Research on the Reliability and Fairness of Opinion Retrieval in Public Topics

Zhuo Chen
Wuhan University
chenzhuo432@whu.edu.cn

Jiawei Liu
Wuhan University
laujames2017@whu.edu.cn

Haotan Liu
Wuhan University
baker-haotanliu@whu.edu.cn

***Abstract*—Neural network models have been widely applied in the field of information retrieval, but their vulnerability has always been a significant concern. In retrieval of public topics, the problems posed by the vulnerability are not only returning inaccurate or irrelevant content, but also returning manipulated opinions. One can distort the original ranking order based on the stance of the retrieved opinions, potentially influencing the searcher's perception of the topic, weakening the reliability of retrieval results and damaging the fairness of opinion ranking. Based on the aforementioned challenges, we combine stance detection methods with existing text ranking manipulation methods to experimentally demonstrate the feasibility and threat of opinion manipulation. Then we design a user experiment in which each participant independently rated the credibility of the target topic based on the unmanipulated or manipulated retrieval results. The experimental result indicates that opinion manipulation can effectively influence people's perceptions of the target topic. Furthermore, we preliminarily propose countermeasures to address the issue of opinion manipulation and build more reliable and fairer retrieval ranking systems.**

I. Introduction

The people's opinion has always been susceptible to the information encountered. Nowadays, we often use search engines or information retrieval tools to obtain the information we need. Narrowly defined Information Retrieval (IR) refers to the use of specific devices and tools, using a series of methods and strategies to search for the required information from a large collection of documents [20]. The information to be retrieved may be a document, an image, a video segment, etc., collectively referred to as candidate items in this paper. Currently, information retrieval tools widely employ pre-trained neural network models, outperforming traditional retrieval models because of their efficiency and effectiveness in recalling and relevance ranking. However, existing research has identified certain reliability defects in neural network models [15, 19, 31, 32]. Adding specific textual perturbations to candidate items can manipulate the relevance ranking of pre-trained neural network retrieval models, resulting in ranking results unrelated to the information need of the user. Moreover, we further find that the vulnerability of information retrieval models or recommendation models can allow people's opinions to be manipulated by unfair rankings. As depicted in Figure 1, the ranking top dominated by opposition opinions makes people's attitude towards the public topic change from neutral to opposing.

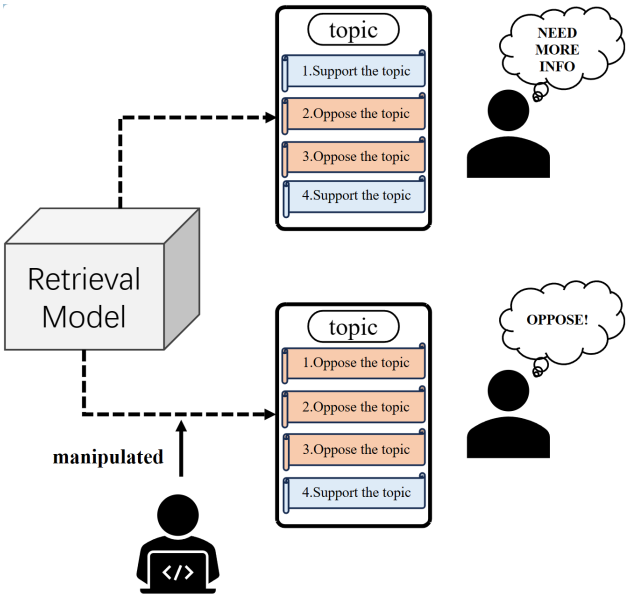

Fig. 1: Compared to the original ranking results, manipulated ranking results by others may alter people's attitudes toward the target topic.

In the scenario of searching for information on public topic, the issues arising from the reliability defect in information retrieval models go beyond returning irrelevant results. It is known that the homogeneity of acquired information within the "information cocoons" has become a scholarly concern for a long time. One form of informational homogeneity refers to individuals noticing content homogeneity phenomenon in the information presented by online media [29]. It is typically because recommendation systems generate "filter bubbles" that provide biased information [24]. Public topics often involve opinions from different stances on the same event. Nevertheless, the information cocoons can lead to the homogenization of user opinions [4, 27]. If the top-ranked results retrieved for a public topic exhibit semantically similar or consistent

Workshop on AI Systems with Confidential Computing (AISCC) 2024
26 February 2024, San Diego, CA, USA
ISBN 979-8-9894372-4-5
https://dx.doi.org/10.14722/aiscc.2024.23010
www.ndss-symposium.org

opinions instead of opinions from different perspectives, the unfair ranking may make users perceive these similar opinions as mainstream. Consequently, users are likely to only consider these opinions as the starting point for thinking and discussion.

Furthermore, if the rank of the retrieved results is based on the candidates items' stances rather than their relevance to the search query, this falls into the category of information manipulation. Rubin et al. [28] suggest that information manipulation is primarily manifested through falsification, exaggeration, concealment, misinformation or hoax. Information manipulation often serves specific interests, transmitting information with inaccuracies or biases to the target audience, thereby misleading their opinions and behaviors. A distinction between information manipulation and the information cocoons lies that the emergence of the information cocoons is driven by personalized needs, while information manipulation generally stems from ulterior motives. Moreover, compared to information cocoons, information manipulation typically targets a large volume of information rather than a few pieces, and it is directed at groups rather than individuals.

The impact of information manipulation is profound, and depending on the manipulative motives, it may pose significant social harm. Epstein et al. [10] find that manipulating search engines to produce biased search results could alter voting outcomes, with a transformational magnitude of at least 20%. As a result, they argue that search engine companies had a significant impact on local and even national politics. Other scholars have also discovered that the ranking of search results has a substantial influence on consumer preference attitudes and behaviours [7, 13]. Information manipulation in retrieval ranking tasks can change the rank of target candidate items, damaging the retrieval reliability and the ranking fairness. In this paper, retrieval reliability refers to robustness, the ability of retrieval systems to output stable and correct results even with abnormal input, and optimizing ranking fairness means that different candidate item groups (e.g., opinions with different stance) will have equal opportunities of exposure [35].

In this paper, we assume that the ranking of opinion candidates in retrieval results potentially affect users' perspectives on public topics, subsequently manipulating public opinion on these controversial subjects. For candidate items containing opinions, the ranking result is manipulated based on items' stances by some methods, with certain types of distortions occur to the ranking result [28]. We refer to this process as opinion manipulation in the ranking scenario.

However, current research on manipulation at the level of opinion stance mostly relies on manual work. Moreover, it is not large enough in scale [5, 12, 21], which is no longer practical given the immense scale of information data and the continuous improvement of computing power. Especially, current research focuses more on opinion manipulation in online media instead of in information retrieval [5, 7, 12, 21, 22, 33], there is limited research on opinion manipulation in information retrieval, and automated manipulation in online media often uses bots which are rarely applied in information retrieval [5, 12, 33]. Additionally, manual manipulation or bot manipulation struggles to identify the reliability weaknesses of information retrieval models, making it hard to effectively assess the retrieval reliability and even enhance model robustness to address opinion manipulation problem.

Based on the assumption and problems mentioned earlier, to explore the feasibility and actual impact of automated opinion manipulation, and enhance the retrieval reliability and ranking fairness in retrieval of public topics, this paper poses the following research questions:

(1) Whether automated manipulation of opinions ranking can be achieved in the retrieval of public topics.

(2) Whether automated manipulation of opinions ranking is significantly impacting users' perceptions of public topics.

(3) How to address retrieval reliability and ranking fairness problems caused by opinion manipulation in retrieval of public topics.

In addressing the aforementioned research questions, this paper focuses on opinion texts related to public topics. Combining stance detection methods and information manipulation techniques, we identify the stance of candidate items and manipulate the positions of these items in the ranking of information retrieval models. The manipulation effects are evaluated on our constructed datasets. Subsequently, the paper conducts user experiments to assess the real-world impact of the proposed opinion manipulation methods on user cognition. We also propose strategies to deal with opinion manipulation, exploring how to provide more reliable and fairer opinion retrieval ranking results.

The subsequent structure of this paper is as follows: the second part provides a brief overview of relevant research on the robustness of information retrieval and stance detection. The third part outlines the basic framework and experimental procedures for implementing opinion manipulation. The fourth part presents the experimental results and analysis of opinion manipulation, demonstrating its feasibility and threat. The fifth part empirically illustrates the impact of opinion manipulation on user cognition. The final part concludes the paper and suggests prospects for future research.

II. Literature Review

There has been a considerable amount of research on manipulating information retrieval results. Scholars have attempted various methods to manipulate information effectively. The fundamental principle of manipulation in the ranking scenario is to enhance the relevance of target ranking candidates to the query, ensuring their top positions in the final ranking. The targets to be manipulated typically include the ranking model, the query, and the candidate items. However, as manipulators find it challenging to modify user queries in practice and are unable to know the details of the ranking model, this paper primarily focuses on manipulating the ranking of candidate items, boosting their rank.

Information manipulation can be classified based on the manipulation conditions. Depending on the manipulator's understanding of the target ranking model, manipulation methods

can be categorized into white-box, gray-box, and black-box approaches [18, 19]. The focus of this paper is to manipulate the ranking results of information retrieval models under the white box setting, as starting from the most basic white box setting can lay the foundation for future research on manipulating opinions closer to the real world. Under white-box setting, manipulators can access information such as the architecture, parameters, and internal data transmission of the information retrieval model, essentially commanding complete knowledge of the model. Therefore, information manipulation under white-box setting is the simplest. Under black-box setting, manipulators can only obtain the output of the target model, with other information being unknown. Gray-box setting falls between white-box setting and black-box setting.

A. Information Retrieval Models and Their Robustness Evaluation

Early information retrieval models adopted traditional models based on exact word matching. These models typically use statistical metrics such as term frequency, inverse document frequency, or combinations of some metrics to calculate relevance scores. A representative example is the BM25(Best-match weighting function implemented in Okapi) algorithm proposed by Robertson in 1994 [26], which comprehensively considers the importance of words, the correlation between words and documents, and the correlation between words and queries. The advantage of such traditional models lies in their processing speed, making the BM25 algorithm widely used in large-scale document retrieval and recall tasks. However, the drawback of traditional models is the inability to match all the occurring words, and they struggle with handling polysemy as well as learning about semantic-level correlations.

Later, people adopted feature-based ranking models trained by supervised training with pre-constructed features. These features can be statistical, text-based, or matching features between documents and queries. The difference between feature-based models and traditional ranking models is that traditional models are generally unsupervised, while feature-based ranking models are trained with supervised learning algorithms such as SVM, decision trees, etc. Feature-based models enhance the retrieval accuracy by supervised training, but their ability to model semantic matching between queries and candidate documents remains insufficient.

People have gradually turned to using neural network models as ranking models because of its powerful modeling capability to explore semantic relationships between queries and candidate documents, resulting more accurate matching. The modeling capacity of neural network is generally proportional to their scale of parameters, so adopting pre-trained models can maintain a strong ranking modeling capability while reducing the cost of fine-tuning. Therefore, pre-trained models have become the mainstream in current information retrieval models. Yang et al. [34] segment candidate items into sentences, employed BERT to model the relevance score for each sentence and then aggregated sentence scores to produce candidate item scores. BERT stands for Bidirectional Encoder Representations from Transformers and it is a pre-trained model that can be fine-tuned with just one additional output layer to perform well on a wide range of tasks, such as question answering and language inference [8]. Nogueira et al. [23] employ the T5 pre-trained model to generate multiple relevant questions for each candidate document and then used these questions to expand the document representation, improving the ranking result.

However, a defect of neural network models is reliability vulnerability, meaning that when input with abnormal inputs, they struggle to consistently produce normal results. In 2014, Szegedy et al. [31] found that applying imperceptible perturbations to a neural network model during a classification task was sufficient to cause classification errors in CV. Later, scholars observed similar phenomenon in NLP(Natural Language Processing). Robin et al. [15] find that inserting perturbed text into original paragraphs significantly distracts computer systems without changing the correct answer or misleading humans. The error caused by perturbation reflects the ability of models to output stable and correct predictions in tackling the imperceptible additive noises, thus helping evaluate the reliability and robustness of neural network models [32]. Many scholars have conducted research on the evaluation of the reliability of neural network models and tried to improve it [9, 18, 19, 30]. Most of the reliability evaluation research approached the issue from the attack perspective, detecting vulnerabilities of target models by perturbation and distortion. Nevertheless, there is comparatively less research from the defense or user perspective.

In NLP, the evaluation of model vulnerabilities from the attack perspective begins with rule-based heuristic methods. Robin et al. [15] employ two methods to construct perturbation texts in reading comprehension and question answering tasks. The first is under black-box setting, they apply word replacement to the target question and create a fake answer to it with predefined types, generating the natural perturbation text via crowd-sourcing. The second is under gray-box setting, the method also constructs perturbation text by word replacement, but it utilizes predicted probability optimization of the target model to get the final perturbation sample. The first method mentioned relies too much on rules and pre-defined patterns while the second generates perturbation text with insufficient semantic fluency. Moreover, these two methods tend to directly construct perturbation text with the words in the question, making the perturbation text easily detectable. The method proposed by Robin et al. to evaluation the vulnerabilities of a model is based on an observation: applying certain perturbations to an image does not change its semantics in CV, while even a slight modification to a text can completely alter its meaning in NLP. During the manipulation process, we expect that the manipulated text retains its original semantics, but it is required to deceive the target model.

Under white-box setting, Ebrahimi et al. [9] utilize an atomic flip operation, which swaps one token for an other, to generate perturbation examples and the method, known

as Hotflip. Hotflip gets rid of reliance on rules, but the perturbation text it generate usually has incomplete semantics and insufficient grammar fluency. While it can deceive the target model, it cannot evade perplexity-based defenses. To further enhance the quality of perturbation text, Song et al. [30] propose an similar method under white-box setting, named Collision, which uses gradient optimization and beam search to produce the perturbation text named collision. The Collision method further imposes a soft constraint on collision generation by integrating a language model, reducing the perplexity of the collision. The method has shown promising results in document retrieval experiments. Inspired by Collisions, Liu et al. [18] propose the Pairwise Anchor-based Trigger (PAT) method under black-box setting. Added the fluency constraint and the next sentence prediction constraint, the method generates perturbation text by optimizing the pairwise loss of top candidates and target candidates with perturbation text. Although the time complexity of PAT has increased compared to previous methods, PAT takes ranking similarity and semantic consistency into account, so its manipulation effect on the retrieval ranking of the target candidates is superior.

This paper mainly adopts Collisions and PAT methods to manipulate the opinions on public topics, evaluating the reliability and fairness of opinion retrieval.

B. Stance Detection

The relevant information retrieved in public topics may include opinions with neutral stance or content without opinions at all, so it is necessary to determine whether there are opinions in the retrieved content and what kind of stances they are, which is essentially a classification task called stance detection. Stance detection are generally divided into two categories: feature-based detection and data-driven machine learning detection [3]. Feature-based detection often requires people to construct features, such as special words used in the controversy, the author's social activities and so on. However, these constructed features, such as special words and bag of words, are difficult to reflect the whole semantic or even the attitude of the text.

Machine learning methods for stance detection can be categorized into four types: supervised learning, unsupervised learning, weakly supervised learning, and transfer learning. In the early days, supervised learning methods for stance detection used SVM, LR, KNN, and other techniques. However, with the widespread use of neural networks, recurrent neural networks including LSTM and GRU have become mainstream stance detection models. Wei-Fan Chen et al. [6] have even employed CNN to extract stance features, further enhancing the effectiveness of stance detection. While supervised learning enables models to effectively learn about knowledge of stance, it requires a large amount of labeled data to achieve satisfactory results. At the same time, the cost of training increases with the growth of data, and there may even be issues like over-fitting. Therefore, other scholars have explored unsupervised learning, weakly supervised learning, and other methods for stance detection. Unsupervised stance detection is either based on user features (such as posting history) or relies on certain rules or grammatical dependencies, but neither aligns with the text-based requirements of this paper. Weakly supervised stance detection typically use a small amount of labeled data to train a simple classifier, which is then used to annotate unlabeled data for further training. However, the method is not ideal for stance detection.

With the remarkable effectiveness of transfer learning, powerful pre-trained models like GPT, BERT, ELMo, etc., have emerged. People increasingly use pre-trained models trained on data from other domains to detect stances. Fang et al. [11] utilized multi-task trained BERT for stance detection on the FNC-1 dataset. Popat et al. [25] enhance the effectiveness of the stance detector by adding consistency constraints on claim and perspective during BERT fine-tuning. Detection methods based on transfer learning make full use of the capabilities of large pre-trained models, reducing training costs. Additionally, fine-tuning on the pre-trained model allows the model's detection performance on target domain to be further improved and BERT achieves the state-of-the-art stance detection performance on many datasets [3]. Therefore, in this paper, we adopt transfer learning by fine-tuning BERT to construct the stance detection model.

III. Opinion Manipulation

This chapter introduces the adopted method for automated opinion manipulation, target dataset, target model, and experimental details.

To achieve automated opinion manipulation in public topics, this paper initially fine-tunes BERT using a portion of the target data to construct a opinion detector. Subsequently, this detector is employed to check whether opinions exist in the candidate items and, if present, to identify the stance of these opinions towards the topic. The task is a three-class classification task ("support", "oppose", "others"), in which the "others" category encompasses three cases: opinions with neutral stance, content without discernible opinion, or the target candidate item is unrelated to the topic. After the detector categorizes the target candidate items, Collisions and PAT are employed separately to generate perturbation texts for target candidate items containing opinions as they are representative ranking manipulation methods. These perturbation texts are inserted at the beginning of the target candidate items, and we rank these items again to evaluate the retrieval reliability and ranking fairness. This process is illustrated in Figure 2.

(1) Collisions. It generates perturbation text using the similarity between the generated text and the query under white-box settings, aiming to make the perturbation text ranked very high even though it is irrelevant to the query. This method utilizes gradient optimization and applies beam search to find the words of the perturbation text, then iteratively repeats these steps until the similarity score converges, generating the final perturbation text in an auto-regressive way. This collision, perturbation text without any constraints, is denoted as T_{aggr}. Semantic soft constraint can be imposed on collision

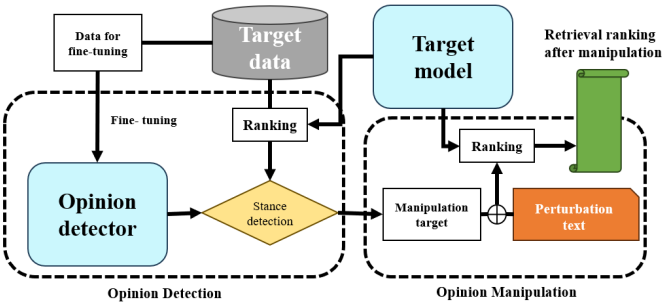

Fig. 2: Opinion manipulation framework.

generation, steering a pre-trained language model(LM) to generate appropriate words, resulting in the low-perplexity text that can evade perplexity filtering. This collision is denoted as T_{nat}.

(2) PAT. Inspired by Collisions, it adopts a pairwise generation paradigm. Given the target query, the target candidate item, and the top candidate item(anchor, used to guide the perturbation text generation), the method utilize gradient optimization of pairwise loss, calculated from the candidate item and the anchor, to find the appropriate representation of a perturbation text. The method also add flunecy constraint and next sentence prediction constraint. By beam search for the words, the final perturbation text, denoted as T_{pat}, is iteratively generated in an auto-regressive way.

The dataset used in this paper is Fake News Challenge (FNC-1). The purpose of the dataset is to explore how AI technologies might be leveraged to solve the fake news problem [1]. The first step towards identifying fake news is to understand others' comments on the topic. Therefore, the primary task of FNC-1 is stance detection. The dataset is derived from news websites, includes a series of news headlines and news articles, along with the opinions of other news organizations on a given news headline. Stance categories include "agree", "disagree", "discuss" and "unrelated".

This paper selects the neural ranking model on NBoost as the target model for experiments. NBoost is a scalable, search-engine-boosting platform developing models to improve the relevance of search results [2]. Nboost sequentially combines Transformer models like BERT and word-based searching engines like Elasticsearch using BM25 to improve domain-specific searching. Using BERT fine-tuned based on the MS-MARCO dataset and TREC-CAR dataset, Nboost can greatly boost search relevance metrics and improve downstream ranking task. There are many research applying Nboost to ranking tasks [16, 18]. Thus, we denote the model in Nboost as BERT-msmarco. On Bing queries dataset and the biomed dataset, BERT-msmarco improves the MRR (Mean Reciprocal Rank) of retrieval ranking results by nearly two times compared to retrieval methods based on BM25 algorithm.

For capabilities of the adversary, we focus on white-box setting where manipulators have full knowledge of the target model, including the model parameters, the model architecture and the score function. Additionally, manipulators can access

TABLE I: Fine-tuning BERT experiment for stance detection.

Model	Setting	Acc	P	R	F1
BERT-1	Epoch=10,lr=1e-5,batch=32	0.942	0.817	0.816	0.816
BERT-2	Epoch=20,lr=1e-5,batch=32	0.958	0.868	0.878	0.872
BERT-3	Epoch=20,lr=2e-5,batch=32	0.972	0.912	0.909	0.910

to queries of public topics and target candidate documents. In the generation of perturbation text, the text length is set to 10, the learning rate is set to 0.05, the number of beams is set to 50 and the maximum number of iterations is set to 2.

The main evaluation metrics used in opinion manipulation include the Average Percentage of target stance (APCT), the Average Percentage Variation of target stance after manipulation (APV), the Average Boost Rank of target stance (ABR), and the Normalized Discounted Cumulative Gain (NDCG).

(1) APCT and APV: Song et al. [30] used the proportion of documents ranking in the top-100 after inserting perturbation text to reflect the success rate of manipulation. We assume that the ranking of opinion candidate items in the retrieval results may influence users' perspectives on public topics. If a certain stance dominate the top candidate items in ranking, users are more likely to be influenced by the stance. Moreover, the larger the proportion of a certain stance is, the more users tend to be influenced by the stance, indicating a more successful opinion manipulation. Therefore, we adopt the Average Percentage of target stance (APCT) after manipulation and the Average Percentage Variation of target stance (APV) after manipulation to reflect the effect of opinion manipulation. Higher APCT value and larger APV value indicate a better effect of opinion manipulation. Specifically, we adopt APCT@K and APV@K to represent the APCT and APV values for the top-K candidate items in the ranking after manipulation. For example, APCT@10 represents the APCT value for the top-10 candidate items in the ranking result list after manipulation.

(2) ABR: The Average Boosted Rank provides a direct representation of the opinion manipulation effect. This metric is commonly used to evaluate the effectiveness of manipulation in information retrieval ranking tasks. Liu et al. employed the average boosted ranks (avg. Boost) metric to assess the manipulation, the greater the average boosted ranks of candidate items due to the perturbation [18], the better the success of the manipulation.

(3) NDCG: Normalized Discounted Cumulative Gain (NDCG) is one of the most effective evaluation metrics in ranking tasks, proposed by Kalervo et al. [14]. It accumulates the scores of all candidate items in the ranking and discounts the score of each candidate item based on its rank, giving higher gains to items ranked closer to the top. To facilitate the comparison of rankings of different queries, the NDCG value is normalized based on the discounted cumulative gain in the most ideal state. NDCG takes several factors into account, including scores of candidate items, ranking positions of candidate items, and numbers of candidate items in rankings of different queries, providing a relatively comprehensive

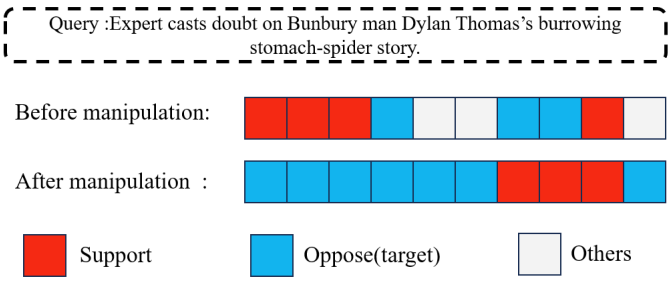

Fig. 3: Opinion manipulation illustration.

reflection of ranking quality. In this study, the NDCG metric is modified slightly for opinion manipulation. Candidate item scores are not given based on the relevance but are instead based on the target stance, which means the highest score is given to candidates with a stance aligning with the target stance, lower score is given to candidates with a neutral stance or no stance, and the lowest score is given to candidates with a stance opposite to the target stance. This rewards those manipulations that achieve higher rankings for the target stance as much as possible. In the experiments, candidates supporting the target stance receive a score of 3, candidates with a neutral stance or no stance receive a score of 1, and candidates with a stance opposite to the target stance receive a score of 0. In this paper, NDCGV stands for the Normalized Discounted Cumulative Gain Variation of target stance after manipulation. NDCGV@K represents the NDCGV value obtained for the top K candidate items in the ranking.

IV. Manipulation Experiment and Analysis

In retrieval of public topics, to achieve automated opinion manipulation, the first step is to detect opinions and their stances in the text. This paper adopts the transfer learning approach to train a stance detector by supervised fine-tuning BERT. The FNC-1 dataset consisting of news topics and their corresponding news texts with stance labels is divided into two parts. One part includes news texts with both "agree" and "disagree" labels for each topic and it is used for conducting opinion manipulation experiments. The remaining part, in which there may be some topics missing certain type of stance, is utilized for fine-tuning the stance detection model.

In this paper, BERT model is fine-tuned under different parameter settings, and the detection test results are presented in Table 1. The evaluation of detection effectiveness is conducted using precision (P), recall (R), and F1 score(F1). BERT-3, in which "3" is only used to distinguish among BERT models under different parameter settings, is chosen as the opinion detector.

After constructing the opinion detector, we utilize the detector to classify the stances of those opinions in the ranked target data. Stances are categorized as "support", "oppose", and "others", represented by the numbers 1, 0, and 2, respectively in the experiments. After the detector classifies the target data, we employ the Collisions and PAT methods to generate perturbation texts T_{aggr}, T_{nat}, and T_{pat} for target candidate items containing opinions. These perturbation texts are inserted at the beginning of the target candidate items. Subsequently, we re-rank these candidates to evaluate the manipulation, as shown in Table 2.

As shown in Table 2, both Collisions and PAT methods yield ABR and NDCG values greater than zero. It indicates that three types of perturbation texts can signigicantly achieve automated manipulation of opinion retrieval results in public topics, ranking the opinions with the target stance as high as possible. It addresses the first research question of this paper. In Figure 3, we show the case of using the PAT for opinion manipulation in the retrieval of topic "A spider crawls into a man's body". The "query" represents an opinion with a particular stance, and each news candidate, represented as a square, is coloured by its stance towards to the "query". Red stands for "Support", Blue stands for "Oppose". This example illustrates how the opinion manipulation can significantly alter the ranking of search results, the ranking consists of some coloured squares and its left side is the ranking top.

Moreover, it can be observed that different manipulation methods have varying effects on the manipulation of opinion candidates. The perturbation text T_{pat} generated by the PAT demonstrates the best manipulation effect on opinions with different stances in public topics. It achieves the highest average percentage for both top 5 candidate items and top 10 candidate items after opinion manipulation, and it achieves the greatest increase in NDCG values for final ranking. The manipulation effect of T_{nat} is worse than T_{pat}, but its scores on APCT, ABR, and NDCGV indicators are relatively close to those of the T_{pat}. The manipulation effect of T_{aggr} is the worst, with a relatively small increase in the rankings of target candidate items. That is because Collisions generates T_{aggr} solely based on gradient optimization, creating perturbation texts that are usually nonsensical. These texts often contain meaningless characters, limiting its deceptive effect on retrieval ranking models. In contrast, T_{nat} adds semantic constraints, resulting in more fluent perturbation texts with a stronger deceptive effect on retrieval ranking models. PAT goes further on the basis of T_{nat} by taking information from candidate items into account, adding constraints of consistency between the preceding sentences and following sentences, and using anchor candidates to construct perturbation texts by pairwise contrastive learning. T_{pat} exhibits stronger semantic consistency with the query and the target candidate item. Moreover, the gradient-based pairwise learning allows PAT to better identify vulnerabilities in retrieval ranking models. Therefore, T_{pat} demonstrates the best opinion manipulation effect.

V. User Perception Experiment and Analysis

To further answer the second research question of this paper, which investigates whether automated manipulation of opinions ranking is effectively impacting users' perceptions of public topic, we designed a user experiment to explore the practical effects of opinion manipulation.

TABLE II: Opinion manipulation experiment result and bold shows the best performance.

Target Data	Method	APCT@5	APV@5	APCT@10	APV@10	ABR	NDCGV@10
FNC-1: Stance "Oppose"	Collisions (T_{aggr})	0.286	-0.014	0.3	0.021	0.289	0.041
	Collisions (T_{nat})	0.743	0.443	0.521	0.243	4.378	0.333
	PAT (T_{pat})	**0.757**	**0.457**	**0.550**	**0.271**	**5.5**	**0.394**
FNC-1: Stance "Support"	Collisions (T_{aggr})	0.557	0.114	0.464	-0.007	0.116	0.001
	Collisions (T_{nat})	0.743	0.3	0.586	0.114	2.268	0.132
	PAT (T_{pat})	**0.8**	**0.357**	**0.657**	**0.186**	**3.054**	**0.198**

TABLE III: User information credibility rating.

Subject	Information Credibility Topic 1	Topic 2
A Group(before manipulation)	5.8	4.8
B Group(after manipulation)	6.4	8.2

We gathered 10 subjects who are PhD students, and divided them into two groups, i.e., A and B, with each group consisting of 5 people. Participants in both groups were unaware of the purpose of this experiment, and communication was not allowed among the participants during the experiment. We let members of Group A read the search results without manipulation and rated the credibility of the two public topics selected for the experiment. Information credibility is defined as the extent to which one perceives information to be believable [17]. It indicates how trustworthy the participants think the news topic is. At the same time, Group B members read the search results manipulated under the same topics and rated the information credibility of that topics. Then we compared the information credibility participants rated in the specific topic before and after the opinion manipulation. The rating scale ranged from 1 to 9, in which a higher score implied a higher level of credibility in the topic, and vice versa. The rating results are presented in Table 3.

It can be observed that there is a significant difference in the perceived credibility of the same public topics before and after manipulation for Group A and Group B. Moreover, participants in Group B perceived higher credibility for the topics compared to Group A. It is because after inserting perturbation text into the target candidate items, the retrieval model ranked candidate items, which endorsed the authenticity of the public topics, higher at the top, increasing participants' trust in the topic and leading to a change in the stances of their opinions. This experiment answers the second research question, confirming that automated manipulation of opinions ranking is effectively impacting users' perceptions of public topics.

VI. Discussion and Conclusion

In retrieval of public topics, this paper proposes to manipulate the ranking of opinion candidates in search results by exploiting the reliability vulnerabilities of information retrieval models. The aim of the manipulation is to influence users' stances on public topics. In the experiment, it is demonstrated that automated manipulation of opinions ranking is achievable. Additionally, to further clarify whether this manipulation can effectively impact users' perception, we design a user experiment. Different subject groups simulated retrieval for public topics before and after opinion manipulation, providing feedback on the perceived credibility of the topics. The results indicate that manipulating the opinion ranking significantly impacts users' stance on the target topics, which means reliability vulnerability and unfair ranking in opinion retrieval have potential huge risks.

To address the retrieval reliability and ranking fairness problems caused by opinion manipulation in public topics if it happens in reality, our goal should be to fairly display diverse opinions with different stances. Therefore, the retrieval task on public topics should not rely solely on ranking candidates based on semantic relevance to the query. Currently, most perturbations on information retrieval models manipulate rankings by disrupting the relevance judgment of the models. So we can divide the retrieval of public topics into two stages. In the first stage, a certain number of relevant candidate items are obtained based on relevance and form a relevant item set. In the second stage, the stance detector is used to detect the stances of candidate items within the set. The ranking of those candidate items is achieved based on the combination of the candidates' stances and their relevance to the query. We can design algorithms to ensure that opinion candidate items with various stances are able to rank as high as possible. However, this method also brings the issue of reliability in the stance detector for further exploration. Additionally, we can explore clustering methods or other methods to summarize opinions with various stances on public topics and present them to users in visualization ways, aiding users in gaining a more comprehensive understanding of the related stances on public topic.

In conclusion, we investigate the manipulating the ranking list of opinion candidate items to influence users' perspectives on public topics in this paper. The feasibility of manipulation and the effectiveness of its impact on user perception have been explored by empirical experiments. However, the manipulation methods employed in this paper are relatively simple, and there is room for further improvement in the opinion manipulation effects. Thus, the assessment of reliability defect in information retrieval models are not enough. Furthermore, with the widespread use of large language models (LLMs), many information retrieval tasks are now being combined with LLMs, we will delve into opinion manipulation based on LLMs retrieval, comprehensively exploring the retrieval reliability and ranking fairness issues when LLMs is employed for opinion retrieval tasks.

Addressing Potential Ethical Concerns

Our objective is to enhance the resilience of ranking models. Throughout our research, we adhered to the ACM Ethical Code to mitigate any potential harm. It is true that the techniques devised in this paper could be misused to attack current IR systems using triggers or adversarial attacks, resulting in potential short-term negative consequences. Nevertheless, it is not our intention to inflict harm upon ranking models. Instead, we aim to publicly disclose these unintentional flaws, enabling the development of novel defense algorithms to safeguard against them in the future. This approach mirrors the actions of white hat hackers who publicly expose bugs or vulnerabilities in software.

We have demonstrated that automatic information manipulation of information retrieval results can be accomplished which reveals a greater threat than previous manual operation [10]. This indicates our work **provides a long-term benefit** to the community and can help to improve IR systems. More importantly, we **minimized real-world harm** by not exposing any real-world failure or damage to any real users.

References

[1] Fake News Challenge. http://www.fakenewschallenge.org/.

[2] koursaros-ai/nboost, January 2024. original-date: 2019-10-29T20:56:24Z.

[3] Nora Alturayeif, Hamzah Luqman, and Moataz Ahmed. A systematic review of machine learning techniques for stance detection and its applications. *Neural Computing and Applications*, 35(7):5113–5144, March 2023.

[4] Andrei Boutyline and Robb Willer. The social structure of political echo chambers: Variation in ideological homophily in online networks. *Political psychology*, 38(3):551–569, 2017.

[5] Long Chen, Jianguo Chen, and Chunhe Xia. Social network behavior and public opinion manipulation. *Journal of Information Security and Applications*, 64:103060, 2022.

[6] Wei-Fan Chen and Lun-Wei Ku. Utcnn: a deep learning model of stance classificationon on social media text. *arXiv preprint arXiv:1611.03599*, 2016.

[7] Chrysanthos Dellarocas. Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management science*, 52(10):1577–1593, 2006.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*, 2017.

[10] Robert Epstein and Ronald E. Robertson. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521, August 2015. Publisher: Proceedings of the National Academy of Sciences.

[11] Wei Fang, Moin Nadeem, Mitra Mohtarami, and James Glass. Neural multi-task learning for stance prediction. In *Proceedings of the second workshop on fact extraction and verification (FEVER)*, pages 13–19, 2019.

[12] Michelle Forelle, Phil Howard, Andrés Monroy-Hernández, and Saiph Savage. Political bots and the manipulation of public opinion in venezuela. *arXiv preprint arXiv:1507.07109*, 2015.

[13] Anindya Ghose, Panagiotis G Ipeirotis, and Beibei Li. Examining the impact of ranking on consumer behavior and search engine revenue. *Management Science*, 60(7):1632–1654, 2014.

[14] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[15] Robin Jia and Percy Liang. Adversarial Examples for Evaluating Reading Comprehension Systems, July 2017. arXiv:1707.07328 [cs].

[16] Jieh-Sheng Lee and Jieh Hsiang. Prior art search and reranking for generated patent text. *arXiv preprint arXiv:2009.09132*, 2020.

[17] Ruohan Li and Ayoung Suh. Factors influencing information credibility on social media platforms: Evidence from facebook pages. *Procedia computer science*, 72:314–328, 2015.

[18] Jiawei Liu, Yangyang Kang, Di Tang, Kaisong Song, Changlong Sun, Xiaofeng Wang, Wei Lu, and Xiaozhong Liu. Order-disorder: Imitation adversarial attacks for black-box neural ranking models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2025–2039, 2022.

[19] Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. Topic-oriented Adversarial Attacks against Black-box Neural Ranking Models, April 2023. arXiv:2304.14867 [cs].

[20] Christopher D Manning. *An introduction to information retrieval*. Cambridge university press, 2009.

[21] Todor Mihaylov, Georgi Georgiev, and Preslav Nakov. Finding opinion manipulation trolls in news community forums. In *Proceedings of the nineteenth conference on computational natural language learning*, pages 310–314, 2015.

[22] Todor Mihaylov, Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Georgi D Georgiev, and Ivan Kolev Koychev. The dark side of news community forums: Opinion manipulation trolls. *Internet Research*, 28(5):1292–1312, 2018.

[23] Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. From doc2query to doctttttquery. *Online preprint*, 6:2, 2019.

[24] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. penguin UK, 2011.

[25] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. Stancy: Stance classification based on consistency cues. *arXiv preprint arXiv:1910.06048*, 2019.

[26] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. *Nist Special Publication Sp*, 109:109, 1995.

[27] Daniel Röchert, German Neubaum, Björn Ross, Florian Brachten, and Stefan Stieglitz. Opinion-based homogeneity on youtube: Combining sentiment and social network analysis. *Computational Communication Research*, 2(1):81–108, 2020.

[28] Victoria L Rubin and Yimin Chen. Information manipulation classification theory for lis and nlp. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–5, 2012.

[29] Daniel Röchert, Gautam Kishore Shahi, German Neubaum, Björn Ross, and Stefan Stieglitz. The networked context of covid-19 misinformation: Informational homogeneity on youtube at the beginning of the pandemic. *Online Social Networks and Media*, 26:100164, 2021.

[30] Congzheng Song, Alexander M Rush, and Vitaly Shmatikov. Adversarial semantic collisions. *arXiv preprint arXiv:2011.04743*, 2020.

[31] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[32] Wenqi Wang, Run Wang, Lina Wang, Zhibo Wang, and Aoshuang Ye. Towards a robust deep neural network in texts: A survey. *arXiv preprint arXiv:1902.07285*, 2019.

[33] Zixuan Weng and Aijun Lin. Public opinion manipulation on social media: Social network analysis of twitter bots during the covid-19 pandemic. *International journal of environmental research and public health*, 19(24):16376, 2022.

[34] Wei Yang, Haotian Zhang, and Jimmy Lin. Simple applications of bert for ad hoc document retrieval. *arXiv preprint arXiv:1903.10972*, 2019.

[35] Yuying Zhao, Yu Wang, Yunchao Liu, Xueqi Cheng, Charu Aggarwal, and Tyler Derr. Fairness and diversity in recommender systems: a survey. *arXiv preprint arXiv:2307.04644*, 2023.