# Characterization and Detection of Political Manipulation in Online Communities

Sihyung Lee

Department of Information Security, Seoul Women's University, Republic of Korea
sihyunglee@swu.ac.kr

## Abstract

As an increasing number of people share their opinions in online communities, attempts to manipulate these opinions are also rising steadily. In particular, when manipulation is used for political purposes (e.g., to decrease the credibility of a presidential candidate), it can change the outcomes of elections, thus having a long-lasting, negative impact. This work is aimed at detecting such political manipulation. We collect 377K opinions from Internet forums at the peak of a presidential campaign in South Korea. We find various characteristics of political manipulation, including the distortion of opinion polls with numerous IDs and the continuity of the same, strong political inclination throughout manipulative posts. Based on these characteristics, we implement a detection system and evaluate it with the collected data. We demonstrate that this system accurately discovers more than 90% of online accounts involved in political manipulation. We also find clues indicating that greater than 10% of these accounts are collectively controlled by a few manipulative users. We believe that the proposed system can ensure people read trustworthy opinions and also relieve web administrators from manually screening for manipulative activities.

**Keywords**: Political manipulation, Online communities, Machine learning, Collective classification

## 1 Introduction

Online communities refer to Web services where participants can share their opinions, and examples are social networks and discussion forums. These communities have become an essential part of our daily lives, since they are convenient (i.e., communication is possible anytime, anywhere) and fun to use (i.e., interactions are supported through text, pictures, voice, and videos) [1]. Consequently, online communities have a great influence on people's thoughts and behaviors – users often switch their views and perform actions according to what they see in online communities [2]. Other studies show that online communities are considered as valuable and credible as traditional mainstream media (e.g., TV and newspaper [3-4]).

While online communities have gained popularity, they have increasingly been targets of misuse by malicious users, who we call *manipulators*. Manipulators circulate unfair negative opinions about a targeted object to foster discontent, and they also circulate undeserving positive opinions to develop preference for another object. In particular, we focus on the detection of manipulators in the *political* domain – these users disseminate propaganda and manipulate the credibility of a political party or a politician. Such manipulation can lead to serious and long-lasting consequences, such as changing election results [5] and provoking anti-government protests [6].

Several past studies characterize manipulative activities, but they mostly focus on consumer communities and are not always applicable to the political domain. For instance, functionalities are different – a product is often rated based on the five-star rating system, and a major sign of manipulation is the deviation of ratings from the average [7]; such a rating system does not exist in many political discussion forums, and users rather cast votes on political agendas, which monotonically increase far beyond five. In addition, word usage is different – positive or negative connotation is considered important in product reviews [8], whereas political inclination plays a role in propaganda [9]. User behavior is also different – a political discussion tends to create longer threads of opinions than consumer reviews as a debate continues; accordingly, it is not uncommon for a user to post repeatedly on the same thread, which is not as common in consumer communities [10].

To better characterize the peculiarities of political manipulation, we carefully investigate real incidents. In particular, we track large-scale, nationwide manipulative activities during a presidential campaign in South Korea [11-12]. These activities involve several parties, government agencies, as well as the military intelligence of surrounding countries. We monitor two popular discussion forums over a six-month period and collect 370K opinions posted by

[*]Corresponding Author: Sihyung Lee; E-mail: sihyunglee@swu.ac.kr

DOI: 10.3966/160792642018081904016

120K users. This dataset leads us to identify various characteristics of political manipulation. For example, manipulators post opinions at the earliest possible time after the onset of a discussion, and they then rapidly vote on these opinions with multiple IDs. In this way, the opinions can be exposed to a large audience for an extended period of time. The IDs also post opinions at nearly similar times, one after another, and these opinions continue to exhibit the same, strong political inclinations.

We demonstrate that the found characteristics combined with a collective classification can accurately discover more than 90% of manipulators; less than 0.5% of normal users are falsely classified as manipulators. We also find groups of colluding IDs; one of the largest groups includes 80 IDs that are concurrently used to manipulate more than a thousand opinions.

## 2 Related Work

Several sources present the prevalence of political manipulation in online communities. Nationwide manipulations are reported in Italy [4], Russia [6], and South Korea [11-12]; human and computer generated messages flood online communities, influencing elections and obstructing anti-government protests. To effectively distribute political propaganda, manipulators utilize various tactics [9], such as the repetition of the same catch phrases, and the use of numerous IDs to disguise a campaign as grassroots movements. As a result, once published false information can be quickly adopted by the public, within a month, and it is even used as supporting evidence in subsequent debates [4]. In addition, manipulative opinions are shown to change real-world voting behavior of up to 10% of readers [5], which is sufficient to change the results of competitive elections. In fact, many elections are competitive, and their outcomes can be overturned with a few changes in votes. For example, in the 2000 US presidential election, George Bush won Florida's electoral votes by only 537 votes out of almost 6 million casts. If Bush had lost in Florida, he could have lost the entire presidential election [5].

So far, a few studies deal with the detection of political manipulation in online communities. Ratkiewicz et al. [13] identify manipulative opinions by analyzing the diffusion pattern – the way in which political opinions are retweeted. For instance, manipulative opinions are first mentioned by a small set of users and then retweeted repeatedly by nearly the same set of users. The retweet capability does not have an exact equivalent in political discussion forums, but it is similar to posting duplicate opinions. As such, the diffusion pattern can be used in conjunction with the attributes proposed in this work and improve performance. The work by Lee [10] detects political manipulation by measuring the degree of collaboration. It utilizes the fact that effective manipulation requires multiple users to work together for an extended period of time. However, manipulators can evade detection by utilizing a large pool of IDs in turn and leaving few clues with each ID. We address this problem by using relational attributes and iterative classification. For example, we identify a group of IDs as being potentially used by the same manipulator if these IDs are used concurrently to spread similar ideas and if they maintain the same political inclination. In addition, we further investigate anomalies in online polls that are frequent targets of manipulators.

Several other studies focus on product-review sites and find manipulative consumer feedback [14-15]. These studies frequently refer to two attribute sets. One set is concerned with the five-star rating system that appears on most product-review sites. These attributes measure to what extent a user's ratings deviate from the average, as well as how consistent a user's ratings are throughout multiple reviews. The other set characterizes opinion text, such as text length and the number of positive words. Kwon et al. [16] also investigate the timing between successive posts. Manipulative reviews often occur in a burst, and such bursts reoccur multiple times. In summary, prior research often utilizes the five-star rating system, which do not always exist in political discussion forums. Other attributes, such as those related to opinion text, can be used to complement the proposed system.

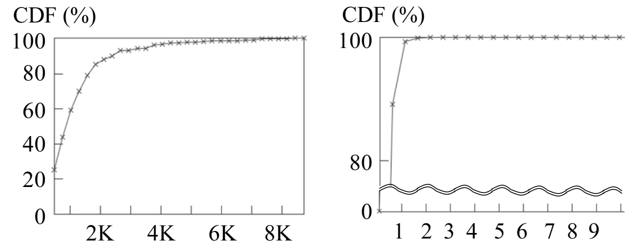
## 3 Description of Dataset

### 3.1 Collection Target and Methodology

To characterize political manipulation in online communities, we build a dataset in which each user is labeled as either a manipulator or a non-manipulator. To this end, we collect data from two popular discussion forums in South Korea – Naver and Nate (news.naver.com and news.nate.com, respectively). These data include 377K opinions written by 121K users. Table 1 summarizes numbers related to the collected data, and Figure 1 shows the cumulative distributions.

**Table 1.** Summary of dataset – # of monthly users found in KoreanClick [17]

| Site | # of topics | # of opinions | # of users |
|---|---|---|---|
| Nate (~17.2 million users per month) | 77 | 30,800 | 10,266 |
| Naver (~36.3 million users per month) | 236 | 346,834 | 111,013 |
| Total | 313 | 377,634 | 121,279 |



(a) # of opinions per topic (b) voting time after onset of discussion (days)

**Figure 1.** Cumulative distributions of collected data

The collection occurs over a six-month period immediately before the last presidential election in South Korea (2012.07.01-12.19). During this period, manipulative opinions culminated to influence the election. Some of these opinions are posted by individuals hired by politicians [18] and others are posted by government agencies and army intelligence, as ordered by the ruling party [19]. These circumstances were further muddled up by North Korean cyber military, who aims to damage the reputation of candidates (as these candidates maintain strong stance against North Korea), and also to instigate anti-government movements [20]. Many of the manipulative opinions in the dataset disparage presidential candidates or their political parties.

In the two forums, a debate begins when a discussion topic is given, and this debate continues as users post opinions, presenting their own views about the topic. A debate may end up converging on one side if this side prevails over the other. After reading an opinion, one can cast a vote on this opinion, and such a vote can be either positive (i.e., agree) or negative (i.e., disagree). The accumulated number of votes helps readers understand how well the opinion is accepted by the large public. To be precise, each opinion is represented by an eight-tuple: {topic, page #, posting time, user ID, screen name, $N_p$ (# of positive votes – recommendation count), $N_n$ (# of negative votes – disagreement count), text}. In Table 2, we present a sample opinion in this eight-tuple. In particular, the two voting numbers, $N_p$ and $N_n$, are recorded on a regular basis (every five minutes ― Using the five-minute interval was sufficient for tracking gradual changes. Collecting faster than the five-minute interval would unnecessarily increase overhead on the websites) in order to track their gradual changes over time and utilize such changes in detecting manipulation. This regular collection continues for one week after the onset of a discussion, since more than 98% of the votes occur within one week, as shown in Figure 1(b). In Table 3, we present a result of such periodic collection for the opinion in Table 2. We list only the times when $N_p$ and $N_n$ increase. Note that $N_p$ rapidly rises within the first five minutes, which can be observed when it is manipulated by automated tools, as shown in Section 4.1. The forums typically provide users with the ability to sort opinions in two different ways: first, by posting time (i.e., the most recent opinion appears at the top of the list) and second, by the number of votes (i.e., the opinion with the largest $N_p – N_n$ appears at the top). In the forums, we choose political topics that are highly engaged with (i.e., 500≥ opinions and 1,000≥ votes), and as a result, a total of 377,634 opinions are collected on 313 topics. A demonstration of the collection process can be found in the following video: https://youtu.be/j2Pbf_1NpKQ..

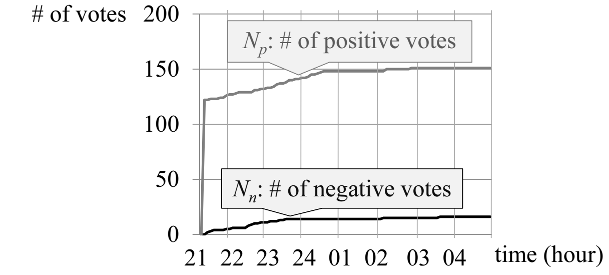
**Table 2.** Sample opinion represented in 8-tuple

| | | | |
|---|---|---|---|
| Topic | Candidate OOO's speech on the economy | | |
| Page # | 1 | Posing Time | 2012-12-13, 21:18:00.283 |
| User ID | User01 | Screen Name | United Korea |
| $N_p$ | 151 | $N_n$ | 16 |
| Text | Can OOO dare to talk about the economy? OOO has long-lasting connection with North Korea. While serving as a minster, OOO provided North Korea with OOO billion dollars, covering it as financial aid to support the poor. However, this money has been used to develop nuclear weapons, which are about to fire at us. Even OOO's father served as a communist soldier during Korean war, … * | | |

*Note.* *The text is translated from Korean into English so that non-Korean readers can understand it.

**Table 3.** Gradual changes of $N_p$ and $N_n$ for opinion in Table 2

| | |
|---|---|
| $N_p$ (time in hh:mm, # of votes) | (21:20, 0) (21:25, 122) (21:35, 123) (21:50, 124) (22:00, 126) (22:05, 127) (22:15, 128) (22:20, 129) (22:45, 131) (22:55, 132) (23:05, 133) (23:15, 134) (23:20, 136) (23:25, 137) (23:35, 138) (23:40, 140) (23:50, 141) (00:00, 142) (00:10, 143) (00:15, 145) (00:25, 146) (00:30, 147) (00:35, 148) (02:15, 150) (02:55, 151) |
| $N_n$ (time in hh:mm, # of votes) | (21:20, 0) (21:30, 2) (21:35, 3) (21:40, 4) (22:00, 5) (22:10, 6) (22:35, 8) (22:40, 9) (22:45, 10) (22:55, 11) (23:10, 12) (23:25, 13) (23:35, 14) (02:10, 15) (03:40, 16) |



## 3.2 Labeling Dataset

After the dataset is collected, we label each user in the dataset as either manipulator or non-manipulator. This labeling is conducted in two stages. In the first stage, we identify opinions deleted after the election and mark the corresponding users as "opinions-deleted." This is because the majority of the deleted opinions are likely to be manipulative – prosecutor's investigation started over the claims that several government agencies were involved in online manipulation, which was followed by a mass deletion of opinions to destroy evidence [11]. We identify deleted opinions by comparing our dataset with the opinions that remain after the investigation (in December 2013). We find that nearly 18% of the opinions are deleted. Although opinions can be deleted by the forums' monitoring team due to policy violations (e.g., the use of abusive words and the infringement of copyright), such deletions amount to 1-3% of the opinions and therefore 18% is not normal.

The second stage of the labeling process is conducted by five judges, to further confirm whether the deleted opinions are manipulative or not. These judges are daily users of the discussion forums, and they gained experience in identifying manipulative accounts over the past 1-3 years; the judges also performed studies on various strategies of political manipulation, such as the spread of unproven rumors [9]. The judges label each user as manipulator or non-manipulator. In a few cases where the judges do not agree on the label, the final label is determined by consensus after having a discussion among the reviewers. The judges are given full access to the database that contains the collected opinions. Consequently, 2,639 users (~2.18%) are labeled as manipulators.

To measure the level of agreement among the judges, we use Fleiss' multi-rater kappa [21]. We obtain κ=0.80, which represents a substantial agreement, based on the scale given by Landis and Koch [22]. To verify our assumption that most deleted opinions are manipulative, we analyze the number of deleted opinions per user in the labeled dataset. More than 85% of manipulators removed a large number of opinions, ranging from 10 to 350, and these opinions comprise 92% of the deleted opinions. In contrast, more than 90% of non-manipulators removed fewer than two opinions, and these deletions appeared to be done by the monitoring team according to the policies.

## 4 Characterization of Manipulators

To identify manipulators, we develop five sets of attributes that can distinguish manipulators from non-manipulators. These attributes comprise the detection model and are our major contributions. We use these attributes with an iterative algorithm to classify users (Section 5).

Table 4 presents the five attribute groups. The first two columns show an attribute's index number and name, respectively, and the last column shows the scope in which this attribute is measured. For instance, attribute #1 represents the maximum number of positive votes for each individual opinion, and attribute #7 represents the total number of positive votes for all opinions written by a user. Sections 4.1-4.3 describe the attributes in detail.

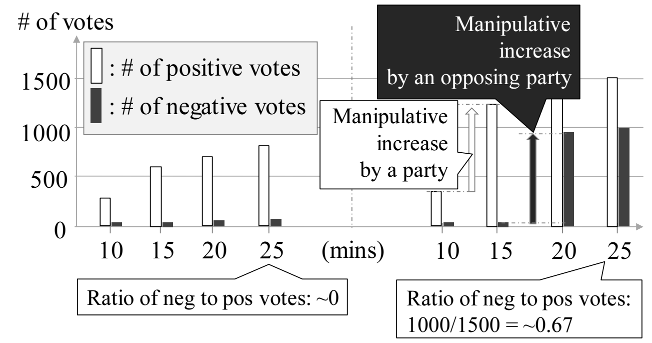
**Table 4.** Attributes that characterize manipulative behaviors

| Index | Attribute Name | Scope |
|---|---|---|
| | *Dynamic Voting Behavior* | |
| 1-3 | max/avg/med # of positive votes | each opinion |
| 4-6 | max/avg/med # of negative votes | each opinion |
| 7-8 | # of positive votes, # of negative votes | all opinions |
| 9-11 | max/avg/med ratio of neg to pos votes | each opinion |
| 12-13 | #/fraction of top-ranked opinions | all opinions |
| 14-16 | max/avg/med # of top-ranked opinions | each topic |
| 17-22 | max/avg/med RGR of pos/neg votes | each opinion |
| 23-26 | max/min/avg/med of earliest posting time | each topic |
| | *Collusion with Multiple IDs* | |
| 27-28 | # of users/manipulators who share common viewpoints | all users |
| 29-30 | # of similar opinions posted by other users/manipulators | all opinions |
| 31-32 | # of opinions/manipulative opinions written within similar time frames | all opinions |
| | *Political Viewpoint* | |
| 33-39 | max/avg/med/total #/fraction of political campaign words | each opinion |
| 40 | # of opinions with campaign words | all opinions |
| 41 | political orientation – strength and consistency | all opinions |
| | *Use of Supporting Evidence* | |
| 42-45 | max/avg/med/total # of digits | each opinion |
| 46-49 | max/avg/med/total # of special chars | each opinion |
| 50 | # of URLs | all opinions |
| 51-54 | max/avg/med/total length of opinion | each opinion |
| | *Dissemination Effort* | |
| 55 | # of political topics where user engaged | all topics |
| 56-59 | max/avg/med/total # of posts | each topic |
| 60 | # of reproduced posts of same user | all opinions |

### 4.1 Attributes Related to Dynamic Voting Behavior

Each post in a political discussion forum can be voted up or down. The more an opinion is voted up, the closer it is to being placed at the top of the list and thus being read by more people. As such, a manipulator would unfairly vote on an opinion to increase or decrease its rank, possibly with the help of a large pool of compromised IDs [23]. In discussion forums, each ID is allowed to vote only once on a post, so multiple IDs are needed to quickly increase the votes. The first attribute group characterizes such voting behaviors.

In Table 4, attributes #1-8 represent the frequency of positive and negative votes. The number of votes monotonically increases and does not decrease, so more manipulation attempts would lead to higher numbers. Attributes #9-11 measure the ratio of negative votes to positive votes on the same opinion. This ratio becomes a large value, typically when the opinion becomes a battleground between two opposing parties. Figure 2 presents an example of such a situation, a sample from our dataset. When one party casts a massive number of positive votes on an opinion and thus promotes its rank, the other party tries to demote the opinion by quickly voting it down.



(a) no manipulation occurred (b) manipulation occurred

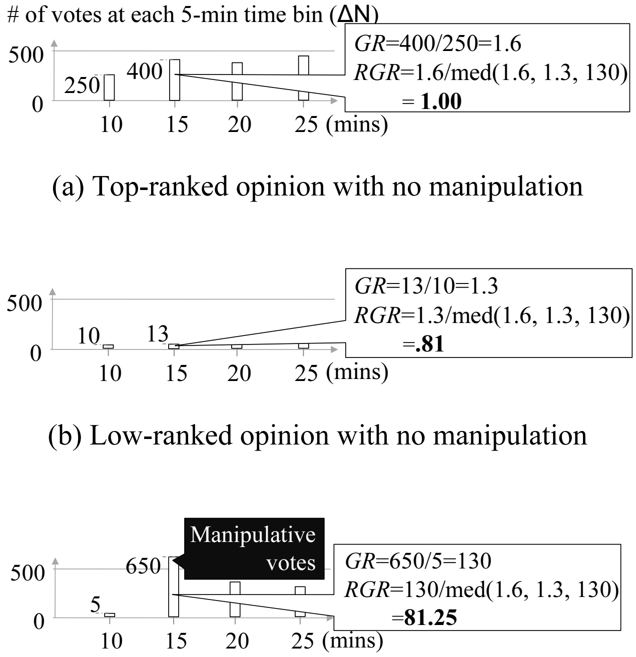
**Figure 2.** Ratio between # of pos/neg votes for two opinions

Attributes #12-16 characterize the number of top-ranked opinions posted by the same user. An opinion is top-ranked if it is shown in the first page of the list. A manipulator strives to gain the top rank to more widely expose ideas to the public. Meanwhile, many opinions of the same manipulator can be promoted together to the top rank. One user having several top-ranked opinions is unlikely to occur purely by chance (i.e., it is unlikely to occur without manipulation), considering that only a handful of opinions are top-ranked among thousands of opinions.

Attributes #17-22 measure the rate at which an opinion is voted. When votes are manipulated with automated tools and many pre-arranged IDs, the number of votes often rapidly increases momentarily. In particular, the voting rate on the target opinion is much higher than that of other opinions, so the target opinion can be top-ranked in advance of other opinions. We denote such a voting rate, relative to other opinions, as the Relative Growth Rate (RGR). RGR is computed according to the following equations:

| | |
|---|---|
| $\Delta N(o,i)$ | # of votes on opinion $o$ in i$^{th}$ time bin |
| $GR(o,i)=$<br>$\Delta N(o,i) \div \Delta N(o,i\text{-}1)$ | voting rate, relative to previous time bin |
| $RGR(o,i)=$<br>$GR(o,i) \div median_o(GR(o,i))$ | RGR – voting rate, relative to other opinions |

Each time bin is a five-minute interval, as we collect the number of votes every five minutes, as shown in Section 3.1. For each time bin, the RGR of an opinion is computed by the ratio of its voting rate to the median for all other opinions posted on the same topic. A large increase in the number of votes compared with the increase in other opinions would result in a large RGR. We demonstrate the RGR for three sample opinions in Figure 3. When no manipulation occurs, the numbers do not show large differences over time; therefore, the RGR is low for both a high-profile opinion (Figure 3(a)) and a low-profile opinion (Figure 3(b)). In contrast, when manipulation occurs (Figure 3(c)), a surge appears in the number of votes at the moment of manipulation, leading to a large RGR. In particular, the RGR of 81.25 means that the increase is nearly 80 times larger than that for other opinions.



(a) Top-ranked opinion with no manipulation

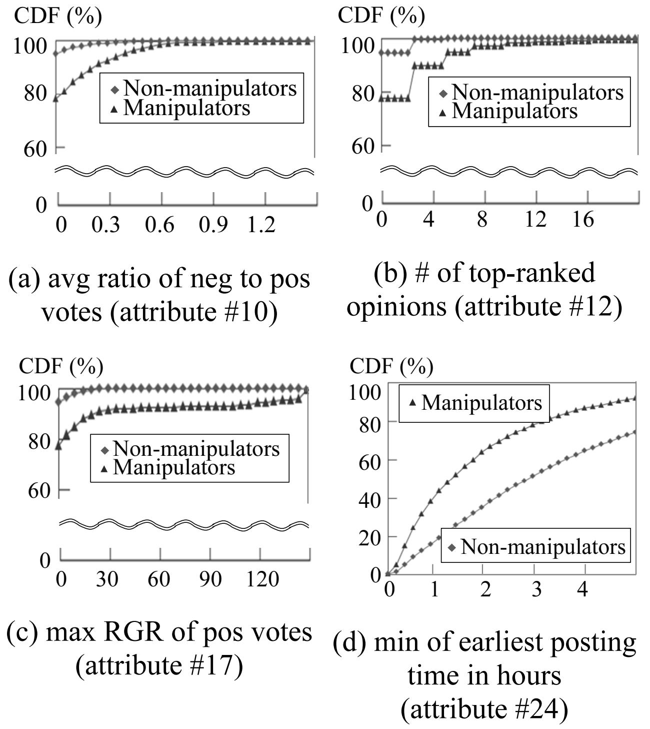
(b) Low-ranked opinion with no manipulation

(c) Opinion with # of votes manipulated

**Figure 3.** RGR for three opinions

The last four attributes of the first group, attributes #23-26 characterize how early a user posts their first opinion after the onset of a discussion. The earlier an opinion is posted, the more often people will read this opinion and vote on it; therefore, the more likely the opinion will be top-ranked.

Figure 4 illustrates the cumulative distributions of four representative attributes. The CDF(%) in the vertical axis denotes the accumulated percentage of users who exhibit the designated characteristics. In the cdf, the larger the gap between manipulators and non-manipulators, the more effective a discriminator the attribute is. In this section, each attribute is examined separately, but in a real classification task, the attributes are used together, while supporting one another (as illustrated in Section 5.4). Figure 4(a) shows that manipulative opinions tend to have negative votes comparable to positive votes, which is a result of manipulation by multiple parties. Figure 4(b) demonstrates that manipulative opinions are more likely to be top-ranked than non-manipulative opinions. In an extreme case, a total of 22 opinions written by the same manipulative account are simultaneously top-ranked. To be top-ranked, manipulative opinions receive votes at a much higher rate than other opinions, as shown in Figure 4(c) (e.g., 30-100 times faster). Lastly, Figure 4(d) shows that manipulative opinions are posted at earlier times than non-manipulative opinions. For example, more than 40% of manipulative opinions are concentrated within the first hour. We further analyze these early-posted manipulative opinions and find that nearly 15% of these opinions are quickly voted up within a few minutes and become top-ranked. In contrast, the posting times of non-manipulative opinions are more evenly distributed over 1-8 hours. To summarize, we present and analyze the attributes related to voting behaviors. These attributes appear to show noticeable differences between manipulators and non-manipulators and are thus expected to be effective discriminators.



(a) avg ratio of neg to pos votes (attribute #10)

(b) # of top-ranked opinions (attribute #12)

(c) max RGR of pos votes (attribute #17)

(d) min of earliest posting time in hours (attribute #24)

**Figure 4.** cdf of 4 attributes related to dynamic voting behavior

## 4.2 Attributes Related to Collusion with Multiple IDs

The second attribute group characterizes correlations among multiple user IDs, particularly the possibility of IDs being used together for the same political campaign. Such a correlation occurs when manipulators collaborate as a group and also when a manipulator utilizes a set of IDs. In particular, multiple IDs are used (1) to shape an opinion as if it reflects a majority view (e.g., by unfairly boosting the recommendation counts, as shown in Section 4.1) and also (2) to evade detection by performing the least amount of manipulation with each individual ID [10]. The proposed attributes help monitor this type of manipulation, thus decreasing the likelihood of avoiding detection.

Attributes #27-30 represent the level of similarity in opinion contents among different users. Manipulators in the same group often reproduce each other's opinions to more effectively spread propaganda to the wider public. Figure 5 presents an example in which similar copies are posted by multiple user accounts. Each vertical bar represents one opinion, and its y and x positions represent the corresponding account and posting time, respectively. In total, 13 similar opinions are posted by five different IDs, one after another. From the viewpoint of User03, four other users, Users04-07, post opinions similar to those of User03 (i.e., attribute #27 for User03 is 4), and these four users post a total of 11 similar opinions (i.e., attribute #29 is 11). By measuring the proposed attributes, one can see the extent of manipulative activity over multiple IDs. This contrasts with the case where each ID is evaluated individually. For example, if we monitor the activity of User03 alone, we can find only two similar copies, which may not be sufficient for confirming manipulation.

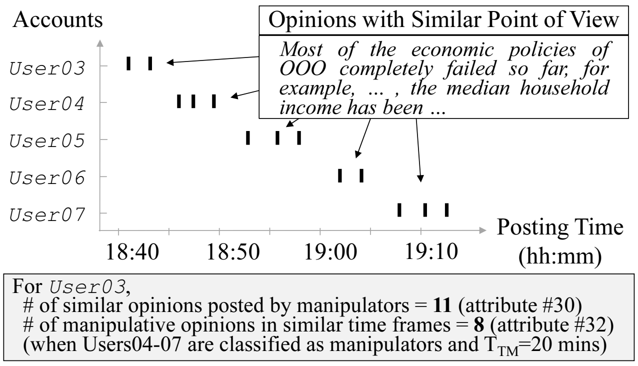

Accounts
Opinions with Similar Point of View
Most of the economic policies of OOO completely failed so far, for example, ... , the median household income has been ...
User03
User04
User05
User06
User07
18:40
18:50
19:00
19:10
Posting Time (hh:mm)
For User03,
# of similar opinions posted by manipulators = 11 (attribute #30)
# of manipulative opinions in similar time frames = 8 (attribute #32)
(when Users04-07 are classified as manipulators and T_TM=20 mins)

**Figure 5.** Correlation among multiple manipulative accounts

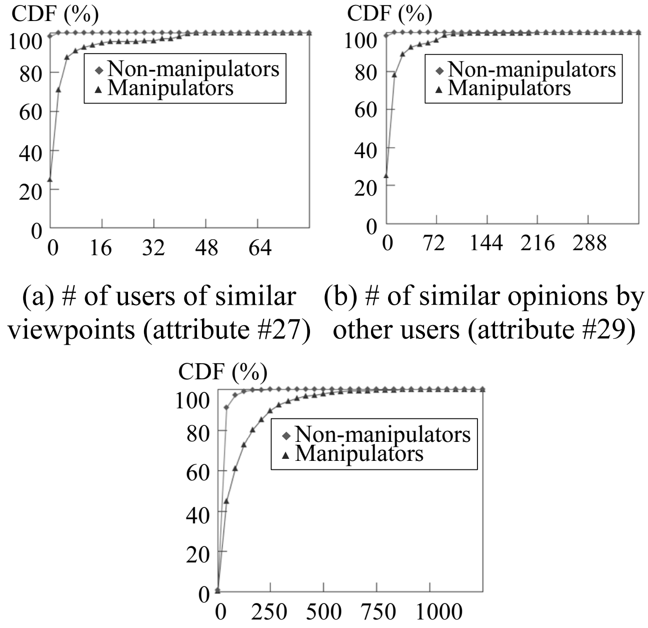
Some of the reproduced opinions, as illustrated in Figure 5, are near-duplicates, but they are often rephrased with slightly different words, while preserving the same point of view. As such, we measure the similarity based on two criteria. First, we use the Jaccard Coefficient (*JC*) [24]:

$$JC(o1,o2) = |W(o1) \cap W(o2)| \div |W(o1) \cup W(o2)|$$

*W(o)* denotes the set of distinct terms used in opinion *o*, and *JC* is computed to evaluate the similarity between two opinions, *o1* and *o2*. If *JC* is equal to or greater than a predefined threshold $T_{JC}$, then we conclude that *o1* and *o2* are similar. The second similarity criterion concerns URLs in opinion text. URLs are used to provide more details in different pages, and they are also linked to video clips. We consider two opinions to have similar objectives if they refer to the same URL [25]. Shortened URLs are expanded, so the final URLs are compared.

Attributes #31-32 characterize the temporal proximity among opinions of different users. Manipulators tend to work concurrently if they are collaborating on the same political campaign. To measure the degree of concurrent work for one user, we count the number of opinions from other users posted at similar times. In particular, we consider two opinions to have been published within similar time frames if these opinions are written within $T_{TM}$ minutes of each other. Figure 5 presents an example of computing attribute #32, assuming that $T_{TM}$ = 20 minutes.

Figure 6 illustrates the cumulative distributions of three representative attributes. Figure 6(a) and Figure (b) indicate that colluders participating in orchestrated manipulation tend to post numerous opinions of similar viewpoint, and Figure 6(c) shows that they also concentrate their efforts by posting the opinions simultaneously. The largest manipulative campaign found in the dataset utilizes 80 different IDs to post more than 300 similar opinions. Non-manipulators, however, rarely post similar opinions to other users. In summary, the proposed attributes model the correlation among users both in terms of opinion content and posting time. Compared with modeling each individual independently, the proposed attributes better illustrate the extent of manipulation with multiple IDs, i.e., a set of manipulator IDs can be identified together, even when each ID does not exhibit clear signs of manipulation.



(a) # of users of similar viewpoints (attribute #27)

(b) # of similar opinions by other users (attribute #29)

(c) # of manipulative opinions in similar time frames (attribute #32)

**Figure 6.** cdf of 3 attributes related to collusion with IDs

## 4.3 Attributes Related to Political Text and Efforts

The rest of the attribute groups model individual users' behavior in the way they compose opinions. These attributes characterize opinion text from political viewpoint, as well as the amount of effort needed to develop such text. For example, manipulators often maintain a strong and consistent political inclination throughout their posts. They also present specific numbers and references to strengthen their arguments.

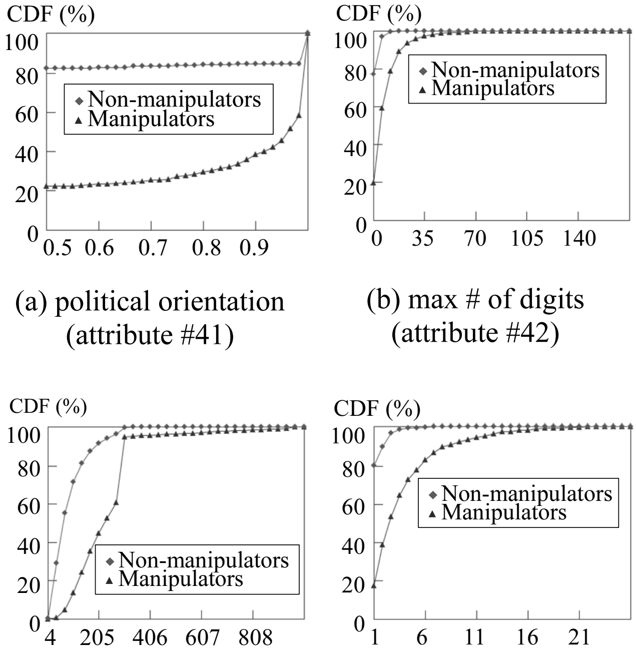
Attributes #33-40 measure the frequency of political campaign words (terms that are utilized in political campaigns). Becker [9] and Trent et al. [26] show that political campaigns tend to use particular words that are known to be effective in influencing people. In particular, we count the terms that are used more often by manipulators than non-manipulators. These terms are collected from several sources, including interviews with former manipulators, as well as articles that report on political manipulation [27]. The terms consist of 62 words, including 39 words that are often used by Liberals (left-wing words) and 23 words that are often used by Conservatives (right-wing words). Many of these words defame politicians or political parties. For example, the right-wing word "빨갱이" literally means red people, but it is more often used to disparage left-wing politicians who do not take a strong stance against North-Korean communists.

Attribute #41, political orientation, is measured by the proportion of campaign words on the predominant side (either the right wing or left wing). For example, if a user utilizes ten campaign words throughout the posts and if nine of these are left-wing words, then the predominant side is the left and the proportion is 9/10 = 0.9. Such a large value indicates that the user is strongly inclined to one side. When an equal number of campaign words are used for each side or when no campaign words are used, the attribute becomes 0.5, indicating a nearly neutral stance.

The rest of the attributes characterize the amount of efforts users put into their posts to better persuade the large public. A logical argument with sufficient proofs is convincing. In this regard, attributes #42-49 evaluate the use of digits and special characters (i.e., characters that are neither digits nor letters); digits are often used as supporting evidence (e.g., statistics about arguments and the date of events), and special characters are used to highlight arguments or to organize them in a list (e.g., bullet points). Attribute #50 indicates the use of hyperlinks, which are used to provide more details about arguments. Attributes #51-54 measure the numbers of characters, since opinions with a list of proofs tend to be lengthy.

Repeated exposure to propaganda can create false beliefs that are rarely corrected once adopted by an individual [28]. In this regard, attribute #55 measures the number of discussion topics on which a user has ever posted opinions, and attributes #56-59 measure the number of posts ever made by a user. To better propagate an idea, it is reproduced in as many places as possible. Attribute #60 represents the level of similarity among a series of opinions posted by the same user. We determine the similarity based on the *JC* and reference to the same URL, as shown in Section 4.2.

Figure 7 illustrates the cumulative distribution of four representative attributes. Figure 7(a) represents the political inclination of users – manipulator opinions are more inclined to one side, and this inclination appears throughout their posts. Figure 7(b) and Figure 7(c) show that manipulative opinions are lengthy and contain many digits, indicating that manipulators strive to reinforce their arguments by providing more evidence. In contrast, 80% of non-manipulators post a single opinion per topic, and they rarely use digits. Figure 7(d) shows that manipulators tend to post multiple opinions on the same topic, and a certain series of opinions reaches more than twenty. To summarize, manipulators exert a large amount of effort posting numerous, deliberate opinions, which is shown in the way they compose and distribute opinions.



(a) political orientation (attribute #41)

(b) max # of digits (attribute #42)

(c) avg length of opinion (attribute #52)

(d) med # of posts (attribute #58)

**Figure 7.** cdf of 4 attributes related to political text and effort

## 5 Evaluation of Detection Method

We evaluate how effectively the proposed system identifies manipulators. We leverage a supervised learning algorithm, which analyzes users based on the attributes discussed in the previous section and then classifies them into manipulators and non-manipulators. To this end, we first estimate parameters that are necessary for calculating the proposed attributes (Section 5.1). We then select a subset of attributes that are shown to be more effective discriminators (Section 5.2). The selected attributes are then used to classify users in the dataset (Section 5.3-5.5). Each subsection uses a part of the dataset as shown in Figure 8.

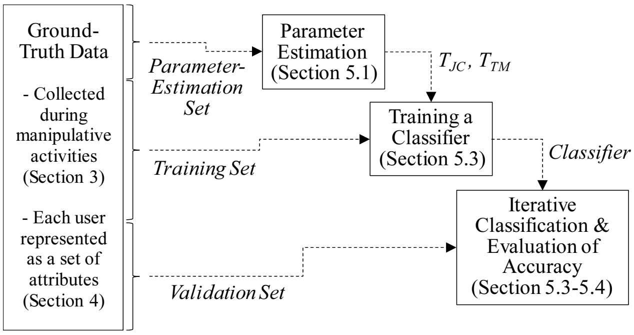

Ground-Truth Data
- Collected during manipulative activities (Section 3)
- Each user represented as a set of attributes (Section 4)
Parameter-Estimation Set
Training Set
Validation Set
Parameter Estimation (Section 5.1)
$T_{JC}$, $T_{TM}$
Training a Classifier (Section 5.3)
Classifier
Iterative Classification & Evaluation of Accuracy (Section 5.3-5.4)

**Figure 8.** Flow diagram of proposed method

### 5.1 Parameter Estimation

Our user-behavior model has two parameters to estimate, $T_{JC}$ and $T_{TM}$. $T_{JC}$ is used to estimate the similarity of opinions (attributes #27-30, #60); two opinions are considered similar if their $JC \geq T_{JC}$. $T_{TM}$ is used to estimate the degree of concurrent work (attributes #31-32); if two opinions (of different IDs) are posted within $T_{TM}$ minutes, then we assume that they are written by colluders.

To estimate the two parameters, we experiment with our dataset to determine the values that yield the best tradeoff between false positives and false negatives. We first build a dataset, *E*, for parameter estimation, by randomly sampling 12,128 users (~10%) from our original dataset. The rest of the users are used for the evaluation in Sections 5.2-5.5. Let $\theta$ denote the two parameters. We estimate $\theta$ by maximizing the likelihood of the set *E*: $\text{argmax}_{\theta} \sum_i \text{P}(User_i = l_i \mid \theta)$. $User_i$ refers to the classification outcome of a user in *E*, and $l_i$ is the label of this user. To calculate $\text{P}(User_i = l_i)$, each user is represented by a vector of *N* attributes $[\text{a}_1 \ldots \text{a}_\text{N}]$. Since these attributes model different characteristics, we assume that the attributes are independent and let $\text{P}(User_i = l_i) = \prod_\text{j} \text{P}(\text{a}_\text{j} = l_i)$. To compute $\text{P}(\text{a}_\text{j} = l_i)$, we discretize the values that $\text{a}_\text{j}$ can have into *k* intervals $\{\text{a}_\text{j}^1 \ldots \text{a}_\text{j}^\text{k}\}$ [29]; if $\text{a}_\text{j}$ belongs to the interval $\text{a}_\text{j}^\text{m}$, then $\text{P}(\text{a}_\text{j} = l_i) = \text{P}(\text{a}_\text{j}^\text{m} = l_i)$. Lastly, $\text{P}(\text{a}_\text{j}^\text{m} = l_i)$ is the proportion of users whose label is $l_i$ and whose $\text{a}_\text{j}$ belongs to $\text{a}_\text{j}^\text{m}$. The final estimated values are $T_{JC}$=0.55 and $T_{TM}$=21 (mins).

We perform further studies to better understand the meanings of the estimated values. $T_{JC}$ =0.55 corresponds with two-thirds of text overlap between two opinions. $T_{TM}$=21 is found in consecutive opinions posted by the same ID – 98% of such series are separated by less than 21 minutes. This result indicates that if multiple IDs post numerous opinions within $T_{TM}$ minutes, then these IDs are possibly controlled by the same, single user.

## 5.2 Statistical Validation and Attribute Selection

We propose a set of 60 attributes to model user behaviors, as shown in Section 4. We expect that many of these attributes help differentiate manipulators from non-manipulators, but some attributes may only add noise, decreasing classification accuracy. We therefore validate each attribute's predictive power and choose a subset of attributes that would yield the best results.

To measure the predictive power of the attributes, we utilize Information Gain (*IG*) and $\chi^2$ statistics [30]. We show the *IG* of the proposed attributes in Figure 9 (the results with $\chi^2$ are similar). The attributes are categorized into nine groups. In general, each group contains one to several attributes with significant predictive power. Among the attributes related to voting behavior (attributes #1-22), the maximum growth rate of votes (RGR, attribute #17) appears to be the most important discriminator. Among the attributes related to posting time (i.e., attributes #23-26), only the minimum posting time (attribute #24) stands out, compared with the maximum, average, and median. Taken together, the results reveal an often-used manipulation strategy – posting at the earliest possible time and then quickly increasing recommendation counts, so that the manipulative opinion remains top-ranked from the beginning and is seen by a large population. The attributes related to the number of negative votes are generally not effective discriminators (attributes #4-6, 8, and 21-22), although the ratio between negative and positive votes does seem to be useful (attributes #9-11). For the rest of the attributes (#27-60), most appear to be discriminating factors, including the use of multiple IDs (#27-32), the strength and consistency of political views (#33-41), the degree of using supporting evidence (#42-54), and the extent to which propaganda is circulated (#55-60). As a result of the evaluation, we exclude twelve attributes that have much less predictive power than other attributes (i.e., IG <0.1) and are therefore not expected to help increase the classification accuracy. These attributes are #3-8, 21-22, 23, 25-26, and 39.

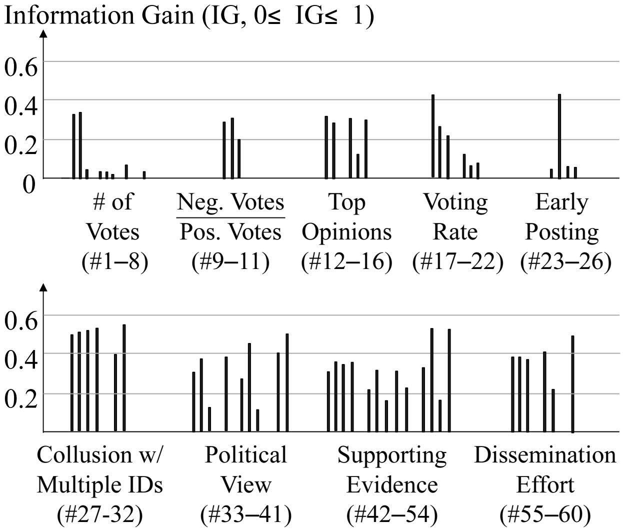


**Figure 9.** Information Gain (*IG*) of 60 attributes in Table 4

## 5.3 Classification Method and Experimental Setup

To classify users with the proposed attributes, we leverage Collective Classification methods [31], the classification of instances considering relationship among multiple instances. This type of correlation is represented by the second group of attributes (i.e., attributes #27-32). For example, attribute #28, the number of manipulators with similar viewpoints, is determined by the labels of other users. We do not directly utilize traditional classification techniques, such as the Support Vector Machine [32], since these classifiers do not consider correlations and assume independence among instances. In Collective Classification methods, multiple iterations are used to gradually obtain the final assignments of attributes and labels. In each iteration, labels are determined based on the current best estimates of attributes, and these estimates converge to desired values as the iteration continues.

Among various Collective Classification algorithms, we choose to use the Iterative Classification Algorithm (ICA). While most Collective Classification algorithms continue to iterate until the convergence criterion is achieved, ICA converges within an order of magnitude fewer iterations (in our experiment, within ten iterations) and does not iterate infinitely. We also find that ICA is as accurate as other algorithms. In addition to ICA, we experimented with Gibbs Sampling, Loopy Belief Propagation, and Mean-Field Labeling but did not observe significant differences in accuracy results. The choice of local classifiers (we tried Naive Bayes, Logistic Regression, and Random Forest) did not much improve the results as well.

We use a 10-fold cross validation to evaluate the accuracy of the proposed method. To this end, we divide the dataset into ten equal-sized subsets. Nine subsets are used to train the classifier, and the other subset is used for validation. We then perform this process ten times, with each subset used once for validation. In addition, the entire 10-fold cross validation is repeated ten times, with different seeds used to shuffle the data. This produces a hundred different results. We present an average of these hundred runs.

Since our dataset contains more non-manipulators than manipulators (i.e., 106,776 non-manipulators vs. 2,375 manipulators, This set equals to our initial dataset (121,279 users) minus the set used for parameter estimation (12,128 users).), the classifiers can bias their decision toward the majority class, as this would allow the classifier to lower overall error [33]. To reduce this type of bias, we re-sample the training set so that the two classes are of equal size; we then train the classifier. The validation set, however, is not re-sampled in order to maintain the original distribution [34].

## 5.4 Classification Accuracy

Figure 10 presents the classification results. We evaluate the accuracy using various subsets of the attribute groups in Table 4 – *DV*, *CM*, and *TE*. The last three groups are represented together by *TE*, as they share common properties (related to the way individual opinions are composed). We also evaluate the combinations of these subsets. The classifiers are evaluated by the areas under their ROC curves (AUC). AUC is an appropriate measure when class imbalance is present. When a single-attribute group is used, the classifiers achieve 80-90% of AUC, and as combinations are used, the AUC improves, reaching up to 98.9% (when all groups are used together, as shown by the black bar at the far right). This means that the attribute groups complement each other. The *DV* and *TE* attributes identify individual manipulators whose behavior manifests in voting behavior and opinion text, respectively. Starting from these manipulators, the *CM* attributes (with the help of ICA) progressively discover collaborators, even when each user instance does not show clear signs of manipulation. Such collaborators amount to 12% of all the manipulators. Many of these collaborators use similar screen names (e.g., future001 and future002); this indicates that the corresponding IDs potentially belong to the same user, masking their real identity.

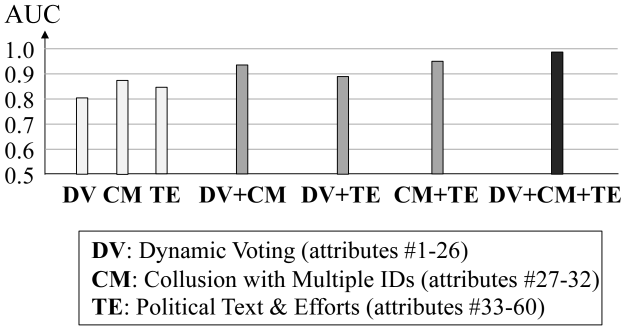


**Figure 10.** AUC for attribute groups and combinations

Table 5 presents more detailed results for when all attribute groups are used together. The vertical axis lists the two classes in the original collection, and the horizontal axis lists the two classification outcomes. The number at each intersection represents a percentage relative to the total number of users in the corresponding class. For instance, out of 2,375 manipulators, 93.94% are correctly discovered, while the rest (6.06%) are not. Similarly, out of 106,776 non-manipulators, 99.54% are correctly classified as non-manipulators, and 0.46% are misclassified as manipulators. This 0.46% means that the proposed system can significantly suppress false alarms, which is critical particularly when non-manipulators constitute a large portion of the entire population. Overall, the proposed attributes and detection methods can correctly classify users in more than 90% of cases.

**Table 5.** Classification results by the proposed attributes

| | | Classified As | |
|---|---|---|---|
| | | Manipulator | Non-manipulator |
| True | Manipulator<br>(2,375 users) | **93.94%** | 6.06% |
| | Non-manipulator<br>(106,776 users) | 0.46% | **99.54%** |

Among the manipulators, 6.06% are erroneously classified as non-manipulators. These manipulators post only a few opinions and do not clearly show correlations with other manipulators. To more accurately detect these manipulators, we would need additional clues from the discussion forums, such as IP addresses. For example, multiple IDs logged in from the same IP can be further analyzed to find additional correlations in text and time. Among the non-manipulators, 0.46% are misclassified as manipulators. Our investigation reveals that more than 75% of these users post opinions unrelated to the discussion topics and their text appears to be advertising products and services. Since this type of post is deemed to be spamming activity, it would be useful to identify and remove them.

## 5.5 Comparison with Previous Works

We compare the proposed system with two state-of-the-art systems that detect manipulation. The first system, the group-based system [14], is a recent work that utilizes correlations among users. Users who post on common topics comprise a potentially manipulative group. Such a group is modeled as one instance, assigned attributes that characterize group activities (e.g., the number of common topics), and then classified together according to the assigned attributes. The second system for comparison is the effort-based system [10]. Its main attributes include the degree of efforts made by users (e.g., the number of consecutive posts), assuming that manipulators work hard to quickly influence a large audience.

Table 6 summarizes the two system's attributes that are applicable to our dataset and are thus used in the evaluation. Attributes #101-104 are computed for each group of users, and attributes #105~114 for each individual user. The attributes that are not applicable are related to functions that do not exist in political discussion forums (e.g., ratings based on the five-star system).

**Table 6.** Major attributes of the two existing systems

| Index | Attribute Name | Group based | Effort based |
|---|---|---|---|
| | *Group Behavior* | | |
| 101 | # of users in a group | V | |
| 102 | # of topics where group members commonly posted opinions | V | |
| 103 | similarity of opinions written by different group members | V | |
| 104 | proximity in posting times of different group members | V | |
| | *Individual Behavior and Opinion Text* | | |
| 105-107 | # of opinions | V | V |
| | # of reproduced opinions | V | V |
| | # of top-ranked opinions | | V |
| 108 | amount of time spent on consecutive opinions | | V |
| 109-112 | # of positive/negative words | V | |
| | # of subjective words | V | |
| | # of first/second-person words | V | |
| | # of campaign words | | V |
| 113 | avg length of opinion | V | V |
| 114 | # of opinions that contain URLs | | V |

Table 7 presents the classification results of the two existing systems compared with the proposed system. The misclassification rate in the existing systems is roughly 10% higher than that in the proposed system. We list two major reasons as follows. First, the existing systems correlate users as belonging to the same group only if they post opinions on multiple common topics. However, manipulators often utilize different IDs when moving between different discussions, so the existing systems fail to correlate such IDs. The proposed system more effectively identifies this type of ID rotation by correlating users based on various criteria, such as the similarity of opinion texts, use of the same URLs, and closeness in posting times. Second, the user-behavior model in the proposed system more carefully considers anomalous voting behaviors (e.g., an extremely high rate of votes on a particular opinion). In analyzing opinion text, the proposed system considers the continuity of the same political inclination. The existing systems do not consider the above-listed characteristics.

**Table 7.** AUC of three classification systems

| Detection System | AUC |
|---|---|
| Group-based System | 86.8% |
| Effort-based System | 87.2% |
| Proposed System | 98.9% |

## 6 Conclusion

We propose a method for detecting users in online communities who spread opinions in an attempt to manipulate people's attitudes toward political issues. The proposed method first models user behavior with 60 attributes, which considers (1) the characteristics of individual users, (2) collusion among multiple users, and (3) the use of multiple IDs by one user. With this model, the method progressively discovers collaborative relationships among users through iterations and then identifies manipulative users based on the discovered relationships and attributes. The proposed method is applied to 370K opinions posted on real political discussion forums and detects more than 90% of manipulators. The proposed method can be used to identify manipulators in domains other than the political domain, and we are currently evaluating its potential. For example, the attributes that model manipulative voting behaviors can be used to detect manipulation in the number of views and likes in online social media. We also plan to apply anomaly detection methods, i.e., to model normal behaviors and to identify cases that highly deviate from the normal.

## Acknowledgments

This work was supported by a sabbatical year research grant from Seoul Women's University (2018). This research was also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2017R1C1B5014992).

## References

[1] M. Fraser, S. Dutta, *Throwing Sheep in the Boardroom: How Online Social Networking Will Transform Your Life, Work and World*, Wiley, 2008.

[2] D. Centola, The Spread of Behavior in an Online Social Network Experiment, *Science*, Vol. 329, No. 5996, pp. 1194-

1197, September, 2010.

[3] P. Adams, *Grouped: How Small Groups of Friends are the Key to Influence on the Social Web (Voices That Matter)*, New Riders, 2011.

[4] D. Mocanu, L. Rossi, Q. Zhang, M. Karsai, W. Quattrociocchi, Collective Attention in the Age of (Mis) information, *Computing in Human Behavior*, Vol. 51, pp. 1198-1204, January, 2015.

[5] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle, J. H. Fowler, A 61-million-person Experiment in Social Influence and Political Mobilization, *Nature*, Vol. 489, No. 7415, pp. 295-298, September, 2012.

[6] BBC News, *Technology, Russian Twitter Political Protests Swamped by Spam*, BBC, 2012.

[7] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, H. W. Lauw, Detecting Product Review Spammers using Rating Behaviors, *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10)*, Toronto, Canada, 2010, pp. 939-948.

[8] N. Jindal, B. Liu, Opinion Spam and Analysis, *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM'08)*, Palo Alto, CA, 2008, pp. 219-230.

[9] K. Becker, *The Handbook of Political Manipulation*, Conservative Daily News, 2012.

[10] S. Lee, Detection of Political Manipulation in Online Communities Through Measures of Effort and Collaboration, *ACM Transactions on the Web*, Vol. 9, No. 3, Article No. 16, June, 2015.

[11] A. Joy, *Infographic: How South Korean Intelligence Interfered in Election*, koreaBANG, 2013.

[12] K. Koo, *Korean Spy Agency Accused of Influencing Presidential Election*, koreaBANG, 2013.

[13] J. Ratkiewicz, M. D. Conover, M. Meiss, B. Goncalves, A. Flammini, F. Menczer, Detecting and Tracking Political Abuse in Social Media, *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM-11)*, Barcelona, Spain, 2011, pp. 297-304.

[14] C. Xu, J. Zhang, K. Chang, C. Long, Uncovering Collusive Spammers in Chinese Review Websites, *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM'13)*, San Francisco, CA, 2013, pp. 979-988.

[15] A. J. Minnich, N. Chavoshi, A. Mueen, S. Luan, M. Faloutsos, TrueView: Harnessing the Power of Multiple Review Sites, *Proceedings of the 24th International Conference on World Wide Web (WWW'15)*, Florence, Italy, 2015, pp. 787-797.

[16] S. Kwon, M. Cha, K. Jung, W. Chen, Y. Wang, Prominent Features of Rumor Propagation in Online Social Media, *IEEE 13th International Conference on Data Mining (ICDM)*, Dallas, TX, 2013, pp. 1103-1108.

[17] KoreanClick, *Nielsen KoreanClick Syndicated Reports*, Nielsen KoreanCick, 2015.

[18] H. Fawcett, *South Korea's Political Cyber War*, Aljazeera, 2013.

[19] S. Choe, *Prosecutors Detail Attempt to Sway South Korean Election*, The New York Times, 2013.

[20] H. Olsen, *North Korean Weighs in on South Korean Presidential Election*, koreaBANG, 2012.

[21] J. Fleiss, Measuring Nominal Scale Agreement Among Many Raters, *Psychological Bulletin*, Vol. 76, No. 5, pp. 378-382, November, 1971.

[22] J. R. Landis, G. G. Koch, The Measurement of Observer Agreement for Categorical Data, *Biometrics*, Vol. 33, No. 1, pp. 159-174, March, 1977.

[23] Manipulation of Recommendation Counts by the Military and Government Agencies, *Media Today*, http://goo.gl/on9VyJ.

[24] B. Liu, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, Springer, 2011.

[25] R. Ghosh, T. Surachawala, K. Lerman, Entropy-based Classification of Retweeting Activity on Twitter, *Proceedings of KDD Workshop on Social Network Analysis (SNA-KDD)*, San Diego, CA, 2011, pp. 1-10.

[26] J. S. Trent, R. V. Friedenberg, R. E. Denton, *Political Campaign Communication: Principles and Practices*, Rowman & Littlefield Publishers, 2011.

[27] S. Lee, Popular List of Political-campaign Words, http://goo.gl/fBDFEf.

[28] R. K. Garrett, B. E. Weeks, The Promise and Peril of Real-time Corrections to Political Misperceptions, *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW'13),* San Antonio, TX, 2013, pp. 1047-1058.

[29] U. M. Fayyad, K. B. Irani, Multi-level Discretization of Continuous-valued Attributes for Classification Learning, *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (II), IJCAI-93*, Chambéry, France, 1993, pp. 1022-1027.

[30] Y. Yang, J. O. Pedersen, A Comparative Study on Feature Selection in Text Categorization, *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)*, Nashville, TN, 1997, pp. 412-420.

[31] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, T. Eliassi-Rad, Collective Classification in Network Data, *AI Magazine*, Vol. 29, No. 3, pp. 93-106, September, 2008.

[32] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 2000.

[33] G. H. Nguyen, A. Bouzerdoum, S. L. Phung, A Supervised Learning Approach for Imbalanced Data Sets, *19th International Conference on Pattern Recognition (ICPR)*, Tampa, FL, 2008, pp. 1-4.

[34] G. Dupret, M. Koda, Bootstrap Re-sampling for Unbalanced Data in Supervised Learning, *Elsevier European Journal of Operational Research*, Vol. 134, No. 1, pp. 141-156, October, 2001.

## Biography

**Sihyung Lee** received the B.S. and M.S. degrees from KAIST in 2000 and 2004, respectively, and a Ph.D. degree from Carnegie Mellon University in 2010. He is currently an assistant professor at Seoul Women's University. His research interests include security in the Internet and WWW.