

# Suspicious Content and Profile Identification Based on Quantifying Data on Twitter

*Progress report*

**M.Tech Thesis Evaluation-1**

December 2020

*by*

**Surendra Singh Gangwar**

**(2016IPG-107)**

*under the supervision of*

**Dr. Santosh Singh Rathore**



विश्वजीविनामृतं ज्ञानम्

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

## CANDIDATE'S DECLARATION

I hereby assure that I have appropriately checked and certified all the things as proposed in the enrollment and affirmation that my theory/report is in the genuine arrangement as appeared in the norm for thesis.

I likewise announce that the work containing in this report is my own work. I, comprehend that counterfeiting is characterized as any one or blend of the accompanying:

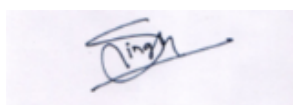
1. To take and pass off (the thoughts or expressions of another) as one's own.
2. To utilize (another's creation) without crediting the source
3. To submit abstract burglary
4. To present as new and unique a thought or item got from a current source.

I appreciate that copyright encroachment incorporates a purposeful showing by the scholarly hoodlum of using someone else's work/considerations absolutely/to some degree and ensuring creation/innovativeness of the work/contemplations. Verbatim copy similarly as close resemblance to some else's work build up copyright encroachment.

I have given due credit to the first creators/sources for all the words, thoughts, outlines, designs, PC programs, tests, results, sites, that are not my unique commitment. I have utilized quotes to recognize verbatim sentences and offered credit to the first creators/sources.

I insist that no bit of my work is plagiarized, and the analyses and results announced in the report/exposition/proposal are not controlled. In case of a grumbling of literary theft and the control of the analyses and results, I will be completely capable and responsible. My personnel supervisor(s) won't be liable for the equivalent.

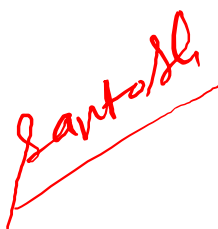
Signature:



Name: Surendra Singh Gangwar

Roll No.: 2016IPG-107

Date: 10-12-2020



## 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 [Dangkesee and Puntheeranurak \(2017\)](#), while others are based on the content-driven features of tweets [Mateen et al. \(2017\)](#). 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

<b>1</b>	<b>Introduction</b>	<b>5</b>
<b>2</b>	<b>Review of key related research</b>	<b>6</b>
<b>3</b>	<b>Objectives</b>	<b>8</b>
<b>4</b>	<b>Methodology</b>	<b>9</b>
4.1	Dataset . . . . .	9
4.1.1	Important Features . . . . .	9
4.2	Tools Required . . . . .	10
4.3	Overview of Architecture . . . . .	11
4.4	Proposed Methods For Training Model . . . . .	12
4.5	Methodology . . . . .	13
4.5.1	Tweets Collection . . . . .	14
4.5.2	Spam labelling . . . . .	14
4.5.3	Evaluating the performance of different Machine learning Models . . . . .	14
4.5.4	Feature extraction . . . . .	14
4.5.5	Hybrid Model . . . . .	15
<b>5</b>	<b>Expected research outcome</b>	<b>16</b>

## List of Figures

1	Overview of Architecture	11
2	Methodology	13
3	Twitter API	14

# 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 [Dangkesee and Puntheeranurak \(2017\)](#). 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

[Raj et al. \(2020\)](#). 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 [Mateen et al. \(2017\)](#). 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 modified 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 word 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 word 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 matrix 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.

Gharge and Chavan (2017) arranged a setup of machine learning tool "Weka", which uses various machine learning algorithms. They have picked SVM as their principle classifier. They introduce a new feature which matches the tweet content with url destination content. They used a random set of 1000 tweets, out of those 1000 tweets 95-97% were classified efficiently.

Gupta and Kaushal (2015) proposed an integrated algorithm that combines the benefits of three distinctive learning algorithms (to be specific Naive Bayes, Clustering, and Decision trees) was implemented. This incorporated calculation classifies a record as Spammer/Non-Spammer with a by and large precision of 87.9%.

Lin and Huang (2013) analyzed the importance of existing features for recognizing spammers on Twitter and utilize two basic yet compelling features (i.e., the URL rate and the collaboration rate) to characterize the Twitter accounts. This study dependent on 26,758 Twitter accounts with 508,403 tweets shows that the classification has precision up to around 0.99 and 0.86 and a higher recall.

Hua and Zhang (2013) proposed an adaptable method to detect spams on Twitter utilizing content, behavioral, and graph-based information and after different investigations, a threshold and associative based model is created. This new model is compared with SVM, and two other existing algorithms using accuracy, precision, recall. The new classifier with a precision of 79.26% is better than SVM with a precision of 69.32%.



### 3 Objectives

The thesis would aim to complete the following objectives:

- Try to develop a system that can classify the tweets as spam or non-spam and user as legit or spam in the real-time.
- 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

The proposed work utilizes the tweets fetched using twitter developer API. Twitter allows its users to fetch twitter data using tweepy library. The Tweepy library required four user credentials like consumer\_key, consumer\_secret, access\_key, access\_secret to send the request over API. We fetched 2000 latest tweets which consists of many features like timestamp, tweet text, username, hashtags, followers count, following count, number of mentions, word count, is retweet etc. All of these features are categorized in content based features and user based features. We created some Hybrid features like reputation of a user and frequency of tweets of a user and following frequency. For labeling the dataset as spam or non-spam we use hybrid features, blacked list URLs and some predefined words in the text. Finally the dataset is prepared for analysing the performances of different machine learning models. Later three different datasets will be created by combining user-content features, user-relation features, user-content-relation features. These dataset will be utilized for our proposed model.

#### 4.1.1 Important Features

- Count of followers and followees: Followers are those users who follow a specific user, while followees are the users which are followed by a specific user. In general, spammers have limited numbers of followers but large followees. Therefore, users with large followees and limited numbers of followers can possibly be considered a spam account.
- URLs: URLs are the connections that direct to some other page on the program. With the improvement of URL shorteners, it has presently become simple to post irrelevant connections on any OSN. This is since URL shorteners hide the original content of the URL, in this way making it hard for detection algorithms to detect malicious URLs. An excessive number of URLs in tweets of a user are an expected pointer of the user being a spammer.
- Spam Words: A record with spam words in pretty much every tweet can be viewed as a spam account. Subsequently, text including spam words can be considered as an significant factor for identifying spammers.
- Replies: Since, data or message sent by a spammer is pointless, thusly individuals once in a while answers to its post. On the other hand, a spammer answers to an enormous number of presents all together on getting seen by numerous individuals. This example can be utilized in the recognition of spammers.
- Hashtags: Hashtags are the novel identifier ("#" trailed by the identifier name) which is utilized to bunch comparative tweets together under a similar name. Spammers utilize an enormous number of hashtags in their posts, with the goal that their post is posted under all the

hashtag classifications and consequently gets wide viewership and is peruse by many.

## **4.2 Tools Required**

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

### 4.3 Overview of Architecture

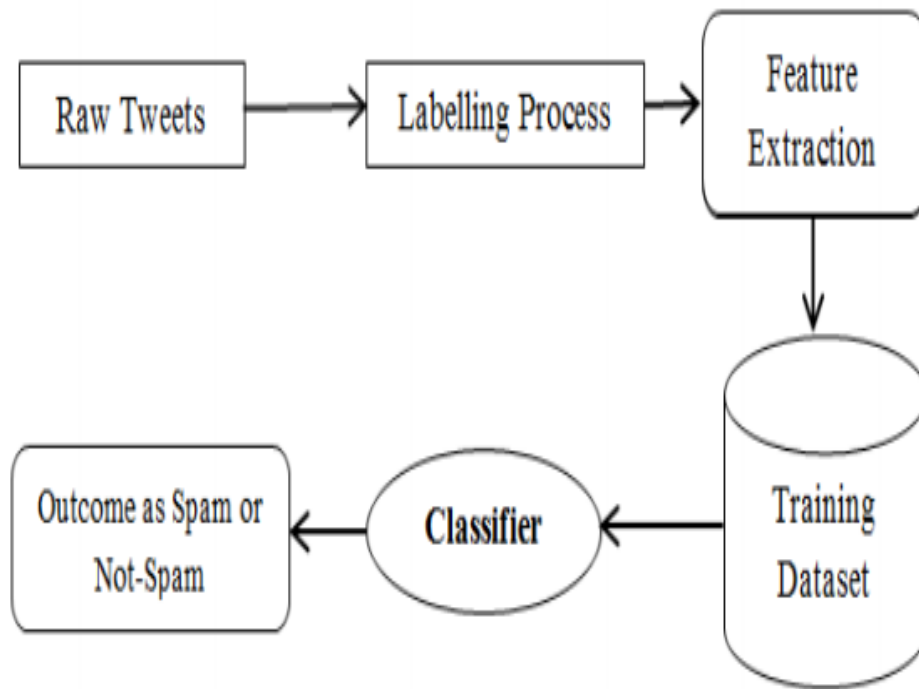


Figure 1: Overview of Architecture

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

## 4.4 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.
- We will perform Feature extraction using n-grams, bag-of-words and afterward term recurrence like tf-idf weighting. We can likewise utilize word2vec and POS labeling to find features.

## 4.5 Methodology

Below Figure represents different steps involved in our proposed work.

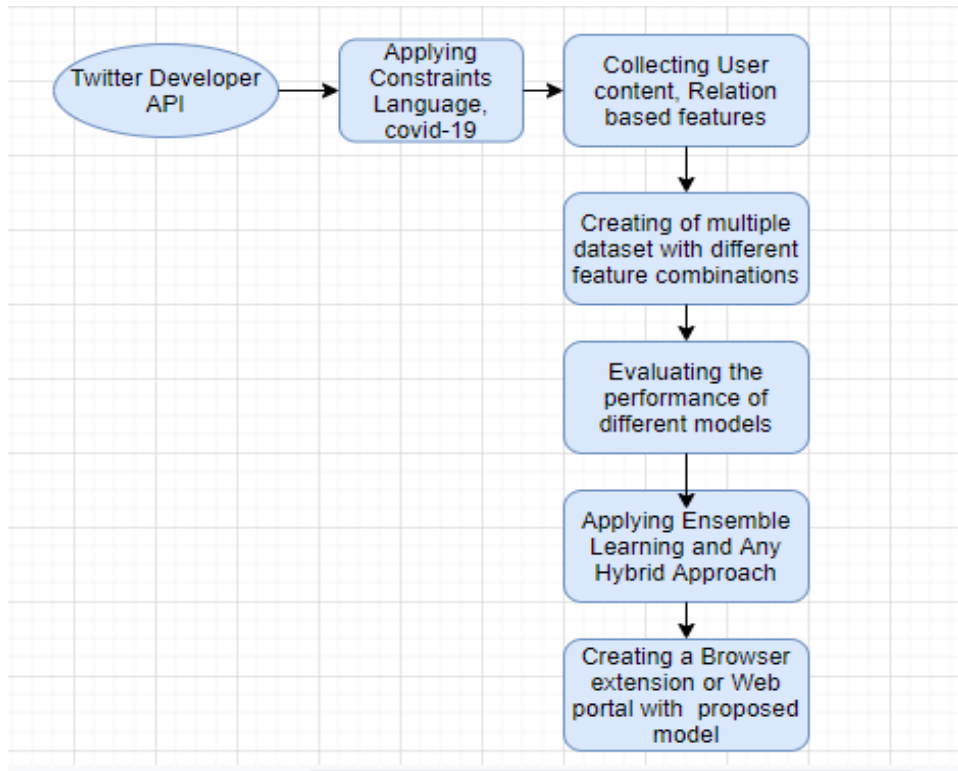


Figure 2: Methodology

### 4.5.1 Tweets Collection

First the framework fetches tweets utilizing twitter dev API The tweets are caught and put away in a particular document arrangement and afterward they are good to go to be examined. Distinctive datasets will be made with a mix of various sorts of features.

```
In [15]: """
INPUTS:
    consumer_key, consumer_secret, access_token, access_token_secret: codes
    telling Twitter that we are authorized to access this data
    hashtag_phrase: the combination of hashtags to search for
OUTPUTS:
    none, simply save the tweet info to a spreadsheet
"""
def search_for_hashtags(consumer_key, consumer_secret, access_token, access_token_secret, hashtag_phrase):
    #create authentication for accessing Twitter
    auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)

    #initialize Tweepy API
    api = tweepy.API(auth)

    #get the name of the spreadsheet we will write to
    fname = '_' + re.findall(r"#[\w+]", hashtag_phrase)

    #open the spreadsheet we will write to
    with open('%s.csv' % (fname), 'w', encoding='utf8') as file:
        #with open('%s.csv' % (fname), 'wb') as file:

        w = csv.writer(file)

        #write header row to spreadsheet
        w.writerow(['timestamp', 'tweet_text', 'username', 'all_hashtags', 'followers_count'])

        #for each tweet matching our hashtags, write relevant info to the spreadsheet
        for tweet in tweepy.Cursor(api.search, q=hashtag_phrase, -filter:retweets, \
                                  lang="en", tweet_mode='extended').items(1500):
            w.writerow([tweet.created_at, tweet.full_text.replace('\n', ' '), tweet.user.screen_name.encode('utf-8')])

In [16]: consumer_key = input('Consumer Key')
consumer_secret = input('Consumer Secret')
access_token = input('Access Token')
access_token_secret = input('Access Token Secret')

hashtag_phrase = input('Hashtag Phrase')

if __name__ == '__main__':
    search_for_hashtags(consumer_key, consumer_secret, access_token, access_token_secret, hashtag_phrase)
```

Figure 3: Twitter API

### 4.5.2 Spam labelling

Initially, All of the tweets are unlabeled, They need to be labeled as spam or non-spam for training purposes. we use hybrid features, blacked list URLs and some predefined words in the text to label them.

### 4.5.3 Evaluating the performance of different Machine learning Models

In this step we will train KNN, Naive Bayes classifier, Decision Tree Classifier and various Ensemble methods like Bagging and Boosting with the prepared dataset and will evaluate the performance of these models in terms of accuracy, precision, recall.

### 4.5.4 Feature extraction

There may be some crucial features which play an important role for classifying the tweets. So in these steps these features will be extracted from the dataset. We will use Bag of words and TF-IDF vectorizer techniques.

- Bag of Words: Bag-of-Words is one of the most major strategies to transform tokens into a group of features. The Bag of word model is utilized in record classification, where each word is utilized as an element for building the classifier.
- TF-IDF Vectorizer:TF-IDF represents frequency-inverse document frequency. It features a particular issue that probably won't be too successive in our corpus yet holds incredible significance.

#### **4.5.5 Hybrid Model**

Finally we will apply hybrid model with different datasets created with combination of different feature types. We will compare the performance of our Hybrid model with different machine learning models and ensemble methods.



## 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] Gharge, S. and Chavan, M.: 2017, An integrated approach for malicious tweets detection using nlp, *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, IEEE, pp. 435–438.
- [4] Gupta, A. and Kaushal, R.: 2015, Improving spam detection in online social networks, *2015 International conference on cognitive computing and information processing (CCIP)*, IEEE, pp. 1–6.
- [5] Hua, W. and Zhang, Y.: 2013, Threshold and associative based classification for social spam profile detection on twitter, *2013 Ninth International Conference on Semantics, Knowledge and Grids*, IEEE, pp. 113–120.
- [6] 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.
- [7] Lin, P.-C. and Huang, P.-M.: 2013, A study of effective features for detecting long-surviving twitter spam accounts, *2013 15th International Conference on Advanced Communications Technology (ICACT)*, IEEE, pp. 841–846.
- [8] 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.
- [9] 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.
- [10] 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.