Implementation of An AI Driven Sentiment Analyzer

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

[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,classification,machine learning, SVM.

1. INTRODUCTION

Sentiment Analysis is understood to be the dominion of being cognizant of as well as analysing sentiments, whether it be in text or voice. The importance of recognizing underlying sentiments from any media plays a crucial role in multiple sectors. If we consider the domain of mental health, sentiment analysis can be used to detect people's emotions and apply for helpful aid if necessary. Another aspect that makes sentiment analysis widely sought after is the business side of things. Consumer data is invaluable to businesses that rely on keeping customers happy. With advancements in technology, the employment of sentiment analysis by organisations is also rapidly keeping up. Analysing consumer data is generally considered a best practice with most companies.This sentiment can be identified from online reviews on websites, forums, as well as social media platforms like Twitter and Reddit.

Sentiment analysis involves categorising the media analysed into positive,negative,neutral etc emotions.With the boom of artificial intelligence, many algorithms have been fine tuned in order to make the analysis more accurate. Some of them are Naive Bayes classifiers,Bayesian Networks, KNN and Support Vector machine algorithms. Apart from artificial intelligence, lexicon based approaches are also employed for analysing sentiments. Lexicon based approaches employ the method of using a dictionary in which labels are placed on the words, stating their polarity level. This is in turn used to detect the sentiments of a text. Artificial intelligence approaches trains a classifier using a dataset and then uses that model for further predictions.

Sentiment analysis has multiple classifications and these can be met based on the purpose of the analysis. 1.Fine-grained Sentiment Analysis: Breaks down a text into groups and analyses each group. Each group is analysed in relation to the others. 2.Aspect-Based Sentiment Analysis: Classifies data by aspect and finds the sentiment associated with each one. By connecting particular sentiments with various characteristics of a good or service, aspect-based sentiment analysis can be used to analyse consumer feedback. 3.Emotion Detection: Emotions like happy,angry,sad etc are detected rather than positive,negative or neutral. 4.Intent Based: A more advanced version, intent based approaches help identify if a sentiment analysis is a fact, opinion,query etc.

With the multiple applications of sentiment analysis, the method is now being considered invaluable. Reduction of costly time and effort in analysing data, and evaluating those projections increases efficiency. [1] states how

sentiment analysis aids in the development of many fields like marketing and research, and can help spearhead more research in these domains.[2] briefs about the various concepts and techniques that can be utilised for performing sentiment analysis.[3],[4], and [5] demonstrate various fields that sentiment analysis can be used in. With sentiment analysis stretching far and wide, more room is opened up for further analysis and research on the topic.

II. METHODOLOGY

In a supervised machine learning approach, the algorithm must first be trained using previously classified data (training data), with which it develops classification rules, and then it classifies the input data (test data), based on these rules. Any supervised machine learning algorithm's performance can be evaluated using test data that has already been pre-classified, and the algorithm's output can then be compared to the pre-classified data.

The Support Vector Machine (SVM) has been selected for sentiment classification. The supervised classifier support vector machine (SVM) algorithm is frequently used to address classification and regression issues.SVM's benefits, such as its capacity to handle huge features, make it effective for text categorization. The fact that SVM is resilient in the presence of a sparse set of examples and that the majority of problems are linearly separable is another benefit. In sentiment analysis, Support Vector Machine has produced promising results. It was created as an upgrade to the maximum margin classifier, which is limited to working with simple linearly separable data. The maximal margin classifier was introduced as an upgrade to the support vector classifier.

SVM generates linearly separated hyperplanes in high-dimensional vector spaces with well-divided feature space. The data points from the two classes are divided into separate areas by these planes. The hyperplane that maximises the distance between the training data points closest to the feature space is always the best hyperplane. Since data points belonging to different classes are rarely clearly distinguished, significant misclassification can appear as a result of linear classification. Since SVM maps the feature space into a higher-dimensional space where non-linear data points are transformed into linearly separable points, it is able to

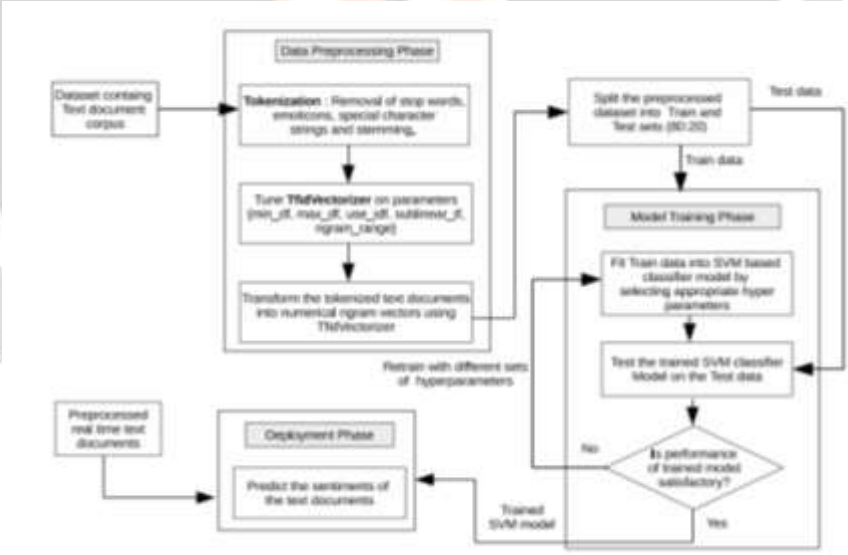

Fig. 1: System Architecture for sentiment analysis process

handle such frequently encountered cases. As a result, they are changed into points that can be separated linearly. As a result, the boundaries between data points in different classes are more distinct.

The high-dimensional space is created by non-linearly expanding the initial space using the kernel function. In addition to linear kernels, polynomial kernels and radial basis function kernels are other varieties of kernel functions. They are each described by the equations below, where k() stands for the kernel function and the result is represented by the product of the two observation vectors xi and x′i. The altered feature space is denoted by the symbol, while the product of the two vectors is denoted by (xi). (x′i).

$$k\left(x_i, x_i'\right) = \sum_{j=1}^{p} x_{ij} x_{ij}'$$

$$k\left(x_i, x_i'\right) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{ij}'\right)^d$$

$$k\left(x_i, x_i'\right) = \exp\left(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{ij}'\right)^2)$$

1. III. IMPLEMENTATION

A modularized architecture, as shown in the figure, forms the backbone of the sentiment analysis framework proposed, encompassing various interdependent components that collaborate to provide an all-encompassing solution.

Each module/phase of the proposed framework is briefly described below:

1)Acquiring the dataset:

It is extremely essential for the nature of a dataset has to be taken into consideration for training a machine learning model. A dataset is a collection of labelled or unlabeled data that is used to train a model to recognize patterns and make predictions based on new data. The quality and quantity of the dataset have a significant impact on the performance of the trained model. By training on a dataset, the model can learn the patterns in the data and generalize them to new data. By training on this dataset, the model can learn the patterns in the data and generalize them to new data.

The dataset chosen for this project was an extensive Amazon Reviews dataset, consisting of over 400,000 consumer reviews for various brands of electronics. This proved to be beneficial for the accuracy of the overall model.

2)Preprocessing the data:

The diagram illustrates that before being input into the SVM-based classifier model, the text documents undergo preprocessing. This involves tokenizing the input text to eliminate stop words, emoticons, and special character strings, and reducing derived words to their stem or root word through stemming. Subsequently, the tokenized word documents are converted into n-gram integer vectors using TfidfVectorizer, allowing them to be used as input for the SVM classifier. A pipeline function is used to perform these subsequent tasks of vectorization and feeding to the model, so that these tasks can be cross-verified.

To calculate the term frequency inverse document frequency (Tf-idf) values of words, the TfidfVectorizer utilises both the inverse document frequency (idf) and term frequency (tf). These Tf-idf values are then used to assign weight or significance to the words during sentiment analysis. The idf of a word 'w' in the text corpus and the Tf-idf of the word 'w' in a particular document 'd' are calculated.

2) Training the model: During the previous phase, a preprocessed and vectorized training dataset is created and used to train the SVM-based classifier model. This approach was preferred over dictionary-based and NLP-based approaches as it is more efficient. The accuracy of other machine learning models were also studied, and SVM was found to be marginally more accurate. Once trained, the SVM model is evaluated on a testing dataset, and if its performance on the test dataset is unsatisfactory, the classifier's training phase is repeated with a different set of SVM hyperparameters to improve its accuracy.

3) Deploying the classifier: Following the completion of the training phase, the SVM classifier model is deployed to perform sentiment analysis on text documents in real-time. Before the analysis can begin, the documents undergo preprocessing using the methods described in the Preprocessing Phase. The preprocessed documents are then presented as input to the trained SVM model, which accurately categorises the sentiments expressed in the documents as either positive, negative, or neutral.

2. IV. EXPERIMENTAL RESULTS

A modularized architecture, as shown in the figure, forms the backbone of the sentiment analysis framework proposed, encompassing various interdependent components that collaborate to provide an all-encompassing solution. The performance of the model was assessed by means of the metrics listed below.

Accuracy: Accuracy refers to the proportion of correct results, comprising both true positives and true negatives, to the overall number of documents that the classifier examined. The formula for accuracy is

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Precision is a measure that denotes the proportion of documents that are accurately classified as positive sentiments by the classifier, relative to the total number of documents that are classified as having positive sentiments.

$$Precision\ (Prec) = \frac{TP}{TP+FP}$$

Recall: Recall is a metric that indicates the proportion of documents that are accurately classified as positive sentiments by the classifier, relative to the true number of documents that contain positive sentiments in the provided text corpus.

F1 Score: $$Recall\ (Rcl) = \frac{TP}{TP+FN}$$ The F1 score is calculated as the harmonic mean of precision and recall.

$$F1\ Score = 2 * \frac{Prec * Rcl}{Prec + Rcl}$$

In the aforementioned formulas, TP refers to true positives, TN refers to true negatives, FP refers to false positives, and FN refers to false negatives.

The proposed sentiment analysis framework's performance was assessed using the Amazon customer reviews dataset [5], where each review is regarded as a document and comprises two features - Content and Label. Each review has a rating between 1 to 5 assigned to it. Positive reviews have a rating of 4 or 5, neutral reviews have a rating of 3 and negative reviews have a rating of 1 or 2. These numerical ratings are used for training the model. The initial step in preparing the dataset involved using the method outlined in Figure 1 to create an n-gram vectorized numeric representation model of the provided text documents. Once the vectorized data is preprocessed, it is divided into training and testing data sets in an 80:20 ratio. The SVM classifier is trained using the training data to classify the sentiments or reviews expressed in the documents as either positive or negative.

The original dataset consisted of different brands of consumer reviews, upon which sentiment analysis was conducted separately. The highest accuracy obtained was 92.6%.

Table 1: Classification report

	precision	recall	f1-score	support
Negative	0.93	0.91	0.92	890
Neutral	0.43	0.97	0.60	114
Positive	0.98	0.93	0.96	2371
accuracy			0.93	3375
macro avg	0.78	0.94	0.82	3375
weighted avg	0.95	0.93	0.93	3375

The resulting test predictions were plotted using a pie chart as shown below.

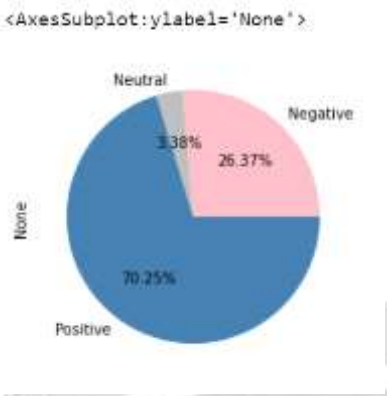

Fig. 2 Distribution of test classification

3. V. CONCLUSION

AI-driven sentiment analysis using SVM algorithm is a powerful tool for understanding and analysing sentiment in text data. The SVM algorithm is an effective machine learning technique for classification tasks, including sentiment analysis.

The advantages of using an AI-driven sentiment analyser with SVM algorithm include its ability to process large volumes of data quickly and efficiently, while maintaining a high degree of accuracy. It is a powerful and promising technology. It may have limitations, such as difficulties in accurately capturing the nuances of human language and sarcasm. Therefore, it is crucial to carefully evaluate the results of the sentiment analysis and to incorporate human judgement and feedback as necessary to ensure the accuracy of the analysis. It is a valuable tool that can help businesses and organisations gain insights into customer sentiment and improve decision-making.

4. 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] Ameen Abdullah Qaid Aqlan.,Dr. Manjula Bairam.,R Lakshman Naik.A Study of Sentiment Analysis: Concepts, Techniques, and Challenges.

[3] Sentiment Analysis on a Set of Movie Reviews Using Deep Learning Techniques

[4] Automatic recognition of self-reported and perceived emotions

[5] The COVID-19 outbreak: social media sentiment analysis of public reactions with a multidimensional perspective

[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