Twitter Sentiment and Sarcasm Analysis

SARTHAK TRIPATHI[1], AYUSHMAN PANDEY[2], ANMOL CHITRANSH[3], PROF S. P MEDHANE[4]
[1, 2, 3, 4] Bharati Vidyapeeth (Deemed to be University) College of Engineering Pune, India

Abstract- As we all are seeing that social media is now the biggest platform for showcasing our views and also our opinions in the support of any thought. But nowadays it is seen that most of the people didn't get the tone of tweet and can misunderstand it. So now there is a biggest requirement to filter out the tone of our tweet i.e., we can classify our tweets into positive, negative or neutral. This project is basically based on the fact that we can use machine learning to filter out our tweets .We will extract the data from twitter and will store it in csv file, pre-process the data, then tokenize the data and using feature extraction will extract our features .Then using different machine learning algorithms such as Support Vector Machine (SVM) and Naïve-Bayes Algorithm, polarities will be assigned to features which lies between -1 and 1.Based on value of polarity tweets will be classified as positive, negative or neutral. Machine Learning Algorithms such as Support Vector Machine, Naive Bayes Algorithm are used for sentiment classification. Support Vector Machine works mainly by investigating information and characterizing the components for calculation whereas Naive Bayes Algorithm works mainly using Bayes Theorem which is highly dependent on closeness of features. Many other tools such as Twitter Sentiment, SentiStrength etc. can be implemented but the best accuracy was given by Naïve-Bayes Classifier.

I. INTRODUCTION

On Social media platforms such as Facebook, Twitter and LinkedIn have gained a lot of importance in the past few years since they provide an easy and accessible mode for the people to communicate in large numbers without any restrictions. Each of these platforms has their own set of characteristics. For example, any user on Facebook can write their reviews and opinions as they like without any length constraints but there exists a reciprocated relationship between one another on the network. But, on Twitter, there are no such relationships between two users. Every user can express their thoughts on anyone or any proposal irrespective of the fact that they may or may not be connected to each other. Also, Twitter only allows each tweet to be of at most 140 characters. These features are important factors which makes twitter an appropriate platform to perform sentiment analysis and predicting and analyzing the public opinions.

In this project, we utilize the massive twitter data such as hashtags, user profile and number of tweets to analyze the data and return the related sentiment. Machine Learning classifiers such as Naïve-Bayes and Support Vector Machine are used to examine and produce the final prediction of sentiment.

II. LITERATURE SURVEY

To investigate any machine learning model, there are numerous fundamental research articles that one should read to acquire a basic grasp of what has already been done in the area and how the current model can perform at least as if not better than, the prior models.

- Twitter Sentiment Analysis [1]

This paper proposes various machine learning algorithms and their accuracy, so that the best algorithm can be used to work with. When it comes to classifying sentiment in tweets, machine learning approaches perform admirably.

However, the authors had a lot of other suggestions for how to improve our accuracy. The following is a list of possible improvements.

For training dataset, the authors collected messages that contained the emoticons :) and :(via the Twitter API. They built the naïve bayes classifier from scratch

- Naïve Bayes

Naive Bayes is a simple classification model. It is simple and effective for text categorization. In the project, multinomial Naive Bayes is used. It is assumed that each feature is conditionally independent of the others in the class. That is, where c represents a certain class and t represents the text to be classified. P(c) and P(t) are the prior probability of this class and this text, respectively. And P(t | c) is the likelihood of the text appearing given this class. In this situation, the value of class c might be either POSITIVE or NEGATIVE, and t is simply a sentence.

- Feature selection

Hundreds of thousands of sentences are included in the training set. However, it remains a huge amount of features for the training set. It is beneficial to remove certain unnecessary features. T three distinct feature selection algorithms were used.

1. Frequency-based feature selection
2. mutual information
3. X^2 Feature selection

The Naive Bayes Classifier was expanded to accommodate three classes: positive, neutral, and negative.

It is quite difficult to collect a big number of neutral tweets. For the training data, they simply classified any tweet without an emoji as neutral. This is obviously a false assumption, but they wanted to see how the tests turned out.

they manually categorized 33 tweets as neutral for the test data.

The outcomes were disastrous. The classifier only achieved 40% accuracy. This is most likely because of the noisy training data for the neutral class.

- Twitter Sentiment Analysis: The Good the Bad and the OMG! [2]

This paper investigates the application of linguistic functions for detecting the sentiment of Twitter messages. They examine the usefulness of present lexical sources in addition to functions that seize data approximately the casual and innovative language utilized in microblogging. And take a supervised method to the hassle however leverage current hashtags withinside the Twitter information for constructing education records.

- Hash tagged data set

The dataset with hashtags is a subset of the Edinburgh Twitter corpus. The Edinburgh corpus contains 97 million tweets gathered over a two-month period. To generate the hashtagged data collection, we first remove duplicate tweets, non-English tweets, and tweets without hashtags. We analyse the distribution of hashtags in the remaining collection (about 4 million) and expect to uncover groupings of frequent hashtags that are indicative of good, negative, and neutral messages. These hashtags are used to filter tweets for development and training purposes. Table 2 shows the top 15 hashtags in the Edinburgh corpus. In addition to the well-known hashtags that are part of the Twitter folksonomy (e.g., #followfriday, #musicmonday) We come across hashtags that appear to reflect message polarity: #fail, #omgthatsotrue, #iloveitwhen, and so on. To choose the final collection of messages for the HASH dataset, we look for hashtags that appear at least 1,000 times in the Edinburgh corpus. We chose the top hashtags that we thought would be most effective for detecting good, negative, and neutral tweets. Table 3 lists these hashtags. Messages with these hashtags were included in the final dataset, and each message's polarity is defined by its hashtag.

- Lexicon capabilities

Words indexed the MPQA subjectivity lexicon (Wilson, Wiebe, and Hoffmann 2009) are tagged with their earlier polarity: positive, negative, or neutral. They create 3 capabilities primarily based totally at the presence of any phrases from the lexicon.

Part-of-speech capabilities

For every tweet they have got capabilities for counts of the quantity of verbs, adverbs, adjectives, nouns, and another elements of speech.

- Micro-running a blog functions

They create binary functions that seize the presence of positive, negative, and impartial emoticons and abbreviations and the presence of intensifiers (e.g., all-caps and person repetitions). For the emoticons and abbreviations, they use the Internet Lingo Dictionary (Wasden 2006) and diverse net slang dictionaries to be had online

Twitter sentiment analysis tests reveal that part-of-speech characteristics may be ineffective for sentiment analysis in the microblogging arena. More research is needed to identify whether the POS features are just of poor quality as a result of the tagger's output, or whether POS features are simply less effective for sentiment analysis in this domain.

The presence of intensifiers, as positive/negative/neutral emoticons and abbreviations was definitely the most useful when combined with elements from an existing sentiment lexicon.

Using hashtags to collect training data was beneficial, as was collecting data based on positive and negative emoticons. However, which approach generates superior training data and whether the two sources of training are equivalent.

- The State-of-the-Art in Twitter Sentiment Analysis:

- A Review and Benchmark Evaluation [3]

Twitter has grown in popularity as a significant social media site, piquing the interest of sentiment analysis specialists. Despite this emphasis, state-of-the-art Twitter sentiment analysis techniques perform poorly, with stated classification accuracies often falling below 70%, limiting the use of the obtained sentiment data.

In this study, the authors investigate the specific obstacles that Twitter sentiment analysis poses, and we explore the literature to see how different methodologies have tackled these issues. The authors conduct a benchmark study of 28 top academic and commercial systems in tweet sentiment classification over five separate data sets to assess the state-of-the-art in Twitter sentiment analysis.

They conduct an error analysis to determine the root causes of common categorization errors. They use chosen systems in an event detection to advance the evaluation.

TSA research is guided by two main goals. The initial line of investigation focuses on using TSA to acquire insights into various commercial or social concerns, predict critical indicators, or watch Twitter for breaking news or events. Recognizing the importance of data obtained through accurate TSA, the second line of study focuses on inventing and innovating better TSA methodologies.

Finally, they presented recommendations to assist the design of the next generation of approaches by summarizing the important themes and conclusions from the review and benchmark evaluation.

- Twitter Sentiment Analysis on Coronavirus: Machine Learning Approach [4]

The examination of data to identify feelings using algorithms that allow us to determine the positive or negative emotions that people have about a topic is a fundamental difficulty in machine learning.

Microblogging and social media are great sources of information, but they are largely utilised to convey personal opinions and beliefs. The authors proposed a sentiment analysis of English tweets during the pandemic COVID-19 in 2020 based on this knowledge.

The Logistic Regression model was used to classify the tweets as positive or negative, with a classification accuracy of 78.5 percent.

The research's major goal is to use machine learning algorithms and natural language processing techniques to determine if public opinion is good or negative. Even though the research revealed despite the diversity of perspectives, it appears that most individuals are still optimistic about the epidemic.

March is the sole month in which negative thoughts predominated, and it is also the month when the COVID-19 sickness first appeared.

was designated a pandemic, and many countries began to take precautions and follow safety regulations. It corresponds to an increase in happy thoughts in conclusion, 54 % were positive sentiments, with 46% of users expressing negative emotions.

Due to the vast amount of information available on the internet, which might be erroneous at times, it is required to collect and analyse data using specific methodologies.

In this paper, they also used a technique known as web scraping.

- Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis [5]

They developed a novel method for lexicon-based sentiment analysis of Twitter messages in this paper. The new approach normalises sentiment, allowing us to acquire sentiment intensity rather than a positive/negative decision. In order to improve the algorithm's performance in circumstances where a message contains mixed sentiment, a new evidence-based combining function was devised. The Stanford Twitter test set and the IMDB data collection they’re used in the analysis. The results show that the two novel functions help the usual lexicon-based sentiment analysis system perform better. It's worth noting that the strategy is better suited to short messages like theists. When used with long documents, the approach outperformed the competition.

Current lexicon-based and learning-based sentiment analysis methodologies face new challenges due to the specific properties of Twitter data. They offered an innovative solution to the challenges in this research. An First, a Twitter-specific augmented lexicon-based method was used to do sentiment analysis. Additional information was obtained using a Chi-square test on the output. Tweets with strong opinions could be discovered. The sentiment is then assigned using a binary sentiment classifier. Polarities to the newly discovered opportunistic tweets, whose lexicon-based training data is provided method. Experiments show that the planned method is quite successful and promising.

- Benchmarking Twitter Sentiment Analysis Tools [6]

The search option is the second way to create a Twitter competitor analysis. This refers to the number of Twitter mentions the account has gotten. Monitoring a Twitter account is simple and yields useful information. In reality, this will generate a report containing all of the theists sent about the account. What is the purpose of this? This will allow us to determine the impact, the reach, quantity of activity, and a variety of other KPIs associated with that Twitter account. It's now time to learn how Twitter can assist us in confronting the Twitter dataset and obtaining Twitter profile analytics.

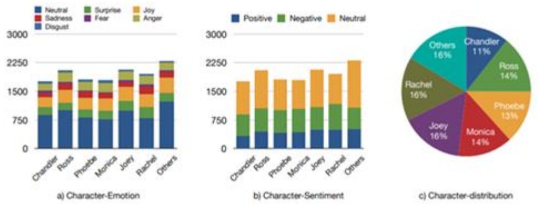
Their findings have significant ramifications for a variety of stakeholders. Researchers in social media analytics can utilise the findings to make more educated decisions about the technologies to use for a project. They might also consider the advantages and disadvantages of employing stand-alone vs workbench tools. The results of the mistake analysis can be used by NLP and text mining researchers and developers to improve future commercial stand-alone tools.

Industry executives can gain a greater understanding of the underlying text analytics' strengths and shortcomings, as they’ll as how they may affect the quality and dependability of social media inputs utilised in decision-making. Some of the labelled test beds with error analysis annotations have been made publicly available through LRE Map as an extra resource. This will make future benchmarking easier.

- MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations [7]

ERC, or Emotional Recognition in Conversation, has numerous applications. There was no database of a comparable larger scale for an emotional discourse with more than two speakers each dialogue prior to this work. The most intriguing aspect of this model was that it included around 13000 utterances and 1433 sentences from the TV series Friends. This dataset assisted several future algorithms in labelling words and sentences with emotion and intent/sentiment labels.

- The MELD dataset has actually evolved from the Emotion Lines dataset developed by Chen at al. (2018).
- It was established from this dataset that a TV show can be a perfect start to creating a dataset simply by scouring through the subtitles of the show and extracting the data needed for analysis by segregation of text in the corpus through timestamps.
- The Baseline Models, include text-CNN (which doesn't use context for its evaluation) and bcLSTM which, on the other hand, uses context with the help of bidirectional RNN.
- The model additionally analyzed the role of context, shift in emotion, and future directions.



a) Character-Emotion

b) Character-Sentiment

c) Character-distribution

- A Study on Sentiment Analysis Techniques of Twitter Data [8]

- In their research work, Abdullah ALsaeedi and Mohammad Zubair Khan described many strategies for sentimental analysis of various Tweets. They gathered information by utilising the Twitter API.
- They used classification techniques such as SVM, Naive Bayes, Maximum Entropy, and others, as well as sentence level and document level sentiment analysis.
- They also talked about sentiment analysis ensemble models and lexical techniques.
- Their results suggest that SVM is superior for sentiment analysis since it provides more accuracy than any other technique when several features are used.
- Nave-Bayes and SVM have attained an accuracy of almost 80%.
- They also discovered that ensemble and hybrid algorithms outperform supervised machine learning techniques.

- Sentiment Research on Twitter Data [9]

The monitoring and analysis of social network data is becoming more common. People's views are not only heard, but they also have a direct impact on key issues and shape political and logical thinking. In 2019, Google CEO Sundar Pichai was questioned in the United States Congress on why the word "idiot" when searched on Google brings up images of then-POTUS Donald Trump. Clearly, there has been some misreading of meaning that has occurred online in recent years. As a result, analysing tweets is vital and critical to understanding people's perspectives.

- This paper leads to the following conclusions:

- Twitter sentiment analysis was created to analyse public opinion on a specific tweet or hashtag.
- Input is provided, such as a username or a hashtag. The tweet is then downloaded from Twitter information and subjected to feature extraction.
- When correct pre-processing is used, an economical feature vector is generated by doing feature extraction in two phases.
- As illustrated in Fig 2.3, the Twitter-specific parameters are extracted and added to the feature vector at the start.
- Following that, these options are separated from tweets, and feature extraction is completed as if it were done on traditional text. The accuracy is assessed using the Naive Bayes Classifier.

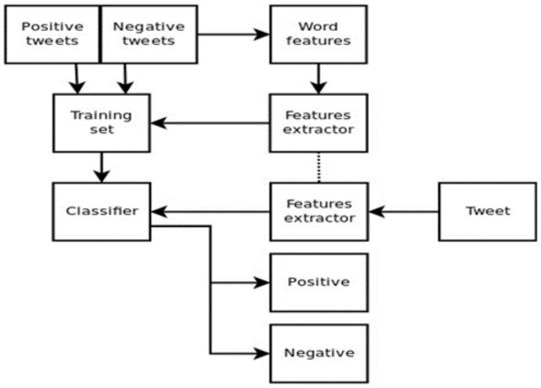


Fig 2.3: Sentiment analysis using Naive-Bayes Classifier [5]

- A Real-Time Twitter Sentiment Analysis and Visualization System: TwiSent[10]

- In their research work, Mamta and Ela Kumar employed hashtags and keywords and then studied how individuals reacted to these hashtags and keywords.
- They conducted analysis in six stages: data collection, preprocessing, feature extraction, sentiment identification, sentiment classification, and output representation.
- They represented the final result using pie charts and a trend graph.
- As demonstrated in Fig 2.4, they utilised hashtags such as #Journalism and #CWC19.
- For sentimental analysis of these datasets, they used a lexicon-based technique.
- The Journalism dataset has been divided into three categories, as seen in the graphic.
- For the #Journalism dataset, 31.6 percent of tweets are negative, 17.7 percent are positive, and the rest are neutral.

- Summary of Literature Survey:

After referring to the papers that were mentioned, we are able to conclude that various algorithms were used for classifying the sentiments but the algorithm that provided the best accuracy was Support Vector Machine. We will be using Support Vector machine for the intelligent model. For the comparison we will be using Naive-Bayes algorithm and also in order to show which algorithm will give more accurate and efficient classification. Lastly, this research gave us an insight about what all work has been done in the area of Sentiment analysis and how we can improve our work in order to give better results to our users.

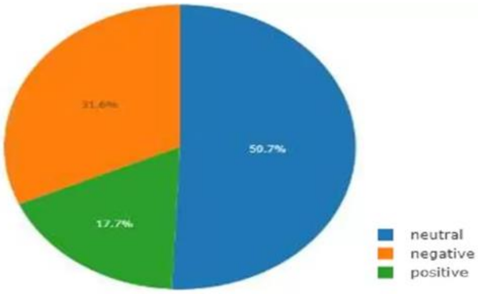


Fig 2.5: Pie chart representation for hashtags like #Journalism[8]

III. PROBLEM STATEMENT

Social media platforms such as Facebook, Twitter and LinkedIn have gained a lot of importance in the past few years since they provide an easy and accessible mode for the people to communicate in large numbers without any restrictions. Each of these platforms has their own set of characteristics. For example, any user on Facebook can write their reviews and opinions as they like without any length constraints but there exists a reciprocated relationship between one another on the network. But, on Twitter, there are no such relationships between two users. Every user can express their thoughts on anyone or any proposal irrespective of the fact that they may or may not be connected to each other. Also, Twitter only allows each tweet to be of at most 140 characters. These features are important factors which makes twitter an appropriate platform to perform sentiment analysis and predicting and analyzing the public opinions.

In this project, we utilize the massive twitter data such as hashtags, user profile and number of tweets to analyze the data and return the related sentiment. Machine Learning classifiers such as Naïve-Bayes and Support Vector Machine are used to examine and produce the final prediction of sentiment.

IV. SARCASM DETECTION

Irony has been part of our language for many years. It's the opposite of what you're saying and usually means having a clear voice tone in a fun way. If you think everyone can understand the irony, you're wrong. Understanding irony depends on your language skills and the knowledge of the minds of others. But what about computers? Is it possible to train a machine learning model that can detect if a sentence is ironic? Yes, it is! Therefore, if you want to learn how to recognize irony with machine learning, this article is useful. This article describes the task of machine learning irony detection using Python. irony detection by machine learning Irony means interesting by being the exact opposite of what you are saying. It has been part of all human language for many years. Today, it is also used in headlines and various other social media platforms and is gaining attention. Irony detection is a natural language processing and binary classification task. You can use the sarcastic and nonsarcastic sentence datasets found in Kaggle to train machine learning models that recognize whether a sentence is sarcastic.

V. ALGORITHM

The purpose of this project is to perform sentiment analysis of twitter streams and threads using Classification algorithms such as Naïve-Bayes Algorithm and SVM and tweepy library in Python. The analyzer predicts the intention behind the tweet and classifies the tweet as a positive, negative or neutral opinion.

In this Project, Support Vector Machine is used as the machine learning classifier that takes the tokenized data as input and determines the polarity value of the tweet. Based on this polarity value the classifier further decides the sentiment as positive, negative or neutral. After designing the analyzer, a User Interface (UI) will enable the people to use the tool in an easy and effective way. The UI helps the user to access the system. It takes a single tweet/a tweet thread / hashtags as input and predicts the sentiment behind the tweet as output.

Scope of the Project

- There is an immense scope of machine learning and data mining algorithms in the data extraction and classification of information on the internet.
- This tool can be used by the government organizations to track what political opinions do people hold on a decision.
- The customer experience team of various brands can analyze the consumer's reviews of the product and utilize them to make further modifications.
- Marketing strategies of big corporate firms can be formulated based on the analysis of trending tweets.

- Challenges and Open Problems:

Aggregating the distinct Opinions during a statement. E.g - A sentence having two or additional opinion phrases for some options.

Context based mostly opinion phrases for features

E.g - High + value - Negative Opinion
High + Quality - Positive Opinion

Words have the choice which means within your domain

Being "competitive" in most conditions isn't taken into thought a very exceptional trait. However, being competitive whereas you're a ahead in football is also a wonderful thing. The agreement is even additional distinguished with phrases resembling "killer". AN offender in football who' a "killer", or contains a "killer instinct" could be a superb athlete. However, currently now not several people in general can assume that a "killer" in actual lifestyles could be a superb thing. General motive sentiment analysis engines gets terribly harried on this context.

- Future Scope

The model developed in this project is a sentiment analysis system that predicts the user's sentiment behind a tweet using Multinomial Naïve-Bayes Classification. The system is accessible through the internet and is connected to a UI so that it is easily accessible and the results are clearly understood.

The existing model can be improved by modifying the system to compute a greater polarity range. Currently, it predicts a range of sentiments between 0,1 and 2. By increasing the polarity range, the sentiment can be further classified into more specific emotions. In addition to this, a graphical representation such as a pie chart can be plotted to understand the statistical aspect of the sentiments of the masses.

The proposed model is currently unable to detect sarcasm in texts which is quiet prevalent in tweets nowadays and as a result, sarcasm can make a lot of difference in the entire sentiment of the tweet.

CONCLUSION

1. Performance Evaluation

In this work, a model is developed which performs sentiment analysis on real-time tweets, twitter handles and hashtags on twitter using Multinomial Naïve-

Bayes classification technique. The model is accessible on the web through a UI.

Sentiment analysis is performed on the requested tweets and the output is presented in the form of positive, negative and neutral emotions.

In order to train and test various models, a dataset with around 160 thousand tweets was considered. The Random Forest and Decision Tree model provided good accuracy but could not handle large datasets. XGBooster model gave the best accuracy but required high computational power.

Finally, the Naïve-Bayes model resolved the drawbacks by using multinomial classification and required less computation resources. Hence, it is concluded that the Naïve-Bayes classifier resulted in a lightweight and faster application.

REFERENCES

[1] Alec Go (alecmgo@stanford.edu) Lei Huang (leirocky@stanford.edu) Richa Bhayani (richab86@stanford.edu) "Twitter Sentiment Analysis'' June 6, 2009 Error! Hyperlink reference not valid.

[2] Efthymios Kouloumpis (epistimos@i-sieve.com) Theresa Wilson(taw@jhu.edu) Johanna Moore(j.moore@ed.ac.uk) "Twitter Sentiment Analysis: The Good the Bad and the OMG!" https://ojs.aaai.org/index.php/ICWSM/article/view/14185/14034

[3] David Zimbra, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. "The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation." ACM Trans. Manage. Inf. Syst. 9, 2, Article 5 (August 2018), https://doi.org/10.1145/3185045

[4] Cristian R. Machuca, Cristian Gallardo and Renato M. Toasa Published under licence by IOP Publishing Ltd Journal of Physics: Conference Series, Volume 1828, 2020 International Symposium on Automation, Information and Computing (ISAIC 2020) 2-4 December 2020, Beijing, China Citation Cristian R. Machuca et al 2021 J. Phys.: Conf. Ser. 1828 012104 https://iopscience.iop.org/article/10.1088/1742-6596/1828/1/012104/meta

[5] Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, Bing Liu Hewlett-Packard Laboratories mails: {riddhiman.ghosh, mohamed.dekhil, meichun.hsu} @hp.com "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis" University of Illinois at Chicago 1501 Page Mill Rd., Palo Alto, CA 851 S. Morgan St., Chicago, LL {lzhang3, liub}@cs.uic.edu https://www.hpl.hp.com/techreports/2011/HPL-2011-89.pdf

[6] Ahmed Abbasi, Ammar Hassan, Milan Dhar E-mail: abbasi@comm.virginia.edu, mah9tg@virginia.edu, msd4ah@virginia.edu "Benchmarking Twitter Sentiment Analysis Tools" University of Virginia Charlottesville, Virginia, USA http://www.lrec-conf.org/proceedings/lrec2014/pdf/483_Paper.pdf

[7] Soujanya Poria , Devamanyu Hazarika, Navonil Majumder , Gautam Naik, Erik Cambria, Rada Mihalceaι Information Systems Technology and Design, SUTD, Singapore ΦSchool of Computing, National University of Singapore, Singapore Centro de Investigacion en Computaci ´ on, Instituto Polit ´ ecnico Nacional, Mexico ´ Computer Science & Engineering, Nanyang Technological University, Singapore Computer Science & Engineering, University of Michigan, USA,2019. "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations"

[8] Abdullah Alsaeedi1 , Mohammad Zubair Khan2 Department of Computer Science, College of Computer Science and Engineering Taibah University, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 2, 2019. "A Study on Sentiment Analysis Techniques of Twitter Data"

[9] Brahmananda Reddy, D.N.Vasundhara, P. Subhash, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S11, September 2019 "Sentiment Research on Twitter Data"

[10] Mamta1, Ela Kumar2, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 06 Issue: 06 | June 2019. “A Real-Time Twitter Sentiment Analysis and Visualization System: TwiSent”