International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 10 | Oct 2020 www.irjet.net p-ISSN: 2395-0072

Twitter Sentiment Analysis using Machine Learning: A Review

Amar Kumar [1], Dr. Avinash Sharma[2]

[1]M.Tech Scholar, Department of CSE, MGI, Bhopal (M.P), India
[2]Head & Associate Professor, Department of CSE, MGI, Bhopal (M.P), India

---

ABSTRACT: The basic knowledge required to do sentiment analysis of Twitter is discussed in this review paper. Sentiment Analysis can be viewed as field of text mining, natural language processing. Thus, we can study sentiment analysis in various aspects. This paper presents levels of sentiment analysis, approaches to do sentiment analysis, methodologies for doing it, and features to be extracted from text and the applications. Twitter is a microblogging service to which if sentiment analysis done one has to follow explicit path. Thus, this paper puts overview about tweets extraction, their pre-processing and their sentiment analysis.

Keywords: *Sentiment Analysis, StopWord, Tokens, Features, Training & Testing Data, Model or classifier, Naïve Bayes, SVM, SKlearn*

I. INTRODUCTION

According to recent data from the social media tracking company Technorati, four out of every five internet users use social media in some form. This includes Friendship Network, Blogging and Micro-Blogging Site, Content and Video Sharing Site etc. It is worth noting that after the World Wide Web (here only referred to as the Web), it has now completely changed into a more interactive and co-creative web. This allows a large number of users to contribute in different forms. The fact is that even those who are almost novices of web publishing techniques, they are making content on the web. Indeed, the value of a website is now determined largely by its user base, which, in turn, determines the amount of data available on it. It might be right to say that the data is new Intel.

One such interesting form of user contributions on the web is review. Many sites on the web allow users to write their own experiences or opinions about a product or service as a review. The web is now full of user-interviews for various items for mobile phones, leisure trips and hotel services to movie reviews etc. It is interesting to see that these reviews not only express opinions of a group of users but are also a valuable source.

To exploit collective intelligence [1] For example, a user looking for a hotel in a particular tourist city might like to go through reviews of hotels available in the city before deciding to book one of them. Or users wishing to buy a special model of a digital camera can first see the reviews posted by many other users about that camera before making a purchase decision. It not only helps the user to get more and relevant information about different products and services on a mouse click, but also helps to reach a more informed decision. Sometimes users like to write their own experiences about a product or service as a blog post rather than a clear review. However, in both cases the data is literally literal.

Popular sites like Carwale.com, imdb.com are now full of user reviews, in this case respectively, reviews of cars and movies. And users who write on these sites are a diverse group, those who have recently purchased a product or have used a service for those who are regular on them. A glance at the Internet Movie Database website (www.imdb.com) will surely show that this can be useful when someone is interested in movies, which is produced and released in any part of the world. Similarly, posting on blog sites shows a large number of users' opinions. Although a blog post is a relatively difficult source for emotion analysis, because it often does not include explicit statements which can be exploited for emotion. Many times, they contain more factual or relevant information and they are not considered to the desired extent. However, they are still a tremendous source of user feedback and ideas and should be exploited for emotion-oriented and other useful analysis.

Predictive (To guess) analytics enclose a number of mechanisms from stats, data mining and game theory that find out current and historical facts to make guesses about future events. The variety of techniques is sometimes divided in three ways: predictive models, descriptive models and decision models.

Predictive models explain for sure relationships and some patterns that usually edge to a certain behaviour, point to fraud, predict system failures, and so many. By explaining the explanatory variables, we can find out or predict results in the dependent variables.

Descriptive models explain for creating partition or segment; generally, it is used to classify (find out) customers based on for instance (behaviour of customers in different locations) socio-demographic characteristics, life cycle, profits, required product and many more. Where predictive models focus on a specific (individual) event or behaviour, descriptive models identify as many different (general) relationships as possible.

1.1 Machine Learning Techniques used in Sentiment Analysis

Since we know that in recent days Machine Learning algorithm play important roles in different industries. In this section we worked for sentiment analysis or finding polarity from movies reviews dataset. We know that we have number of algorithms to solve our problem out of them we are explain some algorithms:

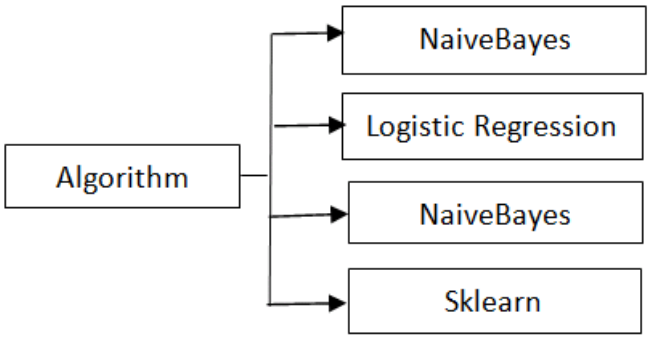

Figure 1: Used Algorithms in Sentiment Analysis

1.2 Overview of Sentiment Analysis

Sentiment analysis is a mechanism or process to compute the sentence whether it is +ve, -ve and neutral. In other ways we can say that it is opinion mining, by this mechanism we can detect the attitude of any speaker. Question arises in your mind where sentiment analysis will play important roles. We explaining some major domains where sentiment analysis is using.

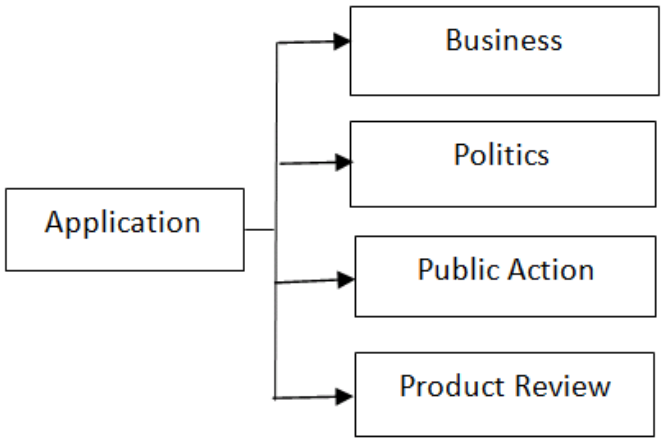

Figure 2: Area where Sentiment Analysis works

Note: Sentiment analysis also used to monitor social activity or phenomena, in recent days every product is launched finally by taking product review.

1.3 Required Framework for Sentiment Analysis

For working with sentiment analysis, we have to download nltk module. In this module we have to download all the required folder. We have to follow following steps to install all required modules.

II. RELATED WORK

Many different methods are used for processing of text in emotional analysis. The purpose is to build verbal chains, learn machine and have a lot more useful approach. Other statistical approaches, domain knowledge-driven analysis can be done. Such approaches proved very beneficial in the work of emotion analysis. Work has been completed by researchers in many different languages like Thai, Nepali, Bengali, Malayalam etc., but very little work has been done in Hindi language. The results provided by the processing of emotion analysis are also very time-saving and accurate. The very first work was done in Hindi, Marathi and Bengali. But at this time the level of work in Hindi is not very appreciable. Therefore, the requirement of the same as the result of various surveys has been felt.

In this paper the authors used the lexicon method for classification so that the proposed algorithm could be compared with the UGGram presence method. Positive and negative words are counted again to select one [2].

In this paper the author, the better news spirit analysis method. In this, the news sentiment is analysed by cutting out the title and text separately and through intensive study of Chinese news, and two different algorithms are applied in both areas. A neutral news assessment method has been proposed for the title part and a subjective sentence recognition algorithm is used for news lessons. In the end, a different weight is used to calculate the final sentiment value for the news and writing spirit. [3]

Here the authors proposed a strategy, which claims to be a fall policy for Hindi language. Their strategies are followed in three ways: In-language analysis, machine translation, resource-based statement analysis. He developed a Hindi Sentinbernet (HSWN) by changing the words of English wordnet by his Hindi counterparts. Finally, 78.14% of them have been accurately [5]

In this paper the authors proposed a system in which it has been assured that this is the first thing that has been done to detect the emotion in Nepali texts and the results show that the approach of learning machine demonstrates better than the rule-based approach and accuracy Makes a lot of impact on System efficiency They developed Nepali Sentiment Corpus and Nepali Centigrade Net. First of all, an approach is seen where a lexical resource developed by Bhavankos is developed. Secondly, an approach to training the text classifier based on machine learning is worried by them [6].

III. TWITTER

3.1 Definition

The word 'micro' in microblogging specifies the limitation of content of the opinion expressed on it. A twitter user can compose at max 140 characters per each tweet. A tweet is not only a simple text message but it is a combination of text data and Meta data associated with the tweet. These attributes are the features of tweets. They express the content of the tweet or what is that tweet about. The Metadata can be utilized to find out the domain of the tweet. The Metadata of tweet are some entities and places. These entities include user mentions, hashtags, URLs, and media Users, Twitter userID. RT stands for retweet,'@' followed by a user identifier report the user, and'#' followed by a word characterizes a hashtag. Work on the Twitter in this paper is limited up to text data.

3.2 Twitter Features

For Opinion Retrieval following features can be useful:

1) Twitter Specific Features

a) URL

Many tweets share a link along with the introduction to the links. The sharing link is initiated as URL. Presence of URL, gives its feature value as 1, else is 0.

b) Mention

In a tweet when user want to refer to another user he can write his name starting with @ symbol. It is called as Mention and it also represented as "@username". If tweet encloses mention the binary feature representing it will have value 1, else is 0.

c) Recency

When the query is fired to get a tweet, it is better to get most recent tweet about that matter. Thus Recency feature measures the age of tweet in seconds after its generation.

d) Hashtag

It is a word starting with # symbol. It refers to a word about the content of text or indicating the topic of tweet. The binary feature value gives the answer of whether the tweet contains hashtag or not.

e) Emoticons

These are facial expressions pictorially characterized using punctuation and letters; they express the user's mood.

f) Retweet

A tweet can be just a statement made by a user, or could be a reply to another tweet. Retweets are marked with either "RT" followed by '@user id' or "via @user id". Retweet is considered the feature that has made Twitter a new medium of information dissemination as well as direct communication.

g) Singleton

If a tweet has no reply or a retweet,

IV. SENTIMENT ANALYSIS IN TWITTER

Sentiment analysis is all about extracting opinion from the text. There are various aspects, reasons, orientation of extracting these emotions according to reason behind the analysis. Event detection, location detection etc. tasks can be done on tweets. When this task is accomplished on twitter data, the framework or architecture to do sentiment analysis varies according to what type of result one wants to achieve from the tweets. One more important factor behind the varying nature of flow of twitter sentiment analysis is use of different methodology-gies and techniques. Many times, researchers derive their own framework or flow to do sentiment analysis to increase efficiency of the result. Some of common steps in twitter sentiment analysis and the keywords in it are defined below:

4.1 Pre-processing

Despite of these generalized orientation of framework of twitter sentiment analysis, we can frame up this topic into the following workflow. Thus, the generalized steps involved in this framework are as follows:

Before starting sentiment analysis, the data pre-processing need to be done.

4.1.1 Removal of Non-English Tweets

When the tweets are extracted from big datasets like TREC or Clue web dataset, it contains English as well as non-English tweets. Therefore, we have to run language identification on each tweet, and have to delete from our collection all tweets that are assigned a 0-probability of being English.

4.2 Feature Selection

4.2.1 Lexicon Features

Based on the subjectivity of the word we can classify the words into positive, negative and neutral lexicons. We have to compare each word with predefined wordnet libraries.

4.2.2 Part-of-speech Features

Parts-of speech features i.e. nouns, adverbs, adjectives, etc. in each tweet are tagged.

4.2.3 Micro-blogging Features

By creating binary features, we can detect the presence of positive, negative, and neutral emotions. By the presence of abbreviations and intensifiers we can classify tweets in positive, negative and neutral. Online available slang dictionaries can be used for emotions and ab-abbreviations [11].

4.2.4 Steps to Extract Features

4.2.4.1 Case Normalization

In this step entire document is converted into lowercase.

4.2.4.2 Tokenization

Tokenization is splitting up the systems of text into personal terms or tokens. This procedure can take many

4.1.2 Removal of Re-tweets

We have to delete any text that followed an RT token (as well as the RT token itself), since such text typically corresponds to quoted (retweeted) material.

4.1.3 Conversion to ASCII

Many tweets contain unusual or non-standard characters, which can be problematic for down-stream pro- cessing. To address these issues, we have to use a com- bination of BeautifulSoup5 and Unidecode6 to convert and transliterate all tweets to ASCII.

4.1.4 Removal of Empty Tweets

After completing all of the other pre-processing, we have to delete any empty tweets.

4.1.5 Restoration of Abbreviations

We can restore popular abbreviations used in the tweets, to their corresponding original forms using a lexicon of abbreviations (e.g. “wknd" to “week- end"). Punctuations are kept since people often express sentiment with tokens such as “:)”, “:-)". These emotions can also be used for sentiment classification [10].

Types, according to the terminology being examined. For English, effective tokenization technique is to use white space and punctuation as token delimiters.

4.2.4.3 Stemming (Snowball)

Stemming is the procedure of decreasing relevant to- kens into a single type of token. This procedure contains the recognition and elimination of suffixes, prefixes, and unsuitable pluralization.

4.2.4.4 Generate n-Grams

Character n-grams are ‘n’ nearby figures from a given feedback sequence. For example, a 3- gram of a phrase ’FORM’ would be ‘_ _ F’, ’_FO’, ‘FOR’, ‘ORM’, ‘RM_’,

‘M_ _’. N-grams of dimension one are known as ‘uni- gram’, two dimensional grams are known as ‘bigram’, three-dimensional grams are known as ‘trigram’. And for the rest dimensions it is called as n-grams [11].

V. CONCLUSION

We studied number of papers and find that the basic knowledge required to do sentiment analysis of Twitter is well stated in this review paper. What is Sentiment Analysis with respect to levels of sentiment analysis, what are the approaches to do sentiment analysis, methodologies for sentiment analysis, features to be extracted from text and the applications where it can be utilized is mentioned hierarchically. If we want to do Twitter’s sentiment analysis we need to know about the twitter, about extracting the tweets, its structure, their meaning. This paper gives brief notion of tweets. When one wants to do sentiment analysis of tweets, he has to do it in a specialized aspect of sentiment analysis. So the brief knowledge about Twitter Sentiment Analysis is given in this paper. Different methods and techniques are discussed in a comparative manner. The accuracy/ result of each method enables us to imagine the efficiency of applied technique in respective circumstances.

REFERENCES

[1] Bing Liu, Sentiment analysis and opinion mining, Synthesis Lectures on Hu- man Language Technologies 5 (2012), no. 1, 1-167.

[2]Shamanth Kumar, Huan Liu, Fred Morstatter, , “Twitter Data Analytics”, Mor- gan & Claypool Publishers, Springer, 2014.

[3]Peter Korenek, Marián Simko, “Sentiment Analysis on Microblog Utilizing Appraisal Theory”, Published at Springer pp. 667–702, 2013.

[4]Matthew A. Russell, “Mining the Social Web, Second Edition”, 2nd Edition of O’Reilly Media, May 2013.

[5]Dudhat Ankitkumar, Prof. R. R. Badre, Prof. Mayura Kinikar, “A Survey on Sentiment Analysis and Opinion Mining”, International Journal of Innova- tive Research in Computer and Communication Engineering Vol. 2, Issue 11, November 2014.

[6]Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal , 1093–1113, 2014.

[7] Erik Cambria, Amir Hussain, "Sentic Computing Techniques, Tools, and Appli- cations", Published at Springer May 9, 2012.

[8] Rupali P. Jondhale,Manisha P. Mali"Study on Distinct Approaches for Senti- ment Analysis", International Journal of Computer Applications (0975 – 8887) Volume 111 – No 17, February 2015.

[9]Zhunchen Luo, Miles Osborne, Ting Wang, "An effective approach to tweets opinion retrieval", Published at Springer, World Wide Web DOI, 2013.

[10] Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, Bing Liu, "Combining Lexicon-based and Learning-based Methods for Twitter Senti- ment Analysis", HP Laboratories HPL-2011-89.

[11] Efthymios Kouloumpis, TheresaWilson, Johanna Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!", Proceedings of the Fifth Interna- tional AAAI Conference on Weblogs and Social Media.

[12] Dhiraj Gurkhe, Niraj Pal, Rishit Bhatia, "Effective Sentiment Analysis of Social Media Datasets using Naive Bayesian Classification", International Journal Computer Applications (0975 8887), Volume 99 - No. 13, August 2014.