Development Engineering 3 (2018) 218–233

Contents lists available at ScienceDirect

Development Engineering

journal homepage: www.elsevier.com/locate/deveng

A framework for detecting injected influence attacks on microblog websites using change detection techniques

Vishnu S. Pendyala*, Yuhong Liu, Silvia M. Figueira

Santa Clara University, USA

ARTICLE INFO

Keywords:
Cumulative sum
Discrete Kalman Filter
Sentiment analysis
Online Social Networks
Twitter bot
Attack injection
Microblogging

ABSTRACT

Presidential elections can impact world peace, global economics, and overall well-being. Recent news indicates that fraud on the Web has played a substantial role in elections, particularly in developing countries in South America and the public discourse, in general. To protect the trustworthiness of the Web, in this paper, we present a novel framework using statistical techniques to help detect veiled Web fraud attacks in Online Social Networks (OSN). Specific examples are used to demonstrate how some statistical techniques, such as the Kalman Filter and the modified CUSUM, can be applied to detect various attack scenarios. A hybrid data set, consisting of both real user tweets collected from Twitter and simulated fake tweets is constructed for testing purposes. The efficacy of the proposed framework has been verified by computing metrics, such as Precision, Recall, and Area Under the ROC curve. The algorithms achieved up to 99.9% accuracy in some scenarios and are over 80% accurate for most of the other scenarios.

1. Introduction

Social media has played an important role in Presidential elections in the United States as far back as 2008, during Barack Obama's election campaign (Cogburn and Espinoza-Vasquez, 2011). Its influence is so powerful that the political cyber hacker, Andres Sepulveda once said (Laquintano and Vee, 2017), “When I realized that people believe what the Internet says more than reality, I discovered that I had the power to make people believe almost anything.” His conviction turned out to be disastrous to the developing nations in Latin America. He proved he was right by faking social media accounts and using them to fabricate trends to sway the results of various Presidential elections in South American countries. This type of attack is also named as injected influence attack, which can be defined as the activity of posting malicious microblogs, often but not always, using automated means in order to hijack the opinions of the other users. The project described in this paper is an effort to detect and prevent such malicious attacks in the future.

Studies such as (Pak and Paroubek, 2010) have shown that Twitter corpus adheres to the Zipf law and can be used for opinion mining using sentiment analysis. Therefore the corpus of microblogs can be modeled as a Zipfian distribution. A Zipfian distribution is a type of discrete power law probability distribution. In an extensive survey on opinion mining and sentiment analysis (Liu, 2012), Liu discussed the statistical characteristics of the sentiment scores of opinion corpora.

Most of the current approaches, such as (Vosoughi, 2015), treat the problem of detection of rumors on Twitter as a classification task and use machine learning algorithms, such as SVM and Naive Bayes to train the classifier. The existing work, however, fails to sufficiently exploit the underlying characteristic of the problem, which is the fluctuation in the opinions expressed in the microblogs. In addition, as smart attackers often mimic normal users' behavior patterns to prevent themselves from being detected, machine learning algorithms which focus on specific patterns may often not perform well. For instance, the solutions presented in Vosoughi (2015) were successful in identifying around 70% rumors correctly, as compared to the results from the solution presented in this paper.

In this work, we propose to detect anomaly from another angle - changes based on the hypothesis that to influence the public opinions, the fake microblogs generated by malicious attackers will inevitably cause changes in the normal opinions. The framework we propose in this paper using change detection techniques utilizes this underlying characteristic to accomplish the task of discerning the fake tweets from the real ones, which is expected to result in better accuracy than the routine Machine Learning approaches.

Please note that occasionally, normal users' sentiments may be shifted due to the release of some startling news. However, such shifts can be easily validated by cross-checking other information sources,

* Corresponding author.
E-mail addresses: vpendyala@scu.edu (V.S. Pendyala), yhliu@scu.edu (Y. Liu), sfigueira@scu.edu (S.M. Figueira).

https://doi.org/10.1016/j.deveng.2018.08.002
Received 20 February 2018; Received in revised form 19 August 2018; Accepted 19 August 2018
Available online 29 August 2018

such as recent released news on other media. For brevity and focus, we do not include the cross-checking in this work and leave it as a future direction. The experimental setup for this paper is designed such that common injected attack scenarios are covered.

In addition, to cover the few cases where opinion shifts caused by injected attacks cannot be distinguished from those caused by genuine opinion shifts, the framework presented in this paper can be further strengthened by correlating and corroborating the results with those obtained by other techniques described in related literature. This paper is hoped to initiate a substantial discussion on solving an important problem that today's democracies, such as in the developing countries in South America, as highlighted earlier and the world in general is facing.

Cumulative Sum (i.e. CUSUM) is a statistical analysis technique (Page, 1954) used for change detection. The technique was successfully used on numerical data for anomaly detection in feedback-based online reputation systems (Liu and Sun, 2010). A Kalman Filter is typically used to predict values of data using a recursive algorithm. A series of observed values differing substantially from predicted values can indicate change.

Given that the sentiment scores from a microblog corpora can be modeled as a probability distribution, we hypothesize that statistical change detection techniques using CUSUM, modified for our purpose and a Kalman Filter can be applied to the sentiment scores of a microblog corpora to identify opinion shifts and therefore hacker attacks. Sentiment scores of microblogs on any given topic can be expected to be reasonably predictable. Delirious posts are commonly identified by humans, based on their out-of-whack sentiment. Machines can achieve the same result once the sentiment is quantified as a numerical score. Change detection techniques help in this process.

For this work, we chose the 2009 Iran elections as the domain, for reasons discussed in one of the following sections. We scored the tweets on the topic using automated sentiment analysis. We then applied our modified CUSUM (mCUSUM) and Discrete Kalman Filter algorithms to study the fluctuations in the sentiments. There was no apparent adverse impact noticed to the extent of suspecting an attack. We then injected a number of tweets with negative sentiment to simulate several scenarios. We repeatedly applied the two techniques to analyze the tweet sentiments in the various scenarios and were successful in detecting the injected tweets with an impressive accuracy of around 90%. For simplicity, we use the words, microblogs and tweets interchangeably, but the techniques discussed in this paper are by no means restricted for use with Twitter.

To the best of our knowledge and literature survey, the paper is unique in efficiently and successfully proposing a framework to use Change Detection techniques such as modified CUSUM (mCUSUM) and Kalman Filter with sentiment analysis to detect OSN hacker attacks intended to influence voters in microblogs. Specifically, we have modified the basic CUSUM in a novel way to make it better fit our application scenario. We tested both the mCUSUM and Kalman Filter Change Detection techniques, applied each one to various scenarios, and compared the results. The results confirmed our hypothesis, so the techniques explored here can possibly be extended to solve similar problems after quantifying the sentiment trends or other critical aspects of information. Moreover, based on the testing results, we further propose a comprehensive way to perform anomaly detection by integrating mCUSUM and Kalman Filter in a flexible way. The model we use is generic enough to be implemented as a framework.

Last but not least, due to the wide adoption of Web information, protecting a trustworthy Web is also essential for other domains. For example, there are a number of humanitarian projects that have been made possible by the Web. One such possible application is Web-based medical diagnosis using the techniques presented in Pendyala et al. (2014) and Pendyala and Figueira (2017). These rely heavily on the truthfulness of the underlying data, which is not entirely tamper-proof. The use of CUSUM, modified for our purposes, and Discrete Kalman Filtering techniques presented in this paper for detecting hacker attacks is hoped to pave way for detection of tampering of critical data such as in the medical domain as well.

The rest of the paper is organized as follows. Section 2 presents existing literature. The next section, section 3 discusses the design aspects, detailing the framework, the approach, the techniques, and the algorithms. Section 4 provides the experiment details and results. Section 5 concludes the paper with a discussion on the results and future directions.

2. Related work

There are a number of papers on the topics related to the influence of OSNs, anomaly detection, and misinformation containment areas, which are closely related to our project. We list some of the interesting ones in the following subsections.

2.1. Influence of social media

The first set of papers we examined relate to whether microblogs make a difference to the outcome of public opinion and decision making. To evaluate the influence of the tweets, we need a quantitative measure of their sentiment. Choy et al. did a sentiment analysis of the tweets related to the Singapore Presidential elections (Choy et al., 2011) to estimate the number of votes each candidate will get. Their work proves the important role that tweets play in elections to the high office and proves the correlation between the sentiment analysis of the tweets and the election results. Similarly, the authors of Tumasjan et al. (2010) examine the role tweets played in the German elections. One of their conclusions is that even mentioning the party name in the tweet has non-trivial impact. The more the number of mentions, the higher the chances of winning the elections.

In an extensive analysis of the Twitter corpus on two significant topics in the recent past, “Brexit” and “Trump”, Hall et al. (2018) lead us to a powerful conclusion, “Society might well need to quickly determine new ethical boundaries around the use of social media data analysis during election campaigns, or AI could determine who our next leaders will be.” Using analytical methods such as Sentiment Analysis, Temporal Profiling, LDA Topic Modeling, and visualization artifacts, such as Network Structure, they scrutinize the role that social media played in the two important referendums: Brexit and US Presidential elections.

Using quantitative analysis, authors of Jin et al. (2014a) portray how rumors spread on Twitter during the Ebola crisis in Africa, highlighting the impact lies on microblogging websites have had on developing countries. More literature survey shows that there is predominant evidence that there is substantial impact of the social media posts on the public opinion, supporting the purpose of this paper, which is to detect malicious use of microblogging during crucial events like elections. Burns and Eltham in their highly cited work (Burns and Eltham, 2009), examine the impact of Twitter on the Iran Election crisis that provides a sociological prelude to the technical discussion in this paper.

2.2. Anomaly detection

Other researchers have approached the topic of anomaly detection. The discussion on detecting certain type of security breach events such as “Sarah Palin's email account was hacked” from the tweets using semi-supervised learning methods in Ritter et al. (2015) gives good insights into the process of examining and making sense of odd-sounding tweets. Singh et al. (2014) propose ways to identify malicious users using five different classifiers and compare the results. They conclude that Random Forest resulted in highest accuracy. To improve the accuracy of our prediction, we may consider using Random Forest to extend our work in the future to confirm that the attacks we detect in