Consumer Approach to AI Driven Sentiment Analyzer: A Literature Review

KAVYA P[1], ANANYA M HEGDE[2], AKANKSHA GILIYAL[3], GEETHIKA S[4], DR POOJA NAYAK S[5]

[12345] kavyatmk2001@gmail.com, anu.mhegde@gmail.com, akankshagiliyal3@gmail.com, geethikasampath11@gmail.com, pooja-ise@dsatm.edu.in

[1234] Student, Department of Information science and engineering, DSATM, Bangalore-88, Karnataka

[5] Faculty, Department of Information science and engineering, DSATM, Bangalore-88, Karnataka

Abstract*—Customers are valued by a business not just for their financial impact, but also for how satisfied they are with the service they receive and it is subjective. Positive word-of-mouth is disseminated by satisfied consumers, and negative word-of-mouth by disappointed ones. Due to subjectivity, it is vital to examine a variety of perspectives rather than just one that conveys one person's subjective viewpoint. In addition to the abundance of sources, the volume of data makes it impossible to manually sort through them to find the underlying trends, issues, or reasons of (dis)satisfaction. Sentiment analysis is a potent tool that enables users to both extract the necessary data and aggregate the overall sentiments of the reviews. For completing this goal, a number of strategies have gained attention in recent years. This paper examines the various Sentiment Analysis strategies of machine learning such as K-NN classifier, Naive Bayes classifier, Support Vector Machine (SVM), and Neural Networks.

***Keywords*—sentiment-analysis, consumer, classification, machine learning, KNN, Naive Bayes, SVM.**

1. I. INTRODUCTION

The past few years have witnessed a drastic increase in the use of the Internet, especially in the domains of social media and e-commerce. This has led to consumers using a plethora of online resources to express their opinion on a diverse range of topics. Take Twitter, for example. It is one of the most popular platforms to voice opinions, and has claimed that there are over 200 million Tweets or posts per day [16]. Thus, the behaviour of consumers towards various information available with respect to expressing opinions on products and services has changed significantly.

Customers now rely heavily on online word-of-mouth reviews while making purchase decisions. This has resulted in a vast variety of opinionated data that can be exploited by organisations which are user-centric in order to gain useful insight on their customer demographic. Private enterprises, being producers, also increasingly exploit social media resources to commercialise their goods and services, as a marketing strategy. These platforms are a generous source of consumer feedback, as well as a fruitful way of advertising. A lot of consumers are also hesitant to fill out surveys or feedback forms, but find it easy to express thoughts freely on online platforms, which in turn colours opinions of other people.

These organisations put heavy emphasis on any potential feedback they observe, so that they can optimise their business operations, whilst also maintaining an online reputation. This is crucial to the company's image, and may affect the way the consumers interact with them.[14]

Sentiment analysis can be seen as the process of determining the underlying information that is of subjective material, usually opinions or sentiment, in a body of text. It is also known as opinion mining, and sees the use of NLP or Natural Language Processing. The stream of computer science, or more precisely, the stream of AI, that deals with enabling computers to gain the ability to understand and interpret the language spoken by humans is known as Natural language processing (NLP).

Polarity is a measure that can be used to categorise the sentiment present in comments, feedback, or criticism, which serve as valuable indications for a variety of applications [3]. It is frequently used to determine an overall positivity or negativity of a review based on its polarity. For instance:

1) Positive Sentiment expressed in a subjective statement: "I love when apples are sweet"—Based on the sentiment threshold value of the word "love," we may infer that this statement expresses positive sentiment toward the movie Mary Kom. As a result, the threshold value for the word "love" is positive.

2) Negative sentiment expressed in subjective statements: The phrase "Sour apples are very distasteful" expresses disapproval of the fruit. The sentiment threshold value of the word "distasteful" allows us to determine that the person is unsatisfied. As a result, the threshold value for the word "distasteful" is negative.

For sentiment analysis on formal texts, a variety of methods—including machine learning (ML) techniques, sentiment lexicons, hybrid approaches, etc.—have proven beneficial. However, it is necessary to investigate their efficacy for extracting sentiment from Internet data.

The polarity of the words or phrases in a particular text is used in lexicon-based techniques to determine the sentiment of that text. A lexicon (dictionary) of words with polarity ascribed to them is needed for this strategy. SentiWordNet, General Inquirer Lexicon3, Loughran McDonald Lexicon, Opinion Lexicon, AFINN Lexicon, and NRC-Hashtag are a few examples of the lexicons that are now in use. Machine learning approaches for classifying text, on the other hand, make use of supervised and unsupervised algorithms that analyse data that has already been classified as positive, negative, or neutral; they then extract features that represent the distinctions between the classes, and they infer a function that can be used to categorise newly discovered examples. Some commonly used learning algorithms to classify text are Naive Bayes, Support Vector Machines (SVMs), Decision Trees, and K-NN Classifier, among others. Barbosa et al. has reported better results for SVMs [5] while Pak et al. observed better performance using Naive Bayes [6].

1. II. RELATED WORK

Natural language processing (NLP)'s study area, sentiment analysis, deals with the process of identifying and extracting sentiment and opinion from text as well as classifying that sentiment. Sentiment analysis investigates the opinions, assessments, and feelings and attitudes toward people, groups, things, movies, problems, events, etc[10].

User reviews that are easily accessible. Even while user reviews can greatly aid in enhancing the accessibility of even well-known apps [9], about 98.76% of users fail to provide app stores with comments on accessibility issues. Surprisingly, 1% or so of mobile app users provide reviews on accessibility to aid in app enhancements in the future.Eler et al research .'s [9] used 214,053 mobile app reviews to find accessibility input. The dataset mentioned is used in our investigation. Only 2,663 mobile app reviews from the accessibility-focused research were found following the manual check, therefore they were the ones we used for our study.The dataset mentioned is used in our investigation. Only 2,663 mobile app reviews from the accessibility-focused research were found following the manual check, therefore they were the ones we used for our study. One of the few studies to investigate the preliminary dataset created by Eler et al. using sentiment analysis is ours.

Textual classification is used. Depending on their goals, many researchers have utilised different taxonomies to categorise their reviews [9]. As an illustration, some studies group their ratings according to types of feedback including complaints, problem reports, and future feature suggestions. Many of them, meanwhile, fail to address or even focus on accessibility.Numerous earlier studies employed predetermined keywords to categorise papers, which is a departure from automatic classification methods. In contrast to Ratzinger et al. [9] who utilised 13 keywords, Eler et al. used 213 keywords to evaluate user reviews. In contrast to past studies, ours employs sentiment analysis to comprehend app users' perceptions of the accessibility of the apps in order to comprehend the users' emotions (positive, negative, or neutral) when reporting on accessibility. AlOmar et al. conducted a study that is comparable to ours by using automated machine learning to examine accessibility user reviews. In this work, we assess accessibility user reviews in a chosen database using sentiment analysis. This study is the first to employ such a strategy that we are aware of [9].

Consumer insight is the study of the psyche and emotions of the consumer [16]. It might be the revolution in customer relationship management or customer experience management from direct marketing to database marketing. Businesses can use this knowledge to change their marketing tactics, enhance operations, and enhance interactions. Good customer relationship management is built on solid consumer intelligence.

Over time, various models have been developed through the use of sentiment analysis. Machine learning classifiers have been created to generate polarity of a feeling at the textual level . The semantic organisation of phrases can be examined using a method that uses the Pointwise Mutual Information (PMI) score. For the sentiment analysis of data such as video and audio, a variety of methodologies have been utilised and developed over time [12]. Customer reviews have been analysed using NLP-based algorithms to produce feature-based

summaries. The polarity of expression has been generated and the opinion on a particular product has been mined from the web using NLP techniques. In order to determine the polarity of a statement using sentiment analysis at the phrase level, a machine learning technique has been presented. For the purpose of extracting summaries for customer reviews that are based on features, an unsupervised system has been created. To determine whether a noun can be utilised as a feature of a product evaluation, the PMI score was determined. Using two distinct corpora, Latent Semantic Analysis (LSA) and PMI are examined as scoring methods, and the results show that LSA is more accurate at classifying semantic orientation [12].

2. III. PROCESS OF CONSUMER SENTIMENT ANALYSIS

Data Gathering is the first step in processing consumer sentiment. In order to analyse and classify the data, we must first collect it from the appropriate channels. This could include Twitter,scraping reviews

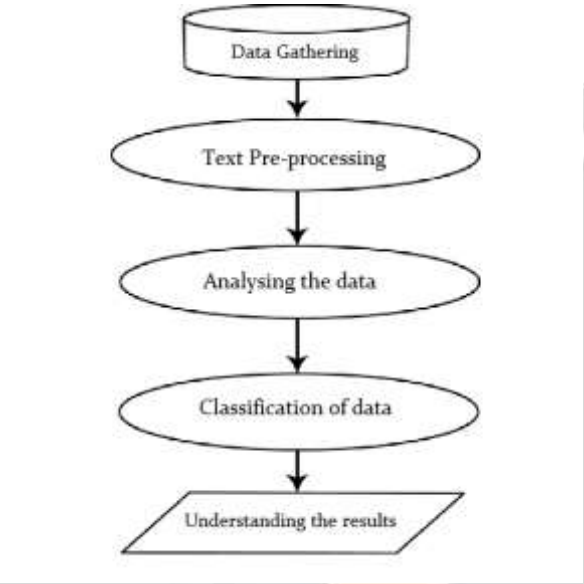


Fig. 1. Sentiment analysis process

on websites, Google reviews,etc. This would involve searching for the particular brand or company reviews being analysed. This data, after collection must be put together in a form that suits the data mining techniques well.

TABLE I: Statistics of the dataset.

Number of Apps	701
App Categories	15
All Reviews	214,053
Accessibility Reviews	2,663

Text Cleaning is the next step implemented in processing data. Text cleaning can involve removing punctuations as well as conjunctions and emojis in sentences that do not serve much purpose in the analysis and subsequent classification of data. The data can be converted entirely to lowercase and numbers as well as extra spaces are excluded.

• Tokenization: This method involves dividing the original text into tokens devoid of white space. Tokenization of the app reviews involves dividing them into a constituent set of words.

• Lemmatization: is a process in which a word's suffix is either changed or eliminated to reveal the word's original form. Additionally, it lowers the number of distinct instances of related words. The suggested method implements the above mentioned methodology to process words in their canonical form beforehand in order to decrease the number of unique occurrences of related text tokens.

•Stop-Word Removal: Words that do not contribute to the classification process, e.g., am, the, etc., are removed.

•Case Normalisation: The entire text must be converted to lowercase since similar-sounding words in different font cases, such as "accessibility" and "accessibility," must be treated identically. It can commonly be referred to as a type of data cleansing that can be used to prevent the repeating of features that are just different, case-sensitivity wise. The uppercase letter "D" can be used by a user to identify himself as "Deaf" in the context of accessibility-related reviews in order to portray his cultural identity. Since our classifier is binary, it will showcase the same classification outcome for "Deaf" and "deaf," the case normalisation will be secure, and user expressions will be prevented from being overruled.
•Noise removal: In this step, any noise that could impair classification performance or confuse the model during learning is eliminated. Numerical data, email addresses, and special characters are among the noise kinds that are eliminated in this step.

Analysis and classification of data is achieved using various sentiment analysis algorithms.The information can be categorised into more complicated emotions like anger, sadness, etc. or into a more general spectrum of "positive" and "negative" attitudes.he algorithms identify and categorise opinions using a sentiment library.

Understanding of the results should be acquired at the end of the sentiment analysis process wherein the data should be appropriately grouped according to the categories specified. The timeline of the sentiments is crucial in order to observe when a classification of sentiments had more precedence. This could include having vastly negative reviews during a certain period or positive sentiments observed in a different period since launch or release.

3. IV. TECHNIQUES

In this section, we will be surveying various approaches of sentiment analysis using Machine Learning — mainly K-NN classifier, Naive Bayes classifier, Support Vector Machine (SVM),and Neural Networks.

4. 1) *K - Nearest Neighbour Classifier*

This machine learning technique is a supervised training method that is commonly used for classification and regression. The Nearest Neighbour algorithm is one that classifies objects based on the proximity of the object to the training data. This idea is extended by the KNN classifier by considering k of the nearest neighbourhood points. Larger k values aid in lessening the impacts of noisy points in the training data set, and cross-validation is frequently used to choose k. [7] However, larger the size of k, smaller is the distinction between groups. The accuracy of the KNN classifier is determined by whether or not features that are pertinent to the given classification are present, or if the object to be classified is irrelevant. As there are no prior assumptions made before using this algorithm, it is known as non-parametric. This approach is efficient for grouping since it is relatively straightforward, simple to describe, able to train on noisy data, and durable.
For sentiment analysis, this technique can be paired with a method to calculate the weight or polarity of the measure to classify objects, and generally gives results with low error rate. In one experiment conducted, the credit approval testing data set for different values of k was classified, and it was found that for k=5 they got the highest rate of correct classification. At k=5, the (%) of error classification was 9.45 (%) [7].

5. 2) *Naive Bayes Classifier*

Used in probability theory and statistics, Beyes' theorem takes into account previous instances of conditions that may be related to an event, and illustrates the probability that the event takes place.
Bayes' theorem can be stated mathematically as the equation below:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

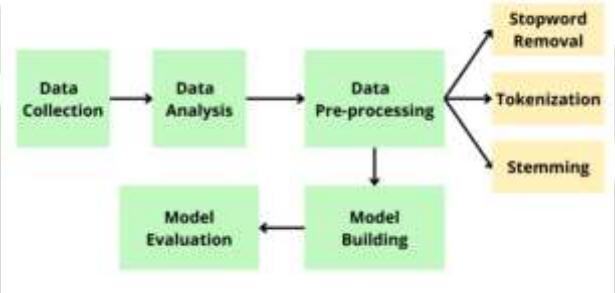
where A and B are events and $P(B) \neq 0$

- P(A|B) is the probability of event A occurring when B is true. It is called the conditional probability or posterior probability of A given B.
- P(B|A) the probability of event B happening when A is true. It can also be explicated as the likelihood of A occurring given B, since P(B|A) = L(A|B). [13]

Naive Bayes Classifier is an extension of this theorem, which is used for building classifiers. When applied to problem cases, which are represented as a collection of feature values, these models assign problem cases a class label that is selected from a small pool of possible labels. There isn't just one technique for training classifiers like this, but rather a section of algorithms built on the presumption that the value of the features are independent of the value of every other feature, given the class variable. [13]
In the case of sentiment analysis, using prior data such as the table below, a Naive Bayes model will attempt to determine how these sentiments are categorised.

TRUE SENTIMENT	TEXT
POSITIVE	The food was good.
POSITIVE	The food tasted really good
POSITIVE	The service was good
POSITIVE	The price was reasonable
NEGATIVE	The location was not good

In this example, it may conclude that the word “good” is highly probabilistic of having positive connotations. Probabilistic values such as these will then be used to assign positive or negative classes to the test data. [10]



3) Support Vector Machine (SVM) Algorithm

A commonly used machine learning algorithm for classification and regression problems is the Support Vector Machine algorithm, or SVM algorithm. This algorithm is used for the categorization of objects, where it creates boundaries for each category in order to segregate the objects into relevant classes. These boundaries are called the best decision boundary or hyperplane. The points or vectors at the extremities are chosen by the SVM algorithm, and are called support vectors, which gives the algorithm its name.

Data that can be separated into two classes by a single straight line are used for linear SVM. The classifier used is referred to as a Linear SVM classifier, and this type of data is known as linearly separable data.

For non-linearly separated data, non-linear SVM is utilised. A dataset is deemed non-linear if it cannot be categorised along a straight line, and the classifier used is referred to as a Non-linear SVM classifier.

A variety of approaches may be used to incorporate this technique in the sentiment analysis process.In one experiment, the objects were classified as positive or negative using a clustering and SVM classification combination. In comparison to the other algorithms, this experiment had the best performance, at 90.99%., and was the highest as compared to the other algorithms used. [17]

4) Neural Networks

We can integrate sentiment analysis tools into deep learning models to fully utilise their capability. Deep learning is a branch of machine learning that mimics how the human brain functions by using "artificial neural networks" to interpret data. Algorithms can be used in a sequential proceeding of events to solve complex problems. This is done in Deep learning, which is a hierarchical machine learning that allows you to process enormous capacity of data accurately and with minimum input from humans. Machine learning models are capable of amazing feats after they have been properly taught to efficiently teach themselves. The deep learning model can perform sentiment analysis on a large scale,ranging from social media posts to online customer reviews,etc.,

The following theoretical aspects explain why neural networks are advantageous. In the first place, neural networks are driven by data, and are able to adapt by themselves depending on the data, not requiring any direct operational directions for the model that is underlying. Second, neural networks may approximate any function with random precision, making them universal functional approximations. This underlying function has to be

accurately identified, and is without a question, most significant since every classification technique looks for a functional relationship between the group membership and the properties of the item.[13]

6. V. CONCLUSION

Customers' attitudes and behaviours are greatly influenced by social media platforms and websites with online product/service evaluations. These platforms also boost customer confidence in the brand, directly affect which rival customers choose, and affect the process of acquiring new clients. A thorough analysis of user-generated content, particularly the analysis of the emotions concealed therein, can give businesses essential information that will help them grow, as it frequently includes feedback on products, services, or the company itself in the form of expressed opinions and attitudes. Unstructured content from social media sites can be evaluated by companies using sentiment analysis, including assertions of the prevailing viewpoint, attitudes, and emotions aimed at a certain entity (e.g. a specific product or a product characteristic).This study has examined a number of sentiment analysis techniques and its many levels of sentiment analysis. Our ultimate goal is to create a Sentiment Analysis model that can effectively categorise different reviews. In this article, we explored a few machine learning approaches, including the K-NN classifier, the Naive Bayes classifier, the Support Vector Machine (SVM), and neural networks.

7. VI. REFERENCES

[1] D. V. Lindberg and H. K. H. Lee, “Optimization under constraints by applying an asymmetric entropy measure,” *J. Comput. Graph. Statist.*, vol. 24, no. 2, pp. 379–393, Jun. 2015, doi: 10.1080/10618600.2014.901225.

[2] B. Rieder, *Engines of Order: A Mechanology of Algorithmic Techniques*. Amsterdam, Netherlands: Amsterdam Univ. Press, 2020.

[3] Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment analysis of review datasets using naive bayes and k-nn classifiers. arXiv preprint arXiv:1610.09982.

[4] Stone, M., Bond, A., and Foss, B.: CONSUMER INSIGHT How to use data and market research to get closer to your customer: Kogan Page, 2008.

[5] Barbosa, L. and Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING ’10, pages 36–44, Stroudsburg, PA, USA. Association for Computational Linguistics.

[6] Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10), Valletta, Malta. European Language Resources Association (ELRA).

[7] Islam, M. J., Wu, Q. J., Ahmadi, M., & Sid-Ahmed, M. A. (2007, November). Investigating the performance of naive-bayes classifiers and k-nearest neighbour classifiers. In 2007 international conference on convergence information technology (ICCIT 2007) (pp. 1541-1546). IEEE.

[8] Pranali Borele , Dilipkumar A. Borikar .An Approach to Sentiment Analysis using Artificial Neural Network with Comparative Analysis of Different Techniques

[9] W. Aljedaani, F. Rustam, S. Ludi, A. Ouni and M. W. Mkaouer, "Learning Sentiment Analysis for Accessibility User Reviews," 2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW), 2021, pp. 239-246, doi: 10.1109/ASEW52652.2021.00053.

[10] Jagdale, Rajkumar & Shirsath, Vishal & Deshmukh, Sachin. (2019). Sentiment Analysis on Product Reviews Using Machine Learning Techniques: Proceeding of CISC 2017. 10.1007/978-981-13-0617-4_61.

[11] Y. Woldemariam, "Sentiment analysis in a cross-media analysis framework," 2016 IEEE International Conference on Big Data Analysis (ICBDA), 2016, pp. 1-5, doi: 10.1109/ICBDA.2016.7509790.

[12] S. Gupta, S. Lakra and M. Kaur, "Sentiment Analysis using Partial Textual Entailment," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019, pp. 51-55, doi: 10.1109/COMITCon.2019.8862241.

[13] Garg, B. (2013). Design and development of naïve bayes classifier.

[14] Olivera Grljević,aZita Bošnjak.Sentiment Analysis Of Customer Data

[15] Gil-Pita, R., & Yao, X. (2008). Evolving edited k-nearest neighbor classifiers. International Journal of Neural Systems, 18(06), 459-467.

[16] Chamlertwat, W., Bhattarakosol, P., Rungkasiri, T., & Haruechaiyasak, C. (2012). Discovering Consumer Insight from Twitter via Sentiment Analysis. J. Univers. Comput. Sci., 18(8), 973-992.

[17] Kumari, U., Sharma, A. K., & Soni, D. (2017, August). Sentiment analysis of smart phone product review using SVM classification technique. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) (pp. 1469-1474). IEEE