Sentiment Analysis on Twitter Data: A Survey

Abhilash Mittal
Computer Science and Engineering Department
Delhi Technological University
New Delhi, India.
abhilashmittal23@gmail.com

Sanjay Patidar
Computer Science and Engineering Department
Delhi Technological University
New Delhi, India.
sanjaypatidar@dtu.ac.in

ABSTRACT

Twitter is the popular micro blogging site where thousands of people exchange their thoughts daily in the form of tweets. The characteristics of tweet is to be short and simple way of expressions. Though this paper will focus on sentiment analysis of twitter data. The research area of sentiment analysis are text data mining and NLP. In different form we can perform the sentiment analysis on twitter data. This research paper will focus on techniques of sentiment analysis where we will perform how to extract tweets from twitter. Eventually we will compare different sentiment analysis techniques and also the approaches containing twitter dataset.

CCS Concepts

• Information systems→Database management system engines.

Keywords

Social media; sentiment analysis; twitter data; text mining.

1. INTRODUCTION

Now a days twitter, facebook, whatsapp are getting so much attention from people and also they are getting very much popular among people. Sentiment analysis provides many opportunities to develop a new application. in the industrial field, sentiment analysis has big effect, like government organization and big companies, their desire is to know about what people think about their product, their market value. the aim of sentiment analysis is to find out the mood, behaviour and opinion of person from texts. for the sentiment analysis purpose, social networking used the various sentiment analysis techniques to take the public data. Sentiment analysis widely used in various domain such as finance, economics ,defence , politics. The data available on the social networking sites can be unstructured and structured. almost 80% data on the internet is unstructured. Sentiment analysis techniques are used to find out the people opinion on social media. Twitter is also a huge platform in that different idea, thought, opinion are presented and exchanged. It does not matter where people came from, what religious opinions they hold, rich or poor, educated or uneducated, they comment, compliment, discuss, argue,insist.

2. SENTIMENT ANALYSIS

Sentiment analysis is the area of study where we deliberate the people's behavior related to any topic, about any chronicle. It is highly produces the big problem zone. It is having different tasks and multifarious names e.g. sentiment mining, subjectivity analysis, affect analysis, review mining etc.

2.1 Level of analysis

There are three different levels of sentiment analysis which are as follows:

2.1.1 Document Level Analysis

At this document level the task is to determine the the overall opinion of the document. It assumes that sentiment analysis at document level expresses opinions at a single entity[12].

2.1.2 Sentence Level Analysis

At this level the task is to determine the whether each and every sentences expressed a positive or negative or neutral opinion.

2.1.3 Entity/Aspect Level Analysis

It is also called feature level analysis. At this level we can find what people likes or dislikes but on the sentence level and document level we can't find what people likes or dislikes. This level perform finer-grained analysis.[11]

3. TWITTER

it is our aim to classify the tweets in various sentiment classes to perform sentiment analysis on twitter data. In this research field, to train a model several approaches have developed which is also used for testing to check its efficiency. It is challenging to perform sentiment analysis on twitter data. There are some reasons defined for this.

Bound tweet Sizesize of the tweets is bounded i.e. it is having only 280 characters which generate intensive statement.

By Using slang wordsthese slang words are bit different from English words. If we used slang words in our sentence then that slang words make an approach outmoded.

Twitter features twitter is permissive to use of URL's, hashtag(#) and user reference. In the twitter feature different processes are used to compare with other different words.

User diversity there are various ways to convey their opinion, people's used different language between the tweets while others used encore words or symbols to convey their emotions. [9]

4. SENTIMENTANALYSIS ON TWITTER DATA

ICCCM 2019, July 27–29, 2019, Bangkok, Thailand
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7195-7/19/07…$15.00
https://doi.org/10.1145/3348445.3348466

This system basically consist of four stage named as: tweet retrieval, tweet pre-processing, classification algorithm and evaluation.

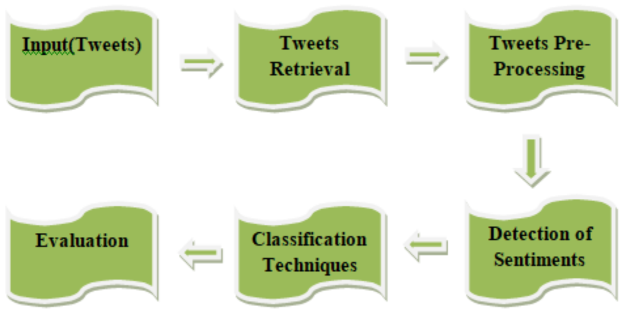

Figure 1. Working of Sentiment analysis in Twitter

Input (Keyword) in the beginning we will choose the subject, with related to that subject we will gather the tweet then after on those tweets we will perform the sentiment analysis.

Tweets Retrieval In this step, tweets will be retrieved, it can be in any form such as, unstructured, structured and semi structured . In sentiment analysis research can be done by the use of different programming language like python or R we can collect the tweets.

Pre-processing in the tweet pre-processing, data is filtered by removing the irrelevant data, inconsistent data and noisy data. Following tasks have to be performed while pre-processing:

- Removal of re-tweets(twitter data sets)
- Removing special characters, URLs, numbers and punctuations etc.
- Removing stopwords
- Stemming
- Tokenization

Sentiment Detection There are various application of sentiment analysis where it is mandatory to find out the sentiment like tweet classification and tweet mining.

In sentiment analysis our task is to find the polarity of the given word. It may be positive, negative or it may be neutral. By the use of different lexicons we can identify the polarity e.g. Bing Lui sentiment lexicons, sentiWord Net etc. which will estimate sentiment score and its strength etc. [11].

- Sentiment classification algorithm

Sentiment analysis can be classified into two approach i.e supervised learning and unsupervised learning. In supervised learning, na ïvebayes, SVM and maximum entropy are used to perform the sentiment analysis whereas in unsupervised learning lexicon based, corpus based and dictionary based are used to perform the sentiment analysis. Furthermore Accuracy of classifier is depends on which training and testing dataset is used for which classification method

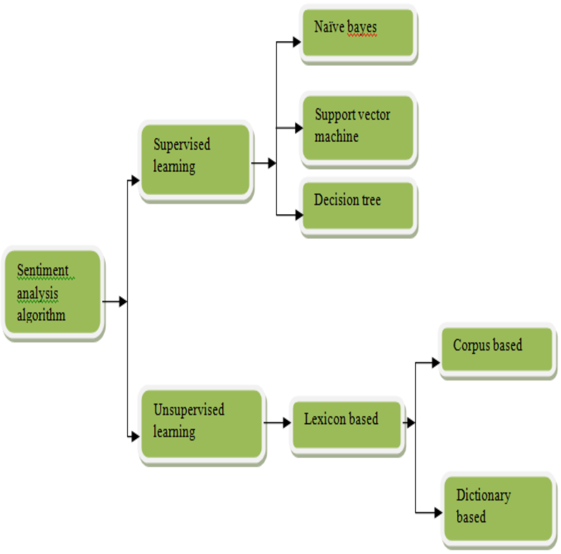

Figure2. Sentiment Analysis Techniques

Evaluation in sentiment analysis final output is analyzed to take decision whether we should opt it or not and then the result are displayed in term of bar graph, pie chart and line graph.

5. LITERATURE SURVEY

For mining the data, there are different text mining approaches are used.

Dhirajgurkhe, Niraj pal and RishitBhatia proposed how twitter data is processed firstly they collected data from various sources and eliminate those feature which does not contribute to find any polarity and then this data send into the sentiment classification engine i.e. na ïve bayes classification algorithm which will calculate the probabilities i.e. how much data is corrected and predict the sentiment for the given query.[1]

M.bouazizi, T. ohtsuki have proposed the tweets which contain more than one sentiment called as multi class sentiment analysis. Where they have identify the exact sentiment conveyed by the user rather than the whole sentiment of the tweet. To identify this thing they have also used SENTA tool. They proposed an approach, with the help of this approach they have calculated the sentiment score whoever sentiment is having highest score that will be considered this process is called as “Quantification”. [2]

Geetikagautam, DivakarYadav have discussed about customer review classification for which they have used twitter dataset which is already labeled. In this task they have used machine learning based algorithm i.e. na ïve bayes, SVM, maximum entropy. They have worked on Python and NLTK for training the SVM, na ïve bayes, maximum entropy. Na ïve bayes is better techniques in term of accuracy and gives the better result compare to Maximum entropy. We can get the better result with compare to SVM by using the SVM with unigram model. And then further accuracy can be improved by semantic analytic followed by wordNet.[3].

AkshayAmolik, Niketanjivane, MahavirBhandari, Dr. m.venkatesan, a highly suitable model have discussed int his paper which will take the twitter data of upcoming Hollywood and bollywood movies. They are able to this task with the help of classifier and features like SVM and na ïve bayes. Both of them are used for high accuracy but in terms of precision naïve bayes is better than SVM and if we talk about recall then SVM is better than naïve bayes. By increasing the dataset we can increase the classification accuracy.[4]

Subhabrata Mukherjee et al. have discussed a hybrid system named as TwiSent which will resolve problem like spam tweet ,pragmatics, noisy text. Twisent consist of spell checker and pragmatics handler. spell checker finds the noisy text whereas pragmatics handler handles the pragmatics in tweets. Twisent gives better result compare to C-feel-IT system. The accuracy of finding the negative sentiment of Twisentsytem is high the C-Feel-IT.[5].

Dmitry Davidov, Oren Tsur, AriRappoport in this paper they have proposed a supervised sentiment classification structure which works well on twitter data. They have used K nearest neighbor and feature vector. the basic purpose of this framework is to identify and distinguish between sentiment types defined by smiley and tags.[6]

Neethu M S, Rajasree R, the author have used the machine learning techniques in this survey paper to explore the twitter data related to electronic product. They have used feature vector for the tweets classification . they have used three types of classifier i.e. SVM, naïve bayes, maximum entropy, and these classifier were tested using Matlab simulator. SVM and naïve bayes classifier are implemented using built in function. Whereas MaxEnt classifier is used by MaxEntsoftware. So basically the all classifier have nearly the same performance.[7].

Pulkit et al. built and proposed a model which extract tweet from twitter based on the post terror activities. they made their study on terrorist attack which was occurred in URI on 18 September 2016. They considered 59,988 tweet which had taken after the attack. They consider only those tweets which has #UriAttack, #uriattack. #uriattacks. They have used the naïve bayes and SVM to extract the last Re-tweet time and number of Re-tweet[13]

SudarshanSirsat et al. proposed a technique in sentiment analysis on twitter data where they have collected reviews of the product. They have used naïve bayesalgorithm which perform better in term of accuracy and efficiency. They have extracted 200 tweets where the average length of tweet was 70.105. the aim of this research is to identify the characteristic of tweet like how many times the tweet was liked and how many times they have Re-tweet the tweet.[18]

Hetuetal.built and proposed a model in sentiment analysis on twitter data based on anaconda python. They extract the dataset from kaggle in which they classify the people emotions based on positive and negative reviews. This model gives high accuracy on large dataset.[17]

Ali hasan et al. proposed a model using the hybrid approach that comprise sentiment analyzer machine learning. They took only those tweet that is followed by the hashtag(#) and contain the current political trends. Basically this model converts the urdu tweet into English tweet. They took 1690 tweet for training data and 400 for testing the data. They have used the naïve bayes and SVM classifier for training the dataset in weka and building a model. They have used three different libraries to calculate the subjectivity and polarity.[16]

FeddahAlhumaidiAlOtaibi et al. proposed a model by using the supervised and unsupervised algorithm. They wanted to know that which restaurant has more popularity between mcDonald and KFC by using the sentiment analysis. Moreover , they extracted 7000 tweets of both the restaurant by twitter API. The tweet was in English and they used R programming language. Because R programming language can perform big computational task. They have used several machine learning techniques but they found MaxEnt has performed better result compare to other technique. Moreover they have also found KFC have many neutral tweet and McDonald have more positive and negative tweet.[14]

Table 1. Summary of studies selected for review

Author	Datasets description	Techniques	Accuracy
Geetika gautam et.al [3]	Twitter dataset on customer review	SVM Max Entropy, Naïve Bayes Semantic Analysis (WordNet)	85.5% 83.8% 88.2% 89.9%
Seyed-Ali Bahrainian et al. [12]	Twitter data On Smartphones	Unigram feature,Naïve Bayes, MaxEnt,SVM Hybrid Approach	89.78%
Neethu M. S. et al. [7]	Twitter data on Electronic products	Naïve Bayes SVM Max Entropy Essemble	89.5% 90% 90% 90%
Apoorv Agarwal et al.[19]	10,000 manually Annotated Tweets	Unigram Kernel Senti-features Unigram + Senti features Kernel + Senti features	56.58% 60.60% 56.31% 60.50% 60.83%
Dhiraj Gurkhe et al. [1]	Twitter Data	Unigram Bigram Unigram+Bigram	81.2% 15% 67.5%

6. SENTIMENT ANALYSIS ALGORITHM

To classify the text classification problem in sentiment analysis, machine learning is used. In sentiment analysis, mainly there are two categories of learning algorithm.

6.1 Supervised learning

In the field of machine learning, different classification technique are used to classify the unlabeled data. These techniques used different classifier for training the dataset. Example of machine learning classifier na ïve bayes, support vector machine, Decision Tree.These can be classified as supervised machine learning classifier which require training data set as prior. In supervised machine learning we do have several data point that describes features variable and target variable. The aim in supervised learning is to predict the final outcome variable given the predictor variable. The goal of supervised learning is to automate time consuming or expensive manual task.

6.1.1 Na ïve bayes

Na ïve bayes theorem is a classification method with the independent assumption between the predictors. In other words the approach of particular predictor of one class is not connected to closeness of some other class. Let's take an instance ,an apple may be considered a fruit if it is red in color, and if it is round in shape and if its diameter is consider to be three inches approximately. Despite of these feature are dependent on one another or in the presence of another feature. All these independent properties contribute to find the probability of na ïve classifier that this is an apple. Na ïve bayes is beneficial for big data sets and can be build easily.

6.1.2 Support vector machine

It is a supervised machine learning technique. SVM is well known used to perform in sentiment analysis. In SVM important information is represented in two vector where every vector is of size k. there is a classifier that separate the data in such a manner that margin should be maximum. SVM is used in sentiment classification and it perform better than na ïve bayes in term of classification problem.[15]

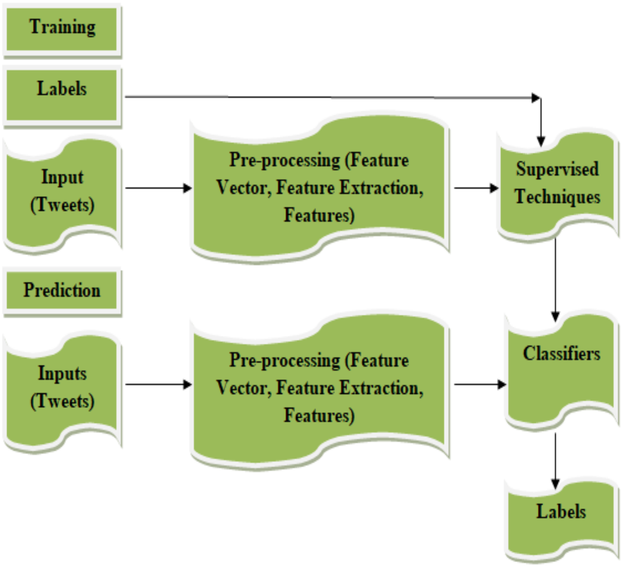
Training
Labels
Input (Tweets)
Pre-processing (Feature Vector, Feature Extraction, Features)
Supervised Techniques
Prediction
Inputs (Tweets)
Pre-processing (Feature Vector, Feature Extraction, Features)
Classifiers
Labels

Figure 3. Supervised Learning Techniques used for Training and prediction of data

6.1.3 Decision Tree

Decision tree is a classification algorithm which comes under the supervised machine learning techniques. It is a graphical representation of all the possible solutions to a decision based on certain condition. Such as for every node there is two path "should we go through it is not" and it will come until we will get our aim. In the field of prediction and classification, decision tree play major role. There are various applications of decision tree. Decision tree is used in product design when there are series of decision and the next outcome depends on the previous outcomes. Another application of decision tree is when the user has some objective he wants to get max. profit or he may want to optimize the cost.

6.2 Unsupervised learning

It is an approach of machine learning. It does not use the information that is neither classified nor labeled. it allows the algorithm to perform on that information without human guidance. The task in the unsupervised learning to grouping of unsorted data according to patterns, differences, similarities without any previous information of data. In unsupervised learning there is no human guidance provided that means there is no training of data given to the machine. Therefore in unlabeled data, machine is unable to find the hidden structure by ourselves.

6.2.1 Lexicon Based

Lexicon based approach comes under the unsupervised learning approach. But it uses a dictionary for antonyms and synonyms of sentiments phrases and words with their corresponding opinionated guidelines. Furthermore lexicon based approach is further classified into two approaches i.e. dictionary and corpus based approach which is generally used by lexicon based to find sentiment.

6.2.1.1 Corpus based

It is mainly used to find the new opinionated words from a corpus by using a list of known sentiment words. And the other main use of corpus based approach is to build a sentiment dictionary with the help of other words. Generally this approach is not dominant compare to dictionary based because it needs dictionary of all English words.

6.2.1.2 Dictionary Based

It is used to compile the opinionated words. Generally dictionaries contains list of positive and negative sentiments. The procedure for the dictionary based approach is very easy. In the first stage we manually collect the list of known positive and negative words. Then the algorithm search for the synonyms and antonyms in the wordNet or online dictionaries for growing this dataset. Later they updated the wordlist if found Followed by next iterations. Then this process continues till we got no words to update the dictionaries. Finally when this process is finished then we clean the list manually.

7. CONCLUSION

The twitter data analysis has been made on various datasets to mine the sentiments or opinions. The area of opinion mining or sentiment analysis are defined in this paper where different authors have compared various techniques on a particular dataset to achieve better accuracy and they have also found that na ïve bayes works faster than other techniques. The biggest drawback of supervised learning techniques is that it does not produces the best result when the datasest is not sufficient. However in lexicon based approach, it is necessary to have all words in dictionary related to opinion, if so not then the performance will be degraded.

8. REFERENCES

[1] Gurkhe D., Pal N. and Rishit B. "Effective Sentiment Analysis of Social Media Datasets using Na ïve Bayesian Classification." (2014).

[2] Bouazizi, M., Ohtsuki, T.: Multi-Class Sentiment Analysis in Twitter: What if Classification is Not the Answer. IEEE Access. 6, 64486-64502 (2018).

[3] Gautam, G., Yadav, D.: Sentiment analysis of twitter data using machine learning approaches and semantic analysis. 2014 Seventh International Conference on Contemporary Computing (IC3). (2014).

[4] Amolik, Akshay, et al. "Twitter sentiment analysis of movie reviews using machine learning techniques." International Journal of Engineering and Technology 7.6 (2016): 1-7.

[5] Mukherjee S., Malu A., Balamurali A.R, Bhattacharyya P."TwiSent: A Multistage System for Analyzing Sentiment in Twitter".

[6] Davidov D., Tsur O., Rappoport A." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys".

[7] Neethu, M., Rajasree, R.: Sentiment analysis in twitter using machine learning techniques. 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT). (2013).

[8] Gupta B., et al. "Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python." (2017).

[9] Jagdale R., Shirsat V., and Deshmukh S. "Sentiment Analysis of Events from Twitter Using Open Source Tool." (2016).

[10] BongirwarV. :" A Survey on Sentence Level Sentiment Analysis" International Journal of Computer Science Trends and Technology (IJCST) – Volume 3 Issue 3, May-June 2015.

[11] Behdenna S., Barigou F. and Belalem G." Document Level Sentiment Analysis: A survey"(2018).

[12] Bahrainian, S., Dengel, A.: Sentiment Analysis and Summarization of Twitter Data. 2013 IEEE 16th International Conference on Computational Science and Engineering. (2013).

[13] PulkitGarg, HimanshuGarg, VirenderRanga "*Sentiment Analysis of the Uri Terror Attack UsingTwitter*" International Conference on Computing, Communication and Automation (ICCCA2017).

[14] Sahar A. El_Rahman, FeddahAlhumaidiAlOtaibi ,Wejdan Abdullah AlShehri " Sentiment Analysis of Twitter Data".

[15] AbdullahAlsaeedi, Mohammad ZubairKhan"A Study on Sentiment Analysis Techniques of Twitter Data" *(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 2, 2019.*

[16] Ali Hasan, Sana Moin, Ahmad Karim and ShahaboddinShamshirband" Machine Learning-Based Sentiment Analysis forTwitter Accounts" 2018 by the authors. Licensee MDPI, Basel, Switzerland.

[17] HetuBhavsar, RichaManglani" Sentiment Analysis of Twitter Data using Python"International Research Journal of Engineering and Technology (IRJET) Mar 2019e-ISSN: 2395-0056 p-ISSN: 2395-0072.

[18] Prof. SudarshanSirsat, Dr.SujataRao, Dr.BhartiWukkadada"Sentiment Analysis on Twitter Data forproduct evaluation" IOSR Journal of Engineering (IOSRJEN) ISSN (e): 2250-3021, ISSN (p): 2278-8719PP 22-25.(2019)

[19] Apoorv Agarwal et al."Sentiment Analysis on Twitter Data" Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 30–38, Portland, Oregon, 23 June 2011. c 2011 Association for Computational Linguistics.