Sentiment Analysis with Twitter Data

Surbhi Goel, Shiva Gangwar, Vaishali Singh, Anand Kumar Srivastava

[a]Department of Computer Science & Engineering
ABES Engineering College, Ghaziabad, U.P., 201009, India

Abstract:

With over 200 million people worldwide and hundreds of millions of tweets posted daily, Twitter is a popular social network.A large fraction of them are highly personal,reflecting the feelings of the individual. Social networks are the primary tools for collecting People's opinions information and feelings on separate topics,since they spend a lot of time on social media every day and share their thoughts. In this technical paper we demonstrate how sentimental analysis is applied and how to link toTwitter and conduct sentimental analysis queries.Weare conducting experiments on various questions from politics to society, and having interesting results. We found that the favorable feeling for tweets is significantly higher and that the existing works are minimal. Our project will create labels for the user tweets available. Our models are to be trained with Logistic Regression, SVM, Multilayer Perceptron's and Naive Bayes. Once all four models have been tested, we will go into a comparative analysis of these models to see which one is better. We will also review, user tweets and their emotional function to provide observations and connections between the different social, economic and psychological factors of the emotional state of the user, which it has tried to convey through a tweet.

Keywords:

- Twitter sentimentanalysis
- Social Networkanalysis
- Sentiment
- Twitter
- Opinion mining
- Social Media
- NLP

1. Introduction:

Opinion and nostalgicmining are important areas for study because it is a difficult task,because of the large number of daily messages on social networks. Over the last couple of years, about 90%of today's statistics have been made available and it is no easy to know a bout this massive data.By example,nostalgicanalyses have many uses in businesses that provide input on goods and allow companies to gain user feedback and social media feedback. This guide has researched perception and nostalgic processing, and explored both various approaches and fields of study.Facebook sentimental research is also being done,but we are primarily focusing on Twitter sentimental analysis in this article.

Another approach can be for abroad ertext to explain the content, to describe it and to highlight whetheritispositive,negative or neutral.The extractive and abstractive approach is two basic ways of producing adescriptive file.Words and phrases from the original text are removed for descriptive purposes in the extracting process. In an abstractive approach, try to learn an internal representation of the text, then create a description more like the human version. Comprehension of text is a big problem to be solved. There are several machine learning methods used, including several supervised and unattended algorithms. Overview can be developed from different approaches. The meaning of phrases within the document could be listed and a presentation can then be rendered for the text based on the numbers of relevance. One more method is known as generative end-to-end simulation. The end-to-end approach performs better in some areas such as object recognition, speech recognition, language translation and request answers.

In this article, the 100 characteristics which we propose will be thoroughly analyzed. Our studies show that add value, but only barely, with Twitter-specific features (hashtags, emoticons, etc).. The most important feature for both grouping activities is the variation of the previous violation of the Their voice and vocabulary marks. We can therefore see that even in a field that is very different from the one they were educated in (Newswire), traditional natural language processing methods are beneficial.We prove that,although it does not require thorough functional engineering, treek ernel models work approximately as well as the best functional-based templates.

For our experiments we use manually registered Twitter data.One of the advantages of these figures is that the tweets are streamed and are a real tweet sample compared with previously used data sets, their language use and content.

Therearealsoother researchers in our new data set. We also present two resources in this paper:

1) A hand annoted emoticons dictionary mapping emoticons to their polarity and

2) An online acronym dictionary with translations in English of more than 5000 often used acronyms.

2. Proposed ApproachFlow

We propose to analyze Twitter posts for users during this project. We use the Logistic Regression, Vector support system, Naive Bayes and MLP model to perform sentiment analysis and classification of emotions and we eventually do a comparative study with these algorithms. After the tweets are emotionally marked, we might assume the tweets are tweets that are positive or negative ones.

3. Problem Identification &Definition

Over recent years social media have become a forum for the sharing of ideas, opinions, information and communication and people now constantly publish everything. User posts usually containtext, photographs, media, and weblinks. Emotions relate to their everyday life experience that they seek to share in social media. The sentiments conveyed here are typically in text form. It is one of the most commonly used sites in social media. In various literature emotion have been described in various categories such as happiness, anger,sadness, etc. We examine user posts on Twitter and collect feelings embedded into these texts. In this plant. After a detailed analysis of the features of an emotional degree (text) message, we can get a range of insights into the emotional and mental condition of ourusers.

4. Literature Survey:

Agarwal et al. Performed Sentiment analysis of twitter data [1]. The researchers analyze Twitter data emotions. The additions are: (1) POS (point of the scale) – prior polarity unique characteristics are added. (2) Use the architecture of the tree kernel to decode the NLP sentence description. Niko et al. Performed Emotion Recognition on Twitter [2]. The subjects of this work core goal were to investigate deep learning for emotion recognition as part of comparative study and preparation in a Unison environment. They created three large collections of tweets with Classifications of emotions of Ekman, Plutchik and POMS.

Actual neural networks actually surpass the baseline of ordinary bag-of-word models. Their experiments indicate that RNNs should be trained in character sequences rather than words. Apart from more precise performance, no preprocessing or tokenization is also needed for this method. The method is language-sensitive and can quickly be translated in other languages. We use character based solutions.

Wen et *al.* Performed Mood detection with tweets [3]. Author used Naive Bayes and SVM models were used for the sorting of tweets into negative and positive emotion.

Emotion analysis was performed at various granularity levels as a NLPtask. It was a mission at the penalty point (HubandLiu,2004;KimandHovy,2004) beginning with a text status categorization (Turney, 2002; Pang and Lee, 2004) and most recently, word status(Wilsonetal.,2005;Agarwaletal.,2009).

Microblog data like Facebook, where people post "Anything" in real-world, raises a number of objections. Some of Twitter's early and later esultsare fromGoetal.(2009),(Bermingham and Smeaton,2010)andPakand Paroubek (2010). Using remote training to obtain feeling-data by Go et al. (2009). They're using tweets ending with optimistic emoticons like :)":-)." They develop models using Naive Bayes, SVM, and complain that Support Vector Machine is over-performing in other classificationsystems, like:"(")"":-)."They use positive and negative emoticons such as::With regards to the apparea, a Unigram, Bigram layout with POS features will be used.All other models are beaten by the unigramm model. In particular, Bigrams and POS features do not help [13].

Paroubek and Pak(2010) which collects data according to one common model of remote learning. We still, though, have another function of classification: contextual and factual. We compile the tweets for contextual evidence in the same manner as Go et al. (2009) using emoticons. We crawl twitter reports of famous newspapers such as "New York Times" and "Washington Examiner,"forobjectiveinformation.Theyreportbothsupportfor POSandbigrams(contrary toGoetal.(2009)findings).Both the methods are primarily based on ngram models, however. Furthermore, search queries have obtained the data used for training and tests and thus biaised. Through comparison, we have apps that make a major improvement from a unigram baseline. We also explore an alternate way of displaying data and report substantial improvements over uni constructing models. Another benefit of this paper is that, without understanding impairments, we record manually annotated results.In contrast to data, our statistics are a random sample of tweets obtained by specific requests.

They evaluated the best result in the linear kernel model after a study on different parameters. One consequence of this work was that the feature aspect must be minimized attentively as the exclusion of the function stops for the best results.

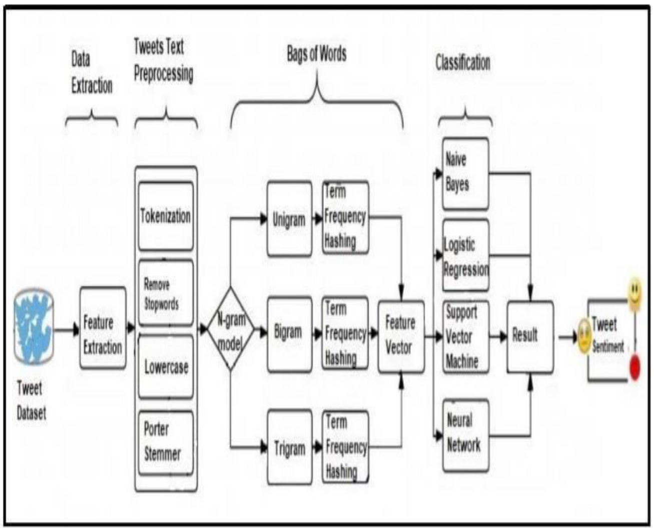


Fig. 1 System Architecture

1. Description of Data:

Twitter is a micro-blog and social service that enables people to post tweets in real world. Tweets are tiny, 140 characters long tweets. Provide users an abbreviation, make orthodox mistakes and use emoticons and others that have different significances because of the nature of this microblog ging service. This will be followed by a brief twitter jargon. Emoticons: the facial expressions expressed by letters and punctuation, and indicate the individual's mood.

Objective:Twitterusersusetheicon"@"onthe microblogtorefertootherpeople.In this way, it immediately warns other consumers. Hashtags: Users normally using hashtags to identify themes. This was done basically to make the tweets more accessible. They receive 11,875 tweets from a private source that have been annotated manually. They have released some of their results. See Recognitions at the end of the paper for details on how to access the results.

By archiving the real-time streaming, they collected the data. During the streaming process, no language, location or other constraints were made. We now have foreign-language tweets. It's compiled. We use Google Translator before the annotation process to translate it into English.Each tweet is defined as negative,positive,junk or neutral by a human annotator. The "junk" mark means that a human annotator cannot comprehend thetweet.A random sample analysis of tweets labelled "junk" revealed that many of these tweets were not correctly translated using Google Translator. For tests we are deleting garbage product tweets. This leaves us with a sample of 8,753 unbalanced tweets.A balanced set of data of 5127 (1709 tweets of negative, positive and neutral categories) is used by stratified samplers.

Twitter Sentimental Analysis

Social networks are a valuable forum to hear the opinions and emotions of individuals on various subjects, since they openly link up and exchange views on social media , including Twitter and Facebook.Various opinion-based information gathering mechanisms seek to collect people's opinions on different topics. The sentiment-aware devices today have many corporate social applications.Since social site networks , particularly Twitter, contain tiny texts, which make it difficult to extract numerous terms and abbreviations from Natural Language Processing systems available, several researchers have been using detailed learning and modeling techniques to extract and reduce the text's polarity[14]. Some most important abbreviations are fb on Facebook, b4 on previous, GMO on oh my god, etc. Therefore, symbolic work is difficult for short texts such as Twitter messages.

Social Media NetworkAnalysis

Analysis of the Social Media Network is an experiement of human interaction and communication on various subjects that has been given greater attention today.Millions of people express their opinions about different topics on social media, such as Facebook and Twitter, daily. It has so many applications, from social science to industry, in various fields of research. Today, Twitter is the most popular social platforms and currently has more than 290 millionfollowers. Twitter is a healthy source to know about choices of the people and emotional research. It is important for each tweet to decide whether the tweet is negative, positive or neutral. One more problemof twitter, is only 140 characters, i.e. Limiting the use of word and functions not in language processing for individual tweets. The text limits of Twitter have recently been extended to 280 features per tweet.

2. Features of our proposed method:

They are suggesting a setoff features for the tests. There are a minimum of 50 types of applications. We measure such characteristics for the entire message and the tweet's last one-third. We have 100 further features in total. These attributes are recognised in the text as sensational characteristics.

The characteristics can be classified into three different categories: First of all, the counts of the different functions are mostly and hence the natural number N is the function value. Secondly, apps with B boolean values. Thirdly, programs with the real R value. Mainly these are applications which record the DAL score.

We are the bag of words, an expletive(exclamation) trademark and a subsidize(capitalized) sentence. Each of them is split into two sub-categories: Polar and Nonpolar. We call a polar function when we measure its previous polarity in either DAL or an emoticonic dictionary.The non-polar band includes all of the other characteristics which are not associated with any previous polarity.Each feature of the Polar and Non-polar is further divided into two main categories: POS and the Other.POS referring to programs that gather word data from speech pieces and other software.

Proposed Approach:

Data Pre-processing

Until implementing data pre-processing, we combine training and testing data set to render a Combi data set. Instead, we remove the twitter handles from the dataset. Upon deleting the handles, we delete the special characters and short words below 3. After extracting special characters, we recover our pre-processed files. We apply to stem to tokenized results. We use the term bag to pick the items. Upon extracting the stemming terms, we mark the processing data based on the apps. They number as follows: negative words labeled as 1 and regular or positive words labeled as 0.

Logistic Regression

Logistic regression is a statistical approach that isused when the dependent variable is binary (dichotomous). Regression algorithms are known as Predictive Analysis so, the logistic regression (Fig. 2) is also a Predictive Analysis. The Accuracy of Logistic Regression on our dataset is95.04.

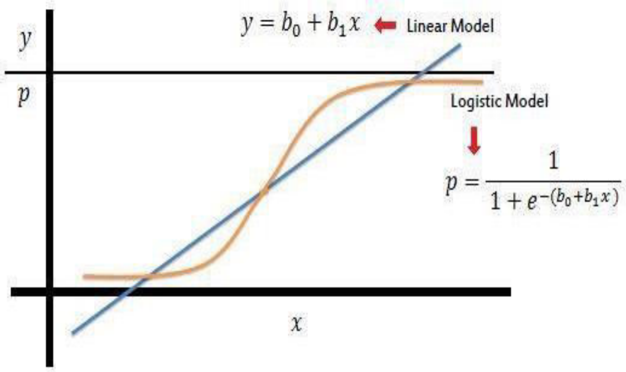
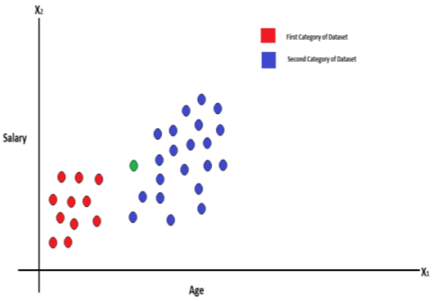


Fig. 2 Logistic Model

Naïve Bayes

The Naive Bayes classifier is an algorithm(Supervised Machine Learning) that uses Bayes theorem.We use variables that are independent from one another to generate.The classification itself has proven to be effective(Fig. 3). Accuracy of Naive Bayes on our dataset is 93.88.

Fig. 3 Naïve Bayes Diagram



SVM(Support Vector machine)

A Support Vector Machine is a Supervised learning algorithm. For classification as well as for regression applications SVM can be used. Most of the time SVM is used for classification purposes only (Fig. 4). This algorithm is based on Hyper-plane which good for the division. The Accuracy of our dataset is 95.22.

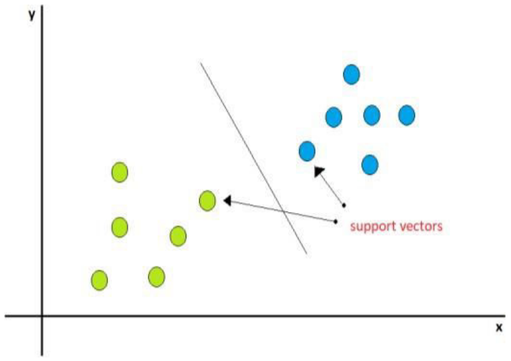


Fig. 4 Support Vector Machine Diagram

Neural Network

We use the Multilayer Perceptron (MLP) model in our neural network. A logistic regression classifier for an MLP can be used in which the data is first transformed with the aid of a trained non-linear transformation. Accuracy of Naïve Bayes on our Dataset is 94.83.

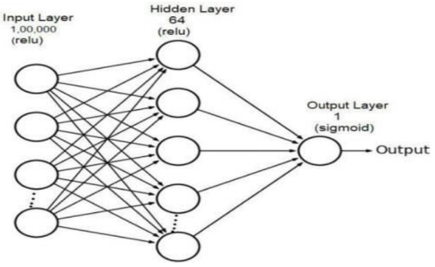


Fig. 5 Neural Network (MLP)

3.Result:

The efficiency of applied models is shown in figure 6. In applying models, SVM has shown
highest efficiency.

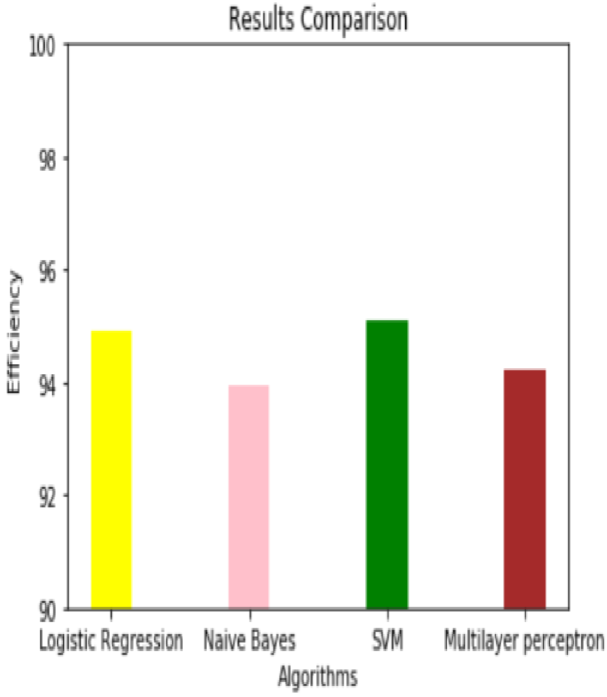


Fig. 6

Dataset training is comprised of 31,962 tweets. It has three columns named id, label, and tweets. We've labeled the tweets in two categories, positive and negative tweets. Support Vector Machine algorithm has shown the highest efficiency means it has correctly labeled the 95.22 percent tweets to their classes.

4,Conclusion:

Twitter sentiment analysis is developed to examine the views of consumers towards the market places critical to success. Our approach suggests that machine based learning approach is more efficient and accurate to analyze a feeling and will be used in conjunction with natural language processing techniques.

References:

[1] Agarwal B. Xie, I. Vovsha, O. Rambow, R. Passonneau, "Twitter Data Sentiment Ana lysis," ACL 2011 Workshop on Social Media Languages, 2011, pp.30–38

[2] Apoorv Agarwal, Fadi Biadsy, and Kathleen Mckeown, 2009. Contextualphrase-level
polarity analysis using lexical affect scoring and syntactic n-grams. Proceedings of the 12th Conference of the ecu Chapter of the ACL (EACL 2014), pages 24–32, March.

[3] Alec Go, Richa Bhayani, and Lei Huang. 2015. Twitter sentiment

classification using distant supervision. Technical report, Stanford.

[4] Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC.

[5]T. Wilson, J. Wiebe, and P. Hoffman. 2015. Recognizing contextual polarity in phrase

level sentiment analysis.ACL.

[6] Niko Colneric and Janez Demsar "Emotion Recognition on Twitter: Comparative Study and Training a Unison Model",IEEE transactions on affective computing.

IEEE,2807817,pages 45-55.

[7]Santdex"[Online][Cited:(Nov-April),(2018-19).]

https://www.youtube.com/user/sentdex

[8] Saedsayad"[Online][Cited:April,(2019).]

https://www.saedsayad.com/logistic_regression.

[9]Codershood"[Online][Cited:April,(2019).] https://www.codershood.info/2019/01/14/naive-bayes-classifier-usingpythonwith-

example

[10] Kdnuggets" [Online] [Cited: April, (2019).] https://www.kdnuggets.com/2016/07//support-vector-machinessimpleexplanation.

[11] Toward science[Online] [Cited: April,(2019).]

https://towardsdatascience.com/activation-functions-neuralnetworks1cbd9f8d91d6

[12] Natural Language Toolkit Project". Natural Language Toolkit. Natural Language

Toolkit. [Online] [Cited: Nov, (2018).] http://www.nltk.org/

[13] Wen Zhang,Geng Zaho and Chenye(Charlie) Zhu "Mood Detection with Tweets",

CS229:Machine Learning fall2014.

[14] Analytics Vidhya" [Online][Cited:Dec(2018).]

https://datahack.analyticsvidhya.com/contest/practice-problem-twittersentiment-

analysis/