IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-8, Issue-1, Jan 2020 E-ISSN: 2347-2693

A Survey on Twitter Dataset Using Sentiment Analysis

B. Nagajothi[1*], R. Jemima Priyadarsini[2]

[1,2]Dept. of Computer Science, Bishop Heber College,Trichy-17, India

**Corresponding Author: jothiibalan8@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v8i1.9397 | Available online at: www.ijcseonline.org

Accepted: 30/Dec/2019, Published: 31/Jan/2020

Abstract -Social networking sites like twitter have millions of people share their thoughts day by day as tweets. As tweet is characteristic short and basic way of expression.There are a number of social networking sites and interrelated mobile applications, and some more are still rising. An enormousquantity of data is generated by these sites daily and this data can be used as a source for differentexamination purposes. People interrelate with each other; share their ideas, opinions, interests and personal information. These user tweet are used for finding the sentiments and also add financial, commercial and social values. though, due to the enormous quantity of user-generated information, analyzing the information manually is an expensive method. Increasing sentiment analysis activity, challenges are being added every day. Automated analytical methods are needed to extract views transmitted in user remarks. Opinion mining is the computational analysis of views transmitted in natural language for decision-making purposes. Preprocessing data play a vital role in getting accurate sentiment analysis results. Extracting opinion target words provide fine-grained analysis on the customer twwets. The labeled data required for training a classifier is expensive and hence to overcome, This paper shows opinion mining analysis types and techniques used to perform extraction of opinions from tweets. A Comparative study on the different techniques and approaches of opinion mining twitter data are dealt with in this survey paper.

***Keywords*:**Sentiment Analysis, Opinion Mining, Social Media, Twitter Data.

I. INTRODUCTION

Individuals and organizations make use for decision-making in order to improve the number of social media on the internet. Each site includes a big number of opinions that make reading and extracting data difficult for the user. The issue can be solved using the sentiment analysis techniques[1] . The principal aim of sentiment assessment is to evaluate and classify feelings and views in the user-generated reviews. There's a wide range of news blogs, twitter, etc.. insocial media on distinct products are available[2]. There's more information on the production of sentiment labels. Sentiment Analysis can summarize the opinions of the information and offer a score. The clients use this according to their requirements. A variety of sentiment analysis and opinion mining apps are available.

In finance, politics, business and government activities the field in which opinion mining is used. Sentiment analysis is used in the field of company in order to detect the customer's interest in its product[3]. Political opinion mining is used to clarify the situation of the politician. Opinion Mining also serves to discover public interest in the government's newly-applied regulations.

Motivation: Current trend is for product reviews to find views and views that are widely accessible in the social media. The outcomes of opinion assessment of the views provided by distinct customers before taking a choice are analysed.

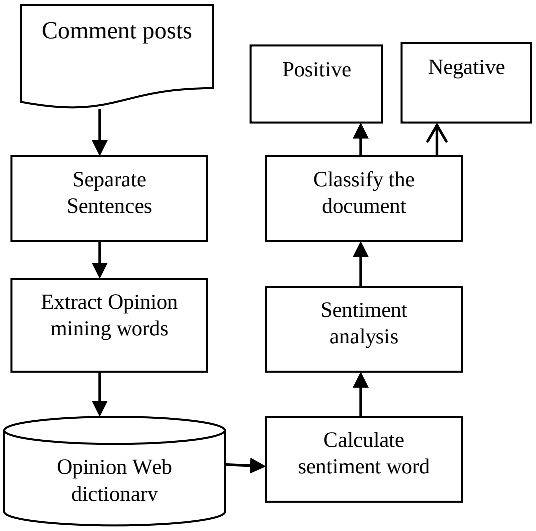

Fig 1: Architecture of Sentiment Analysis.

This enables every client to make a opinion about this item. As large-scale information is accessible, it is difficult to assess the opinions of all users. Hence the assessment of the sentiment is necessary.

The main Objective of sentiment analysis is to classify the sentiment into different categories. The general sentiment analysis architecture is shown in Figure 1. The various levels of sentiment classification are the document level, sentence level and aspect level. Document-level classification is called the classification of each document in a positive or negative class. Each paper is classified as a favorable or negative class, called the classification of sentiment at the document level. It is assumed by this classifier that the document contains the user's opinion about a single object while expressing the sentiment of a document. Aspect sentiment analyzes classify the opinion on a paper if the opinion on various aspects is expressed in a paper.

Classifiers for sentiment that use information from one domain may not be working very accurately if it is used to classify information from another. One of the primary factors is that the sentiment words of a domain may be distinct from the other. This requires domain adjustments to bridge the gaps among domains. The domain used for the training of the classifier is the source domain and the domain we use is the target domain. The benefit of this procedure is that some or no labeled information from the target domain is needed, where labeled information is expensive and unfeasible for manual labeling of opinions for each domain type. The Cross-domain Sentiment Classification is a classification of this sort. When areas of different dimension are added to the subject adaptive sensitivity classifier, heterogeneous domain adaptation is needed.

II. SENTIMENT ANALYSIS

The research region which interprets the views in text mining is known by the name of opinion mining or sentiment analysis against a specific subject, over any event etc. It creates an enormous issue area. In particular, sentiment analysis is divided into mostly three distinct levels, naming and having distinct task types, for example sentiments analysis, opinion extraction, opinion mining, sentiment mining, impact analysis, subjectivity analysis, review mining, etc.

A. Document Level Analysis: This level classifies whether there is a positive sentiment or negative sentiment in the full text. The paper is regarded on a single subject. Texts that include comparative learning can not therefore be regarded at document level.

B. Sentence Level Analysis: Sentence by sentence, the job at this stage is to decide whether each phrase is a positive, negative, or neutral view. Neutral, it is neutral when the phrase gives no view. The assessment of sentence level is linked to the classification of subjectivity. This provides factual data from subjective phrases and views. Well-bad conditions, i.e.

C. Entity/Aspect Level Analysis: Essentials and dislikes are found in both the document and the phrase level analysis. The level of entity / aspect provides the assessment. Level of the entity / aspect was called the level of the function previously. Identification structures are the key job of the entities, and aspect level directly addresses opinion or feeling. It is based on the idea that an attitude and a place for view lies in an opinion.

III. RELATED WORK

There are various text mining approaches used to mine thedata**.**

Dr.R.Jemima Priyadarsini [4] study on information mining classification methods for analysis of liver disease disorder was defined. They evaluated algorithms like C4.5, Naive Bayes, Decision Tree, Support Vector Machine, Back Propagation Neural Network and Classification and Regression Tree Algorithms. This algorithm produces different results on the basis of velocity, precision, efficiency and cost. The C4.5 outcomes in comparison with other algorithms are seen.

Dr.R.Jemima Priyadarsini [5] Presents the Thyroid Disease Technology Diagnosis Survey. These methods minimize the noisy data from the databases for the patient. For the research, the support vector machine is regarded algorithms like Naiv bayes, decision tree, back propagation, etc. The results were based on velocity, precision, efficiency and price in these algorithms. These efficient classification data are also useful in finding patient treatment.

Dr.R.Jemima Priyadarsini [6] The various kinds of clustering methods are profoundly discussed with regard to their merits and merits in distinct perspectives. K-means, DB Scan, Density-Based, Optics and EM algorithms are the multiple clustering methods regarded in this job. All algorithms have been tested with specific information sets and the outcomes have been calculated and displayed. Finally, the research shows the efficient outcome of k-means clustering using effective time complexity.

Lisa Branz [7] an method is created for study purposes to recover, analyze and interpret social networks information. The information is filtered by relevant criteria and analyzed with instruments for sentiment analysis specially adapted to the information source. The strategy is based on two examples of study issues that confirm historical results of cultural and gender variations in feelings. Hypothesis 1: There is a considerable difference between male and female software engineers in favorable feelings in tweets about sports; male presentation is more positive than female when it concerns related topics of sport. . Hypothesis 2: The sentiment expressed in tweets is significantly different from that expressing less sentiment among software engineers from collective cultures and software engineers from free spirit cultures.

Prabhsimran Singh, Ravindra Singh, and Karanjeeet Singh Kalhon, [8] The government's policy examined the demonetization from the point of view of the normal person, by using the sentiment analysis strategy and by using Twitter information, the use of certain hashtag (# demonetization) is gathered. Geolocation-based analysis (status-based tweets are gathered). The feeling assessment API from the cloud-based significance classifies the state into six categories, it's pleased, sad, very sad, very happy, neutral and without any information.

Vamshi Krishna [9] Discusses a fresh model-based strategy for opinion mineralization and the evaluation of sentimental text reviews that are mostly unstructured on the web forums or the social media page. The views of any item, individual, event or an exciting theme are exchanged in clouds in latest years. These views contribute to the choice to choose a product or receive feedback on any subject. Opinion mining and sentiment analyzes are linked in a way that opinion mining analyzes and summarizes expressed views, while sentiment analysis classifies opinions into beneficial and negative texts. For sentimental assessment, extraction is a key issue. A theme for extracting aspect and supporting vector machine learning for classifying sentiment for textual assessments is used in the model suggested in the article. The goal is to mechanize the mining process, views and hidden sentiment.

Xing Fang, Justin Zhan, [10] They have solved the problem of the categorization of feeling polarity, and this is one of the fundamental sentiment assessment issues. Data gathered from Amazon.com are used for online product reviews. This paper provides an examination of both the categorization of sentences and review levels. For this research, software from Scikit-learn is used. Scikit-learn is a software package for open source teaching. Classification techniques for the categorization of Naïve Bayesian and Random Forest and SVM.

GeetikaGautam, DivakarYadav, [11] They contribute to the customer review classification sentiment analysis. Data from twitter already labeled are used for this task. In this paper they used 3 supervised techniques: Naïve-Bayes, Max-entropy, and SVM followed by the semantitic analysis used together with all three methods for the calculation of thesimilarity. They trained and classified the following: naive Bayes, Max-entropy and SVM with Python and NLTK. Naïve-Byes approach produces a better results than the Max-entropy and SVM model with unigramm. Then, if the Word-Net of semantical analysis is applied after the above procedure, the correctness is increased.

Neethu M S, Rajasree R, [12] The information on electronic goods linked to thetwitter using MachineLearning strategy are analyzed in this document. A fresh feature vector exists for the classification of tweets and the opinion of the public on electronic goods. Thus Feature-Vector from 8 characteristics is created. A special keyword, the presence of negation, post tag, and the number of positive keywords, ematicons and negative keywords, number of negative hashtags and number of positive hashtags are all features of this system. Using integrated Matlab functions, Naïve-Bayes and SVM classifiers are implemented. Using the Maximum entropy software, Max-Entropyclassifier is used.

Gupta[13] Aim to examine certain articles on Twitter sentiment analysis studies, to describe adopted methodologies and applied models and to describe a Python-based general strategy.

Anupkant[14] to determine the challenge to the subject in relation to a document's general polarity. Subjective information processing is the primary driver of opinion mining. A sample of more than 785 quotes in the form of a sentence-based compilation was gathered. The data set of 234 opinion quotes was examined for favorable and negative characteristics after excluding neutral citations. It was noted that majority (91.6%) were positive, the proportion of quotes with adverse orientation was 8.4%, and almost 6% had positive and negative polarity opinions. The regression of the program's logistics on the course set has resulted in an accuracy of 0.98 on the classification system ; the model of regression has therefore been used on test information to show favorable and negative opinions.

Perera[15] In order to concentrate on the aspects based on restaurant reviews, they will automatically have a feeling profile of their main characteristics given their set of restaurant views. A different approach to opinion mining which utilizes SentiWordNet, twoword sentences and language rules for opinion mining together.

Angelpreethi[16] Proposed a Opinion mining analyzes a phrase written about a topic in a natural language and categorizes it as positive or negative, depending on the feelings, emotions, views of the person concerned. The views and remarks of users today are increasing daily about goods on the Internet. For other consumers to purchase one item, these remarks are useful. It is almost impossible to investigate this enormous amount of reviews physically. A machine automatically approaches the feeling or opinion division from the reviews, in order to address this complexity.

Hnin[17] The scheme suggested presents an opinion mining language method. This scheme analyzes users ' opinions in Myanmar and carries out the mining of view on the aspect level. Finally, it classifies as positive, negative or neutral aspects / features included in the reviews. The significant job for opinion mining at the level of the aspect is to identify relationships between elements and opinions. Due to casual review writing styles this detection is a major challenge. Especially because of the nature of Burma's language, it is a challenging job of aspect mining on Myanmar reviews. The scheme is therefore primarily aimed at extracting from the review user's syntactic patterns and some language rules the appropriate pairs of elements and opinions.

Zvarevashe[18]to design a sentiment analysis framework for opinion mining for the customer feedback situation of hotels. Many of the accessible hotel reviews datasets are not marked and contain a lot of research work for scientists regarding the preprocessing of text information. In addition, feeling datasets often are delicate and difficult to produce because feelings are feelings such as emotions, attitudes, and views that are frequently rife with idioms. The proposed framework is called the feeling division, which mechanically prepares a feeling dataset for instruction and testing to draw unbiased views of hotel facilities.

IV. OPINION MINING AND SENTIMENT ANALYSIS

Figure2 illustrates the workflow for sentimental assessment. The scheme comprises of four primary components: a data collection module, a data handling module, an output analysis and a classification module.

a)*Twitter Data download*

Download the tweets using Twitter API (https://github.com/aritter/twitter_download). 9684 training and 8987 testing tweets are downloaded.

b) Parser

The parser removes all engaged tweets from the downloaded data. A parser is a compiler or interpreter component that breaks data into smaller elements for easy translation into another language. After removing these we have 7612 tweets for training and 7868 tweets for testing.

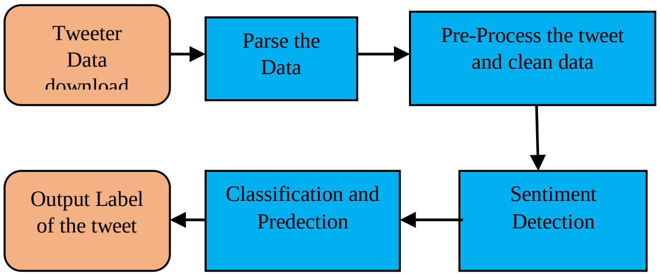
Tweeter Data download
Parse the Data
Pre-Process the tweet and clean data
Sentiment Detection
Classification and Predection
Output Label of the tweet

Fig 2: Work flow for sentiment analysis

c) Pre-processing

Data pre-processing is nothing but filtering the data to remove incomplete noisy and inconsistent data.

Following tasks are involved in the pre-processing task:

- Replace Emoticons by their polarity.
- Remove URLs and Targets.
- Expand acronyms. eg 'brb' to 'be right back'
- Remove stop words
- Tokenization
- Stemming
- Case-folding
- Remove punctuation marks
- Replace sequence of repeating characters eg. 'hellooooo' by 'helloo'

d) Sentiment detection

Sentiment term identification is an important piece of job in several sentiment assessment and opinion mining apps, such as tweet mining, opinion holders ' findings and the tweet classification. The main job of the sentiment analysis is the classification of division of a specified tweet function. It can be categorized into Positive, Negative and Neutral phrases. In three classes, the polarity is i.e. Negative and neutral, positive and negative. The identification of polarity is performed with distinct lexicons, for example. Bing lexicon Luisentiment, SentiWordNet, etc., helps to calculate feeling power, feeling score, etc.

e) Classification Algorithm

In sentiment analysis, two basic methods viz supervised approach to learning and unattended approach to learning are found. Twitter information feel classification isdone with monitored approaches to machine learning, such asNaïve-Bayes, SVM and Maximum Entropy, etc. The classifier effectiveness is determined by the dataset on which techniques are to be used. Trainingdataset is employed to assist classify the sample information when using supervised machine learning approaches to the classification model.

f)Analysis of Output:

The basic purpose of sentiment analysis is to transform unstructured information into meaningful or substantial data. The findings are presented in graph-like graphs, bar graphs, and line graphs after the assessment is concluded.

V.COMPARATIVE STUDY OF THE SENTIMENT ANALYSIS HAVING TWITTER DATASET

The Following table shows the works of various authors on Sentiment Analysis having a Twitter dataset.

Table 1: Summary of Research Articles having Twitter Dataset

S.No	**Author and Year**	**Dataset**	**Techniques**	**Accuracy**
1	Lisa Branz 2018	real-time stream of tweets	Machine learning algorithm	80% 85.9%
2	PrabhsimranSingh 2017	Tweetinvi API	Machine learning algorithm	82.5%
3	Vamshi Krishna 2018	Earthquake Reviews	Support vector machine (SVM)	90.23%
4	Fang	product review data	Python. Naïve Bayesian, Random Forest, and SVM:	73% 80%
5	Geetika Gautam 2014	Customer Review Twitter Dataset	Naive Bayes Maximum Entropy SVM Semantic Analysis (WordNet)	88.2% 83.8% 85.5% 89.9%
6	Neethu M. S. (2013)	Twitter posts about electronic products	Naive Bayes SVM Maximum Entropy Assembled	89.5% 90% 90% 90%

VI. CONCLUSION

In order to mine opinion or opinion, the analysis of Twitter information is carried out in distinct ways. This paper described the notion of feeling assessment and opinion mining for different levels of feeling assessment. This survey studied distinct sentiment analysis methods and sentiment analysis methodologies. On the analysis Twitter feelings, need to learnt about the Twitter and its structure, its importance and tweets. A short concept of tweets is given in this document. This review paper therefore discusses the vital data necessary for a feeling assessment of Twitter well. A literature study demonstrates that accuracy can be improved by using machine learning methods such as SVM, Naïve-Bayes and Maximum Entropy for the semantic assessment of WordNet. In addition, precision with the hybrid method can be improved to 4-5 percent.

REFERENCES

[1] Aaron Smith. 2018. Social Media Use in 2018. Retrieved February 25, 2018, from, http://www.pewinternet.org/2018/03/01/social-media-use-in-**2018**

[2] Mohammad, Saif M., ParinazSobhani, and Svetlana Kiritchenko. "Stance and sentiment in tweets." *ACM Transactions on Internet Technology (TOIT)* 17, no. 3: 26, **2017.**

[3] Alengadan, BletyBabu, and Shamsuddin S. Khan. "Modified aspect/feature based opinion mining for a product ranking system." In 2018 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC), pp. 1-5. IEEE, 2018.

[4] Rajam, K., and R. Jemina Priyadarsini. "A Survey on Diagnosis of Thyroid Disease Using Data Mining Techniques." IJCSMC 5, no. 5: **354-358**, **2016.**

[5] Sindhuja, D., and R. Jemina Priyadarsini. "A survey on classification techniques in data mining for analyzing liver disease disorder." International Journal of Computer Science and Mobile Computing 5, no. 5: **483-488**, **2016.**

[6] Giftson, D. Osmond Niranjan, and R. Jemina Priyadarshini. "An Extensive Analysis On Various Clustering Algorithm In Data Mining."International Journal of Computer Science and Mobile Computing, Vol.**5** Issue.**5**, pg. 273-277, **May- 2016.**

[7] Lisa Branz and Patricia Brockmann. "Poster: Sentiment Analysis of Twitter Data: Towards Filtering, Analyzing and Interpreting Social Network Data". *In DEBS '18: The 12th ACM International Conference on Distributed and Event-based Systems, June 25–29,* **2018.**

[8] Singh, Prabhsimran, Ravinder Singh Sawhney, and Karanjeet Singh Kahlon. "Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by the Indian government." *ICT Express* 4, no. 3 (): **124-129**, **2018.**

[9] Vamshi, Krishna B., Ajeet Kumar Pandey, and Kumar AP Siva. "Topic Model-Based Opinion Mining and Sentiment Analysis." In *2018 International Conference on Computer Communication and Informatics (ICCCI)*, pp. **1-4**. IEEE, **2018.**

[10] Fang, Xing, and Justin Zhan. "Sentiment analysis using product review data." *Journal of Big Data* 2, no. 1: 5, **2015**

[11] Gautam, Geetika, and Divakar Yadav. "Sentiment analysis of Twitter data using machine learning approaches and semantic analysis." In *2014 Seventh International Conference on Contemporary Computing (IC3)*, pp. **437-442**. IEEE, **2014.**

[12] Neethu, M. S., and R. Rajasree. "Sentiment analysis in twitter using machine learning techniques." In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pp. **1-5**. IEEE, **2013.**

[13] Gupta, B., Negi, M., Vishwakarma, K., Rawat, G. and Badhani, P.. Study of Twitter sentiment analysis using machine learning algorithms on Python. *International Journal of Computer Applications*, *165*(9), pp.**0975-8887**, **2017.**

[14] Anupkant, S., PVM Seravana Kumar, NayaniSateesh, and D. Bhanu Mahesh. "Opinion mining on author's citation characteristics of scientific publications." In 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), pp. **348-351**. IEEE, **2017.**

[15] Perera, I. K. C. U., and H. A. Caldera. "Aspect based opinion mining on restaurant reviews." In 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), pp. **542-546**. IEEE, **2017.**

[16] Angelpreethi, A., and S. Britto Ramesh Kumar. "An enhanced architecture for feature based opinion mining from product reviews." In 2017 World Congress on Computing and Communication Technologies (WCCCT), pp. **89-92**. IEEE, **2017.**

[17] Hnin, Cho Cho, NawNaw, and Aung Win. "Aspect Level Opinion Mining for Hotel Reviews in Myanmar Language." In 2018 IEEE International Conference on Agents (ICA), pp. **132-135**. IEEE, **2018.**

[18] Zvarevashe, Kudakwashe, and OludayoO.Olugbara. "A framework for sentiment analysis with opinion mining of hotel reviews." In 2018 Conference on Information Communications Technology and Society (ICTAS), pp. **1-4**. IEEE, **2018.**