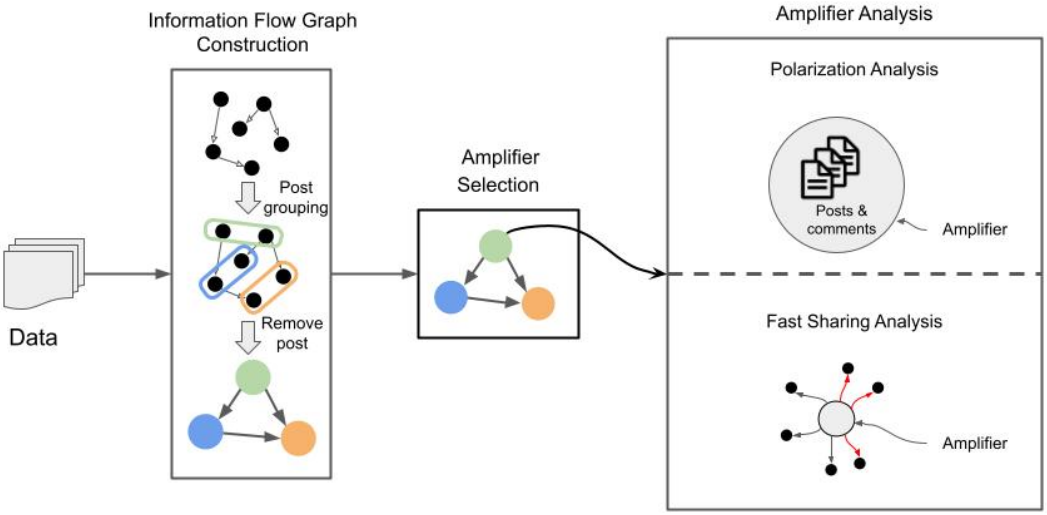


Figure 1: An overview of the methodology framework

were selected for this study: 核四 (Nuclear Plant), 藻礁 (Algae Reef Protection), 萊豬 (Pork Imports), and 綁大選 (Referendum Dates). The data includes information from 2,616 Facebook pages, 295,423 users, 27,345 posts, and 1,462,116 comments. There were a total of 8,558 posts related to the Nuclear Plant, 10,277 posts related to Algae Reef Protection, 12,088 posts related to Pork Imports, and 2,062 posts related to Referendum Dates.

2.1 Construction of the Information Flow Graph

2.1.1 Graphs

In discrete mathematics, graphs model pairwise relationships between objects. Graphs are made of vertices that are connected by edges:

$$G = (V, E) \tag{1}$$

where V is a set of vertices and $E \subseteq \{(x, y)|x, y \in V \; and \; x \neq y\}$ is a set of edges. Depending on the symmetry of the edges, graphs can be undirected graphs whose edges connect vertices symmetrically, or they can be directed graphs whose edges connect vertices asymmetrically.

2.1.2 Representing a Social Network as a Graph

To display social networks as graphs, all users, groups, and pages are represented as vertices. Edges represent the relationships between vertices, and in this case the vertices were connected such that the direction of an edge shows the sharing of posts from the source vertex to the target vertex. Algorithm 1 was used to create the information flow graphs.

This algorithm defines authorial relationships as such: for posts within Facebook groups, groups were always treated as authors of posts. Depending on who shared the post, users or pages were also treated as authors. For posts not in groups, only users or pages were treated as authors.

2.2 Selection of Amplifiers

To isolate instances of potentially coordinated behavior that may impact normal users, the most significant vertices (amplifiers) were selected.

Algorithm 1 Construct information flow graph

$P = \{p_0, p_1, ..., p_{N-1}\}$ is a set of posts
V is the set of groups/users/pages
init $G = (V, \emptyset)$ as an directed graph with empty edges
$i \leftarrow 0$
while $i < N$ **do**
    if p_i is shared from another post p_j **then**
        let a_i be the author of p_i
        let a_j be the author of p_j
        add edge (a_j, a_i) to G
    end if
    $i \leftarrow i + 1$
end while

2.2.1 Posting Time Similarity

For all eleven months of data, posting time vectors with 168 dimensions (7x24) was created. Each dimension represents the proportion of all posts made within that hour for each day of the week.

By computing the cosine similarity of posting time vectors, pages with similar posting time behaviors ($\geq 95\%$) were selected as potential amplifiers. Since the value for each dimension in the posting time vector is non-negative, the scores were between 0 and 1. Given two posting time vectors, u and v, the cosine similarity is formulated as:

$$Similarity(u, v) = cos(u, v) = \frac{u \cdot v}{\|u\|\|v\|} \quad (2)$$

2.2.2 Amplifier Score

An amplifier is a page with posts that were shared much more often than those of other pages. This study operates under the assumption that posts that are shared more often have a greater chance of influencing the opinions of others.

An amplifier score is the median share count of a post from page N. Given an increasing series $Q = (q_1, q_2, ..., q_n)$, which is the share count of page N, the amplifier score is formulated as:

$$score_{amp}(N) = \begin{cases} q_{\frac{n+1}{2}}, & \text{if } n \text{ is odd} \\ \frac{q_{\frac{n}{2}} + q_{\frac{n}{2}+1}}{2}, & \text{if } n \text{ is even} \end{cases} \quad (3)$$

Therefore, pages with scores in the top 20% of those with high cosine similarities were considered to be amplifiers.

2.3 Amplifier Analysis

In this section, one measurement was used to analyze sharing behaviors associated with a given amplifier, and a second was used to analyze the impact of amplifiers on normal users within their audiences.

2.3.1 Polarization

Polarization is defined as the agreement between the contents presented in a post and the content of the comments associated with that post. High agreement in groups and pages indicates significant opinion polarization.

To calculate polarization, an agreement function $f_{agree}(p, c)$ was introduced, where p represents the content of a post, and c represents a single comment associated with p. Outputs of agreement function are **Same**, **Different**, and **Irrelevant**. Given post p and all its comments, set C, the polarization score is formulated as:

$$S_{pol}(p) = \frac{\sum_{c \in C} \mathbb{1}_{same}(p, c)}{\sum_{c \in C}(\mathbb{1}_{same}(p, c) + \mathbb{1}_{different}(p, c))} \quad (4)$$

$$\mathbb{1}_{label}(x, y) = \begin{cases} 1, & \text{if } f_{agree}(x, y) = label \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

$S_{pol}(p)$ measures the percentage of agreement within all comments for post p. Using $S_{pol}(p)$, the polarization score of a set of posts P is computed by:

$$S_{pol}(P) = \frac{\sum_{p \in P} S_{pol}(p)}{|P|} \quad (6)$$

For this study, the agreement function was approximated by a **Deep Neural Network** (DNN). BERT [Devlin et al., 2018], a neural network based on Transformer [Vaswani et al., 2017], was used as an encoder to extract and pass features to a one-layer classifier. HuggingFace Transformer [Wolf et al., 2020] was used to train the model. Our model was initialized from a

pretrained Chinese BERT-wwm [Cui et al., 2020, Cui et al., 2019] model.

Data was collected from January through April, 2021. To train the agreement function, all comments containing only a single character, repeated words, URLs, emojis, or neutral phrases such as "早安 (good morning)," "午安 (good afternoon)," or "hi" were removed. A dataset with 9,485 training examples was finalized from the remaining comments. During preprocessing, text sequences were tokenized by a tokenizer. Chinese text was tokenized into character-based tokens, and English text was tokenized into sub-word units [Sennrich et al., 2015].

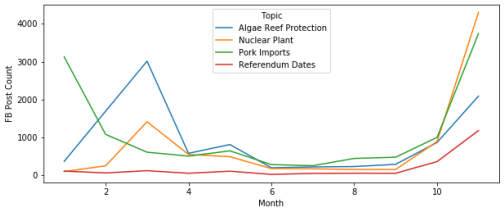
Since BERT's max token length for inputs is 512, the training data with too many tokens in the contents or comments were split into parts of suitable lengths.

Token sequences were truncated by a sliding window with window size w and stride k. First, comments were truncated with $(w, k) = (100, 50)$. Then, for each substring of comments with length m, contents were truncated with $(w, k) = (512 - 2 - m, 128)$ (2 for <cls> and <sep> tokens). After truncation, content tokens $W_{content} = (u_1, u_2, ..., u_n)$ and comment tokens $W_{comment} = (v_1, v_2, ..., v_m)$ were concatenated into $W = (<cls>, u_1, u_2, ...u_n, <sep>, v_1, v_2, ...v_m)$.

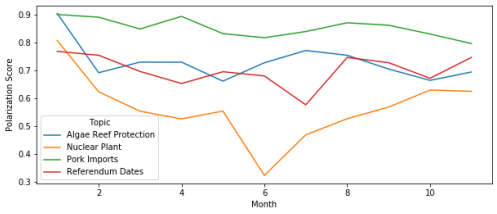
W was first mapped to word embeddings $[\vec{e}_{<cls>}, \vec{e}_{u_1}, ...]$ and then fed into BERT to get the hidden vectors $H = \left[\vec{h}_{<cls>}, \vec{h}_{u_1}, ...\right]$. The vector of the first token $\vec{h}_{<cls>}$ was passed through a one-layer classifier to get the label distribution $\hat{y}$. Cross entropy was chosen as the loss function, where y is the true label.

$$Loss(\mathbf{y}, \hat{\mathbf{y}}) = \sum_i y_i \log \hat{y}_i \tag{7}$$

Five-fold cross validation was used to train five models with the same architecture and hyperparameters. The average validation accuracy of these models was 72.5%. During inference, since an example could be split into several segments after truncation, the class probabilities of all segments were averaged, and the class with the highest probability was selected as the output. Thus, each example had five predictions (one from each model). The final prediction was decided by max voting.



(a)



(b)

Figure 2: (a) Post trends for the four referendum topics (b) Polarization score trends for the four referendum topics

2.3.2 Fast Sharing

In order to uncover potentially manipulative sharing behavior, the rate of sharing content was examined. For each post, shares S posted by the same user were considered to be "fast shared" if the two following conditions were met:

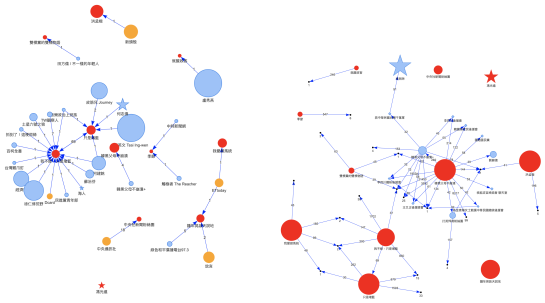
- $|S| \geq K$
- $\frac{t_{|S|} - t_1}{|S|} < T$, t_1 is the time of the first share and $t_{|S|}$ is the time of the last share in S

K and T are predefined thresholds, where $K = 5$ and $T = 5$.

3 Results

This study aims to analyze the role Facebook will play in affecting the upcoming four-topic referendum, which is to be held in Taiwan on December 18, 2021.

As can be seen in Figure 2a, Pork Imports was the most hotly debated topic at the beginning of the year, and then discussions slowed, finally picking up again at the end of the year. The topics of Algae Reef Protection and the Nuclear Plant are both related to energy production for the country, and they trended in similar ways, as each had a



(a) (b)

Figure 3: (a) Upstream information flow through amplifiers (b) Downstream information flow through amplifiers

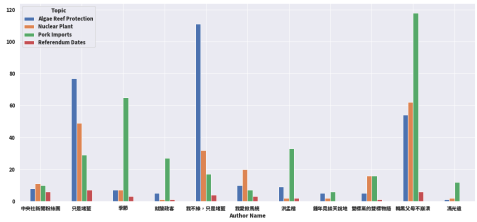


Figure 4: Topic distribution among amplifiers

spike earlier in the year and then dropped off until recently. The Referendum Dates was the least discussed topic and had little fluctuation throughout the year until recently.

To measure the social media "echo chamber effect" on these topics, polarization scores and the trends of overall polarization scores were compared. In Figure 2b, Algae Reef Protection, Pork Imports, and Referendum Dates showed that opinions expressed in the posts and comments in each page had a high degree of alignment; however, the Nuclear Plant topic dipped significantly around June, and its overall polarization score was the lowest of the four topics, indicating greater debate/disagreement in the posts, and a reduction in the echo chamber effect.

Many Facebook pages are used to promote the spread of ideas. This study is concerned with the amplifiers that were discovered, which demonstrated very different posting behaviors from those of normal users, which could be indicative of collaborative behavior. To better understand how information flowed through these amplifiers, the role that they played in the propagation of ideas was examined.

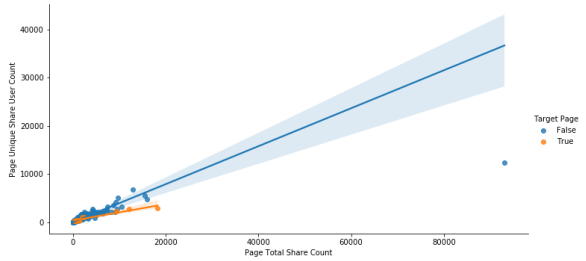
For Figure 3, yellow nodes represent information sources that exist outside of Facebook. Red nodes represent page amplifiers. Blue nodes represent regular users, pages, or groups that shared a post (whether on their own pages or elsewhere). The star nodes represent key opinion leaders (KOL), which are user account pages with many followers. Finally, black nodes represent pages, groups, or users with an average of fewer than 50 likes per post. All the other nodes have an average of 50 or more likes. The number that appears below the black nodes indicates the total number of pages, users, or groups that shared the information. The larger a given node, the more often its page contents have been shared. The numbers above the arrows show how many times that source information was shared.

Figure 3a shows only direct sharing from amplifiers. Figure 3b shows only those nodes that shared a source at least 30 times. All other nodes were omitted for legibility.

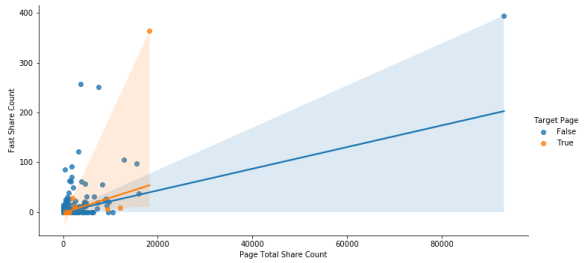
Interestingly, many of the pages with similar posting times and frequencies were media pages (such as news sources). However, when considering the number of times the information on the pages was shared, only one node represents a media outlet. In other words, users, pages, and groups seemed to prefer sharing posts from pages that were not associated with actual news sources.

In Figure 3a it can be seen that amplifiers shared very little information from outside sources. Most of the information they posted was original. The two pages with nearly identical names, 我不綠，只是堵藍 (Not pro DPP, Just Hating on KMT) and 只是堵藍 (Just Hating on KMT), had a high amount of shared information as well as very similar posting times.

In Figure 3b, it can be seen that pages with similar political stances have more edges connecting them. Based on the size and color of the nodes, the web-shaped cluster right of center has more influential pages sharing content. The amplifier in the center of this cluster radiates out in all directions to many nodes. However, the cluster on the bottom left has less influential pages, users, or groups sharing content. Compared to the web-shaped cluster, the three amplifiers in the cluster on the bottom left have few pages, groups, or users sharing all of



(a)



(b)

Figure 5: (a) The relationship between the number of unique users sharing posts and the total share count for each page (b) The relationship between fast share count and page total share count

their information.

To see which topics the amplifiers were most interested in, the distribution of all topics discussed among amplifiers was graphed in Figure 4. Most pages seemed to focus on one or two topics of discussion, typically Algae Reef Protection and Pork Imports. The Referendum Dates topic was the least discussed among all amplifiers.

Each dot in Figure 5a represents a unique Facebook page. The orange dots are amplifiers, and the blue dots represent all the other pages that discussed the four referendum topics between January 1, 2021, and November 30, 2021. The figure shows the relationship between the number of unique users sharing posts and the total share count for each page. Amplifiers have unique users who share their content with much greater frequency than the users of average pages.

As demonstrated in Figure 5b, some amplifiers have a much greater proportion of "fast shares," while other amplifiers do not demonstrate that kind of behavior.

As seen in Figure 6, although there seems to be a number of strategies used to influence opinions (fast share counts, collaborative posting behaviors,

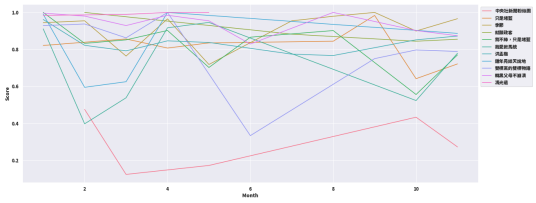


Figure 6: Amplifier polarization score trends

etc.), such efforts do not appear to have had much influence on the polarization scores. Note the very large fluctuations for some of the amplifiers. From Figure 6, interestingly, it can also be observed that the only media associated page on average have a much lower polarization score compared to all other non media associated pages.

4 Discussion

During the buildup to the referendum, the two opposing parties have adopted negative media campaigns to mobilize voter support.

As the divide between the two parties has deepened, their public statements regarding their willingness to cooperate have become increasingly negative. Such strongly opposing stances can cause people to believe that their democracies are not working [Somer and McCoy, 2018].

Even worse, the two sides have adopted non-democratic strategies such as slandering and making false claims about their opponents. Such behavior can lead to decreased public faith in government, even in supposedly consolidated democracies, such as those in North America and Western Europe [Foa and Mounk, 2016].

This study shows evidence of consistent collaborative behavior employed through Facebook. As can be seen in Figure 3, some amplifiers have nearly identical names to other pages, they have very high similarities in their posting times, and they share a tremendous amount of information with each other. The probable goal of such behavior is to increase the polarization of Facebook users and reach wider audiences.

From Figure 2b and Figure 6 it can be seen that most polarization scores were already high at the beginning of the year, suggesting that these

users had already been polarized. One possible reason could be the algorithms used by Facebook, which track user behavior to create targeted advertising for each individual [Powers, 2017]. In this way Facebook promotes ads that their users are more likely to click on, in order to increase revenue. These algorithms thereby promote self-selecting user behavior, which can lead to further polarization and amplification of the echo chamber effect.

In terms of the upcoming Taiwanese referendum, the overall impact of the behaviors documented on Facebook over the last eleven months remains to be seen. Although democracies can benefit from some degree of polarization [Schattschneider, 1975], extreme polarization can have significantly harmful effects on normal democratic processes [Somer and McCoy, 2018]. The authors recommend that users of Facebook and other social media stay vigilant of the potentially harmful influence of polarized media.

References

[Abrams and Fiorina, 2012] Abrams, S. J. and Fiorina, M. P. (2012). The breakdown of representation in american politics.

[Cui et al., 2020] Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., and Hu, G. (2020). Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

[Cui et al., 2019] Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z., Wang, S., and Hu, G. (2019). Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

[Devlin et al., 2018] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

[Fiorina,] Fiorina, M. P. with samuel j. abrams and jeremy c. pope. 2005. culture war? the myth of a polarized america.

[Fiorina et al., 2008] Fiorina, M. P., Abrams, S. A., and Pope, J. C. (2008). Polarization in the american public: Misconceptions and misreadings. *The Journal of Politics*, 70(2):556–560.

[Foa and Mounk, 2016] Foa, R. S. and Mounk, Y. (2016). The danger of deconsolidation: the democratic disconnect. *Journal of democracy*, 27(3):5–17.

[Marshall, 2020] Marshall, W. P. (2020). Internet service provider liability for disseminating false information about voting requirements and procedures. *Ohio St. Tech. LJ*, 16:669.

[Niemi et al., 2001] Niemi, R. G., Weisberg, H. F., and Kimball, D. C. (2001). *Controversies in voting behavior*. Cq Press Washington, DC.

[Powers, 2017] Powers, E. (2017). My news feed is filtered? awareness of news personalization among college students. *Digital Journalism*, 5(10):1315–1335.

[Schattschneider, 1975] Schattschneider, E. E. (1975). *The semisovereign people: A realist's view of democracy in America*. Wadsworth Publishing Company.

[Sennrich et al., 2015] Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

[Somer and McCoy, 2018] Somer, M. and McCoy, J. (2018). Déjà vu? polarization and endangered democracies in the 21st century.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

[Wolf et al., 2020] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.