AN INTELLIGENT SPAM DETECTION USING TEXT MINING

(MONIKA PAWAR, MONAL SONARIKAR, SNEHAL SHARMA)
GUIDED BY - PROF. KIRTI WALKE

(DEPARTMENT OF INFORMATION TECHNOLOGY, SKN-SITS COLLEGE OF ENGINEERING, SAVITRIBAI PHULE PUNE UNIVERSITY, INDIA)

ABSTRACT:

With the growing usages of social media application has main part of our daily life routine. According to oxford dictionary the spam is unsolicited and inapplicable messages send to the internet for large number of peoples. The motivation of spam is to spread the malware and pronunciation. The problem of spam becomes very critical for internet community. Sometimes large amount of web pages & social websites are considered as spam. Many spam techniques were proposed to approach the problem of spamming. The problem is direct affected to the users of the social networking sites such as Gmail, YouTube, Twitter, Facebook, Google+. The social networking sites Facebook and twitter are most familiar to the users. People used twitter in daily life and this reason why spammer most of attack on this social networking sites. In this research with the use of OSN we can classify data into two forms spam and Non spam .We first collect data from OSN application such as twitter by downloading comments. By using different algorithms and different techniques we further check whether the data is spam or non spam.

*We have developed a twitter based spam detection system using twitter comments and URLs. Our system comprises a database as training data set of comments & recognition will be applied on same comments for spam detection. The spam detection shows to drive people off of twitter and onto another social networking site, it will likely violate our spam polic*ies. So our main aim is to develop spam detection system taking "spammer do not take advantage" into mind.

Keywords: *Spam, Non spam, filter keywords, TF-IDF, Cosine similarity, Naïve bayes, Score ranking, Detection*

1. INTRODUCTION

The first measure commercial spam incident started on match 5,1994.Todays world most use services of the social networking sites and emails .In social networking sites google, Facebook, twitter are integral part of their people . The all users of interconnected through social networking sites. Twitter have gained so much popularity as it becomes daily part of life of almost every users to check their profile at least once in a day. Sometimes huge amount of information is easily provided to the user but in many situations or such situations are occurs when all information are not useful to the one user then its said to be “spam”.

In our system we are dealing to detect spam data or spam comment using various algorithms techniques like TF-IDF, Cosine similarity, Naive bayes. With the help of this three algorithms becomes easy to categorize the data into spam and non spam context.

TF-IDF counts how many times each term occurs in each document. Cosine similarity is used in text mining and it gives useful measurance between two documents. Naïve Bayes classifier technique is used to predict the capabilities.

Figure 1: Spam

2. Developed system

In todays world the spam data is increasing tremendously. Hence it is necessary to handle a data in such a way that the people do not face the spam data and will feel safe while browsing the data from social networking site. In our system architecture is developed for the purpose of detection mechanism. In
System architecture users can easily download comment from twitter.

In our system twitter will be the main source of users for downloading comments. Downloaded comments are then parsed by using keyword parser. The function of keyword parser is to analyze string of keywords either in natural language or in computer language for confirmation to the rules
of formal grammer, After parsing the keywords tokenize step is performed. Tokens are used to classify spam and non spam comments using naïve bayes theorem. Using this theorem probability of the comment can be calculated.

We applied various algorithms to understand
and classify the data set. This algorithms are .

1) TF-IDF
2) Cosine similarity
3) Naïve Bayes

We describe TF-IDF:

1) TF-IDF

TF is short form for term frequency and IDF is inverse document frequency. TF-IDF shows how a particular word is important in a collection of document. The proportionality increases as the number of word occur in a given document. Stop-word filtering can be succesfully done by using TF-IDF. Stop-word filtering is used not only in text summarization but also in classification .TF-IDF is the combination of two statistics which is term frequency and inverse document frequency.

2) Cosine similarity

In high dimensional positive spaces cosine similarity are used.Text mining information retrival are some cases in which different dimensions re assigned to each term and documents are characterized by vector. The number of times the term appears in the document will give the value of each dimension correspondingly. Cosine similarity gives useful measure of how similar any two document are likely to be in terms of their subject matter.

3) Naïve Bayes

Since 1950'5 Naïve Bayes classifier provides features of high scalablity It requires number of features linear in the number of predictor in a learning problem.It is also known as Simple Bayes and Independent Bayes.It is a technique used for constructing classifiers and models.It is a family of algorithms which is typically based on a common principle.Naive Bayes is not single algorithm for training classifiers.Naive Bayes classifier are trained efficiently in supervised learning setting. The main advantage of using Naïve Bayes classifier is it requires small amount of training data for estimation of parameters necessary for classification.

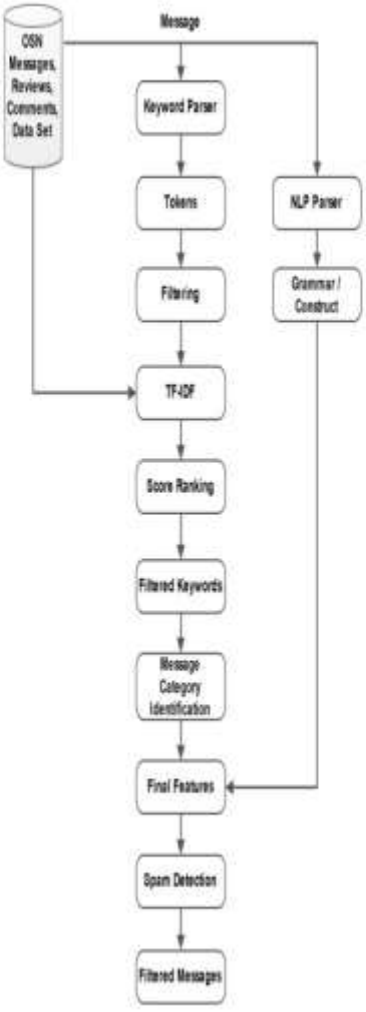

Fig 2. Architecture Diagram

3. FEATURES OF APPLICATION

Fig 3. Homescreen

As shown in figure home screen is having 4 options for user by tapping on appropriate option user will redirected to next screen which is shown in next figure.

Fig 4. Screen after pressing words management button

As shown in figure when user press words management button it will redirect to this tab which is having option of main form.

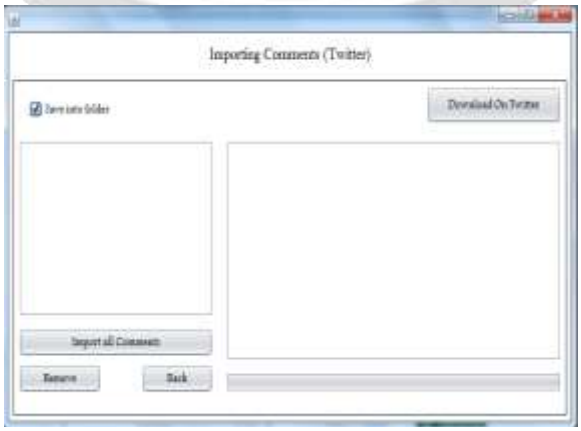

Fig 5. Importing Comments

As shown in above figure when user press on get document management button then comments are imported from twitter.

Fig6. Detect spam or non spam

When user selects the choice detection form shows him actual detection of comment spam or no spam.

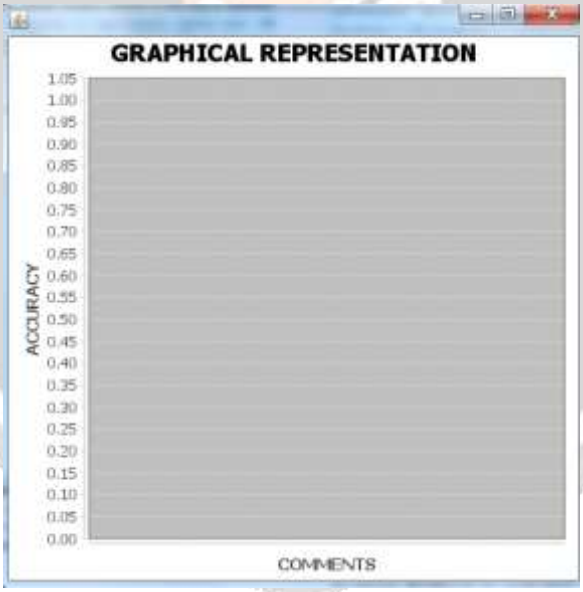

Fig7.Graph

4. ANALYSIS

4.1. OLD SYSTEM: In previous system developer develop a system to identify and characterized set of video attributes that can be used to separate the video spammer from legitimate users .In old system they can predict the spam from data mining model. This data mining model is not specific to classified as a spam or legitimate videos. Clustering approach are used to classify the spam and legitimate videos in old system but naïve bayes and decision tree models are most efficient algorithms for predicting spammers.

4.2. NEW SYSTEM: We develop a new system with some improve feature which reduces the limitations of old system. Mainly the drawback of old system clustering is not better algorithm technique to classify the videos so therefore the new system develop a spam detection system to identify and classify spam data. Naives bayes algorithm are also used in new system for prediction capability. Here TF-IDF and cosine similarity algorithm are also used to detect a spam .Main motivation of new system is users can easily access the social networking sites without any theft..\

5. TEST CASES

We are using unit testing for testing the system we have developed. The objective in unit testing is to isolate a unit and validate its correctness. Automation approach is efficient for achieving the objectives of this testing and it enables the many benefits.
Following test cases were performed on the system developed.

Table1. Testing Activities

Activity	Description	Test Results
Loading Screen Tests	This activity is used to Fetching data from server	Passed
Home Activity Test	This activity is used to Show home Screen	Passed
Words management Test	This activity is used to Filleting of keywords	Passed
Document management Test	This activity is used to import the all document	Passed
Detection form Test	This activity is used to Show actual detection of spam or non spam data	Passed
View graph Test	This activity is used to Show graphical representation.	Passed

6. CONCLUSION:

In our project we conclude that spam detection based on text as well as comments .We will be working on training data set of comments and recognition will be applied on same comments for spam detection. Our own application will be developed. We wont be able to do changes on OSN..

7. FUTURE ENHANCEMENT

Spam data are serious concern so by using this approach we can handle spam data efficiently. They also works on large data and classify this large data into specific format. Major use of this system is to filtered the important data and avoid legitimate data.

8. REFERENCE:

1] H. Aradhye, G. Myers, and J. Herson. "Image analysis for efficient categorization of image-based spam e-mail." In Proc. of the Int'l Conf. on Document Analysis and Recognition (ICDAR), volume 2, pp.914 918, 2005.

2] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri, "Know your neighbors: Web spam detection using the web topology" Int'I ACM SIGIR, pp. 423–430, 2007.

3] L. Gomes, F. Castro, V. Almeida, J. Almeida, R. Almeida, and LBettencourt, "Improving spam detection based on structural similarity", In USENIX Workshop on Steps to Reducing Unwanted Traffic on the Internet (SRUTI), pp.85-91, 2005.