International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 05 | May-2018 www.irjet.net p-ISSN: 2395-0072

Sentiment Analysis of a Topic on Twitter using Tweepy

Dhanush M[1], Ijaz Nizami S[2], Abhijit Patra[3], Pranoy Biswas[4], Gangadhar Immadi[5]

[1]*Student, Department of Information Science & Engineering, New Horizon College of Engineering*
[2,3,4] *Students, Department of Information Science & Engineering, New Horizon College of Engineering*
[5] *Professor, Department of Information Science & Engineering, New Horizon College of Engineering*

Abstract - *In this day and age, assessments and audits are vital for impacting the achievement of a brand, item or administration. A basic request along this line is to foresee the fame of an inclining theme, which can empower an extensive variety of utilizations. While Twitter information is to a great degree enlightening, it shows a testing assignment as a result of its voluminous and scattered nature. This paper is an exhaustive push to plunge into the novel space of performing assumption investigation of individuals' feelings. The reason for this work is to assess a point and furthermore compute what number of individuals have a positive or a negative view, in view of the content audits utilizing estimation examination taking surveys from Twitter.*

***KeyWords*: Popularity Prediction, Sentiment Analysis, Natural Language Processing, Lexical Analysis, Doc2Vec, Tweepy, Textblob**

1. INTRODUCTION

Advertising on social media is one of the most important strategies a company or an organisation opts to promote its product. It has always been so important for these organisations to know how their products are doing or how people are reacting to it. The idea we had in our mind when we started was to be able to predict the polarity of these products or a personality so we can tell if the strategy that they had opted for was successful. Twitter happens to be a great platform for people to go and post about anything that is how they feel or what they think about a certain product or personality. It's indeed a great tool to check what people think about a topic. Hence Sentiment Analysis was the tool fit for this job. It has emotions, attitudes or assessment which considers us that is as we think as humans. Now this is not necessarily that simple always. Contents can be written in different contexts many times.[1] Sarcasm, mockery or ironies are cases of difficult expressions to identify. Doc2Vec proves to be a promising method for that. Determining the polarity of such sentences is still a challenge and yet to be solved. This method can also be used in studying states and customer data on lower scales or reviews o online responses. While an unpopular tweet may not gain much attention, a popular but negative tweet gains more attention similarly a positive popular tweet has a more diverse result.[2]

Twitter proves to be a great source of data for analysing because:

- The API is clean and comes with rich developer tools.
- The data is rich in information and has a data format fit for analysis.
- Twitter data is accessible to anyone with fair usage rights.[3]

The current systems process a large amount of tweets to crush down and give the results. Use of large datasets has become crucial for sentiment analysis especially when we are using supervised-learning based approaches. Users not necessarily express their feelings by tweeting they might retweet or hashtag things that they think is interesting. We have conducted our experiments based on the latest real time tweets, and we have got promising results.[2]

2. LITERATURE SURVEY

Sentiment Analysis can be used to observe the attitude of any Statement made by people responding or reacting to it, Today Sentiment Analysis has reached to the level where it can determine not only the positivity or negativity of a statement but also deal with the different topics and behaviors of these statements.[4] There has been a lot of work recently in the genre of "Sentiment analysis "by multiple researchers. actually the Evolution of this field started by early 2000's. In the beginning it was only able to perform binary classifications, that is assigning positive or negative tags to the comments. There are researches on sentiment analysis that are based on opinions of the users version of summarization system of the product.[1] There has been a lot of effort been put in this field where programmers have applied soft programming approaches, That is usually fuzzylogic and neural networks for sentiment analysis. There are algorithms construct fuzzy domain sentiment ontology tree based on the reviews that includes the extraction of sentimental words or sentences, distinct features of the products and relation amongst features thus precisely predicting the polarity of the reviews in the networking site. By designing membership functions for the process they formulated and standardized the elite process of evaluating the strength of reviewer's opinions in the presence of an adverbial modifier on the social networks. We have taken the data from Twitter using the Tweepy API for analysing the respective data by the use of sentiment analysis. The data has to be cleaned before any preprocessing technique can be applied to the same. The

topic can be mentioned in different contexts like a tweet or a retweet or a hashtag for example. In some other related studies tweets have been carried out to evaluate the preferences of users over a certain class of products. Political Sentiment analysis can be carried out before elections to know which party has the highest chances of getting the majority.

3. METHODOLOGIES

3.1 NLP

Do you understand human emotions? Me neither but turns out machines are getting at this than humans. It's called "Natural language processing". The meaning of this this technique is pretty much its name. It tries to understand a language similar to how we do as humans. With sentiment analysis computers no just understand what we say but also what we meant when we were writing that statement.

We all know that the we talk as humans is very different compared to how we talk to computers.

NLP comes with a flexible model. Adding or removing layers is easy. Better flexibility means suit well to the developing models for understanding complex linguistic structures.

3.2 Doc2Vec

Doc2Vec is a more mature version of Word2Vec. Such Word embeddings are methods that are used for extracting features out of text. So we can input those features in a machine learning model to work with text data. Since a large part of our project deals with text data. We opt for this approach to add feature tags to our data. Word embedding methods also try to save syntactical information for better judgment. Each of the words have a vector along with them. We initialize all the word vectors and take a window size and iterate through ou document.Doc2Vec is a word embedding model made by google. Companies across the world have applied this technique to get info about their product in real time. Once they understand how the customer feels after analysing the reviews and opinions they can make their product more approachable and suitable for normal use. Also build things like recommendation systems or more targeted marketing campaigns for them. Sentiment analysis is still an evolving field in machine learning their have been so many advances and new trends in our day to day speaking and texting styles that the machines just need to keep learning.

3.3 Naive Bayes

Naive Bayes Classifier is a machine learning based algorithm for text classification. We have a data set in this process. Whenever a new value arrives it is compared with the earlier values that it had and calculate the probability and puts it to the class with the highest probability. The Value ranges for classification is decided by the system itself. It is known for its simplicity and effectiveness. Can be used to make fast models and make quick predictions. This algorithm learns the probability of an object to fall into a certain class having certain features. It assumes that the occurrence of a certain feature is independent of the occurrence of other features. The algorithm is based on the Bayes' theorem by thomas bayes that is base for the naive bayes algorithm. The main aim of the naive bayes algorithm is to calculate the conditional probability to classify an object for a class with certain features.

3.4 Logistic Regression

Logistic Regression is a better and more mature way of classifying objects that covers all the negative parts of Linear Regression. The problem with linear regression is it might give the probability result greater than 1 or a negative result that is not possible. So the Logistic regression approach uses a more mature formula for the same.

The conditions that are to be satisfied are

1.It must be always positive(Since p>= 0)

$P = \exp(B_0 + B_1 * x) = e^{(B0 + B1 x)}$

2.Must be less than 1(Since p<=1)

$P = (\exp(B_0 + B_1 * x) = e^{(B0 + B1 x)}) / ((\exp(B_0 + B_1 * x) = e^{(B0 + B1 x)})+1)$ -(1)

Even though the probability expression is not a linear, Simple algebra converts the expression into a linear function.

If we call the LHS as y* and rewrite the same then we can have something very similar to a linear expression.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

Now If we put this y_i value in the equation 1(1)

We get

$P = (\exp(y_i))/(\exp(y_i)+1)$

Since any exponential expression will always give a positive result and dividing a number by a greater number is always gives a number less than 1 so Logistic regression proves to be a better solution for classifying and determining the conditional probability for classification.

4. PROPOSED SYSTEM

Initially, the Twitter API is connected through a python library called tweepy. Tweepy library crawls the required data from the twitter and stores locally. Before storing

locally, the data is divided based on its positivity and negativity by using a python library called Textblob.

The positive data is stored separately and negative data separately.

4.1 Database design:

The database is divided into 2 kinds:

1.Training Data

- Training-positive
- Training-negative

2.Testing Data

- Testing-positive
- Testing-negative

Training and Testing Dataset:

Training-positive and Training-negative consists of 1 million positive and 1 million negative tweets respectively. The data crawled from twitter is stored in Testing-positive and Testing-negative separately.

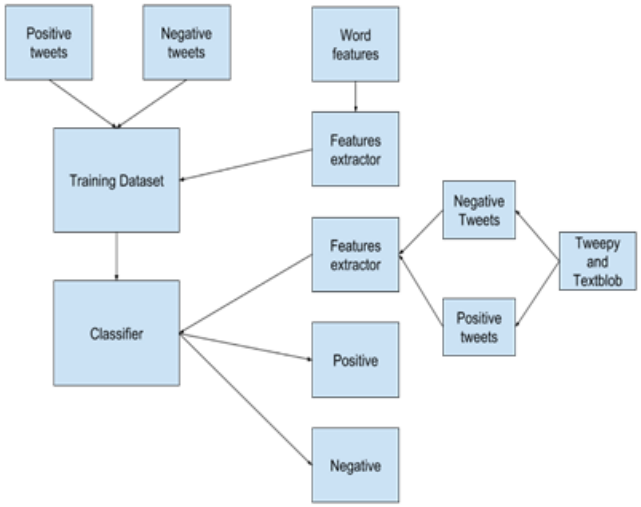

Fig 1. Block Diagram of Proposed System

4.2 Cleaning the data:

Pre-processing is one of the important steps in text mining. We first filter out duplicate tweets, non-English tweets, and tweets that contain and do not contain hashtags. 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.

4.3 Tokenization:

The procedure of encroachment a stream of content into words, images, phrases, or other important components called tokens. The rundown of tokens moves toward becoming contribution for the further handling, for example, parsing or message mining. It parts sentences into words. Literary is just a square of characters toward the start. All procedures all together recuperation require the expressions of the informational index. Consequently, the necessity for a parser is a tokenization of reports. This may sound slight as the content is as of now put away in machine-comprehensible configurations.

4.4 Normalisation

To complete preparing on characteristic words composition, it is basic to perform Normalisation that for the most part includes taking out the accentuation, changing over the whole content into lowercase or capitalized, changing over numbers into words, extending condensings, canonicalization of content, expels prevent words from input content information. Stop words are the word that is consequently excluded from a PC produced concordance or record.

4.5 Apply NLP

One of the real difficulties in regular dialect preparing is instructing PCs to comprehend the way people learn and utilize dialect. Google web crawler construct their machine interpretation innovation in light of NLP profound learning models. This model enables calculations to peruse the content on a website page, assess its importance and make an interpretation of it into another dialect. NLP calculations are commonly in view of machine learning calculations. as a substitute of manual coding vast arrangements of principles, NLP can depend on machine figuring out how to naturally take in these tenets by dissecting an arrangement of illustrations (i.e. a huge corpus, similar to a book, down to a gathering of sentences), and making a statically surmising.

4.6 Sentiment Analysis

Sentiment examination is another essential utilize case for NLP. Utilizing sentiment examination, information researchers can survey remarks via web-based networking media to perceive how their business' image is performing, for instance, or audit notes from client benefit groups to distinguish territories where individuals need the business to perform better.

5. CONCLUSION

Here we present a system for predicting the positive and negative image of any topic or product on twitter using Sentiment Analysis based on text reviews which are getting from Twitter. The advantages of using this system are that it helps in analyzing customer satisfaction of the

product or the opinions of twitter users on some individual or a trending topic and helps to rate prediction based on Twitter tweets.

6 REFERENCES

[1]Geetika Gautam, Divakar yadav - Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis. Department of Computer Science & Engg. Jaypee Institute of Information technology, Noida, India

[2]Yusuf Arslan, Aysenur Birturk, Bekjan Djumabaev, Dilek Kuc ¸uk - ∗Department of Computer Engineering, Middle East Technical University, Ankara, Turkey - Real-Time Lexicon-Based Sentiment Analysis Experiments On Twitter with A Mild (More Information, Less Data)Approach

[3]Wiesław Wolny University of Economics in Katowice, ul. 1 Maja 50, 40-287 Katowice, Poland - Knowledge Gained from Twitter Data

[4]Adyan Marendra Ramadhani, Hong Soon Goo, Department of Management Information Systems - Twitter Sentiment Analysis using Deep Learning Methods.

[5]Tejaswi Kadam, Gaurav Saraf, Vikas Dewadkar, P.J Chate ,"TV Show Popularity Prediction using Sentiment Analysis in Social Network", International Research Journal of Engineering and Technology (IRJET),pp. 1087-1089, November 2017