How to Cite:
Kalaivani, M. S., Jayalakshmi, S., & Priya, R. (2022). Comparative analysis of sentiment classification using machine learning techniques on Twitter data. *International Journal of Health Sciences*, *6*(S2), 8273–8280. https://doi.org/10.53730/ijhs.v6nS2.7098

Comparative analysis of sentiment classification using machine learning techniques on Twitter data

Kalaivani. M. S
PhD, Computer Science at Vels Institute of Science, Technology and Advanced studies (VISTAS), Pallavaram & Assistant Professor in the Department of Computer Applications, Tagore College of Arts and Science, Chromepet, Chennai
Corresponding author email: kalaims2007@gmail.com

Dr. S. Jayalakshmi
Associate Professor at Vel Tech Multi Tech Dr.Rangarajan Dr.Sakunthala Engineering College (VTMT), Chennai
Email: jayalakshmi.research@gmail.com

Dr. R. Priya
Professor and Research Supervisor in the department of computer applications VISTAS (Vels University) Chennai
Email: priyaa.research@gmail.com

***Abstract*---**Sentiment analysis, a track of Natural language processing field, which is used to categorize the online content into positive, negative and neutral comments. In this pandemic situation people order various products through online and there is an option to share their feedback. People refer ratings and reviews of other customers, before buy the product .The ecommerce field has reached next level of growth .The opinion or view of the customers play vital role in the growth of a business. The organization analyzes negative comments and predicts expectation of the customers, to develop their business. With the help of this analysis, effective decisions can be made to manage critical situations in business. In recent years various methods and techniques are used to analyze customer views. Machine learning techniques are well suited for sentiment analysis and achieved effective results .In this paper, support vector machine and Naive Bayes methods are used in sentiment classification with twitter data.

***Keywords*---**sentiment analysis, machine learning techniques, support vector machine, Twitter.

Introduction

The usage of internet has become increased rapidly and the information can be exchanged through internet. Sentiment analysis is one of the techniques of Natural Language processing, which defines about analyzing the opinions or thoughts of people towards a particular product, service etc. This technology plays vital role in result prediction and decision making. Twitter, A Popular micro blog where people record their views or opinions in the format of text. Twitter comments (tweets) and contents can be taken as input for sentiment analysis. Sentiment analysis applied in variety of domains, including financial predictions [Nasukawa, T. Yi, J. 2003, B. Pang and L. Lee,2008], marketing strategies [Cambria, E.; Das, D.; Bandyopadhyay, S.; Feraco, A. 2017], and medicine analysis [Hussein, D.M.E.-D.M. 2018, Sohangir, S.; Wang, D.; Pomeranets, A.; Khoshgoftaar,2018]. Some hypothetical and scientific concerns control the level of accuracy in determining polarity score [Keenan, M.J.S.,2018]. Nowadays, a huge amount of data is available in online such as web pages, newsgroup postings, and on-line news databases. Sentiment analysis is also known as “opinion mining” which is one of the techniques of Natural Language Processing. It is applied in computational linguistics, text mining, examine emotional states etc. It is the elucidation and classification of emotion in from input data using text analysis techniques. Machine learning techniques SVM,Maximum Entrophy,Navie Bayes[Liu, 2015] and Deep learning techniques Convolutional neural networks[Pilsung Kang, Seungwan Seo ,Czangyeob kim,Haedong kim ,2020],Long Short term memory,Recurrent neural network,Deep neural network[M. U. Salur, I. Aydin ,2020] are applied in sentiment analysis and results were obtained. [Abdullah Alsaeedi1, Mohammad Zubair Khan2 ,2019]

In recent years people express Their views and ideas about any product or service through reviews and ratings.It helps to other people to purchase the product. These reviews and feedback can be used by the business people to analyse their product Demand in market and customer expectation.Thus the feedback perform an important role in decision making of customers and as well as business people. These reviews should be processed in an organised way to extract meaningful information from them. sentiment analysis, Analysis peoples reviews are feedback about any product our service To provide a meaningful data. It categorise the online information into positive negative and neutral data. Sentiment analysis categories online data in three different levels.

Document level sentiment: analysis in this level the whole document is considered as a single unit and the polarity score will be allocated to The entire document. the whole document explains about a single theme.
Sentence level sentiment analysis :In this level it sentence as classified with subjectivity and the scores allocated for each sentence the polarity of the sentence as determined as positive negative are neutral here neutral opinion means there is no opinion about product or service.
Aspect level sentiment analysis:This is known as feature level sentiment analysis, which describes about the aspect of the review. Example the screen size of the mobile is small but the clarity of the screen is good. here screen size and clarity of the screen r the aspect of the review. In recent years machine learning techniques are applied for sentiment classification an efficient results were achieved .the

unstructured form of on line data should be Normalised for text processing. this data need to be converted into a proper format to instruct a machine. the machine understand information based on training data and it predicts the results for untrained or testing data.

Importance and Background Study

There are 3 different approaches in machine learning techniques.supervised learning :in this type of learning training and testing data are labelled or already known. The training data set is used by the machine for training and based on the information testing data is labelled. then Training and testing labels are compared to obtain the accuracy of the algorithm.Unsupervised learning :in this type of learning does not have a labelled dataset. Clustering approach is followed to predict expected results. Different parameters are used in this learning, to check the performance of the give an algorithm.Semi supervised learning: in this type , A portion of dataset is labelled and reminding are unlabelled data.By using the labelled data set this approach , Predict the output for unlabelled dataset. The small labelled dataset is trained and based on the results the whole data set is tested and the results are predicted.The Input reviews can be about, book, movie,hotel,or any product or service.The recent analysis tools in the market are able to compact with remarkable capacities to deal with online data. For example, knowing the status of a company and their competitors' products are worth information for business improvement .It provides information for clarifying business and social news, such as product offers [D. Factiva, 2009], stock returns [S. R. Das and M. Y. Chen, 2007],and the results of political decisions [A.Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, 2010].

Twitter is a micro blog, a platform for people to share their views. At 2016, Twitter has more than 313 million dynamic clients inside a given month, including 100 million clients daily [Abdullah Alsaeedi1 , Mohammad Zubair Khan2 ,2019].In the social media context, sentiment analysis and mining opinions are highly challenging tasks, and this is due to the enormous information generated by humans and machines [A.Giachanou and F. Crestani, 2016].Opinions are essential to every human deed, because they are important factors of our practices. Organizations want to know about user's sentiment about their product and services. People use web based social networking sites to express their thoughts. In recent explosions in client-produced content on social sites are introducing unique difficulties in capturing, examining and translating printed content since information is scattered, confused, and divided [M. Kaplan and M. Haenlein, 2010].

Classification Techniques

Classification is the process of categorize the unlabeled data, by using the machine learning approaches. This process needs training data. Effective training for a classifier helps in predicting the results easily. Naive Bayes, Maximum Entropy and Support Vector Machine are some of the supervised machine learning algorithms, which requires training data.

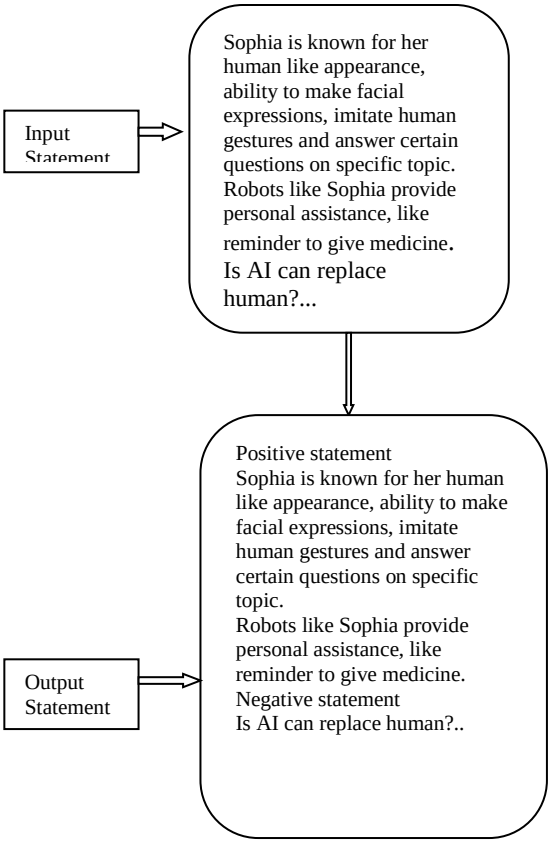

Figure 1. Classification of sentiment analysis.

a) Naive Bayes: This is a classification method that relies on Bayes' Theorem, which takes independence assumptions between the features. It expects that the nearness of a specific element in a class is disconnected to the nearness of some other elements. For example, a fruit might be considered to be an apple which is red in color; its shape is round and its size three inches in breadth. NaiveBayes methods study these properties and predict the result.
b) Maximum Entropy: The Maximum Entropy (MaxEnt) classifier evaluate the conditional distribution of a class marked a given a record using a type of exponential family with one weight for every constraint. This model maximizes the likelihood with maximum entropy, it scaling and quasi-Newton optimization are usually employed to solve the optimization problem.
c) Support Vector Machine: The support vector machine (SVM) is well known classifier for sentiment. SVM investigates information, characterizes choice limits and uses the components for the calculation, which are performed in the input space [Harb, M. Plantié, G. Dray, M. Roche, F. Trousset, and P. Poncelet, 2008]. The vectors in m size, consists the important information. At this point, each datum (expressed as a vector) is ordered into a class. Next, the machine identifies the boundary between the two classes that is

far from any place in the training samples [Pang, L. Lee, and S. Vaithyanathan, 2002].SVM classifier is considered more effective in text classification, when compare with other classifiers.

d) Ensemble methods are mostly used in text classification; it combines the multiple classifiers view, to get accurate result. There are mainly three types of ensemble classifiers defined here, weighted grouping, Meta classifier grouping and fixed grouping. Ensemble methods can be used with the SVM, MNB, random forest, and logistic regression classifiers. Chalothorn and Ellman, demonstrated that The ensemble model could produce superior accuracy of emotion classification compared to a single classifier [T. Chalothom and J. Ellman, 2015]. In their work Ensemble methods provides progress in classification accuracy for all type of datasets. [M. M. Fouad, T. F. Gharib, and A. S. Mashat, 2018].

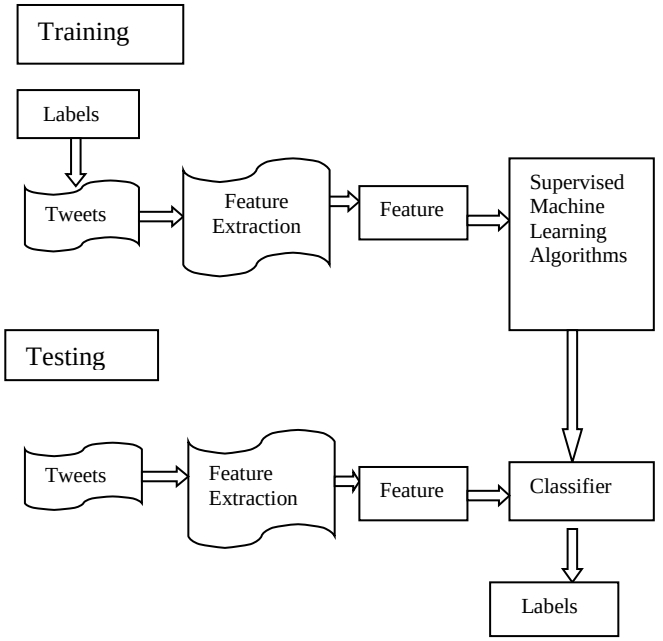

Figure 2. Procedure of supervised machine learning approaches for sentiment analysis.

e) Sentiment Analysis using Supervised Machine Learning Approaches: This method depends on labeled datasets which are given as input to machine learning models during the training process. To obtain the required output, the input datasets are trained to the systems. These datasets are classified as training data set and testing data set. Support vector machine and Naive Bayes classifiers are frequently used in sentiment analysis.

f) Musto [Musto, G. Semeraro, and M. Polignano, 2014] proposed a lexicon-based approach to identify the sentiment of any given tweet T, which began by breaking down the tweet into a number of small-scale phrases X.Hu *et al.* [X. Hu, J. Tang, H. Gao, and H. Liu, 2013] Exploited emotional signals to detect sentiments appearing in social media data. These emotional signals were defined as any information that correlated or was associated with

sentiment polarities. Machine learning, rule-based, and lexicon-based methods are combined as one system. BalageFilho and Pardo[P. BalageFilho and T. Pardo, 2013].the hybrid systems are performed well in classification when compare with individual classifiers. In another model n-gram features combined with dynamic artificial neural network. Unigram, bigram, and trigram features were recognized in sentiment analysis. This model achieved 90% accurate results in negative class. Recently, Zainuddin*et al.* [N. Zainuddin, A. Selamat, and R. Ibrahim, 2017] introduced a framework sentiment analysis, which has two major tasks. The first one about aspect-based feature extraction and second task about aspect based sentiment classification.

Experiments and Results

Data set:Airline review dataset is taken and machine learning algorithms are implemented .The dataset contains more than 9000 reviews from twitter. It has positive and negative reviews. Preprocessing techniques are applied to normalize the dataset. Data Preprocessing:In our research .sentiment classification is done with a twitter dataset .Twitter has more than 200 million users and more than 500 million tweets are shared through online. The text has hash tags, URLs, special symbols which are considered as noise and will be removed in preprocessing phase. Stop word removal, stemming are basic data reprocessing techniques done with our input dataset.Naïve Bayes is a probabilistic classifier, which gives excellent outcomes with text data analysis. In this method, probability P is defined as,

$$P(m|n) = \frac{P(m|n)p(m)}{p(n)}$$

P (m | n) is the probability of class x.
P (m) is the prior probability of class.
P (n | m) is the probability of predictor of the given class.
P (n) is the prior probability of predictor.

Support Vector Machine is an efficient learning algorithm for text classification. It has structured format for input and output data. Text scores are calculated and given to SVM as input for text categorization.In our analysis SVM Algorithm obtained effective result than Naïve Bayes algorithm. We have split the database 75 % for training and 25% for testing.SVM achieved nearly 83% accuracy and Naïve Bayes obtained 76% accuracy in sentiment classification. The below table explains about accuracy, precision values for both algorithms.

Algorithm	Accuracy	Precision
SVM	83.5	89.36
NB	76.56	88.34

Table 1 : Accuracy and Precision values of both Algorithms.

Single reviews are tested to verify the prediction of both methods. Sample Review: “Travelling @ABC gives great happiness”

Algorithm	SVM	Naïve Bayes
Actual result	Positive Review	Positive Review
Predicted Result	Positive Review	Negative Review

Table2: Result prediction of both Algorithms.

The above table explains about implementation of SVM and NB for a sample review.

Conclusion and Future Work

In this paper, we have applied machine learning algorithms for sentiment classification .Airline reviews from twitter, was taken as input dataset. Machine learning algorithms SVM and Naïve bayes methods are implemented and SVM achieved better accuracy than NB. All the problems solved by using by machine learning can be solved by applying deep learning algorithms also. In future sentiment classification can be applied in large datasets which has nearly three lakh reviews. Deep learning Techniques can be implemented with various feature extraction techniques. The NLP technique POS tagging could improve the accuracy of the models. Word embedding Technique WordToVec can be used with GloVe and the performances can be analyzed in the future work.

References

A.Giachanou and F. Crestani, (2016)Like It or Not: A Survey of Twitter Sentiment Analysis Methods," ACM Comput. Surv., vol. 49, no. 2, pp. 1-41,

A.Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe,(2010) "Predicting elections with twitter: What 140 characters reveal about political sentiment," Icwsm, vol. 10, no. 1, pp. 178-185, .

Abdullah Alsaeedi1 , Mohammad Zubair Khan2 ,2019: A Study on sentiment analysis on Twitter data, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 2, pp 63-374

Abdullah Alsaeedi1, Mohammad Zubair Khan2 ,(2019) A Study on Sentiment Analysis Techniques of Twitter Data",*IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 2,* 361 | P a g e www.ijacsa.thesai.org

B. Pang and L. Lee,(2008)Opinion mining and sentiment analysis,'' *Found.Trends Inf. Retr.*, vol. 2, nos. 1_2, pp. 1_135,.

Cambria, E.; Das, D.; Bandyopadhyay, S.; Feraco, A.(2017) A Practical Guide to Sentiment Analysis; Springer: Berlin,Germany, .

D. Factiva, 2009,Quick Study: Direct Correction Established Between Social Meidia Engagement and Strong Financial Performance,PR News, 2009.

Harb, M. Plantié, G. Dray, M. Roche, F. Trousset, and P. Poncelet, (2008)"Web Opinion Mining: How to extract opinions from blogs?," in Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology, ACM, pp. 211-217.

Hussein, D.M.E.-D.M. (2018)A survey on sentiment analysis challenges. J. King Saud Univ. Eng. Sci., 30, pg.no 330–338.

Keenan, M.J.S.,(2018) Advanced Positioning, Flow, and Sentiment Analysis in Commodity Markets; Wiley: Hoboken, NJ, USA, .

Liu,(2015) *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge, U.K.: Cambridge Univ. Press, .

M. Kaplan and M. Haenlein, (2010)Users of the world, unite! The challenges and opportunities of Social Media," Business horizons, vol. 53, no. 1, pp. 59-68,

M. M. Fouad, T. F. Gharib, and A. S. Mashat, (2018)Efficient Twitter Sentiment Analysis System with Feature Selection and lassifier Ensemble, in International Conference on Advanced Machine Learning Technologies and Applications, : Springer, pp. 516-527.

M. U. Salur, I. Aydin ,(2020)Novel Hybrid Deep Learning Model for Sentiment Classification,IEEE Access,Vol.8.March .

Musto, G. Semeraro, and M. Polignano, (2014)A comparison of lexicon-based approaches for sentiment analysis of microblog posts," Information Filtering and Retrieval, vol. 59, .

N. Zainuddin, A. Selamat, and R. Ibrahim, (2017)Hybrid sentiment classification on twitter aspect-based sentiment analysis," Applied Intelligence, pp. 1-15, .

Nasukawa, T. Yi, J. (2003),Sentiment analysis Capturing favorability using natural language processing.In Proceedings of the 2nd International Conference on Knowledge Capture, Austin, TX, USA, 4–6 December, pp. 70–77

P. BalageFilho and T. Pardo,(2013) "NILC_USP: A hybrid system for sentiment analysis in twitter messages," in Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation , vol. 2, pp. 568-572.

Pang, L. Lee, and S. Vaithyanathan, (2002)Thumbs up: sentiment classification using machine learning techniques," in Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, : Association for Computational Linguistics, pp. 79-86.

Pilsung Kang, Seungwan Seo ,Czangyeob kim,Haedong kim ,(2020) Comparative Study of Deep Learning-Based Sentiment Classification ,IEEE Access,Vol .8. January.

S. R. Das and M. Y. Chen,(2007) Yahoo! for Amazon: Sentiment extraction from small talk on the web, Management science, vol. 53, no. 9, pp. 1375-1388, .

Sohangir, S.; Wang, D.; Pomeranets, A.; Khoshgoftaar,(2018) T.M. Big Data: Deep Learning for financial sentiment analysis. J. Big Data .

T. Chalothom and J. Ellman, (2015)Simple Approaches of Sentiment Analysis via Ensemble Learning," Berlin, Heidelberg,: Springer Berlin Heidelberg, pp. 631-639.

X. Hu, J. Tang, H. Gao, and H. Liu,(2013) "Unsupervised sentiment analysis with emotional signals," in Proceedings of the 22nd international conference on World Wide Web, : ACM, pp. 607-618.