Deep About Tweet - A Sentiment Visualizer

Yash Pandya
Department of Information & Technology,
L. J Institute of Engineering & Technology, Ahmedabad, India
yashp6149@gmail.com

***Abstract* :-** Twitter gets new user every second, We can find users attitude towards various topics and visualize it based on attitude/emotions. We can find links in tweets that user uses frequently and also type of source used several times. We also find words,hashtags and from different hashtags we can find geolocation that most of tweets with same hashtags comes from, common topics and tags associated with tweet and cluster them using Data Mining concept in word cloud, we can make a chart of tweets vs time which helps to understand the users attitude with respect to time, we will make an API for our system so that it can be used in various system or applications if needed.

*****

I. Introduction

There are millions of tweets done every day. It becomes necessary to analyze those tweets and determine the category to which they belong! So we made a portal that helps us analyze this content. We call it "Deep about Tweet". As the name itself suggests it gives us a complete information about a type of tweets (along with #hashtags).It uses pie charts to display the statistical data. A simple decent way to organize your tweets according to your mood. Use of basic as well as advanced technologies to make it complete.

II. Abstract

Twitter is an online microblogging and social- networking platform. Over 250 million tweets per day, we hope to achieve a reflection of public sentiment by analyzing the emotion expressed in the tweets, it is Important for many applications such as firms trying to find out the response of their products in the market, predicting political elections and predicting phenomena like a stock exchange.

The aim of this project is to develop a functional classifier for accurate and automatic emotion based classification of an unknown tweet stream. A similar project is made by a company which has around 65% accuracy. For us, at first it was based on a complete sentence which limited the accuracy, but by converting sentences to words, we helped accuracy escalated to 78%.

III. Limitation of an Existing System

There are several systems that can work on that but it classifies tweets based on their polarity thus we can identify it is positive, negative or neutral with low accuracy. There is no mechanism used to collect frequently used hashtags. IN many sentiment analysis systems there is no functionality for clustering most used topics/tags and filter it based on users.

IV. Objective

Our system we help finding sentiment based on emotion categorized by sad, happy, joy, anger, anticipation & more. We can find frequent topics that are most used on twitter & their location. We can find common tags that are most tweeted. We can find people who frequently tweet about same tag(clustering the tweets based on tags).We can find Links used in tweets & also the type of device used to tweet. Time Chart of a user is also obtained. Word Cloud can be prepared.

V. Feasibility

The feasibility refers to the chances of a doable project.

- Technical Feasibility: With the help of advanced technologies like python and Machine learning as useful frameworks like Django it becomes possible to implement such a concept.
- Legal Feasibility: As the free Twitter API is available, it is feasible to make its use.
- Economic Feasibility: Well this portal would ultimately help firms for the analyses of their projects, as nowadays most of the complaints and reviews are raised by tweets.
- Scheduling Feasibility: It is necessary to understand the timeline of the project. Whether if not completed in time, would it fail?
- Operational Feasibility: Analyses if the project plan satisfies the requirements mentioned in requirement analysis.

VI. Dataset

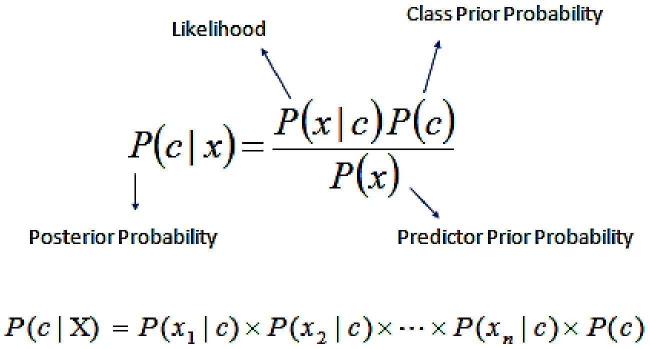
Our data was gathered from kaggle, an archive of tweets which provides information including the tweet's content, the number of retweets, and the author's demographic information. We filtered the set of English tweets based on the hashtags and words We then collected the resulting tweets through categories of engagement, potential reach, and recentness. Each category provided 250 tweets, resulting in

750 tweets per sample. We repeated this processes five - seven times between
11/12/2017 and 12/10/2017, collecting a total of 3,750 tweets. We randomly distributed these tweets into a 70-30 split for our training and test data. The final result was a training set of 2,625 tweets and a test set of 1,125 tweets. In order to run the data into our Naïve Bayes and algorithms, we had to first process the data as a matrix of word occurrences. We created a matrix corresponding to our tweet example, with dimensions of the number of tweets we were processing and the number of tokens in our vocabulary. Each entry in the matrix corresponded to the number of given tokens (corresponding to the column) found in the content of a given tweet (corresponding to the row). The resulting engagement of each tweet was stored in an array. The engagement was calculated as to whether or not a tweet passed a certain thresh.

VII. METHODS

we have used Naive Bayes to predict the result in terms of the label of several emotions.

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.



$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

Figure 1. Naive Bayes Method

Additionally, we have used Mean method to identify and analyses the timely nature of the person and represented it with a bar chart. it helps to calculate averages to help make sense of this data. One such average is called the Mean, sometimes also known as the Arithmetic Mean. Mean is the measure of central tendency most commonly used. Mean is equal to the sum of all the values of a collection of data divided by the number of values in the data.

$$U = \sum_{i=1}^{P_{total}} \int_{t=0}^{\infty} I_i(t)\ dt$$

Figure 2. Arithmetic Mean Method

VIII. RESULT

- Based on words that a sentence contains, emotion is decided and represented in the form of Pie Chart. above figure shows the different classification rate through which system was tested and final outcome of the system is also attached.

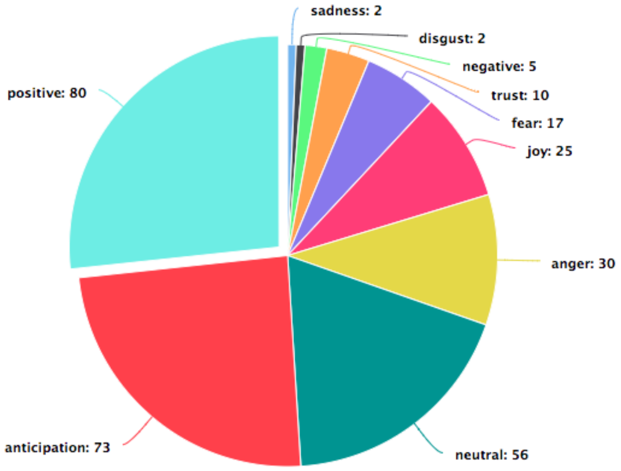


Figure 3. Emotion analysis

- An image composed of words used in a tweet or subject, in which the size of each word indicates its frequency or importance.

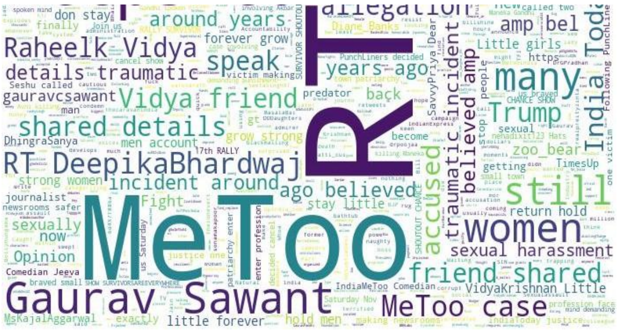


Figure 4. Word analysis (word cloud)

- Based on the user device that Twitter is accessed, a Pie Chart is plotted & various source devices are represented.

- REST API's so it can be easily usable with any kind of languages or easy to make some other applications with use of these APIs for data collecting.
- Tweets are classified based on their originating location.
- At what time interval the user uses Twitter most frequently is represented on a bar graph.

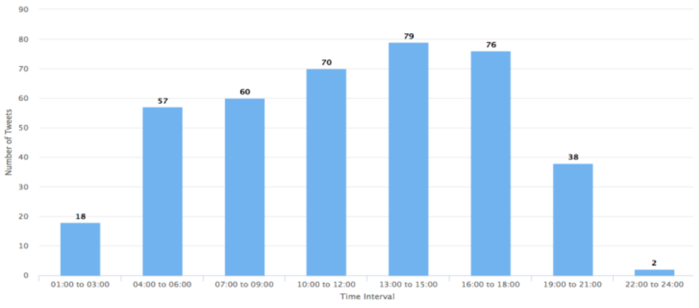


Figure 5. Timechart

IX. CONCLUSION

The task of sentiment analysis, especially in the domain of microblogging, is still in the developing stage and far from complete. So we propose a couple of ideas which we feel are worth exploring in the future and may result in further improved performance. One more feature we that is worth exploring is whether the information about the relative position of a word in a tweet has any effect on the performance of the classifier. Analyzing the public sentiment is important for many applications such as firms trying to find out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like a stock exchange, so the product we are developing will be used for the same.

X. FUTURE ENHANCEMENT

Current feature extraction method assigns the same weight for each word and ignores the structure of the sentence or the context. To solve this issue, we will look into learning vector representations of the words. Also, we want to know deeper in the logic gap between words on a psychological level. Naive method isn’t particularly compelling. We thought that the limited space permitted in tweets would allow for purely statistically based techniques, however, that assumption seems to not hold this class of problem, we found poor performance. This indicates that our text representation isn’t particularly good. a more intelligent method for feature selection could lead to better results. the most informative results from another method.

DECLARATION

AVAILABILITY OF DATA AND MATERIAL

All relevant data and material are presented in the main paper.

COMPETING INTERESTS

The authors declare that they have no competing interests.

ACKNOWLEDGMENTS

Authors acknowledge the support provided by the Department of Information Technology, L.J Institute of Engineering & Technology, Ahmedabad, Gujarat, India.

REFERENCES

1. Patel, R., Shah, N. 2017. Big Data Analytics Applications in Government Sector of India. International Journal of Scientific Research in Engineering. 1(1), 50-54
2. West, D.M., 2012. Data Mining, Data Analytics, and Web Dashboards, Gov. Stud. Brook. US Reuters.
3. . Breese, J.S., Heckerman, D., Kadie, C. 1998.Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence.
4. Joachims, Thorsten. “Text Categorization with Support Vector Machines: Learning with Many Relevant Features.” Proceedings of the European Conference on Machine Learning (ECML), 1998,
5. Santiago, Cassandra, and Doug Criss. “An activist, a little girl and the heartbreaking origin of 'Me too'.” CNN, Cable News Network, 17 Oct. 2017,
6. Zacharek, Dockterman, et al. “The Silence Breakers.” Time, TIME magazine, 2017
7. Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." Proceedings of the workshop on languages in social media. Association for Computational Linguistics, 2011
8. Bifet, Albert, and Eibe Frank. "Sentiment knowledge discovery in twitter streaming data." International Conference on Discovery Science. Springer Berlin Heidelberg, 2010.
9. Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREc. Vol. 10. 2010.