JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Sentiment Analysis Using Twitter

[1] Satyam Gupta, [2] Riya Sharma, [3] Rani Kumari, [4] Archith. R. A, [5] Ankit Paul, [6] Dr. M.S. Nidhya
[1]UG Scholar, [2]UG Scholar, [3]UG Scholar, [4]UG Scholar, [5]UG Scholar, [6]Associate Professor
[1,2,3,4,5,6]School of CS and IT,
[1,2,3,4,5,6]Jain (Deemed-to-be University), Bengaluru, India

Abstract : The advent of web technology and its growth has brought a large volume of data present on the internet for the people using it, and huge amounts of information is being generated. The net is becoming a platform where people can learn online, exchange ideas, and share opinions. Social networking sites like LinkedIn (US), Instagram (US), and Twitter (US) are swiftly gaining recognition from people from all corners of the world. This is because these sites permit people to express and share their views about various subjects.

They do so by making posts on these social media websites. These opinions are often as simple as making an observation about the weather or complex essays about how they feel about the newest MacBook. When an opinion gains traction and is liked or shared thousands of times, we understand that plenty of individuals resonate with the post and therefore, the opinion. Thus, if a person or an entity wants to understand the general opinion of the public before making a career or marketing decision, they can do so by analysing the posts on social media.

In this paper, we are using Twitter as the social media platform to decode the user's opinion regarding the topic of data science, and extract the user’s sentiment in a logical and structured manner.

***IndexTerms* – Opinion Mining, Analysing Sentiments, Data Science Field, Social Media.**

I. INTRODUCTION

In this age, the web has changed the way people express their views and sentiments. They do so through online forums, blogs, reviewing sites, social networking sites, etc. These web communities present an interactive medium where consumers can influence and inform each other.

Social networking sites generate a gigantic amount of data that is sentiment-rich through a variety of tweets, comments, reviews, etc. Moreover, these sites can provide a stage for enterprises to communicate with their consumers for the purpose of advertisement. People mostly rely on user-generated content over these websites for making market decisions. For e.g., if someone wants to shop for a product or wants to use any service, then they first find its reviews online, and discuss it on social media before making a choice. The quantity of content generated by users is simply too vast for a standard user to decipher. So, there's a demand to automate this. Thus, various sentiment analysis techniques are widely used. Sentiment analysis tells the user whether the knowledge about the merchandise is up to the mark or not before they make the choice to buy. Marketers and firms use this analysis data to know about the user opinions regarding their products and services so that they can improve it.

Sentiment analysis is often defined as a process that automates the mining of attitudes, opinions, views, and emotions from text, speech, and tweets. Sentiment analysis involves classifying opinions in text into categories like "positive" or "negative" or "neutral". It is also mentioned as subjectivity analysis, opinion mining, and appraisal extraction.

Sentiment Analysis may be a term that has many tasks like sentiment extraction, sentiment classification, subjectivity classification, summarization of opinions, or opinion spam detection, among others. It aims to analyse the public sentiment, be it criticism or acclaim, towards elements like products, individuals, topics, organizations, and services.

In this paper, we have analysed the sentiments regarding the field “Data science” through Twitter.

II. LITERATURE REVIEW

Various scholars have put in a lot of effort in the topic of "Sentiment Analysis on Twitter" in recent years. Sentiment Analysis was developed for binary classification, which assigns reviews or opinions to bipolar classes such as negative or positive exclusively, when it was still in its infancy.

Pak and Paroubek**[1]** established a strategy for categorising tweets into three categories: objective, positive, and negative. They developed a Twitter data set by using the Twitter Application Programming Interface (API) to collect tweets and automatically annotate them with emoticons. A sentiment classifier based on the multinomial Naive Bayes approach that incorporates features like Part of Speech (POS) tags and Ngrams was constructed using this dataset. The training set they utilised was less effective since it only included tweets with emoticons.

To classify tweets, Parikh and Movassate **[2]** used two models: the Maximum Entropy model and a Naive Bayes bigram model. They discovered that the Maximum Entropy model performed poorly compared to Naive Bayes classifiers.

Go and L.Huang **[3]** proposed employing distant supervision to solve sentiment analysis with Twitter data, with their training set consisting of tweets with emoticons that acted as noisy labels. Support Vector Machines (SVM), MaxEnt, and Naive Bayes are used to create models. POS, bigrams, and unigrams made up their feature space. They discovered that SVM outperformed other models and that the unigram feature was more effective.

L. Barbosa and J. Feng **[4]** developed a two-phase automatic sentiment analysis technique for Twitter classification. They classified tweets as subjective or objective, and then classified the subjective tweets as negative or positive in the next phase. Exclamation marks, hashtags, links, retweets, and punctuation were all used in the feature space. They were used in conjunction with properties such as POS and words prior polarity.

Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau **[5]** proposed a three-step paradigm for categorising sentiment into neutral, negative, and positive categories. They experimented with a feature-based model, a tree kernel-based model, and a unigram model, among others. There are 100 features in the feature-based model. Tweets were represented by a tree kernel-based model. Over 10,000 characteristics are used in the unigram model. They came to the conclusion that features that combine the prior polarity of words with their parts of speech (pos) tags are extremely relevant and play a crucial role in classification. The tree kernel-based model was found to outperform the other two models.

Using punctuation, single words, n-grams, and patterns as different feature types, Dmitry Davidov and Ari Rappoport **[6]** suggested a method to use Twitter user-defined hashtags in tweets as a classification of sentiment type, which is then integrated into a single feature vector for sentiment classification. They constructed a feature vector for each example in the training and test sets and used the K-Nearest Neighbor approach to assign sentiment labels.

III. METHODOLOGY

We follow these steps to perform sentiment analysis on Twitter Data:

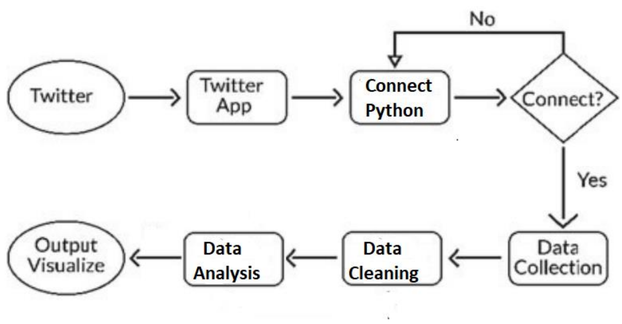

Figure 1: This data flow diagram denotes the various processes taking place in Twitter Sentiment Analysis.

3.1 Creating a dataset

A Twitter developer account is necessary for creating the Twitter app. By creating the Twitter app, we are given a consumer secret, consumer key, access token secret, and access token. These keys are necessary to access the tweets using the module, tweepy.

Next, we set up MongoDB and the Jupyter Notebook. We then run the program for the collection of tweets. Our keyword will be 'Data science'. These tweets with the hashtag 'Datascience' will be stored in the MongoDB database.

As the tweets are now stored in JavaScript Object Notation (JSON) format and there is too much information, we will extract what we need into a Comma-Separated Valued (CSV) file.

3.2 Processing the tweets

We will first partially clean the tweets using the Regex module to remove the special characters, links, retweets, extra spaces, etc.

Next, we will correct the spelling mistakes by first separating the incorrect words. Then we compare each incorrect word with ones from the custom dictionary and thus, incorrect words that very closely match the ones from the dictionary are converted to the dictionary words. For example, 'Datascien' will become 'Data science'.

3.3 Analysing the cleaned data

Here we use Textblob to calculate the sentiment of the tweet. It is always between -1 to 1, where -1 to 0 is a negative tweet, 0 is neutral, and 0 to +1 is a positive tweet.
We will then plot a pie chart from these readings:

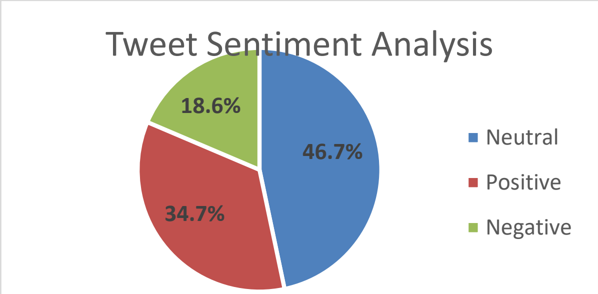

Figure 2: This pie chart shows the percentage of tweets that are positive, negative, or neutral in the 5000 tweets that we have analysed with the help of Textblob

Now we will also plot a word cloud. It is a visual representation of word frequency. The more frequently a term appears within the text, the larger it appears in the image generated. Word clouds are used as a simple tool to understand the focus of these tweets.

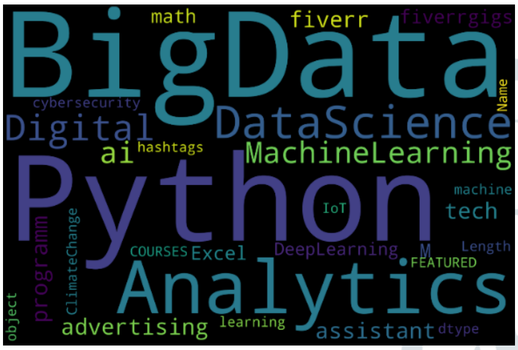

Figure 3: This word cloud is a visual representation of word frequency in the cleaned tweets. The higher the frequency, the bigger is the word

IV. CONCLUSION

Considering the present growth of technology around the globe, people are relying more and more on computers and the web. They are getting accustomed to buying products online instead of visiting a store. Now, before buying anything, people go through the product reviews. If the review is good, their chances of buying the product increase and vice versa. Similarly, if a selected career path doesn't get a positive reception, fewer people will try to enter it. For example, many game developers hold the opinion that their career path lacks stability. This can help spread awareness among students planning to pursue that career path. Same with different types of products. If a mobile phone was heavily advertised yet is receiving a lot of criticism on social media websites, people will want to investigate before buying it. So, the only way for corporations to succeed in making the product that users actually want, will be to analyse public attitudes and sentiment. Our study attempted to ascertain public attitudes by focusing on the efficiency of the current method of analysis and generating findings by reducing search constraints. In our paper, we may get results for a variety of keywords. Our goal is to see a steady improvement in the accuracy of the algorithms used in sentiment analysis.

REFERENCES

[1] A.Pak and P. Paroubek. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining in Proceedings of the Seventh Conference on International Language Resources and Evaluation, 1320-1326
[2] R. Parikh and M. Movassate. 2009. Sentiment Analysis of User- Generated Twitter Updates using Various Classification Techniques, CS224N Final Report.
[3] Go, R. Bhayani, L.Huang. 2009. Twitter Sentiment Classification Using Distant Supervision. Stanford University, Technical Paper.
[4] L. Barbosa, J. Feng. 2010. Robust Sentiment Detection on Twitter from Biased and Noisy Data. Coling 2010: Poster Volume, 36-44.
[5] Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau. 2011. Sentiment Analysis of Twitter Data, in Proceedings of the ACL 2011 Workshop on Languages in Social Media, 30-38.
[6] Dmitry Davidov, Ari Rappoport. 2010. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. Beijing. Coling 2010: Poster Volume, 241-249.
[7] Weber CT, Syed S. 2019. Interdisciplinary optimism? Sentiment analysis of Twitter data, R. Soc. open science, 6: 190473.
[8] Po-Wei Liang, Bi-Ru Dai. 2013. Opinion Mining on Social Media Data, IEEE 14th International Conference on Mobile Data Management, Milan, Italy, 91-96.
[9] Pablo Gamallo, Marc os Garcia. 2014. Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets, 8th international workshop on Semantic Evaluation, Dublin, Ireland, 171-175.
[10] Bromham L, Dinnage R, Hua X. 2016. Interdisciplinary research has consistently lower funding success, in Nature 534, 684–687
[11] Mohammadi E, Thelwall M, Kwasny M, Holmes KL. 2018. Academic information on Twitter: a user survey, PLoS ONE 13.

[12] Bhatti, U. and Hanif. M. 2010. Validity of Capital Assets Pricing Model.Evidence from KSE-Pakistan.European Journal of Economics, Finance and Administrative Science, 3 (20).

[13] R. Xia, C. Zong, and S. Li. 2011. Ensemble of feature sets and classification algorithms for sentiment classification, Information Sciences: an International Journal, 181(6): 1138–1152.

[14] Zhunchen Luo, Miles Osborne, Ting Wang. 2013. An effective approach to tweets opinion retrieval, Springer Journal on World Wide Web.

[15] Froehlich HE, Gentry RR, Rust MB, Grimm D, Halpern BS. 2017. Public perceptions of aquaculture: evaluating spatiotemporal patterns of sentiment around the world, PLoS ONE 12.