Twitter Spammer Detection

Ashwini Bhangare[1], Smita Ghodke[2], Kamini Walunj[3], Utkarsha Yewale[4]

[1234]*(Student, Dept. of Computer Engineering, MIT college of* Engineering*, Maharashtra, India)*

---***---

Abstract*: Twitter is one such popular network where the short message communication (called tweets) has enticed a large number of users. Spammer tweets pose either as advertisements, scams and help perpetrate phishing attacks or the spread of malware through the embedded URLs.*

In this system initiate features which take advantage of the behavioral-entropy, profile characteristics, spam analysis for spammer's detection in tweets. By taking a supervised approach to the problem, but leverage existing hash tags in the Twitter data for building training data.

In this, fetch twitters tweets for a particular hash tag. Each hash tag may have 1000 of comments and new comments are added every minute, in order to handle so many tweets we are using twiter4j API and perform preprocessing by removing quotes, hash symbols and spam analysis through URL, Number of Unique Mentions (NuMn), Unsolicited Mentions (UIMn), Duplicate Domain Names (DuDn) techniques and googlesafebrowsing API.

Key Word *:***Number of Unique Mentions (NuMn), Unsolicited Mentions (UIMn), Duplicate Domain Names (DuDn) techniques and googlesafebrowsing API.**

1. INTRODUCTION

Online social networks (OSNs), such as Twitter, Facebook, and some enterprise social network, have become extremely popular in the last few years. Individuals spend vast amounts of time in OSNs making friends with people who they are familiar with or interested in. Twitter, which was founded in 2006, has become one of the most popular micro blogging service sites. Nowadays, 200 million Twitter users generate over 400 million new tweets per day.

The popularity of Twitter attracts more and more spammers. Spammers drive unnecessary tweets to twitter users to promote websites or services, which are harmful to normal users. In order to stop spammers, researchers have proposed a number of mechanisms. The focus of recent works is on the application of machine learning techniques into Twitter spam detection. However, tweets are retrieved in a streaming way, and Twitter provides the Streaming API for developers and researchers to access public tweets in real time. There lacks a performance evaluation of existing machine learning-based streaming spam detection methods. This system introduce features which exploit the behavioral-entropy, profile characteristics, spam analysis for spammer's detection in tweets. We take a supervised approach to the problem, but leverage existing hashtags in the Twitter data for building training data.

Twitter is one such popular network where the short message communication (called tweets) has enticed a large number of users. Spammer tweets pose either as advertisements, scams and help perpetrate phishing attacks or the spread of malware through the embedded URLs.

Spam is a problem throughout the Internet, and Twitter is not immune. In addition, Twitter spam is much more successful compared to email spam. Various methods have been proposed by researchers to deal with Twitter spam, such as identifying spammers based on tweeting history or social attributes, detecting abnormal behavior, and classifying tweet-embedded URLs.

Fetching of twitters tweets for a particular hashtag. Each hashtag may have 1000 of comments and new comments are added every minute, in order to handle so many tweets we are using twiter4j API and perform preprocessing by removing quotes, hash symbols and spam analysis through URL, Number of Unique Mentions (NuMn), Unsolicited Mentions (UIMn), Duplicate Domain Names (DuDn) techniques and googlesafebrowsing API.

2. RELATED WORK

[1]The popularity of Twitter attracts more and more spammers. Spammers send unwanted tweets to Twitter users to promote websites or services, which are harmful to normal users. In order to stop spammers, researchers have proposed a number of mechanisms. There lacks a performance evaluation of existing machine learning-based streaming spam detection methods. In this paper bridged the gap by carrying out a performance evaluation, which was from three different aspects of data, feature, and model. A big ground-truth of over 600 million public tweets was created by using a commercial URL-based security tool. For real-time spam detection, we further extracted 12 lightweight features for tweet representation. Spam detection was then transformed to a binary classification problem in the feature space and can be solved by conventional machine learning algorithms and evaluated the impact of different factors to the spam detection performance, which included spam to no spam ratio, feature discretization, training data size, data sampling, time-related data, and machine learning algorithms. From the results it come to know that spam tweet detection is still a big challenge and a robust detection technique should take into account the three aspects of data, feature, and model.

[2]Twitter has become a target platform on which spammers spread large amounts of harmful information. These

malicious spamming activities have seriously threatened normal users' personal privacy and information security. In this paper, we propose a novel learning model, Supervised Spammer Detection with Social Interaction (SSDSI), which simultaneously integrates social information with content information for detecting spammers on Twitter. Different from the existing methods, the proposed SSDSI takes the frequency of social interaction between users and their neighbors into consideration, which can reflect the real social phenomenon as well as possible. By using matrix factorization technique to induce the latent features of text content. Meanwhile, in order to improve the effectiveness of learning model, we utilize social network information and the label data to guide the latent features learning process. [3] The combined success of social networking sites and smart phones has changed the way people communicate. It is now possible to publish and track contents in real time at any time and from anywhere. The large number of users on social platforms constitutes an unprecedented opportunity for attack for malicious users. Social engineering techniques, spammers, phishing and malicious attacks are examples of threats that can lead to data loss, data theft, identity theft, etc. The detection of suspicious messages or profiles is mainly covered in the literature as a binary and static classification problem. In this paper, we propose a dynamic behavioral framework for identifying suspicious profiles on social networking sites. This approach is based on three indicators: balance, energy and anomaly, synthesized from daily activities of users.

[4]Twitter, with its rising popularity as a micro blogging website, has inevitably attracted the attention of spammers. Spammers use myriad of techniques to evade security mechanisms and post spam messages, which are either unwelcome advertisements for the victim or lure victims in to clicking malicious URLs embedded in spam tweets. In this paper, we recommend several work of fiction features able to distinguishing spam accounts from legitimate accounts. The features analyze the behavioral and content entropy, bait-techniques, and profile vectors characterizing spammers, which are then fed into supervised learning algorithms to generate models for our tool, CATS. Using this system on two real-world Twitter data sets, we observe a 96% detection rate with about 0.8% false positive rate beating state of the art detection approach. Our analysis reveals detection of more than 90% of spammers with less than five tweets and about half of the spammers detected with only a single tweet. Our feature computation has low latency and resource requirement making fast detection feasible. Additionally, we cluster the unknown spammers to identify and understand the prevalent spam campaigns on Twitter.

This paper [5] gives different 12 lightweight features extracted for tweet representation. Spam detection transformed to a binary classification problem in the feature space and can be solved by conventional machine learning algorithms. The auther evaluated the impact of different factors to the spam detection performance, which included spam to non spam ratio, feature discretization, training data size, data sampling, time-related data, and machine learning algorithms. Which show the streaming spam tweet detection.

In this they [6]have used Matrix factorization. The text matrix are adopted to capture the consistency of users' behavior. social regularization based on users' interaction is introduced to distinguish different types of users. So Supervised Spammer Detection method with Social Interaction is proposed, which jointly learn a classifier by using combine text content, social network information and labeled data. [7]This paper uses more advanced strategies, namely coordinated posting behavior, finitestate machine-based spam template, and passive spam.

3. PROPOSED SYSTEM

The system aims to investigate the utility of linguistic features for detecting the spam twitter accounts and tweets. We take a supervised approach to the problem, but leverage existing hash tags in the Twitter data for building training data.

Integrate the System with Twitter

The system will integrate with twitter and able to read the tweets for particular hash tags. The proposed system divides the flow in four different tasks.i.e. Hash tagged data set, Preprocessing, GoogleSafeBrowsing.

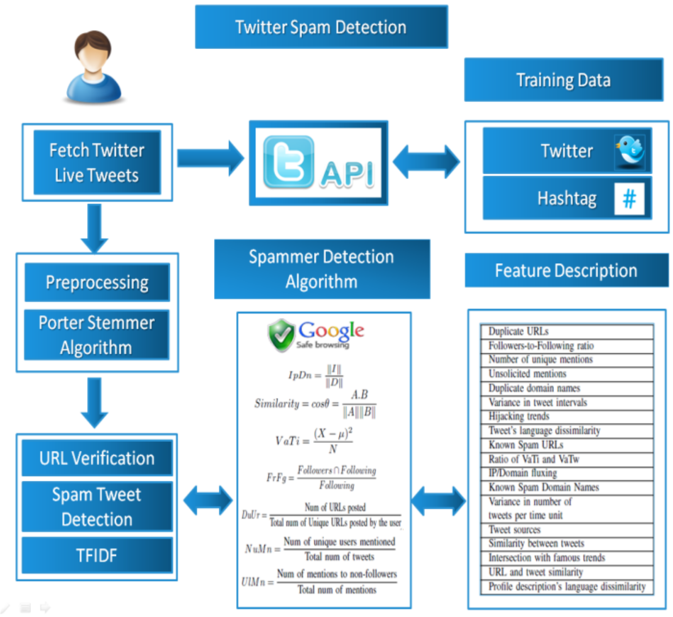

Twitter Spam Detection
Training Data
Fetch Twitter Live Tweets
API
Twitter
Hashtag
#
Preprocessing
Porter Stemmer Algorithm
Spammer Detection Algorithm
Feature Description
Google Safe browsing
$IpDn = \frac{\|I\|}{\|D\|}$
$Similarity = cos\theta = \frac{A.B}{\|A\|\|B\|}$
$VaTi = \frac{(X-\mu)^2}{N}$
$FrFg = \frac{Followers \cap Following}{Following}$
$DuUr = \frac{\text{Num of URLs posted}}{\text{Total num of Unique URLs posted by the user}}$
$NuMn = \frac{\text{Num of unique users mentioned}}{\text{Total num of tweets}}$
$UlMn = \frac{\text{Num of mentions to non-followers}}{\text{Total num of mentions}}$
URL Verification
Spam Tweet Detection
TFIDF
Duplicate URLs
Followers-to-Following ratio
Number of unique mentions
Unsolicited mentions
Duplicate domain names
Variance in tweet intervals
Hijacking trends
Tweet's language dissimilarity
Known Spam URLs
Ratio of VaTi and VaTw
IP/Domain fluxing
Known Spam Domain Names
Variance in number of tweets per time unit
Tweet sources
Similarity between tweets
Intersection with famous trends
URL and tweet similarity
Profile description's language dissimilarity

Fig.1: Architecture diagram for Twitter Spammer Detection

To create the hash tagged data set, first filter out duplicate tweets, non-English tweets, and tweets that do not contain hash tags. From the remaining set (about 4 million), investigate the distribution of hashtags and identify what we hope will be sets of frequent hashtags that are indicative of positive, negative and neutral messages. These hashtags are used to select the tweets that will be used for development and training. Pre-processing is one of the important steps in text mining, Natural Language Processing (NLP) and information retrieval (IR). which gives tokenization, normalization .i.e. remove @,remove #and URL. Data pre-processing is used to extract interesting and non-trivial knowledge from unstructured text data. Information Retrieval is important for deciding which documents in a collection should be retrieved so that we can satisfy a user's need for information.

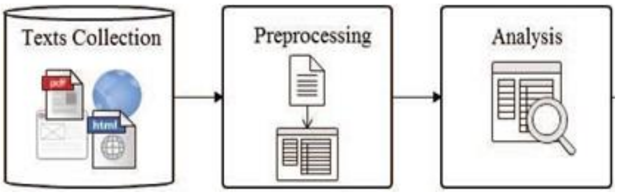


Fig 2: Text Collection

This function facilitates Fetching & Downloading Live Tweets for the particular #hash Tag.

Upload Input Data Set

This function will upload the dataset (tweets downloaded) for a particular #hash Tag.

Pre Processing Techniques

Pre-processing techniques are applied on dataset to get clean data.

Remove @

The first pre-processing technique is remove @ which means it scans the whole document of input dataset and after comparing it with @ it deletes @ from every available comment with @.

Remove URL

The next step of pre-processing is remove URL where the whole input document gets scanned and compared with http:\\... and the comments having URL are deleted.

Remove Stop Words

Further move on to stop word removal being the next step in data pre-processing. Stop word removal exactly means that from the whole statement after scanning it removes the words like and, is, the, etc and only keeps noun and adjective from the statement

GoogleSafeBrowsing:

By verifying URL and sum up all the techniques we get the final conclusion.

a. Calculating overall Unsolicited Mentions

$$\text{NuMn} = \frac{\text{Num of unique users mentioned}}{\text{Total num of tweets}}$$

b. Calculating over all Duplicate Domain Names.

$$\text{DuDn} = \frac{\text{Num of unique domain names in tweets}}{\text{Total num of domain names posted}}$$

c. Calculate overall Variance in Tweet Intervals (VaTi).

$$\text{VaTi} = \frac{(X-\mu)2}{\text{N}}$$

d. Number of Unique Mentions (NuMn)

$$\text{NuMn} = \frac{\text{Num of unique users mentioned}}{\text{Total num of tweets}}$$

4. PROPOSED METHOD

Algorithm Used:

A. Porter stemming algorithm

Porter stemmer' is a method for removing the commoner morphological and in flexional endings from words in English. Following are the steps of this algorithm:-

○ Gets rid of plurals and -ed or -ing suffixes
○ Turns terminal y to i when there is another vowel in the system.
○ Maps double suffixes to single ones: -ization, ational, etc.
○ Deals with suffixes, -full, -ness etc.
○ Takes off -ant, -ence, etc.
Removes a final –e

B. Support Vector Classification Algorithm:

Support vector machine (SVM) proposed by vapnik and cortes have been successfully applied for gender classification problems by many researchers. An SVM classifier is alinear classifier where the separating of the hyper plane is chosen to minimize the expected classification error of the unseen test patterns.

SVM is a strong classifier which can identify two classes. SVM classifies the test image to the class which has the maximum distance to the closest point in the training.

SVM training algorithm built a model that predicts whether the test image falls into this class or another.SVM requires a huge amount of training data to select an effective decision boundary and computational cost is very high even if we restrict ourselves to single pose (frontal) detection. The SVM is a learning algorithm for classification. It tries to find the optimal separating of the hyperplane such that the expected classification error for unseen patterns is minimized.

For linearly non-separable data the input is mapped to high-dimensional feature space where they can be separated by a hyperplane. This projection into high-dimensional feature space is efficiently performed by using kernels. More precisely, given a set of training samples and the corresponding decision values {-1, 1} the SVM aims to find the best separating hyperplane given by the equation WT x+b that maximizes the distance between the two classes.

C.Natural Language Processing(NLP)

Natural Language Processing (NLP) pass on to AI technique of communicating with an intellectual systems using a natural language such as English. Natural Language is mandatory when you want to hear decision from a dialogue based clinical expert system, etc.

Machine learning refers to the ability of computers to learn from data gathered in its previous experience so that can make inferences on its decisions and behavior when it encounters new data. It evolved from the field of pattern recognition and computational learning theory. Whereas programmed systems follow fixed rules for behavior, machine learning systems get a feel for their behavior based on training data and continuously evolving new data it gathers. There are three reasons for studying natural language processing:

You want a computer to communicate with users in their terms; you would rather not force users to learn a new language. This is important for informal users and those users, such as managers and children, who have neither the time nor the inclination to learn new interaction skills.

There is a large storage of information recorded in natural language that could be accessible via computers. Information is continuously generated in many forms such as books, news, business and government reports, and scientific papers, many of which are available online. A system requires proper arrangement of information to process natural language to recover much of the information available on computers.

Problems of AI arise in a very clear and unambiguous form in natural language processing and, thus, it is a good domain in which to experiment with general theories.

The NLP make computers to carry out following tasks with the natural languages humans use. The input and output of an NLP system can be :

- Speech
- Written Text

The main components of NLP as given –

1. Natural Language Understanding (NLU)

Understanding involves the following tasks –

- Map the given input into useful representations.
- Analyzing different aspects of the language.

2. Natural Language Generation (NLG)

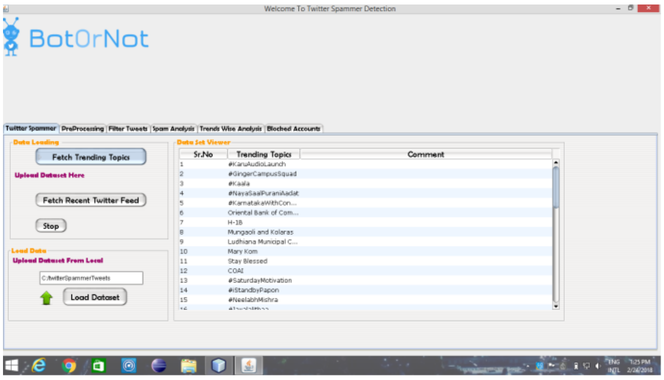
It is the course of creating major phrases and sentences in the form of natural language from some internal depiction.

It involves –

- **Text Scheduling** – It includes retrieving the relevant content from knowledge base.
- **Sentence planning** – It includes choosing required words, forming meaningful phrases, setting tone of the sentence.
- **Text comprehension** – It is mapping sentence plan into sentence structure.

5. IMPLEMENTAION AND RESULTS

This function facilitates Fetching & Downloading Live Tweets for the particular #hash Tag.



Sr.No	Trending Topics	Comment
1	#KaruAudioLaunch	
2	#GingerCampusSquad	
3	#Kaala	
4	#NayaSaalPuraniAadat	
5	#KarnatakaWithCon...	
6	Oriental Bank of Com...	
7	H-1B	
8	Mungaoli and Kolaras	
9	Ludhiana Municipal C...	
10	Mary Kom	
11	Stay Blessed	
12	COAI	
13	#SaturdayMotivation	
14	#iStandbyPapon	
15	#NeelabhMishra	

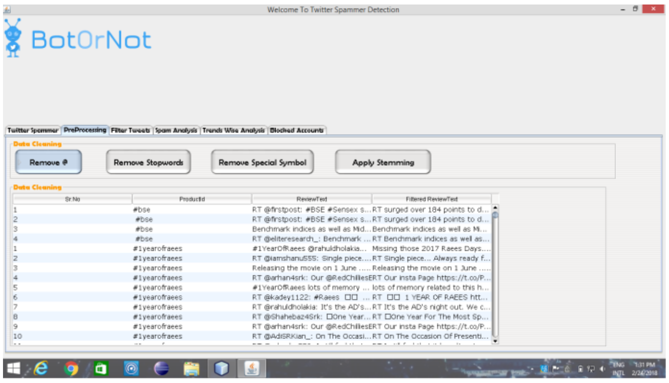
Fig 3: Fetching Trending Topics

Pre Processing Techniques:

Pre-processing techniques are applied on dataset to get clean data.

Remove @:

The first pre-processing technique is remove @ which means it scans the whole document of input dataset and after comparing it with @ it deletes @ from every available comment with @.

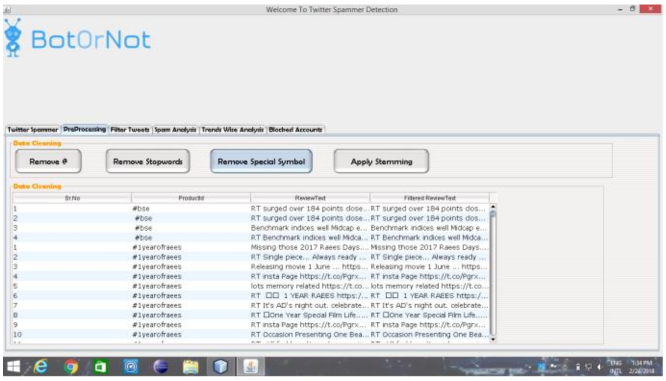


Sr.No	ProductId	ReviewText	Filtered ReviewText
1	#bse	RT @firstpost: #BSE #Sensex s...	RT surged over 184 points to d...
2	#bse	RT @firstpost: #BSE #Sensex s...	RT surged over 184 points to d...
3	#bse	Benchmark indices as well as Mid...	Benchmark indices as well as Mi...
4	#bse	RT @eliteresearch_: Benchmark ...	RT Benchmark indices as well as...
1	#1yearofraees	#1YearOfRaees @rahuldholakia...	Missing those 2017 Raees Days....
2	#1yearofraees	RT @iamshanu555: Single piece....	RT Single piece... Always ready f...
3	#1yearofraees	Releasing the movie on 1 June	Releasing the movie on 1 June .....
4	#1yearofraees	RT @arhan4srk: Our @RedChilliesE	RT Our insta Page https://t.co/P...
5	#1yearofraees	#1YearOfRaees lots of memory ...	lots of memory related to this h...
6	#1yearofraees	RT @kadey1122: #Raees □□ ...	RT □□ 1 YEAR OF RAEES htt...
7	#1yearofraees	RT @rahuldholakia: It's the AD's...	RT It's the AD's night out. We c...
8	#1yearofraees	RT @Shahebaz4Srk: □One Year...	RT □One Year For The Most Sp...
9	#1yearofraees	RT @arhan4srk: Our @RedChilliesE	RT Our insta Page https://t.co/P...
10	#1yearofraees	RT @AdiSRKian_: On The Occasi...	RT On The Occasion Of Presenti...

Fig 4 : Remove @

Remove Special Symbol

The next step of pre-processing is remove special symbol where the whole input document gets scanned and compared with ,",*,?,$ are deleted.

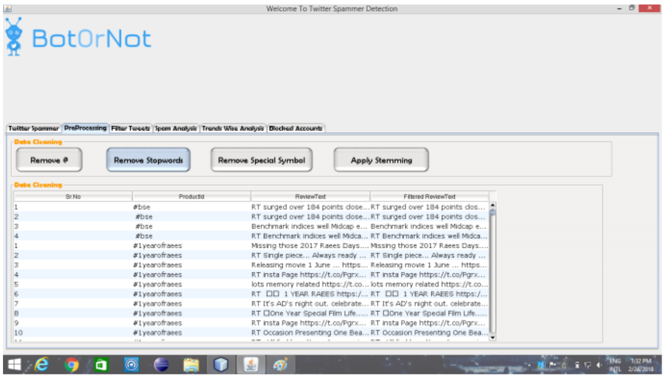


Sr.No	ProductId	ReviewText	Filtered ReviewText
1	#bse	RT surged over 184 points close...	RT surged over 184 points clos...
2	#bse	RT surged over 184 points close...	RT surged over 184 points clos...
3	#bse	Benchmark indices well Midcap e...	Benchmark indices well Midcap e...
4	#bse	RT Benchmark indices well Midca...	RT Benchmark indices well Midca...
1	#1yearofraees	Missing those 2017 Raees Days....	Missing those 2017 Raees Days....
2	#1yearofraees	RT Single piece... Always ready ...	RT Single piece... Always ready ...
3	#1yearofraees	Releasing movie 1 June ... https...	Releasing movie 1 June ... https...
4	#1yearofraees	RT insta Page https://t.co/Pgrx...	RT insta Page https://t.co/Pgrx...
5	#1yearofraees	lots memory related https://t.co...	lots memory related https://t.co...
6	#1yearofraees	RT □□ 1 YEAR RAEES https:/...	RT □□ 1 YEAR RAEES https:/...
7	#1yearofraees	RT It's AD's night out. celebrate...	RT It's AD's night out. celebrate...
8	#1yearofraees	RT □One Year Special Film Life.....	RT □One Year Special Film Life.....
9	#1yearofraees	RT insta Page https://t.co/Pgrx...	RT insta Page https://t.co/Pgrx...
10	#1yearofraees	RT Occasion Presenting One Bea...	RT Occasion Presenting One Bea...

Fig 5. Remove special symbol

Remove Stop Words

Further move on to stop word removal being the next step in data pre-processing. Stop word removal exactly means that from the whole statement after scanning it removes the words like and, is, the, etc and only keeps noun and adjective from the statement



Sr.No	ProductId	ReviewText	Filtered ReviewText
1	#bse	RT surged over 184 points close...	RT surged over 184 points clos...
2	#bse	RT surged over 184 points close...	RT surged over 184 points clos...
3	#bse	Benchmark indices well Midcap e...	Benchmark indices well Midcap e...
4	#bse	RT Benchmark indices well Midca...	RT Benchmark indices well Midca...
1	#1yearofraees	Missing those 2017 Raees Days....	Missing those 2017 Raees Days....
2	#1yearofraees	RT Single piece... Always ready ...	RT Single piece... Always ready ...
3	#1yearofraees	Releasing movie 1 June ... https...	Releasing movie 1 June ... https...
4	#1yearofraees	RT insta Page https://t.co/Pgrx...	RT insta Page https://t.co/Pgrx...
5	#1yearofraees	lots memory related https://t.co...	lots memory related https://t.co...
6	#1yearofraees	RT □□ 1 YEAR RAEES https:/...	RT □□ 1 YEAR RAEES https:/...
7	#1yearofraees	RT It's AD's night out. celebrate...	RT It's AD's night out. celebrate...
8	#1yearofraees	RT □One Year Special Film Life.....	RT □One Year Special Film Life.....
9	#1yearofraees	RT insta Page https://t.co/Pgrx...	RT insta Page https://t.co/Pgrx...
10	#1yearofraees	RT Occasion Presenting One Bea...	RT Occasion Presenting One Bea...

Fig 6: Remove Stop words

Following are the steps of Porter Stemmer Algorithm:-

a. Gets rid of plurals and -ed or -ing suffixes

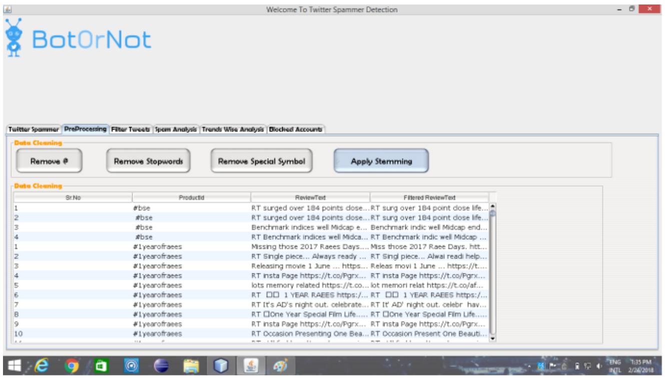
b. Turns terminal y to i when there is another vowel in the stem

Maps double suffixes to single ones: -ization, -ational, etc.

c. Deals with suffixes, -full, -ness etc.

d. Takes off -ant, -ence, etc.

Removes a final –e



Sr.No	ProductId	ReviewText	Filtered ReviewText
1	#bse	RT surged over 184 points close...	RT surg over 184 point close life...
2	#bse	RT surged over 184 points close...	RT surg over 184 point close life...
3	#bse	Benchmark indices well Midcap e...	Benchmark indic well Midcap end...
4	#bse	RT Benchmark indices well Midca...	RT Benchmark indic well Midcap ...
1	#1yearofraees	Missing those 2017 Raees Days....	Miss those 2017 Raee Days. htt...
2	#1yearofraees	RT Single piece... Always ready ...	RT Singl piece... Alwai readi help...
3	#1yearofraees	Releasing movie 1 June ... https...	Releas movi 1 June ... https://t....
4	#1yearofraees	RT insta Page https://t.co/Pgrx...	RT insta Page https://t.co/Pgrx...
5	#1yearofraees	lots memory related https://t.co...	lot memori relat https://t.co/af...
6	#1yearofraees	RT □□ 1 YEAR RAEES https:/...	RT □□ 1 YEAR RAEES https:/...
7	#1yearofraees	RT It's AD's night out. celebrate...	RT It' AD' night out. celebr hav...
8	#1yearofraees	RT □One Year Special Film Life.....	RT □One Year Special Film Life.....
9	#1yearofraees	RT insta Page https://t.co/Pgrx...	RT insta Page https://t.co/Pgrx...
10	#1yearofraees	RT Occasion Presenting One Bea...	RT Occasion Present One Beauti...

Fig 7: Porter Stemming

6. CONCLUSION

In this paper, machine learning approach for spammer analysis and seen that spammer detection has many applications and it is important field to study. Spammer Detection has strong commercial interest because companies or individual want to improve the security on social media.

7. FUTURE SCOPE

Spammer Detection has strong commercial interest because companies or individuals want to improve the security on social media. In future the picture message and location for detecting spammer. Enhancing the detecting model by considering other features and applying network analyzing to improve accuracy in the model

REFERENCES

[1] "CATS: Characterizing Automation of Twitter Spammers",Guofei Gu, Chao Yang. 2013.. In Proc. Of 2013 IEEE Transaction.

[2] "An Evaluation of the Effect of Spam on Twitter Trending Topics",Louis Lei Yu 2013. . In Proc. of SocialCom /EconCom / BioMedCom 2013.

[3] "Design and Evaluation of a Real-Time URL Spam Filtering Service", Kurt Thomas and Chris Grier and Justin Ma. IEEE Security and Privacy, 2011.

[4] “Suspended Accounts In Retrospect: An Analysis of Twitter Spam”,Kurt Thomas, Chris Grier, Vern Paxson, and Dawn Song. ACM Internet MeasurementConference (IMC), 2011.

[5] "A Performance Evaluation of Machine Learning-Based Streaming Spam Tweets Detection",Chao Chen, Jun Zhang, Member, IEEE, Yi Xie, Yang Xiang, Senior Member, IEEE, Wanlei Zhou, Senior Member, IEEE, Mohammad Mehedi Hassan, Abdulhameed AlElaiwi, and Majed Alrubaian,IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS.2016.

[6] "Detecting Spammers on Twitter Based on Content and Social Interaction"Hua SHEN,Xinyue LIU,2015 International Conference on Network and Information Systems for Computers.

[7] Spammers Are Becoming “Smarter” on Twitter",Chao Chen, Jun Zhang, Yang Xiang, and Wanlei Zhou,Jonathan Oliver,IEEE Computer S ociety,March/April 2016.