Detection of Political Manipulation through Unsupervised Learning

Sihyung Lee
Department of Information Security, Seoul Women's University
Seoul 10797 – South Korea
[e-mail: sihyunglee@swu.ac.kr]
*Corresponding author: Sihyung Lee

Received April 10, 2018; revised July 17, 2018; revised August 25, 2018; accepted October 19, 2018; published April 30, 2019

Abstract

Political campaigns circulate manipulative opinions in online communities to implant false beliefs and eventually win elections. Not only is this type of manipulation unfair, it also has long-lasting negative impacts on people's lives. Existing tools detect political manipulation based on a supervised classifier, which is accurate when trained with large labeled data. However, preparing this data becomes an excessive burden and must be repeated often to reflect changing manipulation tactics.

We propose a practical detection system that requires moderate groundwork to achieve a sufficient level of accuracy. The proposed system groups opinions with similar properties into clusters, and then labels a few opinions from each cluster to build a classifier. It also models each opinion with features deduced from raw data with no additional processing. To validate the system, we collected over a million opinions during three nation-wide campaigns in South Korea. The system reduced groundwork from 200K to nearly 200 labeling tasks, and correctly identified over 90% of manipulative opinions. The system also effectively identified transitions in manipulative tactics over time. We suggest that online communities perform periodic audits using the proposed system to highlight manipulative opinions and emerging tactics.

Keywords: Political manipulation, fake news, online communities, WWW, unsupervised learning

This work was supported by a research grant from Seoul Women's University (2019). 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).

http://doi.org/10.3837/tiis.2019.04.005 **ISSN : 1976-7277**

1. Introduction

Political manipulation refers to forcing or persuading people to change their behavior to win political advantage. It is often achieved by presenting false information and deceiving people. Political manipulation dates back over a hundred years, and many tactics have been developed, such as deliberate ommision of details and repetition of a false statement until people eventually believe it [1]. Manipulation results are not trivial, e.g. changing the outcome of nation-wide elections or significant impact on people's livelihood for several years [2].

Traditionally, political manipulation has leveraged television and newspapers to trick the public. However, current manipulation increasingly utilizes online communities—web portals, social networks, and discussion forums—since these communities are widely used as a main source of information and can instantly and widely disseminate information to the public [3]. One recent example is Russian interference in the US presidential election [4]. Numerous opinions were posted in online communities intended to damage the reputation of candidates that maintained a strong stance against Russia. The posts were disguised as being written by US citizens. Previous studies have shown that online manipulation can switch up to 10% of reader's votes, which is sufficient to make a significant difference, particularly in competitive elections [5].

Several methods have been proposed to detect political manipulation in online communities, most based on supervised learning with various features to identify manipulative activity. Ratkiewiz *et al.* [6] utilized that manipulative propaganda is generally repeatedly mentioned by relatively small groups of users. Lee [3] measured the amount of labor required to distribute propaganda (e.g. the number of successive posts within a short period), which appeared to be significant among manipulative users. Although the proposed features are effective discriminators, supervised learning requires large prelabeled training sets (e.g. > 100K instances) to achieve acceptable accuracy, which requires substantial time and effort. Furthermore, manipulative tactics continue to evolve, so new training sets are required on a regular basis.

Our study focused on building a practical system that does not require large labeling effort(s), but still achieves comparable accuracy to existing approaches. The proposed system is based on unsupervised learning. In particular, we group opinions in online communities into clusters, where each cluster contains opinions with similar characteristics. After identifying cluster structures, we determine if the clusters are manipulative by labeling a handful of opinions from each cluster. This reduces labeling effort to only several hundred instances. We can also track changes in manipulative tactics by tracing clusters over time. For example, previously unseen clusters could indicate new manipulation behaviors.

To validate the proposed method, we collected over a million opinions from popular web portals in South Korea during two presidential campaigns and one local election, when manipulative activities are at their peak. We analyzed the collected data to extract a set of features that distinguish manipulative and non-manipulative opinions. Using these features, the proposed method reduced labeling effort from 200K+ instances to approximately 200 instances, while correctly classifying more than 90% of instances. We also compared clustering results of different political campaigns and showed that certain tactics became more common whereas others gradually disappeared.

The remainder of this paper is organized as follows. Section 2 reviews related work and highlights differences from the current study, and Section 3 explains the procedures used to collect and label data. Section 4 proposes the opinion model, including the features that we analyze, and Section 5 describes the clustering method employed and evaluates its efficiency and accuracy. Finally, Section 6 concludes the paper and presents future research directions.

2. Related Work

Political manipulation is widespread in online communities of many countries. In particular, several government agencies and military intelligence units in South Korea led manipulative campaigns during presidential elections [7], spreading pro-government opinions to support ruling party candidates and disparaging anti-government views as being attempts by pro-North Korean forces to disrupt state affairs. Various government officials invovled are currently on trial. Massive numbers of manipulative posts have also been reported in Italian [8], Russian [9], an US [4] nation-wide elections. Millions of people have changed their voting behaviors due to online manipulations [5]. Compounding their direct impact, false beliefs are difficult to subsequently change once they become accepted [10], and continue to reappear as firm evidence in subsequent elections [8].

A few previous studies have explored different features to detect political manipulation, but almost all have trained a classifier based on supervised learning. Ratkiewiz *et al.* [6] model the way an idea propagates through multiple users in Twitter as diffusion patterns, and show that manipulative and non-manipulative ideas have different patterns, e.g. manipulative propaganda is more frequently retweeted by fewer users than non-manipulative ideas. Since retweet capability is not available for the web portals we analyzed, diffusion patterns are not directly applicable to the current study. Lee [3] models the amount of labor and collaboration among users, based on the observation that manipulators tend to work hard in teams to quickly influence the public. Some features are reused in the current study, including the number of opinions consecutively posted by a user. However, we exclude features related to the use of specific phrases, e.g. the number of words frequently used in political campaigns. This is because word usage differs from one campaign to another, hence collecting and preparing words for each campaign requires huge effort, which is inconsistent with one of our key goals to reduce the workload. Overall, previous studies utilized supervised learning, requiring labeling of a large training set. This is not practical, since manipulative tactics continue to evolve, as discussed in Section 5.4.

Although a significant portion of manipulative posts are for political purposes [11], they also have other objectives, most notably commercial reasons, such as product reviews that unfairly support particular products or negatively comment on other products. Previous studies have proposed various features to characterize manipulation in the commercial domain. Since feature importance differs significantly across domains [12], many features proposed in the commercial domain will not apply to the political domain, and vice versa. Features related to the rating system (e.g. the five-star rating system) appear most often. For example, users whose ratings deviate significantly from average ratings are commonly identified as potential manipulators [13]. TrueView [14] goes one step further and compares ratings across multiple sites. Although rating systems are common for product reviews, they are rarely used for political posts. Other studies have analyzed the timing between consecutive posts and shown that bursts of manipulative reviews tend to be posted on the same product over a short period [15], and such bursts reappear multiple times [16]. The current study also proposes features to identify concurrent posts. In contrast to other studies, ClickStream [17] uses unsupervised learning to cluster users with similar social-interaction patterns (e.g. friend request, photo viewing, and instant messaging), and then identifies anomalous clusters. In the web portals we analyzed, users mostly read, write, and approve posts, but social interactions are not of major concern.

3. Data Collection and Annotation

We prepared a set of opinions, labeled as either manipulative or non-manipulative, to analyze manipulative opinion characteristics (Section 4), and hence devise a detection system (Sections 5.1), estimate parameters (Section 5.2), and evaluate the accuracy of the proposed detection system (Sections 5.3−5.4). We first detail the process of collecting opinions (Section 3.1) and describe the method to label the opinions (Section 3.2).

3.1 Collecting Opinions

We collected over one million opinions posted on three popular web portals in South Korea[1], as summarized in **Table 1**. The opinions were collected during three political campaigns, since we expect that manipulations are most likely to occur during these periods to sway voters. Indeed, several government agencies and cyber military units were accused of organized manipulation in online social media during these campaigns [18], as ordered by the ruling party [19]. Several political parties were also charged with circulating fake news to defame opposing party candidates [20], and North Korea was also commonly reported to be involved in manipulative activities to create anti-government movements [21].

Table 1. Collected opinion summary

ID	Collection period	Major events during collection period	Opinions	Manipulative opinions (%)
1	2012.07–2012.12	The 18th presidential election	377,634	63,064 (16.7%)
2	2014.03–2014.07	The 6th local elections	344,237	32,358 (9.4%)
3	2017.04–2017.05	The 19th presidential election	379,093	53,684 (14.2%)

Table 2. Summary of web portals that opinions were collected from. Statistics from KoreanClick [22]

Web portal	Visitors per month	Description
Site 1	~36.3 million	Portal in South Korea, frequently used among all ages. It is famous for providing answers to a diverse range of user queries.
Site 2	~32.8 million	Popular portal in South Korea. Its online-community service is widely used by social groups.
Site 3	~17.2 million	Popular portal among those in twenties and thirties. Users can easily create personal web pages and share news and ideas.

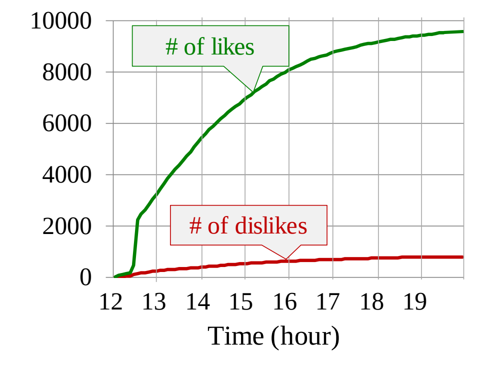
Table 2 summarizes the three sites where opinions were collected. These sites are among the most popular news portals in South Korea, and all sites present news articles grouped into categories; we focused on the politics category. Users can share their thoughts regarding each article by posting opinions, which we collected. We considered articles where users were highly engaged with (i.e., > 1,000 posts). Users can approve (or disapprove) of an opinion by pressing a button, similar to the 'Like' capability on Facebook. Some opinions were approved

[1] Demonstration of the collection process can be found at https://youtu.be/j2Pbf_1NpKQ. We performed collection throughout three years, 2012, 2014, and 2017, and the video presents the process for 2014.

by more than ten thousand users. A handful of opinions with the most approvals are shown at the front page of article, and we call such opinions *top posts*. Controversial opinions often received comparable approvals and disapprovals.

Each collected opinion is represented by a six-tuple, as shown in **Table 3**. Text values were originally written in Korean, translated into English for international readers. Item 1 of the tuple is news article title, on which the opinion is posted. This can be regarded as the subject of the opinion. Items 2−4 are the ID used to log into the portal, time the opinion was posted, and opinion content, respectively. Item 5 is the number of users who approve of the opinion, and item 6 the number who disapprove. Items 5 and 6 were collected regularly (5 minute intervals[2]) to track gradual changes over time and include such changes when detecting manipulation. For example, a sharp increase in user responses, such as 455→2240 (marked in boldface in **Table 3**), would most likely be due to an automated tool with a large pool of stolen IDs [23]. We performed regular collection for one week after posting, since the numbers rarely change beyond this period.

Table 3. Sample of collected opinions

ID	Item	Value	
1	Title	Candidate OOO’s campaign promises	
2	User ID	User-01	
3	Posting Time	2017-0503, 12:33:24	
4	Content	The son of candidate OOO received preference when applying for a job at a government organization. Witness testimony, including the son’s alumni, can be found at http://xxx.xxx/xxx. It would take years of preparation and hard work to land such a job for normal people like me. Even so, OOO pledges to build a fair society, where everyone is treated equally. Furthermore, OOO has been exploiting the victims of the ferry disaster merely for political purposes. OOO is clearly not eligible for the presidency and thus should resign immediately.	
5	# of likes	0, 82, 111, 142, 181, **455**, **2240**, 2463, 2621, 2855, 3053, 3238, 3443, 3650, 3858, 4039, 4213, 4378, 4547, 4732, 4894, 5081, 5259, 5434, …, 9605	
6	# of dislikes	0, 0, 16, 32, 56, 100, 131, 170, 184, 208, 236, 244, 256, 270, 291, 304, 315, 325, 332, 348, 368, 380, 384, 392, 400, 428, 432, 448, 472, 476, …, 892	

3.2 Labeling Opinions

We labeled the collected opinions as either manipulative or non-manipulative in two steps. In the first step, we collected additional evidence to help label each opinion, including (i) whether the opinion was reported by users, (ii) whether the opinion was deleted, and (iii) whether the associated user ID was deleted. We assumed that reported and deleted opinions were more

[2] Collecting numbers at intervals shorter than five minutes did not improve detection accuracy.

likely to be manipulative, according to the following reasons. Users can report abusive opinions, which are then reviewed by surveillance team and selectively removed if they are found to violate portal policies (e.g. libel, profanity, copyright infringement, unapproved advertisement, etc.) [24]. Multiple violations can lead to deletion of the associated user ID. Opinions and IDs are also deleted to remove the evidence of illegal activities. In fact, mass deletions occurred during the election campaigns in **Table 1**, when investigation started over the claims that several government organizations were involved in manipulation [25]. We found that approximately 15−20% of opinions were deleted during the study periods, which was unusual, with only 1−3% of opinions deleted in non-election periods. Thus, an opinion or ID being reported or deleted was a significant indicator of it being manipuative.

To obtain the evidence for (i)-(iii), we re-visited the portal sites three months after collection and checked the status of each collected opinion. We uniquely identified a particular opinion by the combination of posting time and user ID, and then confirmed whether the opinion and ID were flagged as reported or deleted[3].

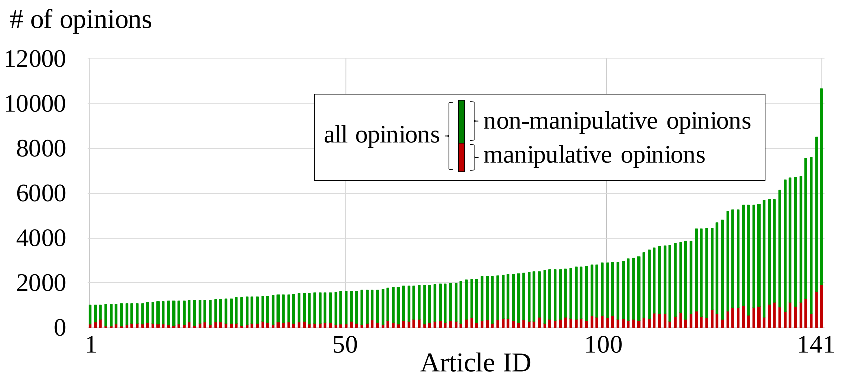


Fig. 1. Number of opinions per article in increasing order

In the second step of the labeling process, we labeled each opinion as manipulative or non-manipulative based on all available of the evidence, and expert assessment. This process was performed by five judges, each of whom had more than two years experience monitoring opinions in online social media, and were familiar with various manipulation strategies [1]. The judges were provided full access to the database of collected opinions and were able to run most SQL queries (e.g. they could list all opinions written by a paricular user ID and count the number of deleted opinions). Each judge labeled the opinion as manipulative or non-manipulative, and the final label was determined as the majority. Overall, 13.5% of the opinions were labeled as manipulative. **Fig. 1** illustrates the distribution of manipulative opinions for 141 articles for the third collection period in **Table 1**. The vertical bars represent the number of opinions in an article, and green and red segments correspond to the number of non-manipulative and manipulative opinions, respectively. Up to 20% of opinions were manipulative for some articles. Distributions were similar in the other collection periods.

We calculated Fleiss' multi-rater kappa (κ) [26] to analyze label quality. This measure shows the strength of agreement among multiple judges. We obtained κ=0.77, which represents substantial agreement according to the Landis and Koch scale [27]. We also verified the assumption that most deleted opinions were manipulative by analyzing the

[3] Deleted opinions remain on the sites, but are flagged as deleted and contents are hidden.

composition of deleted opinions in the labled dataset. Almost 88% of deleted opinions were judged to be manipulative, with the remainder comprising advertisements or the use of abusive language.

4. Opinion Modeling

We build an opinion model that characterizes abnormalities and thus can distinguish manipulative from non-manipulative opinions. This model is defined as a tuple of 78 features, as shown in **Table 4**. **Table 5** shows the terms and symbols used throughout Sections 4 and 5.

Let *Oi* denote a modeled opinion posted on article *Ai* by user *Ui*. The features are divided into two categories.

1. features 1–18 characterize opinion *Oi*, and
2. features 19–78 characterize author *Ui* over all their opinions.

These categories complement each other, for example suppose the first group raises suspicion that *Oi* was manipulated (e.g. it was immediately approved by a thousand user accounts), this can be verified by the second group (e.g. *Ui*'s opinions tended to be approved much more quickly than other users' opinions). Sections 4.1 and 4.2 explain the two categories, in detail.

Table 4. Features that characterize an opinion

ID	Feature	Description
Opinion characteristics (Section 4.1)		
1	time of day	time slot when *Oi* is posted, in hours (i.e., 0.00−23.99)
2	time after article post	posting time of *Oi* minus publication time of *Ai*
3	content length	text length of *Oi*, in bytes
4	# of URLs	# of web links in *Oi*
5	# of numerals	# of numerals in *Oi* (consecutive digits counted as one)
6	# of special characters	# of special characters in *Oi* (except those in URLs)
7	max Δ for likes	max growth in approvals on *Oi* over all collection intervals
8	time of max Δ for likes	time when max growth is seen in approvals
9	# of likes	# of approvals at final collection interval
10	max Δ for dislikes	max growth in disapprovals on *Oi* over all collection intervals
11	time of max Δ for dislikes	time when max growth is seen in disapprovals
12	# of dislikes	# of disapprovals at final collection interval
13	# of similar posts – own	# of *Ui*'s opinions with similar text to *Oi*
14	# of similar posts – others	# of opinions by other users with text similar to *Oi*
15	# of users with similar posts	# of users who post text similar to *Oi*
16	# of posts on same article	# of *Ui*'s opinions posted on same article *Ai*
17	# of *top posts*	# of *Ui*'s opinions shown at front page of article *Ai*
18	# of concurrent posts	# of *Ui*'s opinions posted at similar times to *Oi*
Author characteristics (Section 4.2)		
19	# of posts	# of opinions that *Ui* has ever written, over all articles
20	# of articles	# of distinct articles where *Ui* has ever posted opinions
21−24	# of posts per article*	# of *Ui*'s opinions on a single article
25−28	# of *top posts**	# of opinions by *Ui* shown at front page
29	# of articles with *top posts*	# of articles where *Ui*'s opinion is shown at front page
30	fraction of articles with *top posts*	fraction of articles where *Ui*'s opinion is shown at front page feature 29 divided by feature 20
31−34	time of day*	time slot when *Ui* posts opinions, in hours (i.e., 0.00−23.99)

35−38	time after article post*	posting time of *Ui*'s opinion minus publication time of article
39−42	content length*	text length of *Ui*'s opinion, in bytes
43−46	# of URLs*	# of web links in *Ui*'s opinion
47−50	# of numerals*	# of numerals in *Ui*'s opinion
51−54	# of special characters*	# of special characters in *Ui*'s opinion
55−58	max Δ for likes*	max growth in approvals on *Ui*'s opinion
59−62	time of max Δ for likes*	time when max growth is seen in approvals on *Ui*'s opinion
63−66	# of likes*	# of approvals on *Ui*'s opinion at final collection interval
67−70	max Δ for dislikes*	max growth in disapprovals on *Ui*'s opinion
71−74	time of max Δ for dislikes*	time when max growth is seen in disapprovals on *Ui*'s opinion
75−78	# of dislikes*	# of disapprovals on *Ui*'s opinion at final collection interval

* represents four features: maximum, average, median, and minimum per article, counted over all articles.

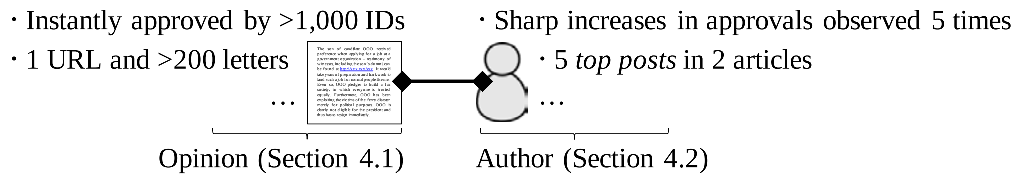

· Instantly approved by >1,000 IDs
· 1 URL and >200 letters
…
· Sharp increases in approvals observed 5 times
· 5 *top posts* in 2 articles
…
Opinion (Section 4.1)
Author (Section 4.2)

Table 5. Terms and symbols used in this article

Symbol	Description	Sections
Oi	Opinion	4, 4.1, 4.2
Ai	Article on which opinion *Oi* is posted	4, 4.1
Ui	User that posts opinion *Oi*	4, 4.1, 4.2
Top posts	Opinions shown on the first page when sorted, i.e., received highest numbers of approvals (fewer than 20 opinions, in general)	3.1, 4, 4.1, 4.2
JC	Jaccard coefficient, to measure similarity in opinion texts	4.1
T_{JC}	Threshold to confirm similarity in opinion texts ($JC > T_{JC}$)	4.1, 5.2
T_{TM}	Threshold to confirm temporal proximity of opinions	4.1, 5.2
W(Oi)	Set of distinct terms used in opinion *Oi*	4.1
K	Number of clusters that *K*-means clustering produces	5.1, 5.2
S	Number of seeds used to color clusters	5.1, 5.2

4.1 Opinion Characteristics

The first feature category (features 1−18) depicts opinion characteristics. Let *Oi* be posted on article *Ai* by user *Ui*. We performed a preliminary analysis of each feature's discriminative power in detecting manipulative opinions, as shown in **Fig. 2**[4], where the vertical axes show the cumulative percentage of opinions that exhibit the designated features. Large gaps between manipulative and non-manipulative opinions imply the feature is an effective discriminator.

Features 1−2 relate to the time when *Oi* is posted. Feature 1 is the absolute time, and feature 2 is the time relative to the publication time of *Ai*. For example, the sample opinion in **Table 3** was written at 12:33:24, and let us suppose the article was published at 11:33:24. Feature 1 = 12.56, represented as hours (12.56 ≈ 12 + 33/60 + 24/3600); and feature 2 = 1.00,

[4] In this section, we examine each feature separately, but in real classification task, the features are used together. We analyze such cases in Section 5.3. Some features are shown together as a sum in **Fig. 2**, when such a combination leads to a better discriminator, e.g. sum of features 4 and 5 (# of URLs + numerals) in **Fig. 2-(b)**.

since the opinion was written one hour after the article was published. Feature 1 captures situations where manipulators prefer to work at particular hours. Feature 2 represents how soon *Oi* is posted after *Ai* is made public; earlier *Oi* is more likely to be read by many users and become a *top post*. **Fig. 2-(a)** shows that manipulative opinions tend to be posted earlier than non-manipulative ones, and more than 50% of manipulative posts are made within the first hour after article publication.

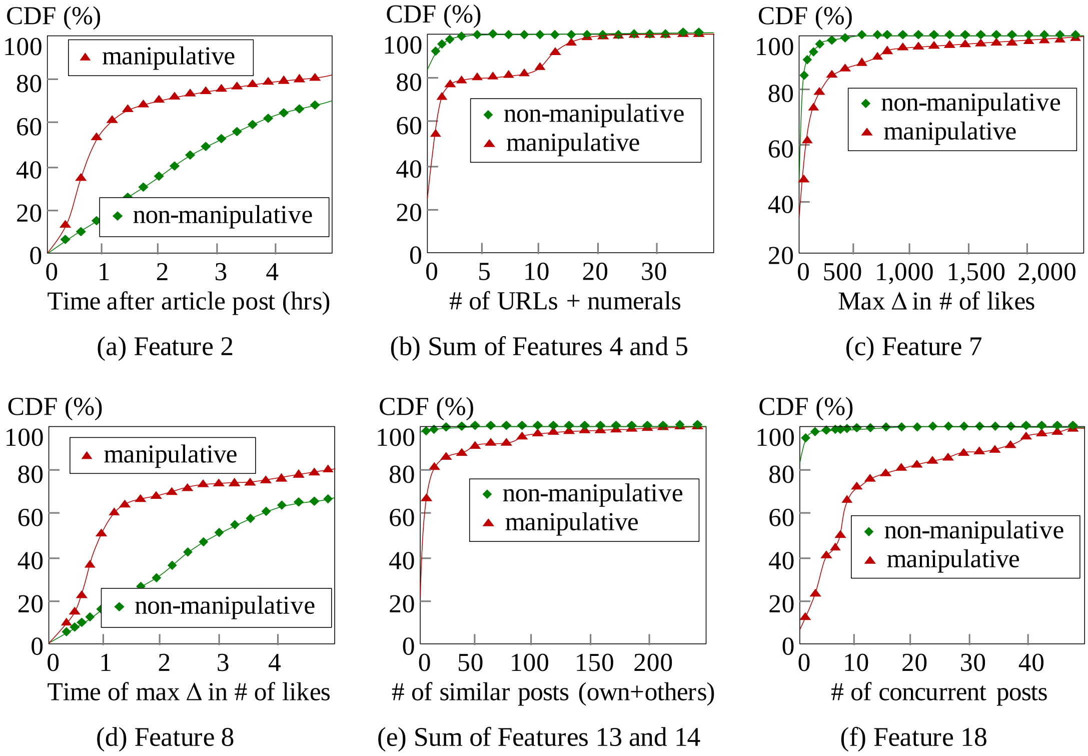


Fig. 2. Cumulative distribution function (CDF) of selective features that characterize an opinion

Features 3−6 model *Oi*'s text, including how many letters, URLs, digits, and special characters are present. For example, the sample opinion in **Table 3** contains 550 letters, 1 URL, and no digits or special characters. Therefore, features 3−6 = 550, 1, 0, and 0, respectively. Manipulative opinions tend to be lengthy, providing various supporting arguments to appear trustworthy, which often include references such as URLs to supplemental documents and videos, and specific numbers from surveys and research articles. The arguments also tend to be enumerated and highlighted using special characters. **Fig. 2-(b)** shows the cumulative distribution of opinions with varying numbers of references. Manipulative opinions tend to use more references than non-manipulative opinions, and nearly 20% of manipulative posts use more than 5 such references.

Features 7−9 represent approval behavior for *Oi*, and features 10−12 characterize disapprovals. Approvals are common manipulation targets—numerous approvals lead to a *top post*, which is seen by a large number of users, increasing the chance of influence. Feature 7 is the maximum growth rate of approvals. For the sample opinion in **Table 3**, maximum growth occurred when the number increased from 455→2240, hence feature 7 = 1785 (2240 - 455). Feature 8 is the time for maximum growth relative to *Ai*'s publication. Hence, if the observed maximum growth occurred 1 hour 30 minutes after publication, feature 8 = 1.50 (in hours). Feature 9 is the number of approvals in the final collection interval. Hence, for the opinion in **Table 3**, feature 9 = 9605. **Figs. 2-(c)** and **2-(d)** show cumulative distributions for features 7 and 8, respectively. Manipulative opinions are approved by more users and at earlier times than non-manipulative opinions. This translates to an effective manipulation strategy: create a *top post* as early as possible, as these tend to remain at the top.

Features 13−18 measure the extent of opinions that share certain characteristics with *Oi*. Manipulators publish a series of opinions to effectively spread propaganda, and these opinions often have various similarities (e.g. duplicate words and similar posting times). Feature 13 is the number of opinions with text similar to *Oi*. We consider that two opinions *Oi* and *Oj* have similar text if the Jaccard Coefficient (*JC*) [28] is greater than a predefined threshold T_{JC}. *JC* can be expressed as

$$JC(Oi, Oj) = |\ W(Oi) \cap W(Oj)\ | \div |\ W(Oi) \cup W(Oj)\ |,$$

where *W(Oi)* denotes the set of distinct terms used in *Oi*. Therefore, $JC > T_{JC}$ means a substantial fraction of *Oi* is reused in *Oj*, with a few words and phrases switched. In contrast to feature 13, which counts similar opinions written by *Ui*, feature 14 counts those by other users, and feature 15 is the number of distinct users who share similar text to *Oi*. Features 16−18 consider different similarity aspects. Feature 16 is the number of *Ui*'s opinions on the same article *Ai*, and among such opinions, feature 17 counts *top posts*. Feature 18 is the number of *Ui*'s opinions posted at similar times to *Oi* (regardless of the articles the opinions are posted to). We consider two opinions *Oi* and *Oj* are written at similar times if their posting times differ less than a predefined threshold T_{TM}. **Figs. 2-(e)** and **2-(f)** show cumulative distributions for similar text and posting times, respectively. Approximately 70% of manipulative opinions are probably reproduced from other opinions (**Fig. 2-(e)**), and some collections of duplicate opinions contain more than a hundred opinions. Manipulative opinions are often posted at similar times, whereas non-manipulative opinions tend to be posted individually (**Fig. 2-(f)**). These results indicate that manipulative opinions can be clustered according to their posting times and texts[5].

4.2 Author Characteristics

The second feature category (features 19−78) depicts the author's (*Ui*) general behavior over all their published opinions, in contrast to the first category that focuses on the details of one particular opinion. We utilize four metrics to describe general behavior: maximum (max), average, median, and minimum (min). These features are marked with an asterisk(*) in **Table 4**. For example, features 21–24 are the max, average, median, and min posts per article, respectively. Thus, suppose *Ui* wrote 9, 4, 4, and 3 posts on four articles, respectively, then features 21–24 = 9, 5, 4, and 3, respectively.

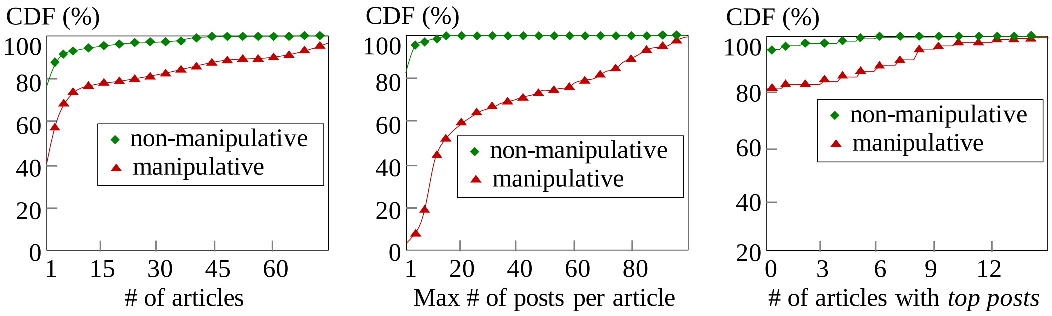
Features 19−24 measure the volume of *Ui*'s opinions to investigate how voracious a writer they are. Feature 19 counts the total number of *Ui*'s opinions over all articles, and feature 20 counts the number of distinct articles where these opinions are written. The opinions are then grouped according to their respective articles, to compute features 21–24 (as detailed above). **Figs. 3-(a)** and **3-(b)** show the cumulative distributions for features 20 and 21. Authors of

[5] Our experiments with the dataset established T_{JC}=0.5 and T_{TM}=20 (mins) as appropriate thresholds for best trade-off between false positives and false negatives. T_{JC}=0.5 corresponds to two-thrids of text overlap between opinions. T_{TM}=20 was found in manipulative opinions consecutively written by the same ID; more than 90% of such series were separated by less than 20 minutes.

manipulative opinions tend to write on multiple articles and post several opinions on each article, sometimes more than 80. In contrast, authors of non-manipulative opinions write on one or a handful of articles, and nearly 80% post a single opinion per article. Thus, being a voracious writer increases the probability *Ui* is manipulative.

Features 25−30 characterize *top posts* among *Ui*’s opinions. We first count the number of *top posts* in each article where *Ui* ever posts opinions, and then compute max, average, median, and min per article (features 25−28, respectively). Feature 29 is the number of articles where *Ui*’s opinions become *top posts*, and feature 30 is the ratio of such articles to the entire articles where *Ui* posts opinions. For example, suppose *Ui* posts 9, 4, 4, and 3 opinions on four articles, and among them 4, 2, 2, and 0 opinions, respectively, become *top posts*. Then features 25−30 = 4, 2, 2, 0, 3, and 0.75, respectively. **Fig. 3-(c)** shows the cumulative distribution for feature 29. Authors of manipulative opinions tend to leave *top posts* in more articles than non-manipulative authors, sometimes in more than 10 articles. Thus, manipulative posts are more likely to become *top posts* than non-manipulative posts, possibly by exploiting the approval system.

The remaining features (31−78) are derived from features 1−12 in the first category and are adjusted to *Ui*’s general behavior. Each of features 1−12 corresponds to four features in the second category, i.e., max, average, median, and min. For example, feature 1 is the posting time of one particular opinion *Oi*, and corresponding features 31−34 are max, average, median, and min posting times for all *Ui*’s opinions, respectively.



(a) Feature 20 (b) Feature 21 (c) Feature 29

Fig. 3. Cumulative distribution function (CDF) of selective features that characterize an author

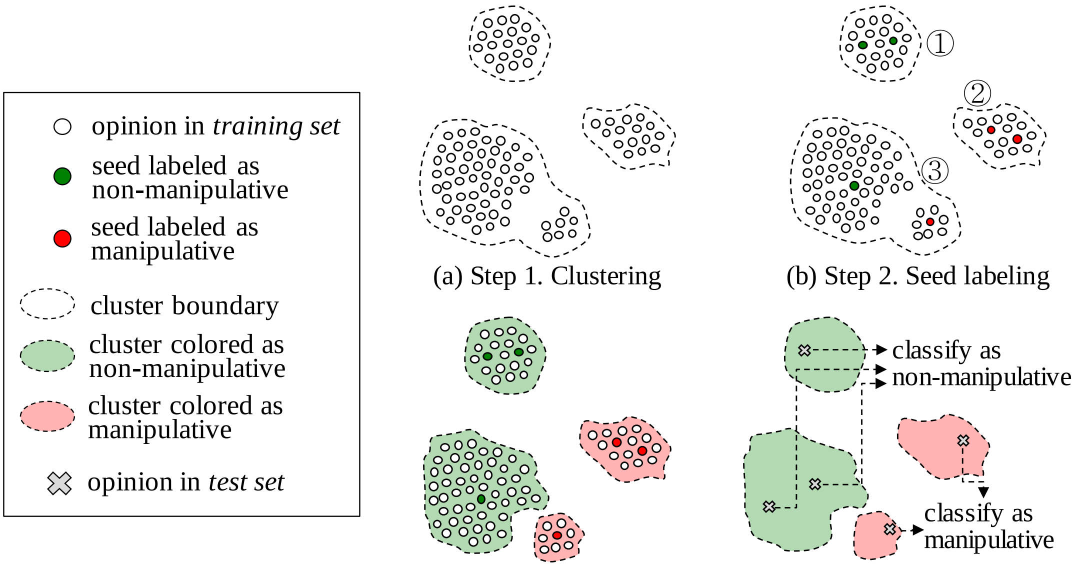
5. Unsupervised Detection of Manipulative Opinions

Based on the model from Section 4, we develop a system to identify manipulative opinions. Section 5.1 describes the details of the proposed system, Section 5.2 estimates system parameters, and Section 5.3 evaluates the resultant system’s accuracy. Section 5.4 applies the proposed system to opinions from three different years and tracks changes in manipulative tactics over the years. Section 5.5 compares the proposed system with conventional classifiers.

5.1 Clustering and Coloring Methods

The proposed system utilizes unsupervised learning, in contrast to previous studies using supervised learning. Supervised learning requires labeling large-scale data on a regular basis, since manipulative opinion characteristics alter over time, whereas unsupervised learning can minimize this effort.

We adopt *K*-means clustering [29][6]. Each opinion comprises an instance for clustering. Opinions with similar features are grouped into the same cluster, and those with dissimilar features are placed into different clusters. Manipulative opinions have distinct behaviors from non-manipulative opinions, as discussed in Section 4, and are likely to form separate clusters. Since not all manipulative and non-manipulative opinions exhibit identical behaviors, we expect multiple clusters of manipulative and non-manipulative opinions. Note that the clustering allows a straightforward interpretation of results; by inspecting the center of cluster (*centroid*), we can understand what opinions compose each cluster and how one cluster contrast with the other clusters.


○ opinion in *training set*
● seed labeled as non-manipulative
● seed labeled as manipulative
cluster boundary
cluster colored as non-manipulative
cluster colored as manipulative
opinion in *test set*
①
②
③
classify as non-manipulative
classify as manipulative

(a) Step 1. Clustering (b) Step 2. Seed labeling

(c) Step 3. Cluster coloring (d) Step 4. Classification

Fig. 4. Overview of clustering and classification process

Fig. 4 shows an overview of how we cluster opinions and how we classify opinions according to clustering results. The *training set* refers to a set of opinions used to build clusters and thus to initialize the proposed system. The *test set* corresponds to unclassified opinions input to the system in real time, which are then classified according to learned clusters. At first, both of the sets are not labeled. Later on in step 2, we label a small subset of the training set. The detailed steps are as follows.

- **Step 1.** We group the opinions in the training set using *K*-means clustering algorithm. The example in **Fig. 4-(a)** assumes that the number of clusters $K = 3$. Euclidean distance is used as the measure of similarity among opinions. We normalize each opinion feature to zero mean and unity standard deviation of one, so that one feature with a large variance does not dominate the similarity measure.
- **Step 2.** We randomly choose a small set of *S* opinions within each cluster and label them as manipulative or non-manipulative (In **Fig. 4-(b)**, S=2). These opinions determine the overall cluster label (*color*) in the next step, and we refer to them as *seed* instances.
- **Step 3.** We color each cluster depending on the seed labels as follows.

[6] We also considered partitioning algorithm METIS and anomaly detection based on Gaussian modeling, but these offered no significant improvement in accuracy.

If all seeds in cluster *C* have homogeneous label *l*, then we color *C* as *l*. **(3-1)**

Otherwise, we perform another round of *K*-means clustering on *C* to divide the opinions into two clusters and repeat this process until all resulting clusters have homogeneous seeds. Then we color the clusters according to step 3-1. **(3-2)**

Fig. 4-(b) shows that clusters ① and ② contain homogeneous seeds, hence they are colored as non-manipulative and manipulative, respectively (**Fig. 4-(c)**). Cluster ③ contains both manipulative and non-manipulative seeds, so it is further divided into two clusters with homogeneous seeds and colored accordingly. The additional clustering reduces the chance of incorrectly coloring a cluster, in case *K* is not sufficiently large and hence manipulative and non-manipulative opinions belong to the same cluster.

- **Step 4.** We classify the opinions in the test set according to cluster color. We compute the distance for unclassified instances to the centroid of each identified cluster, and then assign the label of the nearest cluster. **Fig. 4-(d)** shows three instances on the left are closer to non-manipulative clusters and hence classified as non-manipulative; whereas two instances on the right are closer to manipulative clusters and hence classified as manipulative.

Choice of the parameters *K* and *S* affects the accuracy and efficiency of the proposed method. In Section 5.2, we experiment with different choices and choose proper values.

5.2 Parameter Estimation

The proposed system requires two parameters: *K*, the number of clusters, and *S*, the number of seeds used to color each cluster. The choice of these parameters poses tradeoffs. On one hand, larger *K* allows us to fully separate opinions into their respective clusters and thus to discover all manipulative and non-manipulative clusters; and larger *S* ensures a sufficient number of seeds to accurately color clusters. On the other hand, *K* and *S* need to be small enough to be practical, since we will need to label $K \times S$ seeds (*S* seeds in each cluster), and fewer seeds means less labelling effort required.

To estimate the parameters, we randomly sampled 1/3 of the opinions in Section 3 (i.e., 366,988 opinions), and retained the remaining opinions for classification in Sections 5.3, 5.4, and 5.5. We then varied *K* and *S* and evaluated the resulting clustering quality. As a measure of quality, we used the percentage of homogeneous clusters—clusters that contain opinions of one, homogenous label and thus will be correctly colored. We ran 100 experiments for each *K* and *S* pair, and calculated the average, since the seeds are randomly selected and *K*-means clustering yields slightly different results for each run.

Fig. 5 shows the effect of different *K* and *S* choices. In general, increasing either (or both) leads to more accurate clustering and coloring. When $K = 70$, $S = 3$, i.e., $70 \times 3 = 210$ seeds total, would be sufficient to correctly color 90% of clusters. If $S = 9$, i.e., $70 \times 9 = 630$ seeds, accuracy approaches 99%. In practice, 90% homogeneous clusters, i.e., approximately 200 seeds, is sufficient to achieve acceptable classification accuracy – within the 10% of non-homogeneous clusters, the vast majority of opinions have the same label except a few opinions, so misplaced opinions are much less than 10%. We demonstrate this is the case in the next section. We also illustrate that supervised learning requires far more seeds to achieve the same level of accuracy.

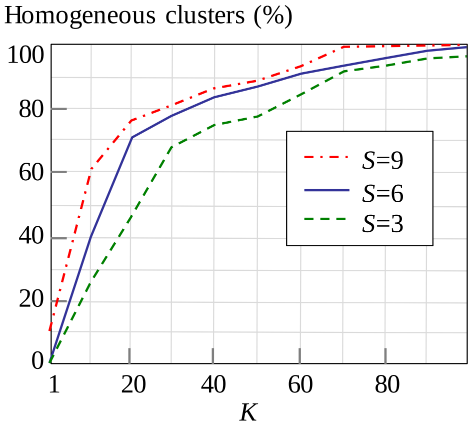


Fig. 5. Number of clusters (*K*) and seeds (*S*) vs. percentage of clusters with homogeneous labels

5.3 Classification Accuracy and Comparison with Supervised Learning

We evaluate the proposed system considering three aspects. First, we measure the accuracy of the system at classifying unknown opinions. Second, we investigate misclassified opinions and present explanations. Finally, we compare the proposed system with supervised learning.

We used the opinions not already used in Section 5.2, i.e., 2/3 of the dataset from Section 3. We randomly sampled half of this dataset (366,988 instances) as training data and the remainder (366,988) as test data. The training set was used to build and color clusters, and we assumed that this set was not yet labeled. Using the training set, we constructed $K = 70$ clusters and labeled $S = 3$ seeds from each cluster, which overall required labeling of slightly more than 200 opinions[7]. According to the clusters, we classified each opinion in the test set and confirmed whether this classification was correct. We ran the experiments 100 times and took the average outcomes.

Table 6. Classification outcomes from the proposed system

		True status	
		Manipulative (49,702)	Non-manipulative (317,286)
Classified As	Manipulative	**96.26%**	1.99%
	Non-manipulative	3.74%	**98.01%**

Table 6 summarizes the classification results. The horizontal axis lists the two classes in the original collection, and the vertical axis lists the two outcomes of classification. The number at each intersection represents a percentage relative to the total number of opinions in the corresponding class. For instance, out of 49,702 manipulative opinions, 96.26% were correctly classified as manipulative, and 3.74% were misclassified as non-manipulative. Similarly, out of 317,286 non-manipulative opinions, 98.01% were correctly classified, while the rest (1.99%) were not. Based on the results, the F1 measure was slightly over 92%.

Among the non-manipulative opinions, 1.99% were erroneously classified as manipulative. Most of these opinions were advertisements for products and/or services, unrelated to the political articles. Such opinions were often copied and massively reproduced, which was also

[7] Further increasing K and S more than doubles labeling effort but only slightly increased the classification accuracy, and the F1 measure remained below 93%.

common in manipulative opinions. Since this type of post is generally considered undesirable, it would be beneficial to detect and remove them. Among the manipulative opinions, 3.74% were incorrectly classified as non-manipulative. These opinions did not show typical characteristics of manipulative opinions, in that the writers appeared to post only a small number of opinions. Closer inspection showed that many of the corresponding user IDs were used together at almost the same time. We believe the tactic of utilizing multiple IDs simultaneously was used to avoid being reported and deleted. We could incorporate additional features to more accurately detect these cases, such as IP addresses the IDs used. For example, IDs that often share IP addresses can be analyzed together as belonging to the same user.

We compared the proposed system with supervised learning. In particular, we aim to answer whether the proposed system can achieve a comparable level of accuracy with supervised learning, while keeping labeling effort at a minimum. As a supervised classifier, we used Adaboost [30][8] with Decision Stumps as its base learner. We trained the classifier using a randomly chosen subset of the training data, gradually increasing the subset from 200 to the entire training set (366,988), and measured classification accuracy using the test data. Note that the supervised classifier requires all training data to be labeled, whereas the proposed method requires labeling a small number of selected instances (i.e., seeds).

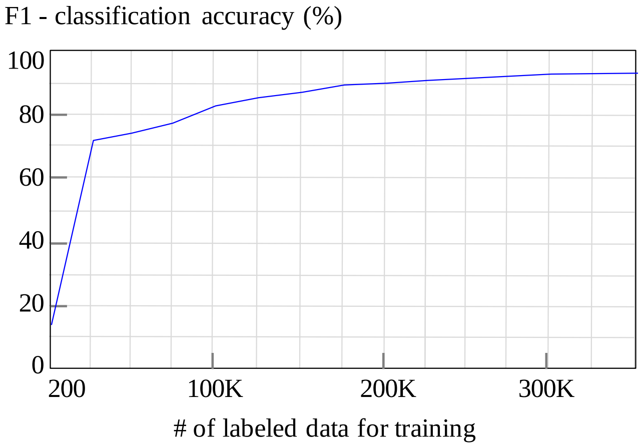


Fig. 6. Supervised classifier accuracy for different training set sizes

Fig. 6 presents the accuracy of the supervised classifier for different sizes of training data. With 200 labled data, the accuracy stayed below 20%. With 25K labeled data, the accuracy went above 70%, and it reached approximately 90% with 200K labeled data. Final accuracy was slightly above 92% when the entire training data were used. This contrasts with the proposed system, which achieved a nearly equivalent level of accuracy (92%) with far fewer labeld data (70×3=210 seed instances). This is because the proposed system selects seed instances with more diverse characteristics—we first cluster instances according to their characteristics and then from each cluster, we equally choose seeds to label. In supervised learning, however, it is not likely that the training data contain instances from each and every cluster unless the data is sufficiently large.

[8] We also considered Support Vector Machine, Random Forest, and Deep Neural Network, as shown in Section 5.5, but the choice of supervised classifier did not significantly alter the results.

5.4 Manipulative Tactics over Time

The opinions collected in Section 3 consist of data from three different years (2012, 2014, and 2017), as shown in **Table 1**. Over these years, new manipulative tactics may have arisen, and other behaviors reduced or disappeared. We analyze such changes over time. We also demonstrate that the proposed system can effectively trace temporal changes.

Fig. 7 shows an overview of our analysis methods. For each year's data, we construct and color clusters, according to the steps 1–3 in Section 5.1. We constructed K = 70 clusters and labeled S = 3 seeds to color each cluster. We then compare the identified clusters of one year with those of the subsequent year. In particular, we aim to discover three patterns as follows.

1. clusters that expanded or newly appeared,
2. clusters that shrunk or disappeared, and
3. clusters that maintained similar scale.

To identify these patterns, we consider two clusters from different years to be the same if their centroids are composed of nearly equivalent features (i.e., the centroids are in close proximity). We then measure the size of the clusters by the proportion of opinions in the cluster to the entire opinions. We focused on the clusters colored as manipulative, as our objective is to characterize and detect manipulation.

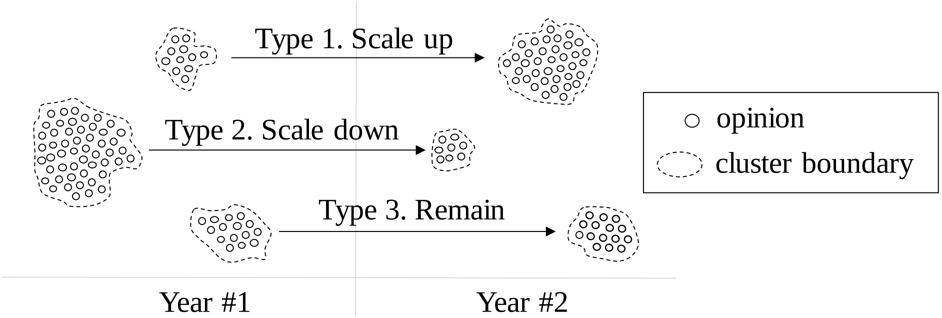


Fig. 7. Cluster analysis over time

Our main findings are as follows.

- Increasingly more manipulative opinions have been posted earlier, i.e., within several minutes after the articles were published. Many of these posts were immediately followed by massive approvals, to preemptively occupy *top-post* positions. Similarly, we observed sharp increases in the number of disapprovals, intended to denigrate opinions of opposing parties. One way to prevent this type of manipulation would be to rate-limit approvals and disapprovals, e.g. less than 100 votes per minute.
- More manipulative IDs have refrained from posting numerous opinions in a short period, reducing the chance of being reported and deleted. An in-depth analysis showed that many such IDs posted opinions at almost the same time, with similar contents; and these IDs often shared a series of letters (e.g. patriot001, patriot002, and patriot003). This indicates that the IDs probably belong to the same manipulative user or group. The proposed system could be expanded to leverage these similarities (i.e., posting time, content, and ID) to identify a group of correlated IDs.
- Manipulative opinions with similar text continued to form large clusters over the years. However, more and more such opinions replaced a subset of words by synonyms rather than

being exact duplicates. The manipulative opinions also became longer (e.g. over two hundred letters) over time. This suggests that automated tools are being used to compose opinions and post them on behalf of humans. The duplicate-opinion finder could be extended to identify word replacements utilizing lexical databases, such as WordNet [31]. Other aspects of opinion text, including URLs, numerals, and special characters, continued to exist in manipulative opinions at broadly constant levels, providing more clues and increasing readability.

To summarize the results, the proposed system identified changes in manipulative tactics by tracing cluster features and sizes over time. As more and more sophisticated tactics have been developed, we recommend web administrators perform periodic audits using the proposed system to track manipulation evolution.

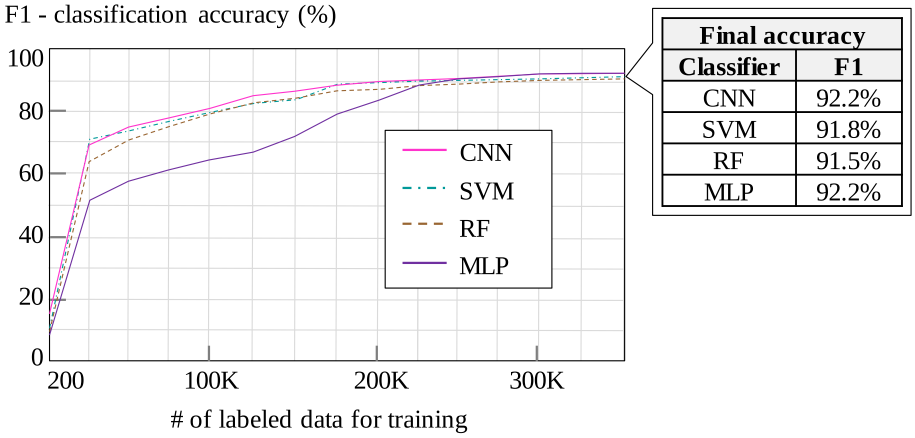
5.5 Comparison with Additional Classifiers

In Section 5.3, we compared the proposed system with an Adaboost classifier. In this section, we show a comparison with four additional classifiers that have been widely used in machine learning applications. These classifiers are Support Vector Machine (SVM) [32], Random Forest (RF) [33], Multi-Layer Perceptron (MLP) [34], and Convolutional Neural Network (CNN) [35].

To implement the SVM classifier, we utilized an open-source library, LIBSVM [36], and its *easy* script that automatically determines the best parameters. To implement the RF classifier, we used Weka package [37]. The MLP and CNN classifiers were built upon TensorFlow library [38]. The MLP and CNN employed 3 and 6 layers, respectively, and their details are summarized in **Table 7**. Further adding layers did not improve classification accuracy but increased computational costs. All four classifiers took the 78 features as input (i.e., a 1×78 matrix) and made predictions over the two class labels, manipulative and non-manipulative. We trained the classifiers using gradually increasing subsets of the training data, and measured classification accuracy using the test data.

Table 7. Dimension of MLP and CNN layers

<table>
<tr><th colspan="7">MLP details</th></tr>
<tr><td>Input</td><td colspan="2">Layer 1</td><td colspan="2">Layer 2</td><td colspan="2">Layer 3</td></tr>
<tr><td>1×78</td><td colspan="2">Fully connected
240 units</td><td colspan="2">Fully connected
240 units</td><td colspan="2">Fully connected
2 units</td></tr>
<tr><th colspan="7">CNN details</th></tr>
<tr><td>Input</td><td>Layer 1</td><td>Layer 2</td><td>Layer 3</td><td>Layer 4</td><td>Layer 5</td><td>Layer 6</td></tr>
<tr><td>1×78</td><td>Convolutional
3 1×15 filters
Stride of 1
↓
Max pooling
1×2 filter
Stride of 2</td><td>Convolutional
5 1×11 filters
Stride of 1
↓
Max pooling
1×2 filter
Stride of 2</td><td>Convolutional
10 1×3 filters
Stride of 1</td><td>Fully connected
40 units</td><td>Fully connected
10 units</td><td>Fully connected
2 units</td></tr>
<tr><td colspan="7">Layer 1 first applies (i) 3 convolutional filters with a dimension of 1×15 and a stride of 1 and then applies (ii) max pooling with a 1×2 filter and a stride of 2. Similarly, Layer 2 applies 5 convolutional filters and max pooling. Layer 3 applies 10 convolutional filters with no pooling. Layers 4 to 6 are fully connected layers, with the last layer connected to softmax units.</td></tr>
</table>



Final accuracy	
Classifier	F1
CNN	92.2%
SVM	91.8%
RF	91.5%
MLP	92.2%

Fig. 8. Accuracy of 4 classifiers with different training set sizes

Fig. 8 presents the accuracy of the four classifiers for different sizes of training data. In all of the classifiers, the accuracy increased as more labeled data were used, and it reached above 90% when greater than 200K labeled data were used. MLP consumed more training data to catch up with the other classifiers, since it had dense connections and thus needed to learn more parameters. The final accuracy was not significantly different from that of the proposed system (92%). Note that the four classifiers require all training data to be labeled, whereas the proposed system requires labeling a small subset of the training data (i.e., 70×3=210 seed instances).

We summarize our major findings as follows:

- In detecting manipulative opinions, the previous classifiers are of limited use in practice, since they require labeling a large number of training data. The proposed system significantly reduces such labeling effort—it first clusters similar instances and then from each cluster, it chooses a small number of instances to label.
- As manipulative tactics can change over time, a new set of data needs to be labeled on a regular basis. In such cases, the proposed system can be repeatedly used without incurring too much labeling costs. In addition, tracing emerging tactics is straightforward since they appear as new clusters, as shown in Section 5.4. Such tracing is not as easy in the previous classifiers because one needs to interpret complex models, e.g. SVM and neural networks.
- The classification accuracy remained below 93%, even though we tried different machine-learning algorithms and large training data. We believe that the accuracy can be improved with the help of additional information. For example, Section 5.3 shows that certain groups of IDs used together did not stand out as manipulative when each ID was only slightly used, and more features (e.g. IP addresses the IDs used) can help detect such cases.

6. Conclusion

We propose a system to inspect opinions in online communities and detect manipulative posts written for political campaigns. Compared to existing tools based on supervised learning, the proposed system accurately discovers manipulative opinions with moderate labeling effort required. This is because (i) the system comprehensively identifies the innate structure of opinions (i.e., clusters of opinions with similar characteristics) and uses this structure to select a small number of samples to label, and (ii) it models opinions with features that clearly separate manipulative and non-manipulative opinions. Moreover, the proposed system can trace emerging manipulative tactics, since such changes appear as new clusters.

We evaluated the proposed system using over a million opinions collected during major political campaigns in South Korea. The system required labeling approximately 200 instances to achieve over 90% classification accuracy, whereas comparable accuracy from supervised approaches would require labeling over 200K instances. We also discovered changes in manipulative tactics, such as early posting followed by approval manipulation, distribution of workload over an increasingly large set of IDs, and widespread use of automatic writing tools.

We believe that the proposed system can provide increased classification accuracy as more opinion related information becomes available. For example, user IP addresses could help identify a group of IDs used together by the same manipulative party. We also plan to apply the proposed system outside the political domain, e.g. to identify manipulative product reviews.

References

[1] K. Becker, “The handbook of political manipulation,” *Conservative Daily News*, 2012. Article (CrossRef Link).

[2] W. H. Riker, “The art of political manipulation,” *Yale University Press*, 1986.

[3] 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 16, 2015. Article (CrossRef Link).

[4] S. Shane and M. Mazzetti, “Inside a 3-year Russian campaign to influence US voters,” *The New York Times*, 2018. Article (CrossRef Link).

[5] R. Bond, et al., “A 61-million-person experiment in social influence and political mobilization,” *Nature*, vol. 489, no. 7415, pp. 295-298, 2012. Article (CrossRef Link).

[6] J. Ratkiewicz, M. D. Conover, M. Meiss, B. Goncalves, A. Flammini, and F. Menczer, “Detecting and tracking political abuse in social media,” in *Proc. of ICWSM*, pp.294-304, 2011.

[7] "South Korea's spy agency admits trying to influence 2012 poll,” *BBC News*, 2017. Article (CrossRef Link).

[8] D. Mocanu, L. Rossi, Q. Zhang, M. Karsai, and W. Quattrociocchi, “Collective attention in the age of misinformation,” *CoRR*, 2014. Article (CrossRef Link).

[9] “Russian Twitter political protests swamped by spam,” *BBC News*, 2012. Article (CrossRef Link).

[10] R. K. Garrett and B. E. Weeks, “The promise and peril of real-time corrections to political misperceptions,” in *Proc. of ACM CSCW*, pp. 1047-1058, 2013. Article (CrossRef Link).

[11] “Manipulation of online public opinion and the battle of Naver's real-time search words,” *The Kyunghyang Shinmun*, 2018. Article (CrossRef Link).

[12] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” *Now Publishers*, 2008.

[13] H. Oh and S. Kim, “Identifying and exploiting trustable users with robust features in online rating systems,” *KSII Tranactions on Internet and Information Systems*, vol. 11, no. 4, pp. 2171-2195, 2017. Article (CrossRef Link).

[14] A. J. Minnich, N. Chavoshi, A. Mueen, S. Luan, M. Faloutsos, “TrueView: harnessing the power of multiple review sites,” in *Proc. of WWW*, pp.787-797, 2015. Article (CrossRef Link).

[15] H. Li, G. Fei, S. Wang, B. Liu, W. Shao, A. Mukherjee, and J. Shao, “Bimodal distribution and co-bursting in review spam detection,” in *Proc. of WWW*, pp.1063-1072, 2017. Article (CrossRef Link).

[16] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, “Prominent features of rumor propagation in online social media,” in *Proc. of IEEE ICDM*, pp. 1103-1108, 2013. Article (CrossRef Link).

[17] G. Wang, T. Konolige, C. Wilson, X. Wang, H. Zheng, and B. Y. Zhao, “You are how you click: clickstream analysis for sybil detection,” in *Proc. of USENIX Security Symposium*, pp.241-256, 2013.

[18] S. Choe, “Prosecutors detail attempt to sway South Korean election,” *The New York Times*, 2013. Article (CrossRef Link).

[19] H. Fawcett, “South Korea’s political cyber war,” *Aljazeera*, 2013. Article (CrossRef Link).

[20] A. Shin, “Opposition party apologizes for spreading fake news on president’s son during election,” *Arirang*, 2017. Article (CrossRef Link).

[21] H. Olsen, “North Korean weighs in on South Korean presidential election,” *KoreaBANG*, 2012. Article (CrossRef Link).

[22] “KoreanClick: Nielsen KoreanClick syndicated reports,” *Nielsen KoreanClick*, 2015. Article (CrossRef Link).

[23] “Manipulation of recommendation counts by military and government agencies,” *Media Today*, 2013. Article (CrossRef Link).

[24] “Responsibility of users for their postings,” *Nate*, 2016. Article (CrossRef Link).

[25] A. Joy, “How South Korean intelligence interfered in election,” *KoreaBANG*, 2013. Article (CrossRef Link).

[26] J. Fleiss, “Measuring nominal scale agreement among many raters,” *Psychological Bulletin*, vol. 76, no. 5, pp. 378-382, 1971. Article (CrossRef Link).

[27] R. Landis and G. Koch, “The measurement of observer agreement for categorical data,” *Biometrics*, vol. 33, no.1, pp. 159-174, 1977. Article (CrossRef Link).

[28] B. Liu, “Web data mining: exploring hyperlinks, contents, and usage data,” *Springer*, 2011. Article (CrossRef Link).

[29] S. P. Lloyd, “Least squares quantization in PCM,” *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129-137, 1982. Article (CrossRef Link).

[30] Y. Freund and R. E. Schapire, “A short introduction to boosting,” *Journal of Japanese Society for Artificial Intelligence*, vol. 14, no. 5, pp. 1-14, 1999.

[31] C. Fellbaum and G. A. Miller, “Wordnet: an electronic lexical database (language, speech, and communication),” *MIT Press*, 1998.

[32] V. Vapnik, “The nature of statistical learning theory,” *Springer*, 2000. Article (CrossRef Link).

[33] L. Breiman, “Random forests,” *Springer Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001. Article (CrossRef Link).

[34] S. Haykin, “Neural networks and learning machines,” *Pearson*, 2009.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, 2017. Article (CrossRef Link).

[36] C. Chang and C. Lin, “LIBSVM-a library for support vector machines,” Retrieved August 24, 2018. Article (CrossRef Link).

[37] “Weka 3: data mining software in Java,” Retrieved August 24, 2018. Article (CrossRef Link).

[38] “TensorFlow: an open source machine learning library for research and production,” Retrieved August 24, 2018. Article (CrossRef Link).

Sihyung Lee is an associate professor at Seoul Women’s University in the Department of Information Security. He received the B.S. (with *summa cum laude*) and M.S. degrees in Electrical Engineering from Korea Advanced Institute of Science and Technology (KAIST) in 2000 and 2004, respectively, and the Ph.D. degree in Electrical and Computer Engineering from Carnegie Mellon University (CMU) in 2010. He then worked at IBM TJ Watson Research Center as a post-doctoral researcher. His research interests include pattern mining from social network traffic and management of large-scale network configurations.