Removing Noise (Opinion Messages) for Fake News Detection in Discussion Forum Using BERT Model

Cheuk Yu Ip(✉), Fu Kay Frankie Li, Yi Anson Lam, and Siu Ming Yiu

University of Hong Kong, Hong Kong, China
{lesterip,fukayli,yiansonlam}@connect.hku.hk, smyiu@cs.hku.hk

Abstract. The exponential growth and widespread of fake news in online media have been causing unprecedented threats to the election result, public hygiene and justice. With ever-growing contents in online media, scrutinizing every single message could be extremely resource intensive, if not impracticable. However, most of the messages are opinion of the authors, not presenting a fact (whether it is fake or true), which contribute a significant portion of noise. This paper suggests a cost-effective approach to identify opinion contents (noise) in discussion forums which cannot be classified as fake or true news. By excluding opinion contents which are not check-worthy in the preprocessing step, the cost of detection could significantly be reduced, especially if voluminous contents are to be dealt with timely. This paper built up an opinion and factual statement dataset in a mixture of officially written Traditional Chinese from the most popular discussion forum in Hong Kong, namely, LIHKG, relating to local Government officials, then used the Bidirectional Encoder Representations from Transformers (BERT) model to identify opinion contents which achieve 98.7% accuracy, and generalized well in public hygiene related contents which the BERT model did not pre-train. This paper further discovered that some of the 15 most active LIHKG users creating discussion threads relating to the local Government officials might be troll accounts with underlying purposes, and assessment on their behavior and sentiments might assist in detecting misinformation.

Keywords: Fact · Opinion · Text classification · Check-worthy · Fake news · Misinformation · Discussion forum · Lihkg · BERT

1 Introduction

Fake news is verifiably false information and is intentionally created to mislead readers [1]. With growing popularity of social media and discussion forums, the public is more likely to be exposed to fake news as malicious spreaders enjoy lower costs and higher anonymity in propagating fake news. The public may also innocently propagate fake news by sharing or retweeting messages containing fake news without knowing the contents are fake. The effect of fake news could be devastating, for instance, the Pizzagate

S. Goel and P. R. Nunes de Souza (Eds.): ICDF2C 2023, LNICST 570, pp. 78–95, 2024.
https://doi.org/10.1007/978-3-031-56580-9_5

conspiracy against Hilary Clinton's campaign chair directly affected the 2016 US Presidential Election result[1]; the fake news exaggerating vaccine risks during the COVID-19 outbreak caused vaccine hesitancy [2]. The tactics of using bot and troll accounts to spread fake news or negative sentiment to maneuver election results or political goal were found in various countries like the US [3], Canada [4] and the UK [5].

In Hong Kong, there were widespread rumors during the Anti-Extradition Law Amendment Bill movement (hereinafter called Anti-ELAB movement) in 2019. For instance, rumors said some passengers were beaten to death by police in Prince Edward railway station on 31th August 2020 while the police were arresting protestors[2]. One major platform where rumors spread was via the LIHKG online discussion forum[3]. LIHKG ranks the 10th highest traffic volume website in Hong Kong[4] and is the most popular discussion forum in Hong Kong offering high anonymity and no censorship. An on-site survey conducted at protest venues in 2021 indicated LIHKG was a major source of protesters receiving movement-related information [6].

With voluminous discussion threads and replies created in the discussion forum, we are interested in filtering the most relevant messages which could potentially be fake news or misinformation and could cause harm to the society. The contribution of this paper includes:

- we built a data set consisting of 2,998 labelled facts and opinion messages extracted from a discussion forum in Traditional Chinese;
- we proposed a methodology to cost-effectively identify and remove purely opinion contents from discuss forums to enhance accuracy and timeliness in fake news detection;
- we showed a model which could be generalized to unseen discussion forum contents.

In addition, we studied the characteristics of the most active LIHKG users creating discussion threads relating to local Government officials, and found over half these accounts were suspended, removed or inactivated. Some of these accounts had coherent behaviors when creating discussion threads and replying to other users' threads. The findings suggested these accounts might be troll accounts stirring up sentiment against the government.

2 Related Work

2.1 Fake News Detection

Over the years, researchers have been devising various methodologies and models to automatically identify and classify fake news. Classical approaches of automatic fact checking include knowledge comparison [7], writing style analysis [8], propagation pattern analysis [9], analysis on the creditability of source [10], or a mixture of the above approaches. Deploying the advanced natural language processing model developed by

[1] https://en.wikipedia.org/wiki/Pizzagate_conspiracy_theory.
[2] https://en.wikipedia.org/wiki/2019_Prince_Edward_station_attack.
[3] https://lihkg.com.
[4] Statistics as in September 2023, https://www.similarweb.com/top-websites/hong-kong.

Google researchers in 2018, namely Bidirectional Encoder Representations from Transformers (BERT) [11], researchers presented that the accuracy of fake news detection was reaching higher than ever, and even up to 99%, by making use of BERT-based model and a basket of characteristics under respective data set and experimental setting [12]. However, most data sets used by the previous research were extracted from sources such as newspaper website, e.g. Reuters [12], NewsGuard [10], knowledge website, e.g. Wikipaedia [7], fact check website, e.g. kaggle.com [12], Sina Community Management Center [9], PolitiFact [10], etc. Some data input into these research models were then labelled as either true news or fake news. For models which could only classify contents as "True" or "Fake" news, inputting purely opinion contents would immediately lead to inaccurate classification, as opinion could not be either true or fake. Therefore, removing noise (e.g. opinion) in data pre-processing plays a crucial role in fake news detection.

2.2 Classification of Fact and Opinion

Fact is objective and is well-supported by available evidence, whereas opinion is either subjective or else which content is not well supported by evidence [13]. Classification of fact and opinion could be the opposite of subjectivity detection problem in which factual or neutral contents were removed during sentiment analysis [14]. Subjectivity detection could be achieved via three approaches, namely: syntactic (by sentence structure), semantic (by meaning), and multi-modal (text plus other attributes like photo and video) [15]. For instance, the sentence "Corruption is just another form of tyranny." is an opinion with negative sentiment, but "'Corruption is just another form of tyranny', said Joe Biden" is a factual narrative describing Joe Biden has made the statement.

Alhindi [16] explored classifying facts and opinion in news classification at documentation level using BERT-base plus recurrent neural network model by exploring the argumentative features (i.e. assumption, common-ground, testimony, statistics, anecdote, and others) and reached 91% F1 score at unseen publishers. Given that contents in a discussion forum are usually written far less systematically and lack sufficient argumentative features, this approach might not be effectively applied to contents in discussion forums. Carrillo-de-Albornoz [17] explored classifying patient-generated contents in online health forums into "experience", "fact" and "opinion" using classical word embeddings and bag-of-words approach and attained on average 70% accuracy in classifying opinion and fact. Blackledge [18] proposed a two-step classification pipeline to remove opinion articles using DeBERTa [19] (Decoding-enhanced BERT with disentangled attention, a model building upon BERT) before putting into a fake news classifier and obtained 10.1% improvement in accuracy. However, the model only made use of 25 factual news and 25 opinion-based news. A larger data set is anticipated for better performance evaluation.

2.3 Check-Worthy Claim Detection

Gencheva, P. [20] trained a model using Support Vector Machines (SVM) and deep feed-forward neural network (FNN) to detect check-worthy claims in political debates based on contextual features, reaction by the moderator and the public. The model precision of 80% in identifying check-worthy sentences. Fatma [21] built a data set

of 23,533 statements extracted from previous US general election Presidential debates between 1960 to 2016 annotated by human, defining the statements into one of the three categories, namely Check-worthy Factual Sentence (CFS) where the general public will be interested in learning about their veracity, Unimportant Factual Sentence (UFS) where claims are factual but not check-worthy, and Non-factual Sentence (NFS) where sentences do not contain any factual claims. Under this categorization, facts are further classified as either CFS and UFS, where NFS is equivalent to opinion in the context of this paper. Konstantinovskiy, L. [22] crowdsourced 28,100 annotations from volunteers, with contents extracted from 6,304 sentences extracted from 14 episodes of UK politician TV shows. Claims that are not check-worthy are classified into 1 out of 7 categories, namely Personal experience, Quantity in the past or present, Correlation or causation, Current laws or rules of operation, Prediction, Other type of claim, and Not a claim. Using bidirectional long-short-term memory (BiLSTM) network and logistic regression as classifier, the model reached an F1 score of 83% in classifying "not a claim" statements, where the performance of classifying other not check-worthy statements is mediocre. Jha, R. [23] used the gate recurrent unit (GRU) to create a model to identify check-worthy facts. The authors used ClaimBuster data set at [21] and created an IndianClaim data set comprising 953 statements annotated as CFS, UFS or NFS. The model reached 92% and 70% F1 score in ClaimBuster data set and IndianClaim data set respectively. In our research, we concur with the view in [22] that whether a sentence is importance could be highly subjective, could vary from context, and could change from time to time. Therefore, our research did not specifically separate CFS and UFS in classification.

To the best of our knowledge, this is the first study on classification of facts and opinion in Chinese using BERT model. The challenges are writing style, sentiment and length of messages in discussion forums varies significantly, which means the model was to classify the contents from few Chinese characters to lengthy paragraphs written by unseen authors. Unlike the sentences in the ClaimBuster data set, which were all from presidential candidates, the tone of which were more consistent and official. Also, the language used in the discussion forum contains a mixture of officially written Traditional Chinese (the Chinese language used in Taiwan area and Hong Kong) and unofficial spoken Cantonese (a dialect of Chinese), which writing style and use of words are deviated from the official language. We are interested to explore the performance of BERT in classifying facts and opinion in Traditional Chinese, and its adaptability to spoken Cantonese.

3 Methodology

Fact-opinion classification can be defined as a multi-class classification problem, i.e. fact, opinion and a mixture of fact and opinion. The problem could be further simplified as a binary classification problem, i.e. containing factual statement and opinion only. For the input, each training sample consists of a thread (with title and contents) to be verified. The output is the predicted label, which can be fact ("**F**"), opinion ("**O**"), or a mixture of fact and opinion ("**M**") in multi-class classification; or simply containing factual statement ("**F**") and opinion only ("**O**") in binary classification.

The design of the model is based on human's behavior in using discussion forums and adversaries' habits in spreading rumors and fake news. Figure 1 shows the methodology of the model.

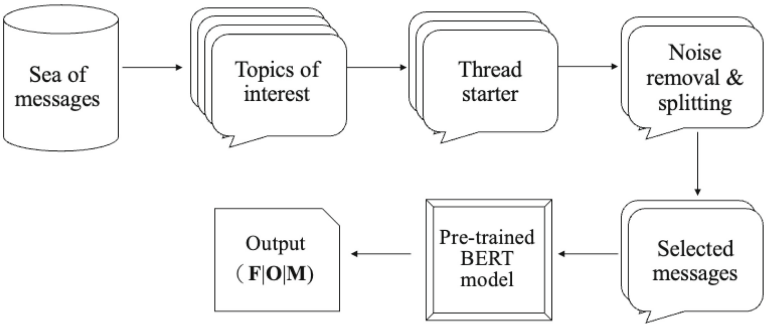

Fig. 1. Data selection and processing flow.

Step 1: Topics of interest only. Empirically, rumors and misinformation mainly surround politics, hygiene and recent popular topics to maximize public attention and create noise in the society. Contents of these topics are most relevant to fake news detection. The list of keywords could consist of a manually selected list plus a dynamically generated frequently discussed word list.

Step 2: Thread starter only. Fake news spreaders tend to disseminate fake news by initiating a new thread with an eye-catching title to mislead as many readers as possible. On the other words, they have little motivation to reply to existing threads unless replying to his own thread to "push" it to higher position of the forum so that it could be more visible by others. Besides, replying to irrelevant contents would be regarded as spamming where the reputation of the message maker could be hampered. Thread starters are therefore most check-worthy.

Step 3: Noise removal and splitting. Unnecessary columns and information, especially URLs, might affect the accuracy of opinion classification. Fake news spreaders might provide the URL of a legitimate news page together with irrelevant or altered message content. The size of the message needs to be split to maintain resource effectiveness. For this paper, as BERT model has maximum size of 512 tokens (i.e. characters), messages with larger contents size are spilt or truncated.

Step 4: Fine-tuning in the pre-trained model. A pool of screened messages is to be manually labelled to set grounds for classification. The classified messages are then tokenized to BERT compatible format, divided into batches and fine-tuned in the pre-trained `bertForSequenceClassification` classifier.

Step 5: Removal of opinion. Remove messages labelled as "**O**" from further analysis of fake news detection, as opinion cannot be fake news. This process avoids mislabeling opinion contents as fake news. Messages labelled as "**M**" could be further split to extract the factual parts for fake news detection at later stage.

4 Data Collection

This study crawled all messages from the LIHKG discussion forum between 7 February 2021 and 31 December 2021. All message threads of the discussion forum could be read freely by the public. The data collection procedure complied with the discussion forum's terms and conditions. From the collected message pool, 77,420 messages were extracted, which contained least one of 12 manually selected keywords representing the Chinese full name or aliases of four main Hong Kong government officials, namely **Official A**, **B**, **C** and **D**[5] in social media. They were the main decision makers in the Hong Kong government during the Anti-ELAB movement. These keywords are most relevant because fake news was widespread during the Anti-ELAB movement against the government [24]. The messages contained the following attributes:

1. `Thread_ID`: Unique identifier of a thread.
2. `Thread_Title`: Title of a thread.
3. `Content`: Body content of a thread, or content of reply in a thread.
4. `Message_ID`: Sequence of the message in a thread. `1` refers to thread starter, `2` refers to the 1st reply in the thread, etc.
5. `Sender_name`: Nickname of the message thread sender. Every LIHKG user could change their nickname freely.
6. `Sender_ID`: Unique ID of a LIHKG user.
7. `Time`: The creation time of a message thread.

Table 1 and 2 show an example of a thread containing factual content and opinion translated in English respectively. Intuitively fake news makers tend to create threads to gain maximum exposure instead of creating reply messages, because thread titles are more visible to discussion forum readers. To testify the hypothesis, we randomly sampled 600 reply messages (i.e. `Message_ID > 1`) for manual classification, and found that only 54 messages were facts or contained factual expression. On the other words, 546 (91%) were opinion or contents without factual expression (e.g. reply with an URL only). The average position of these reply messages was 184 (i.e. `Message_ID = 184`) which means on average a post has 183 replies, indicating that the chance of these messages being visible by a reader was slim compared with thread starter as the reply messages were displayed in chronological order. The value of fack-checking the factuality of reply messages is much less significant.

[5] The names of the officials were intentionally blinded to avoid subjectivity perceived by readers.

Table 1. Sample message thread containing factual content.

Thread_Title	(English translation) Deputy Director of the Hong Kong and Macao Affairs Office (HKMAO) Huang Liuquan: I believe that **Official A** will lead the (Hong Kong) Special Administrative Region Government to unite the community and promote the better development of Hong Kong
Content	A briefing on the Outline of the 14th Five-Year Plan was held at the Central Government Office this morning (23rd). The Deputy Director of the HKMAO, Mr. Wong Liu-kuen, said in his speech that he was glad to see that the Hong Kong Government had resolutely taken up the constitutional responsibility of safeguarding national security and social stability since the implementation of the National Security Law of Hong Kong
Message_ID	1
Sender_name	King Lok WONG
Sender_ID	230530
Time	23/8/2021 14:15

Table 2. Sample message thread containing purely opinion.

Thread_ID	2656971
Thread_Title	(English translation) Hong Kong Confederation of Trade Unions will dissolve next. Which will be the one following?
Content	National Security Department could govern Hong Kong for **Official A**. Could she step down?
Message_ID	1
Sender_name	Treasure life, don't be a prostitute
Sender_ID	115598
Time	11/8/2021 12:36

5 Preprocessing and Labelling

Filtering out the non-reply messages, a total of 6,261 message threads (containing a title and content) remained. During preprocessing, messages containing the aliases of the government officials but referring to other meanings were removed. Also, to remove possible bias in machine learning, all URLs were trimmed from the contents via regular expression.

To fit in the maximum size of 512 tokens in BERT model, messages with title less than 60 Chinese characters were selected. Main contents were limited to 400 Chinese characters, where extra contents were truncated to prevent the system from running out of memory. If the last sentence of the content was partially trimmed such that the sentence was incomplete, the first half fragment of the sentence was removed. To achieve the

optimal training result, all messages were scrutinized again and messages without any of the 12 sets of keywords were removed. As a result, a total of 2,998 messages finally remained.

The 2,998 messages were manually labelled by one trained assistant applying the definitions proposed in [13] to ensure the same standard applied. The labelling was reviewed by another assistant to identify labelling inconsistencies. In the first data set, all 2,998 threads were labelled "**O**" (opinion), "**F**" (factual expression) or "**M**" (a mixture of opinion and factual expression) in a new attribute `Type`. A message was labelled "**M**" even if the majority of the contents were factual with only a minor part being opinion, and vice versa. The classified data set consisted of four attributes, namely `Thread_ID`, `Thread_Title`, `Content` and `Type`. Other attributes, namely `Message_ID`, `Sender_name`, `Sender_ID` and `Time` were not fed into the training model. Table 3 shows the distribution of the data set.

Table 3. Classification of data set into three categories.

Type	Number of Entry	Percentage
Factual expression	1,544	51.5%
Opinion	1,122	37.4%
Mixed	332	11.1%

In the second data set, all messages with a mixture of opinion and factual expressions were also labelled as "**F**" because if the message contained factual expression, any sentence in the message could represent fake news. In short, the 2,998 messages were labelled as "**O**" (opinion only) or "**F**" (containing factual expression) only. The distribution of the data set is as follows (Table 4):

Table 4. Classification of data set into two categories.

Type	Number of Entry	Percentage
Factual expression	1,876	62.58%
Opinion	1,122	37.42%

The experiment was conducted using PyTorch 1.12 with BERT Base Chinese transformer model developed by Hugging Face[6]. The model had been pretrained with a dictionary size of 21,128 tokens at character level and 12 hidden layers, learning rate of 1e−5 and a batch size of 32. The baseline BERT model was used to test the predicted accuracy for 2-category and 3-category classification. As the accuracy fluctuated from time to time, the average accuracy rate of 5 tests was taken. The average predicted accuracy rates were 50.0% and 38.5% respectively as shown in Table 5. Without fine-tuning,

[6] https://huggingface.co/bert-base-chinese.

the baseline model was not robust to classify opinion from factual information, and performed even worse a mixture of factual information and opinion.

Table 5. Predicted accuracy of baseline BERT model.

Data Set	Accuracy
2-category	50.0%
3-category	38.5%

The data set was trained and tested with a proportion of 90% and 10%, i.e. 2,700 threads for training and 298 threads for testing with the same ratio of classification in each category. The `Thread_ID`, `Thread_Title, Content` and `Type` were converted to tensors, and contents of tensors containing `Thread_Title` and `Content` were tokenized . The tensors were then loaded into the BERT model in a mini-batch size of 32 samples. After that, the model was fine-tuned using the `bertForSequence-Classification` classifier. Six epochs were trained. The experiment was carried out using a notebook computer with Apple M1 Max CPU with 10 Core CPU and 32 Core GPU and 32 GB memory. The dataset and codes could be provided for review upon request.

The training accuracy at each epoch is shown in Table 6, items with highest accuracy are bolded. The model performed excellently with 94.7% training accuracy at 3-category data set, and performed better at 2-category data set, i.e. 98.7% accuracy. The training accuracy started to drop at the 6th epoch possibly due to overfitting. The training took 5 h 37 min and 38 s for 3-cateogory set, and 5 h 59 min and 38 s for 2-category data set. The training model was then applied to a testing set consisting of 298 threads which are blinded. Table 7 shows the model reached 91.9% and 96.0% training accuracy at 3-category and 2-category data set respectively, which means the model generalized well at unseen data. For the 3-category testing, the accuracy refers to the total true positive in "**F**", "**O**" and "**M**" out of all testing sets. The precision in Fact refers to the ratio of testing sets correctly predicted "**F**" out of total predicted "**F**"; the recall in Fact refers to the ratio of testing sets correctly predicted "**F**" out of total actual "**F**", F1 score in "**F**" refers to the harmonic means of precision and recall in "**F**", and the like. The precision in Fact and recall in Opinion reached as high as 97.9% and 100.0% respectively. For the 2-category testing, Precision refers to items correctly predicted as Facts. The precision reached 99.4%.

Table 6. Accuracy of each epoch in each classification.

Epoch	3-Catorgory data set	2-Category data set
	Accuracy	Accuracy
1	87.9%	93.5%
2	89.0%	96.6%
3	92.7%	97.0%
4	92.6%	98.3%
5	93.5%	**98.7%**
6	**94.7%**	98.3%
Runtime	5:37:38”	5:59:38”

Table 7. Testing Accuracy, Precision, recall & F1 Score in each classification.

	3-Catorgory data set			2-Category data set
	Fact	Opinion	Mixed	Fact
Accuracy	91.9% (Overall)			96.0%
Precision	97.9%	91.4%	84.9%	99.4%
Recall	88.0%	100.0%	84.9%	93.6%
F1	92.7%	95.5%	84.9%	96.4%
Runtime	8.5 s		8.5 s	

The testing on the 298 items of data set on average took 8.5 s, suggesting the model was resource efficient. The output contained three columns, namely `Thread_Title`, `Content` and `Category`, where `Category` showed the predicted classification, i.e. either “**F**”, “**O**” or “**M**” as the case could be. In both 2-category or 3-category data set, the model performed less satisfactory in differentiating contents with mixture of opinion and factual expression, particularly in the event the contents were dominated by opinion the content will be so classified, and vice versa.

Table 8 shows two examples of inaccurately classified messages translated in English. The first message was manually classified “M” in 3-category dataset as the tone was subjective, plus factually the donation was to encourage vaccination for the government policy instead of solely for **Official A**. The message was classified as “F” in 2-category dataset as it contained at least partial factual information. Both 3-category and 2-category model predicted it as “**O**” (opinion), likely because in most training samples, the majority of the opinion messages tended to be shorter and with a subjective tone. For the second message, the 3-category model incorrectly classified the factual message as “**M**”, but the classification under 2-cateory model was consistent with the manual labelling. The inconsistency in 3-category model could be because most training samples being “**M**”

were mainly commentary in nature, in which the tone and use of word were similar to the questioned message.

Table 8. Testing Accuracy, Precision, recall & F1 Score in each classification.

Thread Title	Content	Classification			
		3-cat		2-cat	
		Lb	Pd	Lb	Pd
[Chance to own a flat!] Vaccination lucky Draw	Vaccination | CC WONG of Sino boosts vaccination for **Official A**. Donated a $10.8M Grand Central unit for lucky draw	M	O	F	O
[**Official A** Plan] Newly approved mortgage insurance peaked in May. Centraline Mortgage: Will keep rising in bull market	**[Official A** Plan] Newly approved mortgage insurance peaked in May. Centraline Mortgage: Will keep rising in bull market. 14:47 2021/06/08 "Since the gov't launched the Home Starter Loan Scheme (aka **Official A** Plan), the demand to buy flats surged. Centraline Mortgage stated newly approved mortgage insurance reached record high of $32.69B (+5.4% by month) and anticipated the figure would keep surging"	F	M	F	F

Lb = Manually labelled classification **Pd** = Classification by the model.

Table 9 summarizes the breakdown of misclassified messages and the sentiment of these messages. For the 3-category data set, 11 messages labelled as "**F**" or "**M**" were misclassified as "**O**", and 11 messages labelled as "**F**" were misclassified as "**M**". For the 2-category data set, 11 out of 12 messages were misclassified from "**F**" or "**M**" as "**O**". Messages wrongly misclassified as "**O**" would be filtered out wrongly and had no chance for fact-checking. Looking into the message sentiment, for both the 2-category and 3-category data sets had 1 positive, 2 neutral and 8 negative sentiment messages being misclassified. The distribution appeared to be aligning with the overall negative social atmosphere in Hong Kong in year 2021 where the mainstream discussions were related to Anti-ELAB movement and COVID pandemic.

Table 9. Summary of Misclassified Messages and Respective Message Sentiments.

3-Catorgory data set				
Manually Labelled as	Classified by the Model as	No. of Misclassified Message / Sentiment of Message		
		Positive	Neutral	Negative
F/M	O	1	2	8
F	M	2	2	7
M	F	0	1	1
Total		**24 Misclassified Messages**		
2-Catorgory data set				
F/M	O	1	2	8
O	F	0	0	1
Total		**12 Misclassified Messages**		

Overall speaking, 37.42% of the contents in politics-related message threads were opinion. The threads tended to quote factual information extracted from news reports and seek for serious discussion in the forum. The ratio of message threads of other topics (e.g. entrainment) being opinion might even be higher as the discussions are more casual and subjective in nature. If these messages were to be put in a binary fake news classifier (i.e. either true or fake news), the classification would never be accurate because the message could neither be true or fake. On the other words, the maximum achievable accuracy could only be $1-37.42\% = 62.58\%$. By reducing the noise, the training accuracy would have, in this case, increased by 37.42%.

6 Ablation Studies

This research also studied the applicability and generalization of the model to reply messages and messages related to other topics. In the first ablation study, 600 political related reply messages were tested in the 3-category and 2-category data set model. The column of `Thread_Title` was removed before testing because they were created by the thread maker instead of the replier. Owing to the commentary nature of the reply messages, only 22 messages (or 3.6%) contained factual expression, 539 messages (or 89.8%) were opinions, 39 messages (or 6.5%) contained a mixture of facts and opinion. Table 10 shows the testing still achieved 95.0% and 95.3% accuracy respectively, implying the model performed well in reply messages.

Table 10. Testing Accuracy, Precision, recall & F1 Score in each classification.

	3-Catorgory data set			2-Category data set
	Fact	Opinion	Mixed	Fact
Accuracy	95.0% (Overall)			95.3%
Precision	97.1%	96.1%	74.1%	95.5%
Recall	50.0%	100.0%	51.3%	61.8%
F1	64.7%	98.0%	60.1%	75.0%

In the second ablation study, a total of 935 messages containing a mixture of thread starter and replies were selected between February and December 2021, each containing one of 6 manually selected keywords representing the Chinese full name and aliases of four main Hong Kong government officials and experts (**Officials E** to **H**) who played key roles in public hygiene policies in Hong Kong during COVID outbreak other than **Officials A** to **D**. Out of the 935 messages, 490 (or 52.2%) contained factual expression, 317 (or 33.9%) were opinions, 128 (or 13.9%) contained a mixture of factual expression and opinion. They were tested in the 3-category and 2-category data set model. Table 11 shows that the testing accuracy still achieved remarkably 91.0% and 96.4% respectively, implying the model generalized well in topics other than the fine-tuned one.

Table 11. Testing Accuracy, Precision, Recall and F1 Score in each classification using COVID related messages.

	3-Catorgory data set			2-Category data set
	Fact	Opinion	Mixed	Fact
Accuracy	91.0% (Overall)			96.4%
Precision	97.1%	93.8%	97.1%	93.8%
Recall	89.2%	99.4%	89.2%	99.4%
F1	93.0%	96.5%	93.0%	96.5%

7 Behavioral Study of the Thread Makers

This paper also explored the behaviors of the top 15 thread creators of the 2,998 selected threads mentioning **Officials A-D**. The 2,998 threads selected for the training data set were created by 762 forum users. The top 15 thread creators created 1,484 (or 49%) threads and is therefore considered influential to the sentiment on government officials. Out of the 1,484 threads, 979 (or 66%) were factual expression, 332 (or 22%) were opinion, 173 (or 12%) were a mixture of factual expression and opinion.

The general behaviors of these 15 thread creators were studied. All thread creation and reply messages (without keyword selection or screening) they made between 7

February 2021 and 31 December 2021 were extracted, which amounted to 149,499 messages. Amongst the messages, 34,036 (or 22.7%) were thread creation, 115,463 (or 77.2%) were replies.

Table 12. Account Statistics of the Top 15 Thread Creators between 7 Feb & 31 Dec 2021.

#	User ID	Total No. of Message	Average Message Per Day	No. of Thread Created	Thread Created Per Day	Thread-to-Message Ratio	Account Status as in Aug 2022
1	25558	25,903	104.4	8,759	35	34%	Inactivated
2	230530	8,367	25.5	2,977	9	36%	Suspended
3	240362	16,236	49.5	4,291	13	26%	Normal
4	130626	8,152	25	3,125	10	38%	Suspended
5	21470	13,787	65	3,624	17	26%	Normal
6	184679	4,613	14	2,791	9	61%	Inactivated
7	212626	4,328	13.2	269	1	6%	Normal
8	42188	2,986	9.1	1,218	4	41%	Suspended
9	5195	2,802	9.5	1,305	4	47%	Suspended
10	79563	13,076	39.9	1,400	4	11%	Normal
11	291865	3,459	10.6	466	1	13%	Normal
12	219007	2,008	6.1	686	2	34%	Suspended
13	19552	32,335	98.6	2,490	8	8%	Suspended
14	20933	3,004	9.2	347	1	12%	Normal
15	70786	8,443	25.7	288	1	3%	Normal
Average		**9,694**	**31**	**2,172**	**7**	**26%**	

Table 12 shows except the top thread creator (#25558) who on average created 35 threads daily, others created no more than 17 threads per day, in which the frequency was reasonable. The frequency of reply might not have significant reference value per se, as a reply could be as short as one emoticon or a word "push". Manually scrutinizing the active pattern of all 15 thread creators, all appeared to have reasonable rest hours. For user #25558, his active hours were normally from 07:00 till 23:00, 7 days a week, with no sign of slowing down during Saturdays, Sundays and public holidays.

In terms of thread contents, most politically related threads created appeared to be copying from newspapers, influencers' social media pages, etc. with source links provided. All thread creators participated in topics other than politics (e.g. showbiz, sports, health care, gaming, gambling, finance) although politics were still most actively participated. Replying with the same content (i.e. spamming) other than using the same emoticons was not common.

In terms of account lifespan, out of the top 15 thread creators, 8 accounts (or 53%) were suspended by the administrator (2), removed by the user (4) or inactivated (2)

as in August 2022 (i.e. 8 months after the data collection period). This showed significant contrast with the top 16th to top 100th thread creators that only 14 accounts (16.5%) were suspended, removed or inactivated. 7 out of 8 invalidated accounts had high thread-to-message ratio, which implies these account owners were proactive in creating message threads (one exception was user #19552 who displayed high self-replying rate as illustrated in Table 13).

Focusing on the 77,420 Hong Kong government official related messages, the top 15 thread creators created 3,815 (4.9%) messages, in which 2,640 (71%) were thread creation and 1,175 (29%) were replies. Table 13 shows that user #25558 created most threads in this context, but 55% of his messages had no reply, and he seldomly boosted his message by replying to himself, leaving the average thread length as short as 2.74 (i.e. on average 1.74 replies per thread created).

Table 13. Characteristics of the Gov't Official Related Threads by the Top 15 Thread Creators.

#	User ID	Total Thread Created	Threads with No Reply (%)	Self-replying Thread (%)	Threads with Others Reply (%)	Average Thread Length
1	25558	898	494 (55%)	130 (14%)	355 (40%)	2.74
2	230530	377	26 (7%)	87 (23%)	348 (92%)	40.55
3	240362	337	1 (0%)	107 (32%)	336 (100%)	66.39
4	130626	171	6 (4%)	83 (49%)	162 (95%)	63.65
5	21470	154	8 (5%)	7 (5%)	146 (95%)	155
6	184679	124	8 (6%)	38 (31%)	111 (90%)	52.8
7	212626	82	0 (0%)	81 (99%)	82 (100%)	161.1
8	42188	60	3 (5%)	19 (32%)	56 (93%)	55.78
9	5195	84	1 (1%)	29 (35%)	83 (99%)	111.33
10	79563	95	6 (6%)	52 (55%)	83 (87%)	26.34
11	291865	56	3 (5%)	33 (59%)	53 (95%)	36.64
12	219007	57	4 (7%)	18 (32%)	53 (93%)	54.24
13	19552	50	0 (0%)	38 (76%)	49 (98%)	48.7
14	20933	63	2 (3%)	47 (75%)	60 (95%)	64
15	70786	32	0 (0%)	32 (100%)	30 (94%)	55.09
Average		**186**	**40 (8%)**	**55 (44%)**	**141 (91%)**	**67**

Manually scrutinizing these 3,815 messages, those factual contents were mainly local news of various topics (e.g. the Government's latest measures against COVID, Legislation Council election, Government officials receiving luxury gifts, etc.) directly copied from local press media and news headlines of online radio or TV programme; those contents representing opinion or a mixture of factual expression or opinion are mainly comments from news commentors, op-eds, appeals and personal comments.

Without strong evidence suggesting the top 15 thread creators were bots or trolls, the short account lifespan raised suspicion as to whether the users had malicious intent. Notwithstanding Lee, F.L.F. [6] opined the Anti-ELAB movement in 2019 was "truly leaderless and decentralized", the possibility of adversaries deploying troll accounts to stir up sentiment against the government could not be eliminated. In the process of fake news detection in a discussion forum, considering the behaviors of an account such as frequency of thread creation, number of self-replying messages, account life spam, thread length, sentiment of contents, etc. could be beneficial in identifying potentially malicious users who might have the tendencies of disseminating misinformation.

8 Limitation and Future Work

The first limitation is the inherent differences in the length of messages, use of words and tone between opinion and factual contents in the messages by discussion forum users. Messages with purely factual information tend to quote contents from news agencies and therefore tend to be longer and written in official traditional Chinese language with objective tone; purely opinion messages tend to be shorter, written in spoken Cantonese and with negative sentiments; messages with a mixture of factual expression and opinion tend to be commentary from news agencies or from key opinion leaders. The same observation applies to contents related to public hygiene policies in the ablation study.

The second limitation is the labelling of messages, especially in 3-category data set model. In certain messages the classification could not be very clear cut. For instance, messages containing exaggerated verbs and adjectives, insulting nicknames, a tag like "Breaking News" etc., might still be labelled as "**F**" instead of "**M**" (mixture), depending on the overall context. In any event, if there have been overwhelming inconsistencies in the labelling, the training and testing accuracy would have been lowered. Should multiple independent labelers be employed in labelling and a reviewing system implemented, the accuracy would have been improved.

The third limitation is the data set to train the classifier is relatively small, and the set of keywords are limited to the names/nicknames of the Government officials only. It is anticipated that with larger data set and wider choices of keywords, the classification accuracy and robustness could be further improved, and the behaviors of the thread creators could be further studied to ascertain its correlation with misinformation dissemination.

9 Conclusion

This study built up a data set in a mixture of Traditional Chinese and spoken Cantonese, for classification of factual and opinion statements relating to Hong Kong government officials, with information extracted from the LIHKG discussion forum. The BERT model we built reached 98.7% accuracy in unseen data and suggested that removing opinion statements could be effectively done before further processing. The study also gave hints on the possibility of detecting user misinformation by taking consideration the behavior of users and the sentiment of the contents they posted.

References

1. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor. Newsl. **19**(1), 22–36 (2017)
2. Garett, R., Young, S.D.: Online misinformation and vaccine hesitancy. Transl. Behav Med. **11**(12), 2194–2199 (2021)
3. Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., Blackburn, J.: Disinformation warfare: understanding state-sponsored trolls on Twitter and their influence on the web. In: Companion Proceedings of the 2019 World Wide Web Conference, pp. 218–226, US (2019)
4. Bellutta, D., King, C., Carley, K. M.: Deceptive accusations and concealed identities as misinformation campaign strategies. Comput. Math. Organ. Theory **27**, 302–323 (2021)
5. Bruno, M., Lambiotte, R., Saracco, F.: Brexit and bots: characterizing the behaviour of automated accounts on Twitter during the UK election. EPJ Data Sci. **11**, 17 (2022)
6. Lee., F.L.F., Liang, H., Cheng, E.W., Tang, G.K.Y., Yuen, S.: Affordances, movement dynamics, and a centralized digital communication platform in a networked movement. Inf. Commun. Soc. **25**(12), 1699–1716 (2021)
7. Ciampaglia, G.L., Shiralkar, P., Rocha, L.M., Bollen, J., Menczer, F., Flammini, A.: Computational fact checking from knowledge networks. PloS one **10**, 6, e0128193 (2015)
8. Zhou, L., Burgoon, J.K., Nunamaker, J.F., Twitchell, D.: Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. Group Decis. Negot. **13**, 81–106 (2004)
9. Wu, K., Yang, S., Zhu, K.Q.: False rumors detection on Sina Weibo by propagation structures in data engineering. In: IEEE 31st International Conference on. IEEE, pp. 651–662, South Korea (2015)
10. Horne, B.D., Nørregaard, J., Adali, S.: Different spirals of sameness: a study of content sharing in mainstream and alternative media. In: International AAAI Conference on Web and Social Media, vol. 13, pp. 257–266, Germany (2019)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, US (2018)
12. Szczepański, M., Pawlicki, M., Kozik, R., Choraś, M.: New explainability method for BERT-based model in fake news detection. Sci. Rep. **11**, 23705 (2021)
13. Corvino, J.: The fact/opinion distinction. Philosophers' Mag. **65**(2), 57–61 (2015)
14. Cambria, E., Poria, S., Gelbukh, A., Thelwall, M.: Sentiment analysis is a big suitcase. IEEE Intell. Syst. **32**, 74–80 (2017)
15. Chaturvedi, I., Cambria, E., Welsch, R.E., Herrera, F.: Distinguishing between facts and opinions for sentiment analysis: survey and challenges. Inf. Fusion **44**, 65–77 (2017)
16. Alhindi, T., Muresan, S., Preotiuc-Pietro, D.: Fact vs. opinion: the role of argumentation features in news classification. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 6139–6149, Spain (2020)
17. Carrillo-de-Albornoz, J., Aker, A., Kurtic, E., Plaza, L.: Beyond opinion classification: extracting facts, opinions and experiences from health forum. PLoS ONE **14**(1), e0209961 (2019)
18. Blackledge, C., Atapour-Abarghouei, A.: Transforming fake news: robust generalisable news classification using transformers. In: IEEE International Conference on Big Data, Virtual (2021)
19. He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: decoding-enhanced BERT with disentangled attention (2020)

20. Gencheva, P., Koychev, I., Marquez, L., Barron-Cedeno, A., Nakov, P.: A Context-aware approach for detecting check-worthy claims in political debates. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, Bulgaria (2017)
21. Arslan, F., Hassan, N., Li, C., Tremayn, M.: A benchmark dataset of check-worthy factual claims. In: The International AAAI Conference on Web and Social Media, US (2020)
22. Konstantinovskiy, L., Price, O., Babakar, M., Zubiaga, A.: Towards automated Factchecking: developing an annotation schema and benchmark for consistent automated claim detection. Digit. Threats Res. Pract. **2**(2), 1–16 (2021)
23. Jha, R., Motwani, E., Singhal, N., Kaushal, R.: Towards automated check-worthy sentence detection using gated recurrent unit. Neural Comput. Appl. **35**, 11337–11357 (2023)
24. Lee, F.L.F.: Social media and the spread of fake news during a social movement: the 2019 anti-ELAB protests in Hong Kong. Commun. Public **5**(3–4), 122–125 (2020)