International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 6, June 2020 ||

Detection of Fake News and Fake Account on Twitter Dataset using SVM and KNN

Nivedita Patil[1], Shruti Mathapati[2], Geeta Navhi[3], Mahantesh Laddi[4]

Student, Department of Computer Engineering, S.G.Balekundri Institute of Technology, Nehru Nagar, Karnataka, India[1]

Student, Department of Computer Engineering, S.G.Balekundri Institute of Technology, Nehru Nagar, Karnataka, India[2]

Student, Department of Computer Engineering, S.G.Balekundri Institute of Technology, Nehru Nagar, Karnataka, India[3]

Assistant Professor, Department of Computer Engineering, S.G.Balekundri Institute of Technology, Nehru Nagar, Karnataka, India[4]

ABSTRACT: Twitter popularity has fostered the emergence of a new spam marketplace 'The services that this market provides include: the sale of fraudulent accounts, affiliate programs that facilitate distributing Twitter spam, as well as a cadre of spammers who execute large scale spam campaigns. It is experienced while surfing on websites. This project explores the identification of fake news which is spread on the social media particularly on twitter. In this paper we present machine learning algorithms we have developed to detect fake news in Twitter. The ease of access and rapid improvements in social media has enabled more than half of the world's population using it in their daily life. These sites not only act as a platform for staying connected with friends and exchanging opinions but also help to share and disseminate information. With the huge availability of information over the Social media platform, People consider them as the primary source of news. Are all the information over the social media credible and trustworthy? With the flexibility of anyone can share anything over the platform, Social Media is more prone to the spread of irrelevant and misleading information. This extensive spread of spam has the potential for extremely negative impacts on individuals and the society. Therefore, detecting the spam or false information on social media has gained attention of many researchers.

KEYWORDS:Twitter , Real News, Fake News, Credibility Scores.

I. INTRODUCTION

Fig 1:The detection of fake accounts percentages

Twitter has become a popular media hub where people can share news, jokes and talk about their moodsand discuss news events. In Twitter users can send Tweets instantly to his/her followers. Also, Tweets can be retrieved using Twitter's real time search engine. The ranking of tweets in this search engine depends on many factors, one of which is the user's number of followers. Twitter's popularity has made it an attractive place for spam and spammers of all types. Spammers have various goals: spreading advertising to generate sales, phishing or simply just compromising the system's reputation. Given that spammers are increasingly arriving on twitter, the

success of real time search services and mining tools lies in the ability to distinguish valuable tweets from the spam storm. There are various ways to fight spam and spammers such as URL blacklists, passive social networking spam traps, manual classification to generate datasets used to train a classifier that later will be used to detect spam and spammers.

So what is Twitter spam? As Twitter describes it in their website [3], Twitter spam is "a variety of prohibited behaviors that violate the Twitter Rules." Those rules include among other things the type of behavior Twitter considers spamming, such as:

- "Posting harmful links (including links to phishing or malware sites) .
- Aggressive following behavior (mass following and mass un-following for attention), particularly by automated means.
- Abusing the @reply or @mention function to post unwanted messages to users .
- Creating multiple accounts (either manually or using
- automated tools)
- Having a small number of followers compared to the number of people one is following.
- Posting repeatedly to trending topics to try to grab attention
- Repeatedly posting duplicate updates
- Posting links with unrelated tweets"(Twitter, n,d.).

Twitter actually fights spammers by suspending their accounts. But in general Online Social Networking sites do not detect and suspend suspicious user accounts quickly. They are not willing to deploy automated methods to detect and remove spam accounts fearing that this will lead to a serious discontentment among users. Thus, they wait until the ,sufficient number of users report a specific account as a spam count to suspend it. However, legitimate users are unwilling to invest time to report spammers. Hence spammers are allowed more time to spread spam .

The spread of rumors and false information among public can have serious impacts on our society. For example during the London riots in 2011[8], Twitter was used as a medium to spread rumors. Rioters spread rumors about certain incidents like London eye being set on fire, police beating up a 16 year old, rioters breaking into McDonalds, rioters attacking London Zoo and the animals being freed, attacking the children's hospital at Birmingham and army being deployed in bank which other users further tweeted, leading to the spread of these rumors. This led to panic in the city and the government had to take immediate steps to halt it. So, it is very important to identify the misleading information spreading across social media for the benefit of the society. Unreliable evidence of hoaxes, conspiracy theories and fake news on social media is so abundant that massive digital misinformation has been ranked among the top global risks for our society.

To mitigate the negative effects caused by fake news/rumor–both to benefit the public and the news ecosystem. It's critical that we develop methods to automatically detect fake news on social media. The Fake News detection on social media is a classification problem that can be solved using various classification algorithms. This classification problem can be solved using algorithms such as Naïve Bayes, SVM, KNN] etc. We will be classifying this problem through all the above mentioned classifiers and also have designed a cascaded classifier by combining few classifiers. We use the Twitter dataset to extract various features and classify it using a relevant classifier to detect the hoax. The various features that are considered to classify are Length of the tweet, Number of words, number of hashtags, Number of retweets, number of swear language words, number of special words, number of URLs. Malicious users spreading rumors on Social media can be tracked down and the further spreading of these fake news can be identified and mined using the above suggested approach in a more effective way.

This paper deals in ranking the tweets based on the credibility during high impact events using SVM, Naive Bayes and K-Nearest Neighbour algorithms. They use the dataset collected from Twitter website by scraping the twitter pages. On feeding the dataset, data is classified using the models derived using different classifiers. As a part of Post-processing steps, performance of the existing output is combined with other classifiers in a cascading way and also ranked the tweets based on the Naive Bayes output. Finally, implemented the maximum voting by feeding each test point to all classifiers. With the data analyzed, 28% of total tweets posted about an event was spam on an average. Around 67% of the total tweets posted about the event contained situational awareness information that was credible.The paper completely deals with detection of spam and non-spam tweets mainly based on the features set. Training the model is done using SVM, Naive Bayes and KNN classifiers.We

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 6, June 2020 ||

concluded to use the following features like number of URLs, number of swear words, number of spam words, length of tweet text, favorites, retweets which seems to be more reliable for the dataset, therefore we chose retweets, favorites and the sum of all other features to detect the class for the tweets dataset using different classifying algorithm. We evaluated its accuracy by comparing with the output of different classifying algorithms such as Naïve Bayes, KNN and SVM. The implementation in the paper resulted in an accuracy of 81%, so we aimed in gaining a much higher accuracy with the above mentioned feature set using only SVM classifying algorithm.

II. RELATED WORK

The prevalence of fake accounts and/or „bots" is continuously evolving, and feature based machine learning detection systems employing highly predictive behaviors provide unique opportunities to develop an understanding of how to discriminate between bots and humans, i.e. between real vs. fake accounts on social media. [8]

Ferrara et al. (2016), in their Taxonomy of Social Bots Detection Systems, have commended the use of machine-learning methods for bot detection based on the identification of highly revealing features that differentiate them from humans (real users). By focusing on differences in behavioural patterns between bots and humans, these features can be easily encoded and adopted by way of machine learning techniques to identify and classify accounts into the category of bot or human based on their observed behaviors.

Creating and using a baseline dataset of verified human and fake follower accounts, Cresci et al.(2015), in their seminal work, have exploited the baseline dataset to train a set of machine-learning classifiers built according to the reviewed rules and features set by media and academia (based on existing literature) respectively. Their results show that while the rules proposed by media are weak in detecting fake followers in a satisfactory fashion, features proposed in the past by academia for spam detection are a lot more accurate and efficient in providing the required results. Building on the most promising features, they have revised these classifiers, developing a Class A classifier, which is lightweight and general, yet itis able to correctly classify more than 95% of the accounts of the original training set, tested for global sensitivity.

A slightly different approach has been used by Zhang and Lu (2016), who proposed a novel method of fake account detection in Weibo, the Chinese counterpart of Twitter. This detection framework is based on the premise of why such accounts exist in the first place, i.e. „to follow their targets enmasse".Hence it has been observed that there is high overlap between the follower lists of the customers of such accounts. They have investigated the top Weibo accounts whose follower lists duplicate or nearly duplicate each other. Using a sample-based approach in their experiment, they found 395 near-duplicates, leading to 11.90 million fake accounts (4.56 % of total users) sending 741.10 million links (9.50 % of the entire edges). They have further characterized four typical structures of spammers, clustering them into 34 groups, and analysing the properties of each group.

One of the early efforts to detect and fight spam and spammers on Twitter was by Fabricio et al. [1]. They manually annotated a sample that consists of 8207 users, 335 of which were spammers and 7852 were non-spammers. They selected 710 of the legitimate users to include in their collection which is twice the number of the labelled spammers. Thus, the total size of their labelled collection was 1065 users. They identified two sets of attributes that distinguish spammers from non-spammers. The first set they called "content attributes" and the second set "user behaviour attributes". The content attributes include properties of the text of the tweets posted by the users. For instance, they captured the maximum, minimum, average and median of the following metrics: number of hash tags per number of words on each tweet, number of URLs per words, number of words of each tweet, number of characters of each tweet, number of URLs on each tweet, number of hash tags on each tweet, number of numeric characters (i.e. 1,2,3) that appear in the text, number of users mentioned in each tweet, number of times the tweet has been retweeted (counted by the presence of "RT @username" in the text).

From the content attribute sets they collected, they inspected 3 attributes that they believed would largely distinguish spammers from non-spammers. Those attributes are 1) Fraction of tweets containing URLs 2) Fraction of tweets containing spam word 3) Average number of tweets that are hashtags. They drew the cumulative distribution function (CDF) for those three attributes, and they found that those attributes indeed distinguish spammers from non-spammers. Since spammers tend to have more tweets containing URLs, more tweets with spam words and higher numbers of hashtags per tweet [1].

They also identified 23 user behavior attributes, some of which are:- number of followers, ratio of followers per followees, number of tweets according to age of user account, number of times the user was mentioned, number of

times the user was replied to, existence of spam words in the user screen name, number of tweets posted per day and per week, etc. Also, they inspected 3 user behavior attributes they believed would distinguish spammers from non-spammers. They drew the CDF for these 3 attributes: number of followers per number of followees, the age of the user account, and the number of tweets received. They found that spammers have a higher ratio of followers to followees and they explain this by the fact that spammers try to follow a large number of users in hope that they will be followed back. They also found that spammers‟ accounts are mostly new since they frequently get blocked and immediately create new accounts. Finally, they reported that non-spammers receive a much larger number of Tweets from their followees compared to spammers (Benevenuto, Magno, Rodrigues, & Almeida, 2010). Fabricio et al. (Benevenuto, Magno, Rodrigues, & Almeida, 2010) used the 39 content attributes and the 23 user behavior attributes to distinguish spammers from non-spammers using a supervised machine learning algorithm which is SVM. They performed 5-fold cross-validations for testing. Their results are presented in the confusion matrix
table below.

TABLE I: FABRICIO *ET AL.* [3] SPAMMER DETECTION ALGORITHM RESULTS

		Predicated Classification	
		Spammer	Non-spammer
True classification	Spammer	70.1%	29.9%
	Non-spammer	3.6%	96.4%

Besides studying the spammer detection problem, they studied the problem of detecting spam tweets [1]. Detecting spam tweets can be very useful for real time search, while detecting spammers is helpful in suspending spammers‟ accounts. To detect spam Tweets, they considered the following attributes: number of words from a list of spam words, number of hashtags per words, number of URLs per words, number of words, number of numeric characters on the text, number of characters, number of URLs, number of hashtags, number of mentions, number of times the tweet has been replied to, and whether the tweet was posted as a reply. They used SVM classifier with these attributes. They used a labeled collection of tweets that are labeled as spam or non-spam. The classifier was able to identify 78.5% of spam and 92.5% of non-spam.

Thomas et al. [3] collected 1.8 billion tweets sent by 32.9 million Twitter accounts. Then, they identified the accounts suspended by Twitter for abusive behavior. They found that 1.1 million accounts were suspended. To verify that the 1.1 million suspended accounts were spam accounts they drew a random sample consisting of 100 accounts. They then analyzed the Tweets posted by those accounts to find common spam keywords, frequent duplicate tweets, tweet content that appears across multiple accounts, the landing page of each tweet‟s URL and the overall posting behavior of each account. They found that 93 accounts were suspended for posting spam and unsolicited product advertisements; 3 accounts were suspended for exclusively retweeting content from major news accounts and the remaining 4 were suspended for aggressive marketing and duplicate posts. Thus they discerned that the majority of the suspended accounts are created by spammers. They found that only 8% of URLs appearing in fraudulent accounts appeared in blacklists. The problem is that those blacklists are used as techniques to identify social network spam which mean social networks should not rely on blacklists to detect spam.

Different researches have been presented to detectfake accounts with different approaches. In this research, we will follow the feature based detection approach [11]. This approach is based on monitoring the behavior of the user such as his number of tweets, retweets, friends, etc. this concept is basedon the confidence that humans usually behave differently than the fakes, therefore, detecting this behavior will lead to the revealing of the fake accounts [12]. In this section, we will demonstrate some of the works that have been presented in this area. Reference [13] has reached an accuracy of 84.5% to detect spammers by identifying 23 attributes, most of these attributes (17 attributes) are demonstrated in Table I
which was mentioned in his research. However, in our research, we have reached more accuracy with smaller set of attributes as will bediscussed in Section III.

In [13], the set of attributes has been minimized by identifying ten attributes for detection, the attributes are presented in Table I. However, as mentioned in this research, the result was not promising for identifying fake accounts with more optimistic perspective that it is able to identify fake tweets with higher accuracy by the support of graph techniques.

Although [14] has presented a minimized set of attributes which contained six attributes, however, it is mentioned that it could only detects determined types of spammers, they are bagger, and poster spammers [15]. In our

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 6, June 2020 ||

approach, we propose minimized set of attributes for detecting all types of spammers. In addition, one of these attributes requires text analysis procedure for finding the similarities among messages which is not required for our proposed approach. Moreover, it is mentioned in [15] that Random Forest algorithm [11] is the
best results for detection for Twitter.

III. PROBLEM STATEMENT

Whenever we start any tracking project for a new twitter dataset or an existing application, the first thing we do is a complexity scan of the complete machine details . We go through basic and advanced settings, check the data quality and consistency – we try to find the fake news and assign the credibility i.e. the credible tweet will be assigned as 1 and non-credible tweet is assigned as -1.As the credibility scores increases that account is considered to be non – spam and the decreased credible scored account is considered as spam account. In this we are going to track some fake news emerging from the dataset.

IV. PROPOSEDWORK

Twitter has been used as the source of dataset. It is one of the major platforms used to spread news every day, and a number of fake news have been spreading over the social media. To collect the dataset, we have chosen the news throughout the world. We tried to collect the data by using the streaming API of twitter but unfortunately twitter will provide only the past seven days of data.

1. Raw data**:**
   - Tweet Text
   - Tweet Date/Time
   - Re-tweets.
   - Location of Tweet
   - User Details
   - No of favorites
2. We have divided the scrapped tweets into two csv files:
   - RawTrainingDataSet.csv
   - RawTestDataSet.csv
3. Data preprocessing, feature selection, feature extraction
   has been performed on both training and test data set. The training dataset is used for modelling the classifiers and
   the model is used for predicting the test data.
4. Tweet Annotation : We manually annotated the raw dataset
   into two categories - Spam and non-SpamWe introduced a new column named Class for representing the category(spam and non-spam) for each tweet in the raw dataset.

- If the tweet is spam, the class is assigned value -1.
- If the tweet is non-spam, the class is assigned value 1.

V. FEASIBILITY ANALYSIS

A feasibility study is conducted to select the best system that meets performance requirement. This entails an identification description, an evaluation of candidate system and the selection of best system for the job. The system required performance is defined by a statement of constraints, the identification of specific system objective and a description of outputs.

The key considerations in feasibility analysis are:

1. Economic Feasibility
2. Technical Feasibility
3. Operational Feasibility

5.1Economic Feasibility

It looks at the financial aspect of the project. It determines whether the management has enough resources and budget to invest in the proposed system and the estimated time for the recovery of cost incurred. It also determines whether it is worthwhile to invest the money in the proposed project. Economic feasibility is determined by the means of cost benefit analysis. The proposed system is economically feasible because the cost involved in purchasing the hardware and the software are within approachable. The personal cost like salaries of employees hired are also nominal, because

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 6, June 2020 ||

working in this system need not required a highly qualified professional. The operating-environment costs are marginal. The less time involved also helped in its economic feasibility.

5.2 Technical Feasibility

It is a measure of the practically of specific technical solution and the availability of technical resources and expertise.

- Programming language: Python.
- Libraries: Numpy, Scipy, sklearn , pandas, matplotlib, Beautiful Soup 4,Requests.
- IDE: Pycharm

The above tools are readily available, easy to work with and widely used for developing .

5.3 Operational Feasibility

The system will be used if it is developed well then be resistance for users that undetermined:

- No major training and new skills are required as it is based on SPIRAL model.
- It will help in the time saving and fast processing and dispersal of user request and application.
- New product will provide all the benefits of present system with better performance.
- User involvement in the building of present system is sought to keep in mind the user specific requirement and needs.
- User will have control over their own information. Important information such as pay slip can be generated at the click of a button.
- Faster and systematic processing of user application approval, allocation of IDs payments, etc. used had greater chances of error due to wrong information entered mistake.

VI. SYSTEM DESIGN

A data-flow diagram (DFD) is a graphical representation of the "flow" of data through an information system. DFDs can also be used for the visualization of data processing (structured design).
On a DFD, data items flow from an external data source or an internal data store to an internal data store or an external data sink, via an internal process.

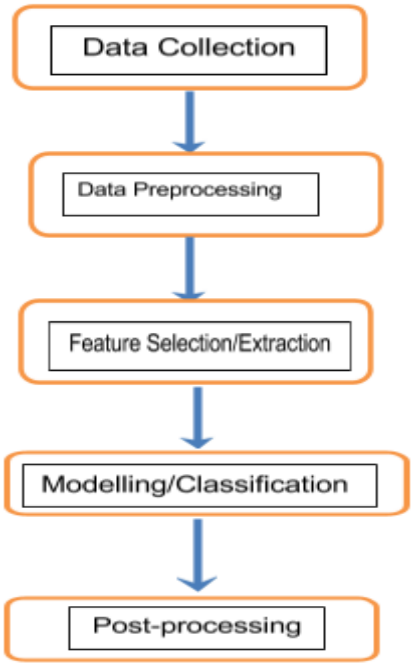
A DFD provides no information about the timing or ordering of processes, or about whether processes will operate in sequence or in parallel. It is therefore quite different from a flowchart, which shows the flow of control through an algorithm, allowing a reader to determine what operations will be performed, in what order, and under what circumstances, but not what kinds of data will be input to and output from the system, nor where the data will come from and go to, nor where the data will be stored (all of which are shown on a DFD).

DFDs help system designers and others during initial analysis stages visualize a current system or one that may be necessary to meet new requirements. Systems analysts prefer working with DFDs, particularly when they require a clear understanding of the boundary between existing systems and postulated systems. DFDs represent the following:

1. External devices sending and receiving data
2. Processes that change that data
3. Data flows themselves
4. Data storage locations

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

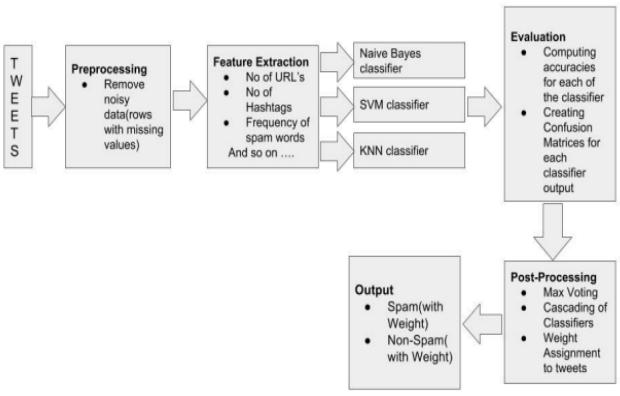
|| Volume 9, Issue 6, June 2020 ||



Dataflow diagram of the project

VII. METHODOLOGY

The implementation of the project is done in python language. Pattern recognition steps followed are – data collection, data preprocessing, feature extraction, feature selection ,modeling, evaluation and then post processing on our dataset and designed the classifier using SVM, KNN and Naive Bayes classification algorithms.As a part of post processing, we have implemented maximum voting, cascading and assigning the credibility weights to each tweets.



Architecture of the project

- Data collection.
- Data Preprocessing.
- Feature Selection\Extraction.
- Classification\Modeling:
  - Support Vector Machine(SVM).
  - Naive Bayes(NB).
  - K-Nearest Neighbour(KNN).
- Post processing:
  - Assigning the weight.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

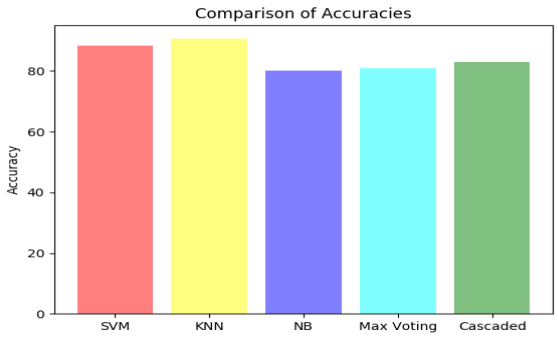
|| Volume 9, Issue 6, June 2020 ||

- Maximum Voting.
- Cascading.

VIII. RESULTS AND ANALYSIS

The outcome of this proposed work is ,the system is trained with the particular data set with the particular categories which include one or more. After training it with the data set the input is given from the trained data set and it will give the result as whether the tweet is fake or not.

- Collect the tweets from the twitter API.
- Train the machine with the tweets collected as dataset from API.
- Test using machine learning algorithm and assign the weights or scores.
- As the credibility score increases the account is considered as fake or not-fake.



IX. CONCLUSION

In this research, we proposed an approach for detecting fake accounts on Twitter social network, the proposed approach was based on determining the effective features for the detection process. The attributes have been collected from different research, they have been filtered by extensive analysis as a first stage, and then the features have been weighted. Different experiments have been conducted to reach the minimum set of attributes with perceiving the best accuracy results..

REFERENCES

[1]F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, 2010, "Detecting spammers on twitter," in Proc. Anti-Abuse and Spam Conference on Collaboration, Electronic messaging, Washington: Redmond

[2]K.Thomas, C.Grier, D.Song, and V.Paxson, "Suspended accounts in retrospect: an analysis twitter spam," in Proc. the 2011 ACM SIGCOMM Conference on Internet Measurement Conference, 2011.

[3]Twitter. How to Report Spam on Twitter. [Online].Available:https://support.twitter.com/articles/64986-how-to-report-spam-on-twitter.

[4]G.Saptarshi, K. Gautam, and G.Niloy, "Spammers' networks within online social networks: A case-study on Twitter,"inProc.World Wide Web Conference 2011, Hyderabad, 2011.

[5]A.Considine. (August 22, 2012). Buying their way to twitterfame.[Online]Available:http://tinyurl.com/n763xe7

[6]C.Malu, D.Umeshwar, H.Meichun, G.Riddhiman, D.Mohamed, L. Yue, and S. Mark, 2011, "LCI: A social channel analysis platform for live customer intelligence,"in Proc. SIGMOD '11 Conf. NY, USA.

[7]Weka 3: Data Mining Software in Java. (2012). [Online].Available:http://www.cs.waikato.ac.nz/ml/weka/

[8]B. Y. E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise ofsocial bots," Commun. ACM, vol. 59 No.7, pp. 96–104, 2016.

[9] S. Cresci, R. Di, M. Petrocchi, A. Spognardi, and M.Tesconi, "Fame for sale : Efficient detection of fake Twitter followers," Decis. Support Syst., vol. 80, 2015,pp. 56–71, October 2015.

[10]Y.Zhang and J. Lu, "Discover millions of fake followers in Weibo,"Soc. Netw. Anal. Min., 2016

[11]YazanBoshmaf et al., "Íntegro: Leveraging Victim Prediction for Robust Fake Account Detection in OSNs," in NDSS '15, 8-11 , SanDiego, CA, USA, February 2015.

[12]VladislavKontsevoi, Naim Lujan, and Adrian Orozco, "Detecting Subversion of Twitter," May 14, 2014.

[13]Fabr´ıcioBenevenuto, Gabriel Magno, Tiago Rodrigues, and Virg´ılio Almeida, "Detecting spammers on twitter," Collaboration, electronic messaging, anti

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 9, Issue 6, June 2020 ||

Abuse and spam conference (CEAS). Vol. 6, 2010.

[14]SuprajaGurajala, Joshua S. White, Brian Hudson, and Jeanna N. Matthews, "Fake Twitter accounts: Profile characteristics obtained using an activity-based pattern detection approach," in SMSociety '15, July 27 - 29, Toronto, ON, Canada, 2015

[15]G. Stringhini, C. Kruegel, and G. Vigna, "Detectingspammers on social networks," in Proceedings of the 26th Annual Computer Security Applications Conference, 2010, pp. 1–9.