Detection of Sentiment Tweet's by using NLP Technique and Naive Bayes Classifier Algorithm

Shubham R.[1] Sanket B.[2] Shubhangi P.[3]
[1,2,3]Department of Computer Engineering
[1,2,3]Bharati Vidyapeeth's College of Engineering Lavale, Pune-410115, India

Abstract— Millions of users share opinions on diverse aspects of life and politics every day using micro blogging over the internet. Microbloging websites are rich sources of data for belief mining and sentiment analysis. In this dissertation work, we focus on using Twitter for sentiment analysis for extracting opinions about events, products, people and use it for understanding the current trends or state of the art. Twitter allows its users a limit of only 140 characters; this restriction forces the user to be concise as well as expressive at the same moment. This ultimately makes twitter an ocean of sentiments. Twitter also provides developer friendly streaming. We scuttle datasets over 4 million tweets by a custom designed crawler for sentiment analysis purpose. We propose a hybrid naïve bayes classifier by integrating an English lexical dictionary to the existing machine learning naïve bayes classifier algorithm. Hybrid naïve bayes classifies the tweets in positive and negative classes respectively. Experimental results demonstrate the superiority of hybrid naïve bayes on multi-sized datasets consisting of variety of keywords over existing approaches yielding >90 percent accuracy in general and 98.59 percent accuracy in the best case. In our research, we worked with English; however, the proposed technique can be used with any other language, provided that language lexicon dictionary.

Keywords: Naïve Bayes Classifier Algorithm, NLP Technique

I. Introduction

Social Networking Sites are becoming popular day by day. As much as Social Networking Sites ease the life of human beings, it also gives arise to various problems faced by users using these networking sites. There are various networking sites like Twitter, Facebook and LinkedIn etc. The major problem faced by users using these various networking sites is of Spam. Spam is any unwanted or prohibited behaviour that directly or indirectly violates the certain rules of any networking site. This paper mainly focuses on Spam Detection in Twitter using Multi-Layer Perceptron Learning. Till date, Twitter has faced various problems due to spam which is created by various spammers to earn and fill their pocket.

Now-s-days, social networking sites are a medium of analytics on a large amount of user data for many companies based on which many prediction models are being built, and recommended systems are been prepared for end users. But this is only the single side of a coin. On the other side there are many such people who are waiting for just a single chance of a single wrong step by the user. Social networking sites have become a platform for the promotion of businesses, classes and may more activities, but some people are using it in a very wrong way. They provide the users with links in a promotional form and redirect them to many misleading and misguiding sites where, the content is either malicious or in such a way that it will change their thoughts negatively. The main problem which comes as "spam" is here, "spam" is a form of content which is irrelevant or unsolicited message which is sent with a purpose of advertising, phishing, spreading malware etc. There arises a big threat of malware. Many people fall prey to spams and are redirected to sites from where a malware is been downloaded to their personal machines, which is continuously sending their valuable information to an unknown server.

The work in this paper deals with detecting such spam tweets and many new features based on language models that help to improve spam detection. The underlying idea of the paper is as follows: we analyze the use of languages in the tweets and the page of where the URL is linked to it. In the scenario of spam the language models differ from each other.

II. Motivation

With the proliferation of Web to applications such as microbloging, forums and social networks, there came reviews, comments, recommendations, ratings and feedbacks generated by users. The user generated content can be about virtually anything including politicians, products, people, events, etc. With the explosion of user generated content came the need by companies, politicians, service providers, social psychologists, analysts and researchers to mine and analyze the content for different uses. The bulk of this user generated content required the use of automated techniques for mining and analyzing. Cases of the bulk user-generated content that have been studied are blogs and product/movie reviews.

Microbloging has become a very popular communication tool among Internet users. Millions of messages are appearing daily in popular web-sites that provide services for microbloging such as Twitter. Users of these services write about their life, share opinions on variety of topics and discuss current issues. Because of a free format of messages and an easy accessibility of microbloging platforms, Internet users tend to shift from traditional communication tools to microbloging services. As more and more users post about products and services they use and express their political and religious views, microbloging web-sites become valuable sources of peoples opinions and sentiments. Such data can be eficiently used for marketing and social studies.

III. Why NLP

Natural language processing (NLP) is the technology dealing with our most ubiquitous product: human language, as it appears in emails, web pages, tweets, product descriptions, newspaper stories, social media, and scienti_c articles, in thousands of languages and varieties. In the past decade, successful natural language processing applications have become part of our everyday experience, from spelling and gram- mar correction in word processors to machine translation on the web, from email spam detection to automatic question answering, from detecting people's opinions about products or services to extracting appointments from your email.

IV. ARCHITECTURE

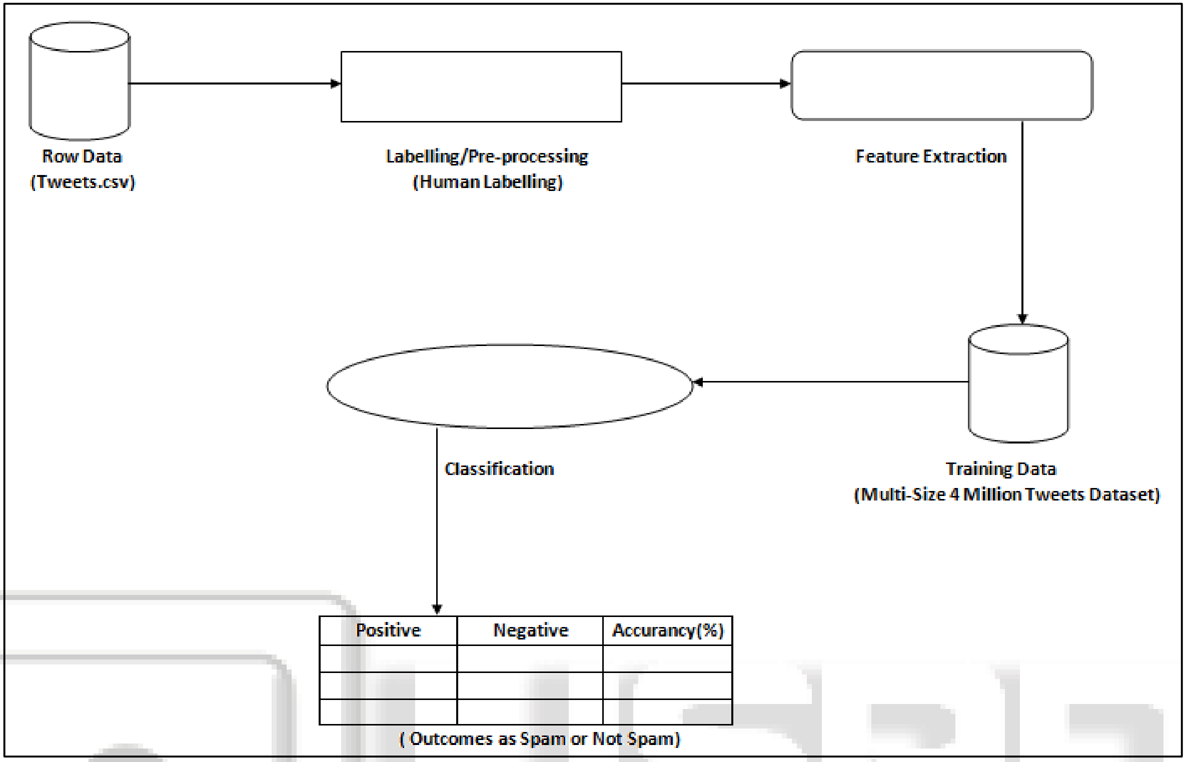

Fig. 1: System Architecture

A. *Data Collection:*

The dataset consists of 5,512 tweets on four different topics (Apple, Google, Microsoft, Twitter). Each tweet was manually labelled by one annotator as either positive, negative, neutral, or irrelevant with respect to the topic. The annotation process resulted in 654 negative, 2,503 neutral, 570 positive and 1,786 irrelevant tweets. The dataset has been used in [3,12,5] for polarity and subjectivity classification of tweets. The Sanders dataset is available at http://www.sananalytics.com/lab

B. *Preprocessing:*

For the purpose of human labelling we made three copies of the tweets so that they can be labelled by four individual sources. This is done so that we can take average opinion of people on the sentiment of the tweet and in this way the noise and inaccuracies in labelling can be minimized.

We labelled the tweets in four classes according to sentiments expressed/observed in the tweets: positive, negative, neutral/objective and ambiguous. We gave the following guidelines to our labellers to help them in the labelling process:

– Positive
– Negative
– Neutral

C. *Feature Extraction:*

The labelled tweets are matched against certain language based features, which will help us to calculate the divergence ratio and decide whether the tweets are spam or not. Each tweet will be represented using natural language processing and content analysis technique.

D. *Training Data:*

To precisely label the text into their respective classes and thus achieve highest possible accuracy, we plan to train the classi_er using pre-labelled twitter data itself. Pre-labelled twitter training data is not available freely, since this year Twitter changed its data privacy policies and it no longer allows open/free sharing of twitter content. However, they mention that using or downloading twitter content for individual research purposes is acceptable. Labelled data Since we do not have direct access to pre-labelled twitter data, we planned to crawl it manually. We crawled various sizes of datasets with various keywords of approximately 4 Million tweets from twitter using a custom python scripted crawler.

The data obtained in such a way is certainly not labelled, so in order to address this issue, we propose to crawl twitter and form two di_erent datasets. The rst one consisting of all the positive sentiment tweets i.e. [\:)", \:-)", \:-D", \:D",\B)"] and the latter one consisting of all the negative sentiment tweets i.e.[\:(",\:-(", \:'(", \X(", \X-("]. Thus, we feed these datasets for the classi_er for training, which function almost similar to hand labelled datasets as used in other sentiment analysis domains**.**

E. Classification:

The classification process can either be performed using the programming method or it can also be done through existing set of tools. We have prepared a setup of Weka (Whitten and Frank 2005) as it contains almost each and every algorithm for machine learning and to carry out data mining tasks. The prepared dataset is evaluated against many of the machine learning algorithms in the Weka tool. The main classifier we followed was the Support vector machine. In previous works this algorithm was used with the default options, it was seen that the algorithm gave the accurate results as compared to other existing classifiers. As the results are not still up-to the mark and the evaluation has to be carried out more precisely modifying the datasets and preparing them according to the requirements.

V. CONCLUSION

Right now we have worked with only the very simplest unigram models; we can improve those models by adding extra information like closeness of the word with a negation word. We could specify a window prior to the word (a window could for example be of 2 or 3 words) under consideration and the effect of negation may be incorporated into the model if it lies within that window. The closer the negation word is to the unigram word whose prior polarity is to be calculated, the more it should affect the polarity. For example if the negation is right next to the word, it may simply reverse the polarity of that word and farther the negation is from the word the more minimized ifs effect should be.

REFERENCES

[1] Sagar Gharge, Manik Chavan "An Integrated approach for sentiment Tweets detection using NLP" 978-1-5090-5297-4/17 2017 IEEE.
[2] Content Based Spam Classification in Twitter using Multilayer Perception Learning 2017 IWEEE.
[3] Detecting Spam Classification on Twitter Using URL Analysis, Natural Language Processing and Machine Learning 2016
[4] Chao Chen, Jun Zhang, Yi Xie, Yang Xiang Wanlei Zhou "A Performance Evaluation of Machine Learning-Based Streaming Spam Tweets Detection" 2329- 924X c 2016 IEEE
[5] Danesh Irani, De Wang, "Click Traffic Analysis of Short URL Spam on Twitter" conference on collaborative computer networking, application and worksharing 2013.
[6] Chao Chen, Jun Zhang, Member, IEEE, "A Performance Evaluation of Machine Learning-Based Streaming Spam Tweets Detection", IEEE Transactions on Computational social systems, vol. 2, no. 3, september 2015.