Suspicious Content and Profile Identification Based on Quantifying Data on Twitter

Progress report

M.Tech Thesis Evaluation-1
October 2017

by

Surendra Singh Gangwar

(2016IPG-107)

under the supervision of

Dr. Santosh Singh Rathore

ABV-INDIAN INSTITUTE OF INFORMATION TECHNOLOGY AND MANAGEMENT GWALIOR-474 010

CANDIDATE'S DECLARATION

I hereby certify that I have properly checked and verified all the items as prescribed in the check-list and ensure that my thesis/report is in proper format as specified in the guideline for thesis preparation.

I also declare that the work containing in this report is my own work. I, understand that plagiarism is defined as any one or combination of the following:

1. To steal and pass off (the ideas or words of another) as one's own
2. To use (another's production) without crediting the source
3. To commit literary theft
4. To present as new and original an idea or product derived from an existing source.

I understand that plagiarism involves an intentional act by the plagiarist of using someone else's work/ideas completely/partially and claiming authorship/originality of the work/ideas. Verbatim copy as well as close resemblance to some else's work constitute plagiarism.

I have given due credit to the original authors/sources for all the words, ideas, diagrams, graphics, computer programmes, experiments, results, websites, that are not my original contribution. I have used quotation marks to identify verbatim sentences and given credit to the original authors/sources.

I affirm that no portion of my work is plagiarized, and the experiments and results reported in the report/dissertation/thesis are not manipulated. In the event of a complaint of plagiarism and the manipulation of the experiments and results, I shall be fully responsible and answerable. My faculty supervisor(s) will not be responsible for the same.

Signature:

Name: Surendra Singh Gangwar
Roll No.: 2016IPG-107
Date: 14-10-2020

ABSTRACT

Online Social networking is the medium of exchanging activities, entertainment and information. Although the social network provides huge benefits to the persons but at the same time also harms with different malicious social activities.This causes our society considerable economic loss and national security even threatens.

Twitter is one of the most famous social media platforms and because of its prevalence, spammers discover this stage to spam with clients. Twitter spamming is all the more compromising in light of the fact that its assortment of crowd, twitter clients length over all divisions of life for example it very well may be the educators or understudies, VIPs or politicians, marketers or clients or even the overall population.Because of URL shorteners, common and informal languages and abbreviations used on social networking sites filtering out the malicious content becomes a challenging problem. Industries and researchers have since used different techniques to eliminate spam content from social networking sites. Some of them are based only on user-based features, while others are based on the content-driven features of tweets. In our work, we will try to make a model that will combine both types of features and create some hybrid features and will also used relation based features to classify the content and users.The benefit of using the function of the tweet content is that we can discern the spam tweets regardless of whether the spammer creates another account that was impractical only with the content-based features of the customer and tweet.Twitter itself uses its methods of filtering suspicious content, but it actually blocks content that has already been delivered to so many users. So we will try to build a strategy that will also block the suspected user. In this work we will perform the classification, evaluation and comparison of various spam separating strategies and sum up the general situation with respect to the exact pace of various existing methodologies.

Keywords: Social network security, Spam Filtering, User-Content features, Relation Features, Word Embeddings, Natural language processing, Ensemble, Machine Learning

Contents

1 Introduction

Currently, applications for social networks are commonly used around the globe, including Twitter, Facebook, Instagram . As the data show, 700 million users are revealed by Instagram on its blog post. Twitter has refreshed its dynamic client numbers over quite a while to 328 million. social platforms are turning into a path for spammers to spread malevolent or irritating messages to ordinary clients. In a tweet user can add text, urls, videos and images. Twitter increases character limit to 280 characters. It allows various functionalities like follow a user, mention, hashtag, reply, retweet. Hashtag is used to categorize a tweet into a special category and all tweets related to that tweet can be read by clicking that tag.

At the point when any remarkable occasion happens, a large number of clients tweet about it and quickly make it a trending subject. These trending themes become the objective of spammers who post tweets consisting of some trademark expressions of the moving point with URL interfaces that lead clients to totally disconnected sites. As tweets normally incorporate abbreviated URL joins, it becomes for the clients to recognize the substance of the URL without stacking the site.

Spammers can have a few thought processes behind spamming, for example, advertise a product to produce exceptional yield on deals, compromising the user's account. Spammers contaminate the continuous pursuit climate, however they additionally affect tweets statistics. To filter out the malicious content becomes a challenging problem because of URLs shorteners, modern and informal languages, and abbreviations used on social networking sites. Spammers influence the users to click a particular URL or to read the content with specific phrases of words.

2 Review of key related research

This section discusses the recent works in the field of spam detection on twitter and their shortcomings.

Kaur et al. (2016) surveyed all the research papers from 2010-2015 and found all the techniques used in these research papers. There are numerous methods utilized by researchers.

User Based Techniques: By analysing these tools we can classify a user as spammer or non-spammer. A user's account includes important information like No. of followers, No. of Following, No. of mentions, Tweets creation time.

Hybrid Features: On the basis of user based features some new features can be derived like Reputation (ratio of followers with following) , Frequency of tweets, Rate at which user follows other users etc.

Content Based Techniques: It analyzes the text properties and decides whether tweets are spam or non- spam. A tweet content has some crucial information like Number of hashtags in comparison to total word count, Users mentioned in a tweet, Number of URLs, count of numerals etc.

Relation Based Features: These features provide more accurate results because these features can't be modifie by the spammer. It uses a connection degree whether a person mentioned a direct friend in a tweet or a mutual friend. They found that hybrid techniques provide better results.

Dangkesee and Puntheeranurak (2017) perform an adaptive classification for spam detection. They used spam world filter and url filter using blacklisted urls. After labeling and preprocessing the data set Naive Bayes classifier is used with 50000 and 10000 tweets. They found that their proposed model outperformed spam world filters by comparing Accuracy, precision, recall, f1-score. They suggest utilization of the safebrowsing instead of url blacklisting for filtering URLs.

Raj et al. (2020) proposed multiple machine learning algorithms to classify tweet content. In KNN(92%), Decision Tree classifier(90%), Random Forest Classifier(93%), Naive Bayes classifier(69%), Random forest outperformed. They suggest that after detecting the tweet as a spam tweet can be deleted.

Song et al. (2011) presented Bagging, SVM, J48, BayesNet with relation based features by creating graph b/w users they used distance and connectivity b/w users.Finally Bagging outperformed with 94.6% true positive and 6.5% false positive. They also highlight one drawback that if any user created a new account and generates a tweet will be added in the spammer category.

Alom et al. (2020) conducted CNN with tweet text and with both tweet text and meta-data features. It utilizes NLP methods like word embeddings, n-grams methods. They convert the text into meatrix before sending it to CNN. Second method gives good accuracy around 93.38% because it combines both the features. This method outperformed other deep learning methods so they suggest to use this in other social networks like linkedin and facebook.

Mateen et al. (2017) proposed a hybrid approach for spam detection in which they used different combinations of approaches like content based and graph based feature and user based with graph based features. Applies J48, decorate and naive bayes classifier with these features. Accuracy for content and graph based features achieved 90% and for user and graph based features achieved 92%. They also find correlated features and remove features with higher correlation.

3 Objectives

The thesis would aim to complete the following objectives:

- Try to develop a system where we can pass real time tweets to classify them as spam or non-spam and user as legit or spam.
- Compare different methods and their performances on the model using different preprocessing techniques.
- Using different combinations of content, user and relation based features with different ensemble based learning.
- Applying proposed model with some constraint on tweets like language and trending topics like covid-19.
- Extend this work to generate the more efficient model by bringing parameter tuning into consideration.

4 Methodology

4.1 Dataset

For our proposed model we will collect all the tweets in the range of 3 months.We will crawl the tweets using Twitter Developer API using Tweepy library. Dataset will be converted into User, content and relation based features. All the tweets will be classified into spam non-spam tweets, later we will try to classify the user as legit user and spam user.

4.2 Tools Required

- Programming Language: Python
- Tweepy, Scikit-Learn, Keras,Pandas, Matplotlib, Numpy, NLTK, Word Embeddings.

4.3 Proposed Methods For Training Model

Models which have been built previously have used either user based features, content based features or relation based features of a tweet etc. we will use the combination of all these three types of features and also filter spam with blacklisted url and safe browsing. We can combine advanced pre-processing techniques to achieve more robust models. We would be using the following pre-processing techniques for building our model:

- Applying constraints on tweets like language specific and trending topics like covid-19.
- Extracting user, content and relation based features and creating three different dataset one with content-relationship, user-relationship, and content, user and relationship based features.
- Feature extraction will be performed with bag-of-words, n-grams and then term frequency like tf-idf weighting. We can also use word2vec and POS tagging to extract the features.

4.4 Brief Architecture

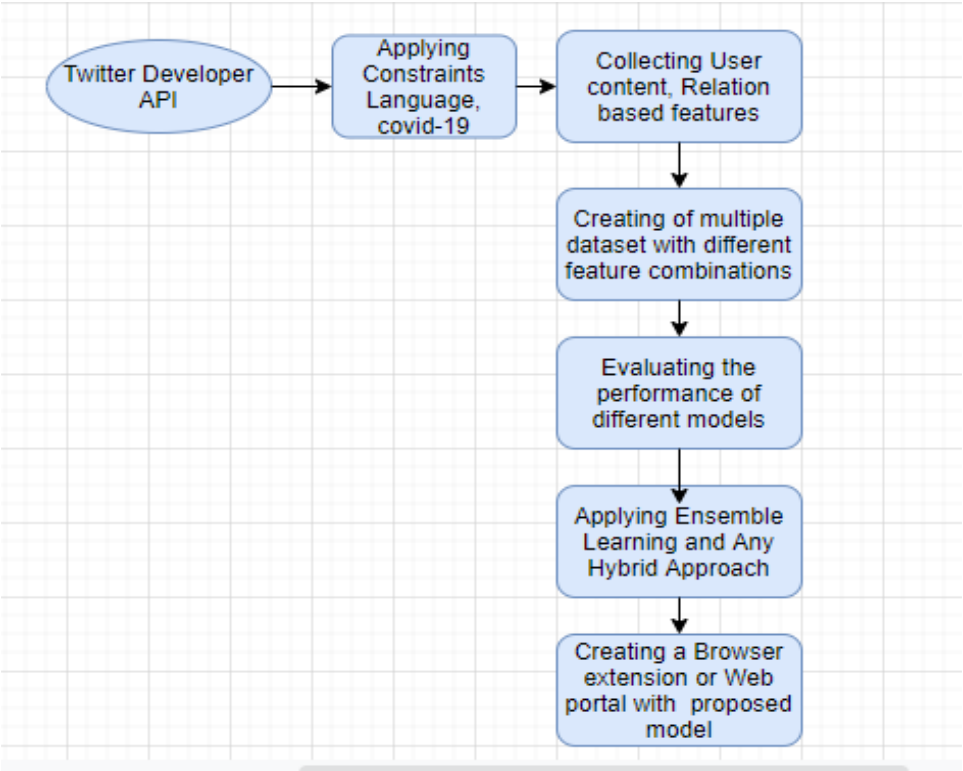

Figure 1: Methodology

We present a brief architecture for developing our model. It shows the various steps which would be involved while developing our Spam detection Model.

5 Expected research outcome

The expected outcomes during the course of research have been discussed below:

- Evaluating the performance of various techniques on datasets with different feature combinations.
- To deliver a robust algorithm for detecting spam tweets with a great accuracy, precision, recall, f1-score.
- Use parameter tuning to come up with even more advanced Architecture for developing robust model.

References

[1] Alom, Z., Carminati, B. and Ferrari, E.: 2020, A deep learning model for twitter spam detection, *Online Social Networks and Media* **18**, 100079.

[2] Dangkesee, T. and Puntheeranurak, S.: 2017, Adaptive classification for spam detection on twitter with specific data, *2017 21st International Computer Science and Engineering Conference (ICSEC)*, IEEE, pp. 1–4.

[3] Kaur, P., Singhal, A. and Kaur, J.: 2016, Spam detection on twitter: A survey, *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, IEEE, pp. 2570–2573.

[4] Mateen, M., Iqbal, M. A., Aleem, M. and Islam, M. A.: 2017, A hybrid approach for spam detection for twitter, *2017 14th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, IEEE, pp. 466–471.

[5] Raj, R. J. R., Srinivasulu, S. and Ashutosh, A.: 2020, A multi-classifier framework for detecting spam and fake spam messages in twitter, *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*, IEEE, pp. 266–270.

[6] Song, J., Lee, S. and Kim, J.: 2011, Spam filtering in twitter using sender-receiver relationship, *International workshop on recent advances in intrusion detection*, Springer, pp. 301–317.