IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Sentiment Analysis and Opinion mining in Twitter Using Logistic Regression with R

Dr. C. Sunitha[1], S. Samyuktha[2], B.Shankaranarayanan[3], S.Aishwarrya [4]
Head of the Dept[1], V MSC SS [2,3,4],
Dept. of Computer Science,
Sri Krishna Arts and Science College,
Coimbatore, India.

**Abstract**--- Over the past few years, there has been a huge growth in microblogging platforms such as Twitter. More tweets are being tweeted per day. There is a large amount of data and it is getting extremely difficult to get the relevant information from the data. There has been a vast number of research on how sentiments are conveyed in a genre such as online reviews about a movie, news articles, and blogs, but how sentiments are conveyed in case of unofficial languages and message-length constraints of microblogging are less analysed. In that case sentiment analysis comes into the picture. The collection of natural language processing that tries to identify and extract the opinion from the review of the respective blogs is called sentiment analysis. It has been proven helpful in many other domains, but will it also prove useful in Twitter?

## I.INTRODUCTION:

Google and Siri have answers to all our questions. But there are a few things it can't answer what the current emotion of a particular human is. But with advancements in sentiment analysis in machine learning our machine is closer to answering these questions. Human emotions and preferences are practically unpredictable and we invented the scientist of psychology and sociology to help us study these things. Both are the scientific study of people, emotion, relationships, and behaviour.

Generally, psychologists will formulate a hypothesis and they would find a subset of people to test it. I'd like to do the same thing in this paper with the help of the best psychological tool out there, Twitter. People around the world have a vast number of reactions and opinions on every topic under the world every second and every day. It is like one big psychological database and it is constantly being updated. We can analyse a large number of texts in seconds with the power of ML.

Microblogging platforms are employed to precise by different people to precise their opinions. Twitter contains a huge number of text posts and it's updated every day. Twitter's audience varies from regular users to company representatives, politicians, celebrities, and even interest groups. Twitter's audience is represented by users of many countries. We separate the tweets as positive, negative, and neutral. While reviews are categorized by formal text patterns and are summarized thoughts of authors, tweets are more casual and restricted to 140 characters of text.

Microblogging nowadays became the main area of communication. Recent research has identified it as online word-of-mouth branding (Jansen et al., 2009). The massive amount of knowledge in microblogging websites makes them a beautiful source of data for opinion mining and sentiment analysis.

## II.SENTIMENT ANALYSIS WORKING PRINCIPLE

Firstly, we might split the input text (i.e. Tweets) into several words or sentences. This process is named tokenization because we are creating small tokens from the big text. We will just calculate the number of times each word shows up once the text is tokenized. This is also called a collection or bag of words. Then we will search the sentiment value for every word from the sentiment lexicon that is all pre-recorded.

Sentiment analyses are useful for analysing the result of the election, customers' opinions on a product, and getting the opinion from movie reviews. This information collected from sentiment analysis is beneficial for companies in creating future decisions. Many traditional approaches in sentiment analysis used a bag of word method. Other techniques such as Naïve Bayes, Maximum Entropy, and Support Vector Machines.

Sentiment analysis may be a broad area of NLP (Natural Language Processing) which deals with the study of opinions, sentiments, and emotions expressed in text. Sentiment Analysis otherwise referred to as Opinion Mining aims at learning people's attitudes, opinions, and emotions towards an entity.

Different approaches that embrace machine learning (ML) techniques, sentiment lexicons, hybrid approaches, etc. are proved helpful for sentiment analysis on formal texts. But their effectiveness for extracting sentiment in microblogging data will need to be explored.

Sentiment analysis within the domain of micro-blogging may be a new research topic so there's still tons of room for further research in this area. A decent quantity of connected previous work has been done on sentiment analysis of user reviews, documents, blogs/articles, and general phrase-level sentiment analysis. These differ from Twitter mainly due to the limit of a hundred and forty characters per tweet which forces the user to precise opinion compressed into a very short text. The results reached in sentiment classification using supervised learning techniques like Naive Bayes and Support Vector Machines, but the manual labelling required for the supervised approach is extremely expensive. Some work has been done on unsupervised and semi-supervised approaches, and there's tons of room for improvement. Various researchers testing new features and classification techniques typically simply compare their results to baseline performance. There's a requirement for proper and formal comparisons between these results arrived through different features and classification techniques to pick the simplest features and most effective classification techniques for particular applications.

We apply the subsequent Machine Learning algorithms for this second classification to reach the simplest result:

- K-Means Clustering
- Support Vector Machine
- Logistic Regression
- KNN
- Naive Bayes
- Rule-Based Classifiers

Sentiment Analysis helps to prioritize customer requests so that it becomes more efficient and better customer service. For instance, when a customer complains about a few services, an algorithm can identify that and prioritize these messages, so that the sales agent can answer that first. This will increase the satisfaction of the customer and reduce the churn rate.

A Sentiment classifier enables businesses to gradually evaluate social media posts and merchandise reviews in real-time. Twitter blogs became the primary port of involving everyone when it involves exchanging information about brands, trends, and products or expressing their own opinion.

Available market prediction sentiment analyses the sentiment of social media, blogs, and news feed toward stocks or brands. It's used as a further feature alongside price data to make better forecasting models.

### II.A.FEATURE MODELLING

An important step in the development of Sentiment Classifier is language modelling. The aim is to bring the text into a structured format that will be statistically assessed within the training process. The two common models are bag-of-words and n-grams.

### II.B.SENTIMENT ANALYSIS USING LOGISTIC REGRESSION:

Our goal is to make a sentiment classifier, which should be ready to classify new text sequences into one among the three sentiment classes i.e., positive, negative, or neutral.

### II.C.PREPROCESSING A TWEET

•Eliminate handles and URLs
•Tokenize the string into words.
• Remove stop words like "and, is, a, on, etc."
• Stemming - convert every word to its stem. Sort of a dancer, dancing, dancing, becomes 'dance'. Porter's stemmer method is used to require care of this.
• Convert all of your words to small letters.

That is done by creating a function that will take tweets and their labels as input, undergo every tweet, pre-process them, count the occurrence of each word within the data set and make a frequency dictionary.
The squeeze function is important or the list finishes up with one element.

### III.A.SIGMOID FUNCTION

Logistic regression makes use of the sigmoid function which outputs a probability between zero and one. The sigmoid function with some weight parameter θ and a few input x^{(i)}x(i) is defined as follows:-

$h(x^{(i)}, \theta) = 1/(1 + e^{(-\theta^T * x^{(i)})})$.

The sigmoid function gives values between -1 and 1 hence we will classify the predictions counting on a specific cut-off. (say : 0.5)

Note that as $(\theta^T)x(i)$ gets closer and closer to $-\infty$ the denominator of the sigmoid function gets larger and bigger and as a result, the sigmoid gets closer to 0. On the opposite hand, $(\theta^T)x(i)$ gets closer and closer to $\infty$ the denominator of the sigmoid function and gets closer to 1 and as a result the sigmoid also gets closer to 1.

### III.B.COST FUNCTION AND GRADIENT DESCENT

The logistic regression cost function is outlined as
$J(\theta)=(-1/m)*\sum i=1$ to $m[y(i)\log(h(x(i),\theta)+(1-y(i))\log(1-h(x(i),\theta))]$

The aim is to scale back cost by improving the theta using the subsequent equation:
$\theta j:=\theta j-\alpha*\partial J(\theta)/\theta j$

Here, α is named the training rate. The above process of creating hypothesis (h) using the sigmoid function and changing the weights (θ) using the derivative of cost function and a selected learning rate is named the Gradient Descent Algorithm.
Initialize the parameter θ, which will use in the sigmoid, then compute the gradient that will update θ, then calculate the value.
Create a function that will extract features from a tweet using the 'freqs' dictionary and the above-defined pre-processing function (process tweet).
Import the information set from nltk and break it into a training set and test set. As all the specified functions are ready we will finally train our model using the training data set and test it on the test data set J is that the final cost and "theta" are the ultimate weights after training the model.

To see it before testing on the test data set two more functions which, given a tweet, will predict the result using the 'freqs' dictionary and theta. The second function will use the predict function and supply the accuracy of the model on the given testing data set.

## III.C.COMPLICATIONS

Is sentiment analysis that simple? Well, not quite. The cases described thus far were chosen to be very simple. However, human language is extremely complex, and lots of peculiarities make it harder in practice to spot the sentiment during a sentence or paragraph. Here are some examples:

- Inversions: "this product isn't so great"
- Typos: "I love this product!"
- Comparisons: "Product a is best than product z".
- Expression of pros and cons during a text passage: "An advantage is that. But on the opposite hand…"
- Unknown vocabulary: "This product is simply whoopie!"
- Missing words: "How are you able to not this product?"

## IV.EXPERIMENTAL RESULTS

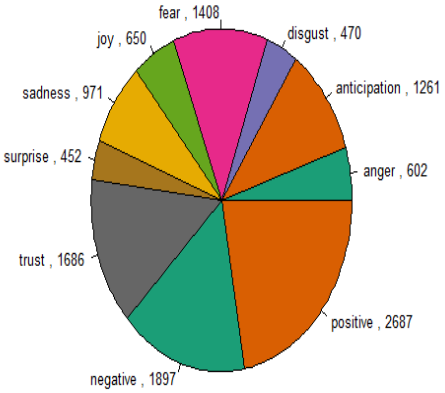
**TF-IDF and Sentiments Result**: Sentiment classifier gives the accuracy of sentiments towards particular topics. Proposed system uses LDA for topic level search which makes the BLR to select more accurate features for sentiment polarity detection. Below is the tabular representation of TF-IDF, positive score, negative score, strongly positive score, strongly negative score and neutral score.

Table 4.1: Experimental Results

| Emotion | Score |
|---|---|
| Anger | 602 |
| Anticipation | 1261 |
| Disgust | 470 |
| Fear | 1408 |
| Joy | 650 |
| Sadness | 971 |
| Surprise | 452 |
| Trust | 1686 |
| **Negative Score** | **1897** |
| **Positive Score** | **2687** |

The below graph explains Sentiment result accuracy in percentage. For searched topics total number of tweets collected, Positive tweets, negative tweets, neutral tweets are calculated.

Figure 1: Emotions and Sentiment nrc Scores




## V.CONCLUSION

The applied methodology shows a basic way of classifying tweets into positive or negative categories using logistic regression as a baseline. It also tried to show how language models are related to logistic regression and can produce better results. It would further develop in using this sentiment analysis using voice recognition and facial expressions.

## VI.REFERENCES

[1]https://www.relataly.com/simple-sentiment-analysis-using-naive-bayes-and-logistic-regression/2007/

[2]https://atharva-mashalkar.medium.com/sentiment-analysis-using-logistic-regression-and-naive-bayes-16b806eb4c4b

[3]https://rcciit.org/students_projects/projects/it/2018/GR33.pdf

[4]https://lexitron.nectec.or.th/public/LREC-2010_Malta/pdf/385_Paper.pd