Unveiling Sentiment Analysis: Exploring Techniques and Navigating Challenges

Mani Rautela
Department of CSE
Graphic Era Hill University,
Bhimtal Campus, India-263136

manirautela.200121019@gehu.ac.in

Bhawna Tewari
Department of CSE
Graphic Era Hill University, Bhimtal
Campus, India-263136

Bhawanatewari.200121157@gehu.ac.in

Shobhit Kumar
Department of CSE
Graphic Era Hill University,
Bhimtal Campus, India-263136
shobhitkumar@gehu.ac.in

Amit Mittal
Department of Allied Science
Graphic Era Hill University,
Bhimtal Campus, India-263136
Amitforestry26@gmail.com

Abstract - **Sentiment Analysis is analyzing text and extracting information like negative or positive opinion. It is also known as opinion mining. Due to the increase in social media usage sentiment analysis has become increasingly important as there is a lot of human generated data for example comments, reviews etc. that can be used to understand the sentiments of a person. Large organizations work on sentiment analysis systems. This system helps them in real world data analytics such as customer reviews on their product, brand engagement, surveys, etc. Moreover, this review paper encompasses a concise summary of relevant research papers, providing a holistic view of the advancements in sentiment analysis. By unraveling the techniques and exploring the challenges, this paper aims to contribute to the deeper understanding of sentiment analysis, paving the way for future improvements in this field.**

Keywords - sentiment analysis, opinion mining, Machine Learning, Support Vector Machine, Naïve Bayes, data analytics

I. INTRODUCTION

We are in the 21st century and have experienced rapid increase in digitization over last few decades. People are using the internet for almost every small task that they do manually traditionally *[4]* which has led to generation of massive data. This data can be used to analyze human sentiments by using various technologies and models, which is what we call Sentiment Analysis. It is a vast growing research area as the increase of social network. We always check customer reviews before buying any product from e-commerce website or visiting any place or restaurant by reading those reviews, we make our mind whether to purchase that item or to visit place or not. As we know there are millions of reviews for a single product which makes it difficult for a user to read all those reviews. So, generating an opinion by all reviews plays a major role *[4]*. Here comes the role of sentiment analysis to reduce this problem.

Negative, Positive, or neutral aspects of a sentence can be easily understood by Sentiment Analysis. There are three levels in which sentiment classification can be done i.e. *[5]*.

a) Document level -the sentences in the document are classified in two classes either positive or negative.

b) Sentence level-It is like document level except it has one more class called neutral class.

c) Feature level- In this feature are identified and extracted from the source data.

Various techniques can be used for sentimental analysis. The two main approaches are one in which classification technique is used to classify text which is known as *Machine Learning Based Approach* and the other one is where text is matched with the positive and negative words in a dictionary to check its polarity this approach is known as Lexicon Based Approach.

II. STEPS INVOLVED IN SENTIMENT ANALYSIS

Before sentiment analysis is performed the sentimental features need to be extracted from the collected data and the data need to be in proper form for further processing *[6]*. The following are the steps involved.

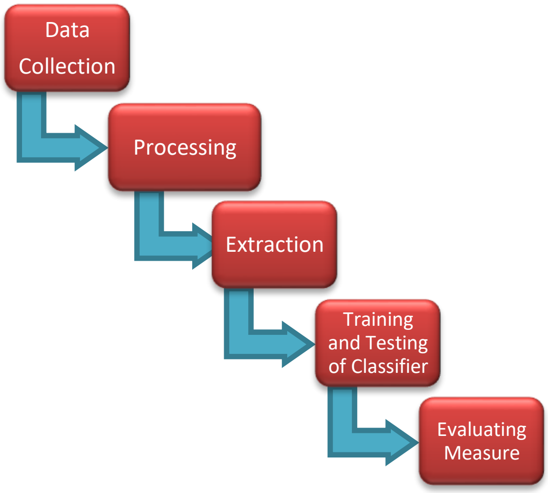

Fig:1 Steps in sentiment analysis

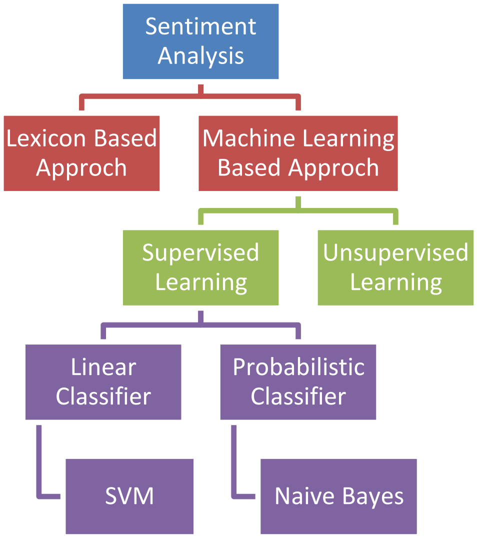

Fig:2 Sentiment Analysis approach

1. *Data Collection*: Data is selected according to the domain of interest and collected from different sources for example, social media site, articles, newspapers etc. The data for analysis is collected through various social media platform (Facebook, twitter, Instagram etc.) and e-commerce website (customer reviews).

2. *Pre-Processing*: pre-processing of data is done as tokenization i.e., splitting a text into words, phrases, other meaningful parts called tokens. Another pre-processing step involve lower case conversion which means all uppercase characters are converted to their respective lowercase *[7]*.

3. *Feature Extraction:* Extraction of features is done after the data is pre-processed which can be done by term presence and frequency, Parts of speech tagging, Negation and opinion words and phrases *[6]*.

4. *Training and Testing of Classifier:* supervised machine learning classifier is selected after the feature extraction. It is explained in section III.

5. *Evaluation Measures:* The output of the classifier is measured by precision, recall, F-measure, and accuracy. It is explained in section IV.

III. SENTIMENT ANALYSIS TECHNIQUES

The major two techniques for analyzing sentiments are: Machine Learning and Lexicon base.

1. Machine learning Approach

There are two types of Machine learning techniques i.e., supervised, and unsupervised learning but here in sentiment analysis, supervised learning techniques are mainly used.

In supervised learning the classifier is trained and tested with the help of training and test data set respectively. There are various supervised learning algorithms but according to our analysis of various research papers on sentiment analysis, commonly used algorithms are SVM (Support Vector Machine), Naïve Bayes [8].

Naïve bayes classifier

Naïve bayes is a supervised classification algorithm. It is a powerful algorithm for predictive modelling. It is based on bayes theorem. where each object is considered to be independent of each other, however the objects are individually dependent on classified objects.

Naïve Bayes formula is:

$$\mathbf{P(P|Q) = P(Q|P) * P(P)/P(Q)}$$

P(P|Q) is the probability of occurring of event P given that event Q has already occurred.

From the equation we can derive:
P (sentiment | sentence) = P(sentiment) P (sentence | sentiment) / P(sentence) *[9]*

Naïve Bayes classifier can work on small datasets and can be trained very fast as compared to other models. It performs well for classification of text as it computes the posterior probability of a class, based on the word's distribution in the document *[6]*.

SVM (Support Vector Machine)

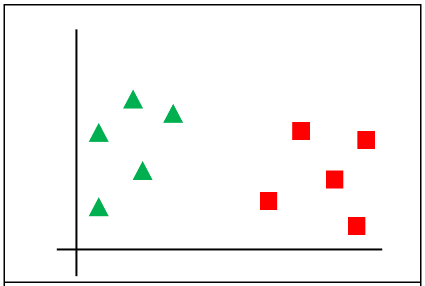

Fig:3 Two different classes

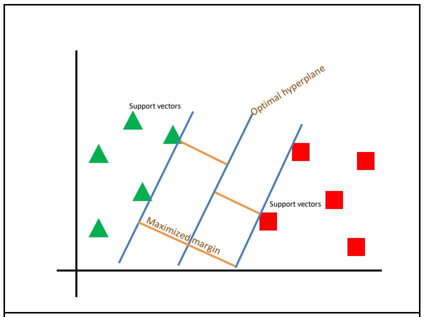

Fig:4 optimal hyperplanes separating two classes.

The output obtained by the SVM process is evaluated for accuracy using Evaluating measures mentioned in section IV *[7]*. SVM is the most widely used algorithm for sentiment analysis with the highest precision of (75.25%) *[15]*.

2. Lexicon Based Approach

This approach determines the polarity or sentiment to classify sentences in three different categories – Positive, Negative, Neutral. There are many open-source lexicons available.

The advantage It is easy to implement and doesn't require expensive training set or data.

The algorithm is a general-purpose supervised learning algorithm that can be used for both classification and regression. It generally works best for classification problem. It works on small as well as complex datasets.

The dataset is divided into classes in hyperplane and then differentiated by finding the best hyperplane performed through SVM classification. Hyper plane is line that differentiates two classes in SVM, the optimal hyperplane is set to be the one for which the support vectors are at maximum distance. The data is transformed with the help of mathematical functions called kernels. Types of Kernal are linear, sigmoid RBF, non-linear, etc.

The disadvantage is that it is not exactly accurate as sometimes few positive words can be used as sarcasm in a sentence for e.g.: "Oh great!! Now I will have to face this", this lexicon approach will consider this as a positive statement whereas its negative.

IV. EVALUATING TECHNIQUES

Cross fold Validation method is used to validate the classifier after it is tested and trained.

The effectiveness of the classifier can be evaluated by:

Terms to be used:

CP: Correct predictions

TP: True Positive

TNP: True Negative Positive

FN: False Negative

1. *Precision:* It measures prediction accuracy

   P=CP/TP+TNP

2. *Recall:* represents model's completeness

   R=CP/TP+FN

3. *Accuracy:* accuracy ranges from 70%to 90% *[6]*

   A=CP/Total Number of Prediction

4. *F-Score:* measured as, *[6]*

   FS=2*Precision*Recall/precision + recall

V. CHALLENGES OF SENTIMENT ANALYSIS TECHNIQUES

1. *Context-Dependent Error:*
   a. *Sarcasm:* - in many sentences positive words can be used as sarcasm (e.g.: oh great!! Now I will have to bare this) in such cases sentiment analysis model can consider the sentence as positive whereas its negative.
   b. Polarity- some sentences cannot be easily classified as positive negative or neutral (for e.g.: his behaviour is not mentionable) in such cases polarity of algorithms is not easily inferred.
2. *Negation Detection:*

In some sentence there may be negation (ie.no not -non-less -dis) it does not mean that the overall sentiment of the sentence is negative (e.g., the experience was not bad).

3. *Multilingual Data-*

   There are various languages used worldwide whereas the sentiment analysis model is primarily trained to analyse words in one language. This causes a problem while conducting e-commerce survey.

4. *Potential Biases in Model Training*

   Afterall the sentiment analyses models are trained by the datasets provided by humans. This means they inevitably reflect human biases in their results.

VI. SUMMARY OF RELATED WORK

AUTHOR(s)	YEAR	FINDINGS
[5]	2013	Sentiment analysis has become popular area of research and SVM has shown high accuracy in sentiment classification.
[9]	2013	Naïve Byes classifier technique is high in analysing the sentimental state of Facebook users.
[7]	2014	Benchmarked datasets were used to train the classifier. The Chi-square feature selection method was found to be effective in improving the accuracy of the data set
[16]	2014	Naïve Bayes and SVM are most used as reference models. There is more research to be made in context-based sentiment analysis. There are some datasets used for reviews in IMDB by algorithm analysis
[4]	2016	This paper proposed that big data can be implemented at industrial level. There are many research papers that are involved in the area of big data application that effectively implements big data analytics in real time and in IOT this becomes very complex.
[6]	2016	Naive Bayes and decision trees are two algorithms that can be used to classify tweets for sentiment analysis. As a result of performance measure Decision Tree shows 100% accuracy, precision, F1-Scoreand recall.
[10]	2017	Different machine learning algorithms were tested on twitter data and Naïve Bayes algorithm performs the best. Hybrid method is recommended for improving accuracy.
[19]	2018	SVM classifier is a better performing classifier in terms of accuracy where result is compared to other feature sets.
[13]	2018	10-fold cross validation and SVM provided best classifying result with better accuracy
[15]	2018	Various machine learning algorithm including supervised and unsupervised algorithm were used for text mining and sentiment analysis for various languages like Spanish and Arabic.
[3]	2018	This paper concluded that numerous data analysis methods were used for sentiment analysis where data was sourced from social networking cites newspapers photos, etc. and were grouped namely: machine learning and NLP

[11]	2018	Twitter is the biggest source of data for sentiment analysis and opinion mining which has applications in industries.
[8]	2019	The important task mentioned in this paper for sentiment analysis is using feature selection for data reduction.
[12]	2020	Semantic clustering and sentiment analysis is used in context-aware to extract user preference and feature of attraction from tourism reviews. When used f-measure the system outperforms.
[14]	2020	It is mentioned that different techniques like NLP and machine learning are used for analysing sentiments.
[20]	2020	In this for twitter sentiment analysis KNN classifier algorithm outperforms Naïve Bayes to compare the twitter sentiment analysis the results of naïve bayes and SVM are compared together for better results.
[1]	2020	Research in Sentiment analysis will grow in near future and SVM and logistic regression combined technique can be used for better performance
[2]	2020	Regression Model was build using HRS. Which predict customer interest in purchasing a particular product based on reviews.
[17]	2021	Sequence neural models a have the best performance but creates complexity for the classifier.in addition CNN also have the best performance with less complexity. For better performance we can develop a less complex model that has good pre-processing and feature extraction techniques.
[18]	2021	Lexicon based techniques work well whereas corpus-based approach is more accurate but lack generalization. The performance of the model is based on the datasets provided to machine learning and deep learning algorithms.
[21]	2022	Machine Learning algorithms can be used to examine infectious disease data and generate report to track various factors, NB classification is used for the analysis and designing reporting system in this paper .
[22]	2022	Flat sentimental and Hierarchical sentiment analysis is used based on the output. In flat sentiment analysis short text resulted in confused and inappropriate output whereas in hierarchical sentiment analysis model successfully extracted tree from short text. The Accuracy and the ability to extract an ideal tree is critically evaluated using hierarchical structure.
[23]	2022	Most papers used naïve bayes and SVM algorithms for sentiment analysis as well as for training datasets. Traditional machine learning models were accurate but due to the feature extraction technique the results where

		satisfactory however deep learning is resource intensive.
[24]	2022	Biometric technology is very exciting to the educational process, people can be identified using biometric technology in a traditional and more effective manner.

VII. CONCLUSION

This paper concludes that the sentiment data is gathered through social media networks (such as twitter, Facebook), e-commerce websites, newspapers, etc. is processed through various techniques that are namely Machine learning and lexicon based. Lexicon Based approach is not commonly used as it does not provide accurate prediction in many cases. Whereas the machine learning approach outperforms the lexicon-based approach. Various machine learning algorithms can be used in which SVM and Naïve Bayes have better accuracy. The workflow of sentiment analysis and challenges faced by the sentiment analysis techniques is also discussed in this paper. Our future view of this paper will be doing experimental research over the topic sentiment analysis where we will be comparing various experimental research with the experiment conducted by us, with the view of accruing better precision and accuracy rate.

Through the summary of related work mentioned above it can be concluded that sentiment analysis has vast scope at present as well as in future. Where various machine learning techniques can be explored such as decision tree, Neural networks etc. The research can be centralized to opinion mining, text mining and social media analytics.

REFERENCES

[1] A. S. J. Anvar and K. P. Krishna Prasad, "A Literature Review on Application of Sentiment Analysis Using Machine Learning Techniques," Int. J. Appl. Eng. Manage. Lett., vol. 4, no. 2, pp. 41–77, 2020.

[2] S. Yi and X. Liu, "Machine learning based customer sentiment analysis for recommending shoppers, shops based on customer's review," Complex & Intelligent System, vol. 1, no. 1, 2020.

[3] M. V., D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis - A review of research topics, venues, and top cited papers," Comput. Sci. Rev., vol. 27, pp. 16–32, 2018.

[4] R. Addo-Tenkorang and P. T. Helo, "Big data application in operations/supply-chain management: A literature review," Comput. Ind. Eng., vol. 101, pp. 528–543, 2016.

[5] S. M. Vohra and T. B. Teraiya, "Comparative Study of Sentiment Analysis Techniques," J. Inf. Knowl. Res. Comput. Eng., vol. 2, no. 2, 2013.

[6] A. P. Anuja and P. Dandannavar, "Application of Machine Learning Techniques to Sentiment Analysis," in 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2016.

[7] Zainuddin and A. Selamat, "Sentiment analysis using Support Vector Machine," in I4CT 2014 - 1st International Conference on Computer, Communications, and Control Technology, Proceedings, 2014, pp. 333–337.

[8] I. S. Ahmad, A. A. Bakar, and M. R. Yaakub, "A review of feature selection in sentiment analysis using information gain and domain specific ontology," Int. J. Adv. Comput. Res., vol. 9, no. 44, 2019.

[9] C. Troussas et al., "Sentiment analysis of Facebook statuses using Naive Bayes Classifier for language learning," in IEEE 2013 Fourth International Conference on Information, Intelligence, Systems and Applications (IISA), 2013.

[10] P. Baid, A. Gupta, and N. Chaplot, "Sentiment Analysis of Movie Reviews Using Machine Learning Techniques," Int. J. Comput. Appl., vol. 179, no. 7, 2017.

[11] S. Shayaa et al., "Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges," IEEE Access, vol. 1, 2018.

[12] Z. Abbasi-Moud, H. Vahdat-Nejad, and J. Sadri, "Tourism Recommendation System Based on Semantic Clustering and Sentiment Analysis," Expert Syst. Appl., vol. 114324, 2020.

[13] T. U. Haque, N. N. Saber, and F. M. Shah, "Sentiment Analysis on Large Scale Amazon Product Reviews," in 2018 IEEE International Conference on Innovative Research and Development (ICIRD), 2018.

[14] S. Malviya et al., "Machine Learning Techniques for Sentiment Analysis: A Review," SAMRIDDHI J. Phys. Sci. Eng. Technol., vol. 12, no. 2, pp. 72–78, 2020.

[15] S. Redhu et al., "Sentiment Analysis Using Text Mining: A Review," Int. J. Data Sci. Technol., vol. 4, no. 2, pp. 49–53, 2018.

[16] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Eng. J., vol. 5, pp. 1093–1113, 2014.

[17] R. R. Subramanian et al., "A Survey on Sentiment Analysis," in 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021.

[18] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," Soc. Netw. Anal. Min., vol. 11, no. 1, 2021.

[19] S. Naz, A. Sharan, and N. Malik, "Sentiment Classification on Twitter Data Using Support Vector Machine," in 2018

IEEE/WIC/ACM International Conference on Web Intelligence (WI), 2018, pp. 676–679.

[20] H. Wisnu, M. Afif, and Y. Ruldevyani, "Sentiment analysis on customer satisfaction of digital payment in Indonesia: A comparative study using KNN and Naïve Bayes," J. Phys. Conf. Ser., vol. 1444, no. 1, 2020.

[21] A. De and S. Mishra, "Augmented intelligence in mental health care: Sentiment analysis and emotion detection with health care perspective," in Augmented Intelligence in Healthcare: A Pragmatic and Integrated Analysis, 2022, pp. 205–235.

[22] A. F. Ibrahim et al., "A study of sentiment analysis approaches in short text," in Digital Transformation Technology: Proceedings of ITAF 2020, 2022.

[23] G. Aadil and S. Dadvandipour, "Traditional or deep learning for sentiment analysis: A review," Multidiszciplináris Tudományok, vol. 12, no. 1, 2022.

[24] A. De and S. Mishra, "Augmented intelligence in mental health care: Sentiment analysis and emotion detection with health care perspective," Augmented Intelligence in Healthcare: A Pragmatic and Integrated Analysis, 2022, pp. 205–235.