A survey on sentiment analysis in social media

[1]Archana Bhusara, [2]Prachi Pancholi
[2]Assistant Professor
[1,2] Department of Computer Engineering,
[1,2] L.D College of Engineering, Ahmedabad.

Abstract: We are living in an era where people use social networking sites like Twitter, Facebook, etc to share information over the internet. And Social media is arguably the richest source of human-generated text input. Opinions, feedback, and critiques provided by internet users reflect attitudes and sentiments towards certain topics, products, or services. Thus, text mining and sentiment analysis have been the focus of research due to the increasing availability of opinion data on social networking websites. Different Machine Learning techniques were used to extract knowledge from the textual text. This survey paper focuses on different Machine Learning techniques used for sentiment analysis.

Index Terms **- Sentiment Analysis, Social Media, Opinion Mining, Machine Learning.**

I. INTRODUCTION

As we know that there are various social media available, Facebook, Instagram, Twitter, and Snapchat are considered as the most crowded ones [1]. And people like to post their thoughts and opinions on social media platforms daily. A massive volume of unstructured and structured text data is being uploaded to the internet due to the rapidly growing extensive web access over the world. Nowadays, there are various social networking sites, social media are available such as Facebook, Twitter, Instagram, etc. People share their opinions with others using these social media. Some companies, start-ups use social media for marketing, customer feedback, and brand management. Currently, it is estimated to have 500 million daily active users of Instagram, 1.79 billion daily active users of Facebook, 310 million monthly active users of LinkedIn, and 373 million monthly active users of Twitter and users can be common people, politicians, industrialists, celebrities, etc. A survey shows that 500 million tweets are posted each day [25]. And all these social media user posts contain unstructured text data and that need to be classified properly to provide any useful information. The Twitter platform may even indirectly influence the traditional media agenda in critical events, social movements, as journalists gather information from tweets and retweet valuable messages shared by users [4]. Twitter is a microblogging service. Twitter is an almost unlimited source used in text classification. There are many characteristics of Twitter tweets. Twitter provides data that can be accessed freely using the Twitter API, making it easier to collect many tweets.

Today, available information contains 81% of unstructured data. Even so, analyzing those unstructured data to discover their hidden patterns is becoming a more challenging task. Data is the basic form of information from which knowledge is formed after being mined and managed. Information is usually stored in text form [3]. The purpose of text mining is to use this unformed or unstructured text data to locate trends, find the hidden pattern or sentiment. Sentiment Analysis can be used to improve customer service and marketing and serves as a measure of social media performance [4]. The text mining techniques used for sentiment analysis are classification, clustering, neural networks, and decision trees. Data pre-processing is necessary before the implementation of techniques. People share their knowledge and information through blogs, posts, and chats by writing in their language. The basic use of text mining methods is to make the text clear to make it easy for anyone to write or search properly.

II. SENTIMENT ANALYSIS

The area of study that identifies and extracts opinions on any event, product, or topic and uses them for the benefit of the business operation is known as sentiment analysis or opinion mining. There are also various names and have different tasks, e.g., sentiment analysis, opinion extraction, opinion mining, sentiment mining, affect analysis, subjectivity analysis, review mining, etc. [2]. The sentiment analysis allows exploring the mindset of the audience members and studies the state of the product from the opposite point of view. This makes sentiment analysis a great use for expanded product analytics, market research, campaign monitoring, brand monitoring, customer services, etc.

Firstly, we need to identify the subjective text and the objective text. Because only the subjective text contains the sentiments. And the objective text contains only information. E.g.:

Subjective: This product is awesome. (sentiment (awesome), so this is subjective.)

Objective: This product comes with free goodies. (There is no sentiment in the statement.)

Sentiment analysis is categorized into three levels:

Document Level:

The document-level analysis is the simplest form of classification. In this, the sentiment is extracted from the entire document, and the opinion is classified based on the overall sentiment of the opinion holder. Classification for the entire document is done as positive or negative. This approach is not suitable for a document that contains opinions about different objects such as forums and blogs [2].

Sentence Level:

The sentence-level analysis is the most fine-grained analysis of the document. Because of this, the polarity is calculated for each sentence as each sentence can have a different opinion. And the opinion classifies into positive, negative, and neutral opinions. Sentence level analysis is related to subjectivity classification. That expresses actual information with sentences that give subjective aspects and opinions [2].

Entity and Aspect level:

Entity and aspect level is used when classification concerns by identifying and extracting product features from the review data. Entity/ Aspect level was earlier also called feature-level analysis. The main goal is to identify the extracted object features and determine whether the opinion commented by a reviewer is positive, negative, or neutral [2].

III. METHODOLOGY

This section of the paper consists of Sentiment Analysis Process flow and different classification techniques used in it to analyze the sentiment. Fig. 1 shows the architecture of the sentiment analysis process.

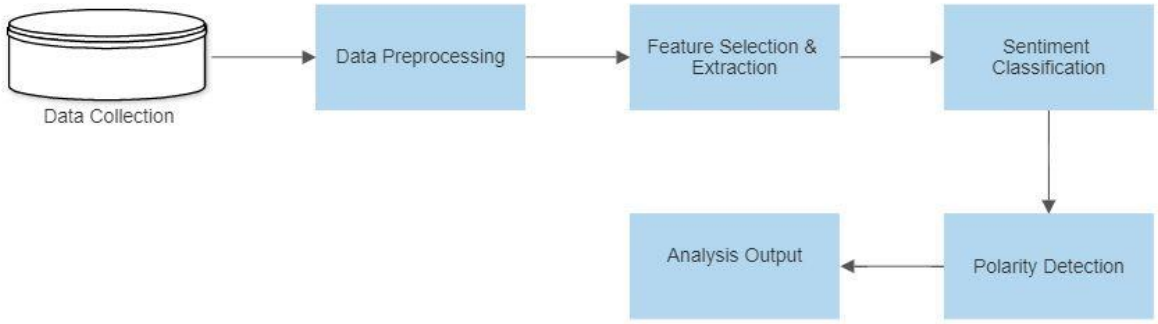

Fig 1: Sentiment Analysis Process Flow

1 Data Collection:

It is the first step of the Sentiment Analysis is to collect data on which sentiment techniques have to be performed. This dataset can be collected from social media such as Twitter, Facebook, Instagram, etc. Without data, the process can not move forward. Because only with this data analysis and classification can be performed.

2 Data Preprocessing:

After the data collection, the data needs to be preprocessed, which is a useful step to remove unwanted noise, duplicate words, and different types of URLs from text data. Stemming, tokenization, stop word removal, and filtering are the techniques that are used to perform the data preprocessing step.

3 Feature Selection & Extraction:

This step is essential to select the best features. After preprocessing, the filtered data is selected to perform extraction techniques for feature extraction. Feature extraction is important for the precision of the model. Different extraction techniques are used in sentiment analysis are POS (part of speech), bag of words, and N-gram, TF-IDF.

4 Sentiment Classification:

Sentiment Analysis uses different classification techniques to classify the extracted features. These are classified into Machine Learning approach Lexicon Based Approach. The Machine Learning approach includes various approaches under supervised and unsupervised learning such as support vector machine (SVM), Naïve Bayes, Decision Tree, Random Forest, and Rule-Based classifier. Fig 2[9] shows the Sentiment Analysis approaches.

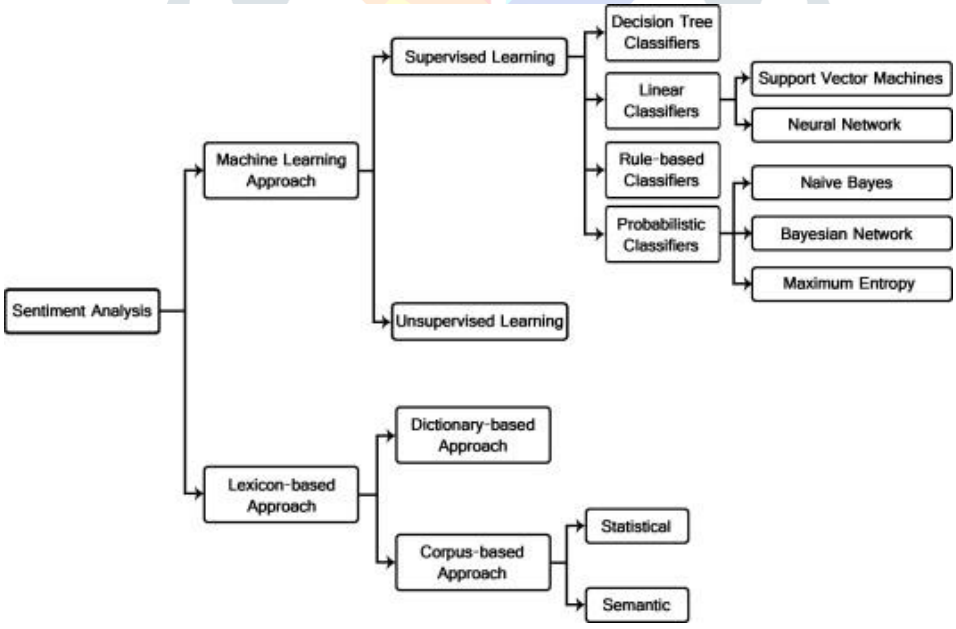

Fig 2: Sentiment Analysis Approach [9]

5 Polarity detection:

After the sentiment classification, polarity detection is done to determine the polarity of the data whether it is positive, negative, or neutral.

6 Analysis Output:

In the final step, after getting the output to analyze the performance, various performance metrics are considered like precision, recall, and accuracy.

IV. Related Work

Several studies implemented different approaches to classify sentiment or opinions. This section highlights the objectives of the studies, methods, and their outcomes.

Tianyi et al. [7] analyzed the opinion of Covid-19 pandemic on Sina Weibo, 2.4 million posts were used as a set of training and testing. For classification SVM, Naïve Bayes, BERT (Bidirectional Encoder Representation for Transformers) has been used in the experiments. Garcia et al. [23] collected 3,155,277 Portuguese tweets on Covid-19. They used machine learning methods to classify the classes into anger, sadness, and fear: Naïve Bayes, Random Forest, SVM, Logistic Regression, and Naïve Bayes gives the better result. Lapoz-chou et al. [17] analyzed 3000 Mexican tweets' opinions on natural disasters (earthquake). And classify using Machine learning methods Naïve Bayes, SVM, Decision Tree. Dhanya et al. [5] collected about 12,500 tweets with opinions about demonetization. Data were classified into positive, negative, and neutral classes using supervised machine learning methods: SVM, Naïve Bayes, Decision Tree. And SVM obtained better results, about 84%. Veny et al. [21] collected 3744 Indonesian tweets on the anti-LGBT issue in Indonesia. And classified these tweets using Naïve Bayes, Random Forest, and Decision Tree, and Naïve Bayes gives the better result. Meylan et al. [12] present a sentiment analysis on Twitter data written in the Indonesian language. The data were classified into positive, negative, and neutral classes for classification they applied K-NN, Naïve Bayes, and SVM. Naïve Bayes gives 80.90% accuracy.

Table 1 shows the summary of some previous works.

References	Objective	Language	Dataset	Algorithm	Accuracy
Tainyan et al.[7]	Covid-19 SA	Chinese	Sina Weibo -2.4 Million	Naïve Bayes(NB) Support Vector Machine (SVM) BERT	68.5% 74.5% 83%
Samuel et al. [8]	Covid-19 Public SA	English	Twitter- 4566	Naïve Bayes Logistic Regression	91.4% 74.29%
Meylan et. al. [12]	Political SA	Indonesian	Twitter- 443	NB, KNN, SVM	NB-80.90%
Dhanya et al. [5]	Demonetization SA	English	Twitter-12,500-tweets	SVM Naïve Bayes Decision Tree	84% 66.25% 63.25%
J.Rexiline et al.[15]	Disaster Response and Recovery SA	English	Twitter-70,817 tweets	Naïve Bayes SVM	-
Khan et al. [19]	Covid-19 Twitter SA	English	Twitter- 50000	Naïve Bayes	-
Lokesh Mandoli et.al. [13]	Twitter SA	English	Twitter 1000- tweets	SVM NB Maximum Entropy	NB- 86%
Hanif et. al. [6]	Customer satisfaction	Indonesian	Instagram-3800	KNN Naïve Bayes	94% 83%
Rosy Indah Permatasari et.al. [10]	Movie Review	Indonesian	Twitter- 350	Naïve Bayes	0.88 F-Measure
Sahar A. El_Rahman et.al. [14]	Food Brand SA	English	Twitter 7000-McD	NB SVM Random Forest Decision Tree Maximum-Entropy	63% 50% 33% 80% 50%
Aljameen et al. [24]	Covid-19 awareness SA	Arabic	Twitter-242525 tweets	Naïve Bayes K-NN SVM	80% 64% 85%
Khushboo et.al. [18]	Product Review	English	Twitter-5000 tweets	Naïve Bayes SVM	81%
Swaranangini et.al. [20]	Recognize Distress	English	Facebook 4731 posts	SVM Naïve Bayes Random Forest	91.49%
Veny Amilia Fitri et.al. [21]	Anti-LGBT SA	Indonesian	Twitter- 3744 tweets	Naïve Bayes Decision Tree Random Forest	86.43% 82% 82%

Ana Reyes-Menendez et.al. [22]	#METOO SA	English	Twitter-31,305 tweets	SVM	-

Fig 2: Sentiment Analysis Approach [9]

V. Analysis

This section shows the most important parameter analysis from the past research papers. Fig 3 graph shows the most data collection platforms. It shows that Twitter is mostly used for data collection.

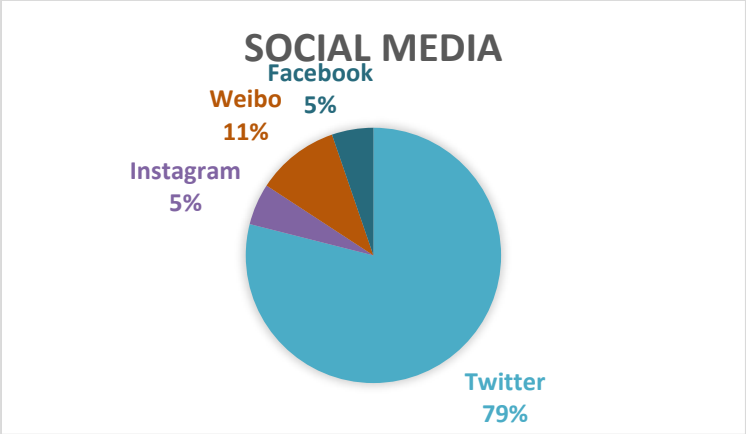

Fig 3: Social Media Platforms for Data Collection

The graph in Fig 4 shows the machine learning algorithms used in sentiment analysis classification. Naïve Bayes is the most frequently used machine learning algorithm. Thereafter, SVM and KNN are often used.

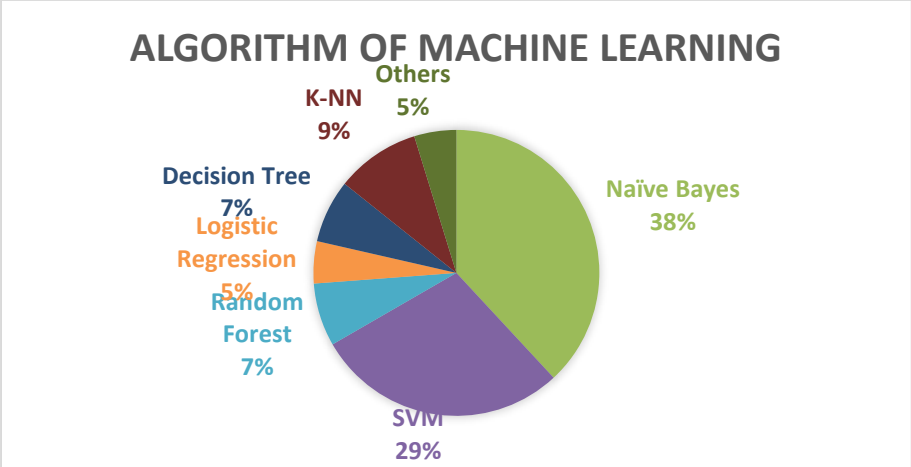

Fig 4: Machine Learning Algorithms

Conclusion:

The Sentiment Analysis plays an important role in decision making in different fields such as political, product and service evaluation, etc. Because of its high value for practical applications, there has been a tremendous growth in research studies and industrial applications. This paper defined the concept of sentiment analysis and opinion mining for various levels of sentiment analysis. And discussed different machine learning approaches and different social media such as Twitter, Facebook, Weibo, etc. for sentiment analysis.

However, most of the study uses Twitter social media datasets. In the future, we will try to use different social media data with the most used machine learning techniques, Naïve Bayes and SVM with better accuracy.

REFERENCES

[1] Said A. Salloum, Mostafa Al-Emran, Azza Abdel Monem, Khaled Shaalan, "A Survey of Text Mining in Social Media: Facebook and Twitter Perspectives", ASTESJ-2017.
[2] Rasika Wagh, Payal Punde, "Survey on Sentiment Analysis using Twitter Dataset", IEEE- ICECA, 2018.
[3] Said A. Salloum, Chaker Mhamdi, Mostafa Al-Emran, Khaled Shaalan1, "Analysis and Classification of Arabic Newspapers' Facebook Pages using Text Mining Techniques", IJITLS-2017
[4] Srishti Vashishtha, Seba Susan, "Fuzzy rule-based unsupervised sentiment analysis from social media posts", ELSEVIER-2019.

[5] N. M. Dhanya, U. C. Harish," Sentiment Analysis of Twitter Data on Demonetization Using Machine Learning Techniques", Springer-2018.
[6] Hanif Sudira, Alifiannisa Lawami Diar, Yova Ruldeviyani," Instagram Sentiment Analysis with Naive Bayes and KNN: Exploring Customer Satisfaction of Digital Payment Services in Indonesia", IEEE-2019.
[7]] Tianyi wang, Ke Lu, Kam Pui chow, and Qing Zhu, "COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model", IEEE-2020
[8] Jim Samuel, G. G. Md. Nawaz Ali, Md. Mokhlesur Rahman, Ek Esawi and Yana Samuel, "COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification", Inf. 2020
[9] Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal (Science Direct), Vol. 5, No. 4, 2014, pp.1093-1113.
[10] Klaifer Garcia, Lilian Berton, "Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA", Elsevier- Applied Soft Computing Journal- 2020.
[10] Rosy Indah Permatasari, M.Ali Fauzi, Putra Pandu Adikara, "Twitter Sentiment Analysis of Movie Reviews using Ensemble Features Based Naïve Bayes" – IEEE-2018
[11] Atiqur Rahman, Md. Sharif Hossen," Sentiment Analysis on Movie Review Data Using Machine Learning Approach" - IEEE 2019
[12] Meylan Wongkar, Apriandy Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter" – IEEE 2019
[13] Lokesh Mandoli, Ruchi Patel, "Twitter Sentiments Analysis Using Machine Learning Methods" – IEEE 2020
[14] Sahar A. El_Rahman, FeddahAlhumaidi AlOtaibi,Wejdan Abdullah AlShehri, "Sentiment Analysis of Twitter Data" – IEEE 2019
[15] J. Rexiline Raginia, P.M. Rubesh Anandb, Vidhyacharan Bhaskarc, "Big data analytics for disaster response and recovery through sentiment analysis", International Journal of Information Management, Volume 42, 2018, Pages 13-24
[16] Ranu Lal Chauhan, " Sentiment Analysis of Pulwama Attack Using Twitter Data"- Springer 2020 pp 119-126
[17] Asdrúbal López-Chau, David Valle-Cruz, and Rodrigo Sandoval-Almazán, Sentiment Analysis of Twitter Data Through Machine Learning Techniques, Springer Switzerland,2020
[18] Khushboo Gajbhiye, Neetesh Gupta, "Real-Time Twitter Sentiment Analysis for Product Reviews Using Naive Bayes Classifier" – Springer 2020, pp. 342-350
[19] Khan, Rizwan & Shrivastava, Piyush & Kapoor, Aashna & Tiwari, Aditi & Mittal, Abhyudaya. (2020). SOCIAL MEDIA ANALYSIS WITH AI: SENTIMENT ANALYSIS TECHNIQUES FOR THE ANALYSIS OF TWITTER COVID-19 DATA. Journal of Critical Reviews. 7. 2020
[20]Swaranangini Sinha, Kanak Saxena, Nisheeth Joshi, "Sentiment Analysis to Recognize Emotional Distress Through Facebook Status Updates", Springer- 2020 pp.799-811.
[21] Maria L. Loureiro a, Maria Allo b, "Sensing climate change and energy issues: Sentiment and emotion analysis with social media in the U.K. and Spain" Elsevier-2020
[21] Veny Amilia Fitri, RachmaditaAndreswari, Muhammad AzaniHasibuan, "Sentiment Analysis of Social Media Twitter with Case of Anti-LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm", Elsevier-2019
[22] Ana Reyes-Menendez a, Jose Ramon Saura a, Ferrao Filipe b "Marketing challenges in the #MeToo era: gaining business insights using an exploratory sentiment analysis", Elsevier-2020
[23] Klaifer Garcia, Lilian Berton, "Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA", Applied Soft Computing Journal- 2020.
[24] Aljameel, S.S.; Alabbad, D.A.; Alzahrani, N.A.; Alqarni, S.M.; Alamoudi, F.A.; Babili, L.M.; Aljaafary, S.K.; Alshamrani, F.M. A Sentiment Analysis Approach to Predict an Individual's Awareness of the Precautionary Procedures to Prevent COVID-19 Outbreaks in Saudi Arabia. Int. J. Environ. Res. Public Health 2021, 18, 218.
[25] Statistics for social media users available at https://www.omnicoreagency.com/instagram-statistics/