TuijinJishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 5 (2023)

Hybrid Sentiment Analysis: A Novel Integration of Corpus-Driven and Machine Learning approaches to perform Sentiment Analysis of Kannada Political Tweets

[1]Shankar R, [2]Suma Swamy

[1]Research Scholar, Sir M. Visvesvaraya Institute of Technology and Visvesvaraya Technological University, Belagavi-590018, India.
[2]Professor, Sir M. Visvesvaraya Institute of Technology and Visvesvaraya Technological University, Belagavi-590018, India.

***Abstract*:** In today's digital age, understanding user opinions holds significant value in gauging product satisfaction and shaping consumer choices. Sentiment Analysis, a subset of opinion mining, focuses on evaluating emotions and viewpoints expressed towards specific subjects. This research dives into the realm of Kannada sentiment analysis, a regional language spoken predominantly in India. The challenge lies in extracting insights from limited Kannada corpora, acquired from https://gadgetloka.com using Python-Beautiful Soup and an API. Manually sifting through vast, unstructured data is a daunting task, prompting the application of an automated system called 'Sentiment Analysis or Opinion Mining.' This system adeptly dissects and extracts user insights from reviews, categorizing them into Positive, Negative, or Neutral sentiments based on their assigned weights. The research involves the analysis of Kannada tweets collected from various political entities and politicians to classify them as Positive, Negative, or Neutral interpretations. This classification holds the potential to assist political parties and politicians in refining their image and credibility. The study leverages Support Vector Machines, a supervised learning algorithm, to optimize the classification process through a set of hyperplanes, yielding enhanced accuracy.
Keywords: *Sentiment Analysis, Hybrid Approaches, Kannada Language, Product Reviews, Political Tweets, Automated Sentiment Analysis, Sentiment Classification, Support Vector Machines.*

1. Introduction

Sentiment Analysis (SA) holds a significant position within the realm of textual analysis due to its extensive potential for application and implementation. SA enables various tasks, such as sentiment forecasting, partisanship detection, text summarization, sentiment breakdown, and more. In India, a diverse linguistic landscape includes languages like Hindi, Kannada, Telugu, Tamil, Malayalam, and others, which have attracted numerous studies and research interests in the field of SA. However, the exploration of Kannada text for sentiment analysis, particularly for product analysis, remains limited.

This paper introduces a case study focusing on mobile product reviews written in Kannada. The feasibility of this study arises from the abundance of user-generated Kannada product reviews available online. SA, as a computational methodology, involves characterizing, identifying, and extracting sentiments, including emotions, opinions, and attitudes, embedded within text, speech, or databases. It leverages principles from Natural Language Processing (NLP) and Machine Learning (ML) algorithms.Artificial intelligence (AI) has emerged as a transformative force in today's world, powering a wide range of applications such as Smart Cities, Autonomous Cars, Personal Assistants, and Smart Classrooms. It has become an integral part of the IT industry, reshaping how we interact with technology. At the heart of AI is Machine Learning (ML), which enables machines to learn and make decisions or predictions based on the knowledge they acquire.

In the realm of ML, machines are provided with data in a machine-readable format and trained to understand the data's structure and corresponding output values. This process involves the use of specific algorithms designed for particular tasks. Once trained, these models can be employed to make real-time predictions for new or dynamic input data.

Machine Learning can be broadly categorized into two types: Supervised Learning and Unsupervised Learning, based on the availability of labeled or unlabeled data. Supervised Learning involves training models with labeled data, while Unsupervised Learning deals with unlabeled data. Examples of Supervised Learning algorithms include Random Forest, Support Vector Machines, Linear and Logistic Regression, and KNN. In contrast, Unsupervised Learning encompasses algorithms like K-Means Clustering, DBSCAN, and Anomaly Detection.

Sentiment Analysis, a crucial technique, involves the analysis and interpretation of sentiments expressed in text. This method is particularly valuable for understanding user opinions, whether it's assessing the quality of a product, gauging movie preferences, or capturing other sentiment-related aspects. By comprehending sentiment, producers can enhance their products and establish better connections with users. In this paper, Support Vector Machines are employed for Sentiment Analysis to analyze and predict the sentiment of given input sentences.

This research employs a corpus-based approach to construct a sentiment lexicon, relying on morphological patterns present in extensive corpora, achieving high accuracy in identifying related words. SA can be viewed as a facet of NLP, differing in that it extracts valuable insights using only a few key phrases from text, making it a powerful tool even without comprehensive context understanding. While some existing research addresses specific applications like reasoning detection, there remains a need for in-depth exploration in both SA and NLP to cater to a broader range of use cases.

2. Literature Survey

The research contributions in Kannada sentiment analysis and corpus generation are multifaceted. S. Parameshwarappa et al. discussed the creation of a Kannada corpus tool for Program Execution and Reporting Language (PERL) from web logs, focusing on raw corpus generation but not addressing sentence tokenization or search methods. Jayashree R explored data retrieval from large datasets, employing sentence-level sentiment classification, yet omitted discussions on multi-label classification and product recommendation based on reviews. Deepamala N applied sentiment analysis techniques, including manually crafted polarity lexicons and word suffix analysis, with comparisons to other machine learning algorithms, revealing lower accuracy due to the manual approach. Lastly, Shankar R, Suma Swamy, and their team conducted an extensive study on Sentiment Analysis in various Indian dialects, offering insights into diverse pedagogies related to opinion mining. These studies collectively enhance our understanding of sentiment analysis in Kannada and its applications, while also highlighting challenges and potential avenues for improvement in this field. Shankar R and Suma Swamy conducted an extensive study on Sentiment Analysis based on Corpora-based classification, utilizing a tokenization approach to categorize sentences as positive, negative, or neutral, although it's a non-machine learning technique with potential disadvantages [1]. Rohini V, Merin Thomas, and Dr. Latha CA employed decision trees in machine learning and a Parts of Speech tagging approach for Sentiment Analysis, achieving a maximum accuracy of approximately 78% after opinion word extraction and polarity assignment [2]. K. M. Anil Kumar, N. Rajasimha, Manovikas Reddy, A. Rajanarayana, and KewalNadgir performed Kannada Transliteration with NLP and Hidden Markov Models, comparing Semantic Algorithms and Machine Learning Algorithms at different linguistic levels (word, phrase, and sentence) [3]. YashaswiniHegde and S.K Padma utilized the Random Forest Ensembling technique to enhance Sentiment Analysis accuracy in Kannada, achieving an improvement from 60% to 72% [4]. Impana and Jagadish explored Cross-Lingual Sentiment Analysis, employing Autoencoder architecture and Bilingually Constrained Recursive Autoencoder (BRAE) models trained on English and Kannada, resulting in promising results [5]. Jayashree R and Srikanta Murthy K conducted Sentiment Analysis using the Naïve Bayesian Algorithm and the Bag of Words model, with dimensionality reduction and stop-word removal, achieving good results at the sentence level, which could potentially be extended to document-level classification [6]. These studies collectively contribute to the field of Sentiment Analysis, employing various techniques, machine learning algorithms, and linguistic approaches to analyze sentiment in the Kannada language, Offering insights and advancements in this domain.

3. Proposed Research Framework

In Figure 1, we present the proposed system for Sentiment Analysis of Mobile Product Reviews in Kannada, which relies on two distinct corpora related to the Kannada language. One corpus contains positive words such as "ಅದ್ಭುತವಾಗಿದೆ" and "ಕೊಳಳಬಹುದು," while the negative corpus contains words like "ಕೊಳ್ಳುವುದುಬೇಡ" and "ಕೆಟ್ಟಮೊಬೈಲ್ವಾಗಿದೆ." The analysis process begins by converting the given sentence into tokens, and each word is compared against the positive and negative corpora. If a word is found in the positive corpus, the positive score is incremented, and if it's found in the negative corpus, the negative score is incremented[7]. When both scores are equal, the review is classified as neutral. Collecting an adequate Kannada corpus is a significant challenge in itself, and reviews of various mobile products are scraped through a web API called Beautiful Soup via Python 3, with data sourced from websites like gadgetloka.com and mobile.gizbot.com/Kannada[8]. The system processes a series of reviews separated by a common separator and the analysis of each review is based on the corpus weights, which represent the counts of corresponding positive and negative words.

To achieve Sentiment Analysis of Mobile Product Reviews in Kannada, the following objectives are established: Building a Kannada Corpora/WordNet, Extracting features from this WordNet, Identifying words and phrases denoting opinions related to the target feature, Classifying the extracted opinions as positive, negative, or neutral reviews[9].Building a Kannada Corpora/Wordnet involves collecting structured text documents in Kannada, which is challenging due to the limited availability of publicly accessible corpora for Indian languages[10]. Researchers often face obstacles in obtaining large and balanced corpora for such languages. To address this issue, we leverage Wikipedia, a free and open-source multilingual encyclopedia project, containing over 11 million articles in Kannada. These articles are created collaboratively, allowing for asynchronous editing and contributions from editors and NLP translators, making it a valuable resource for our corpus-building efforts [11,12].

In summary, the proposed system utilizes two corpora to perform Sentiment Analysis of Mobile Product Reviews in Kannada, with the corpus-building process facilitated by Kannada-language articles on Wikipedia, enabling the extraction of features and opinions from these documents for sentiment classification.

Extracting Features from Reviews: identify attributes that indicate polarity, such as "samsungbahalaakarshakawagide" (ಸ್ಯಾಮ್ಸಂಗ್ಬಹಳಆಕರ್ಷಕವಾಗಿದೆ) or "Iduatyanthaketta mobile aagide" (ಇದುಅತ್ಯಂತಕೆಟ್ಟಮೊಬೈಲ್ಆಗಿದೆ), reviews provide a list of possible attributes. These attributes, when extracted, yield the exact features under consideration[13,14]. For instance, a statement like "ನೋಕಿಯಾN9 ಮೊಬೈಲ್ಅದ್ಭುತವಾಗಿದೆ" denotes a positive opinion about the Nokia N9 mobile device.

Extracting Words and Phrases Denoting Opinions: To discern words and phrases denoting opinions related to the target features, keywords like ಅದ್ಭುತವಾಗಿದೆ, ಕೊಳ್ಳಬೊಹುದು, ಕೊಳ್ಳುವುದುಬೇಡ, ಕೆಟ್ಟಮೊಬೈಲ್ಆಗಿದೆ, etc., are extracted and matched with the corpora for further analysis. Part-of-Speech (POS) tagging aids in accurately matching these words with the correct entries in the database, enabling the determination of the direction of opinion expressed in the product review[15]. This approach proves beneficial when dealing with sentences containing distributed emotions and multiple attributes.

Classifying Extracted Reviews: In this phase, reviews are classified as positive, negative, or neutral based on the extracted content. For example, "ಕೆಟ್ಟಮೊಬೈಲ್ಆಗಿದೆ" would be classified as a negative opinion, while "ಅದ್ಭುತವಾಗಿದೆ" would be categorized as a positive opinion. However, there are distinct cases where clear-cut classification may not be possible, resulting in a vaguer score that complicates value assignment. Some texts may not fit neatly into the positive or negative classes and are instead classified as neutral. Additionally, certain tokens may target the operator rather than the product itself, further challenging their classification. These complexities underscore the intricacies of sentiment analysis in Kannada product reviews.

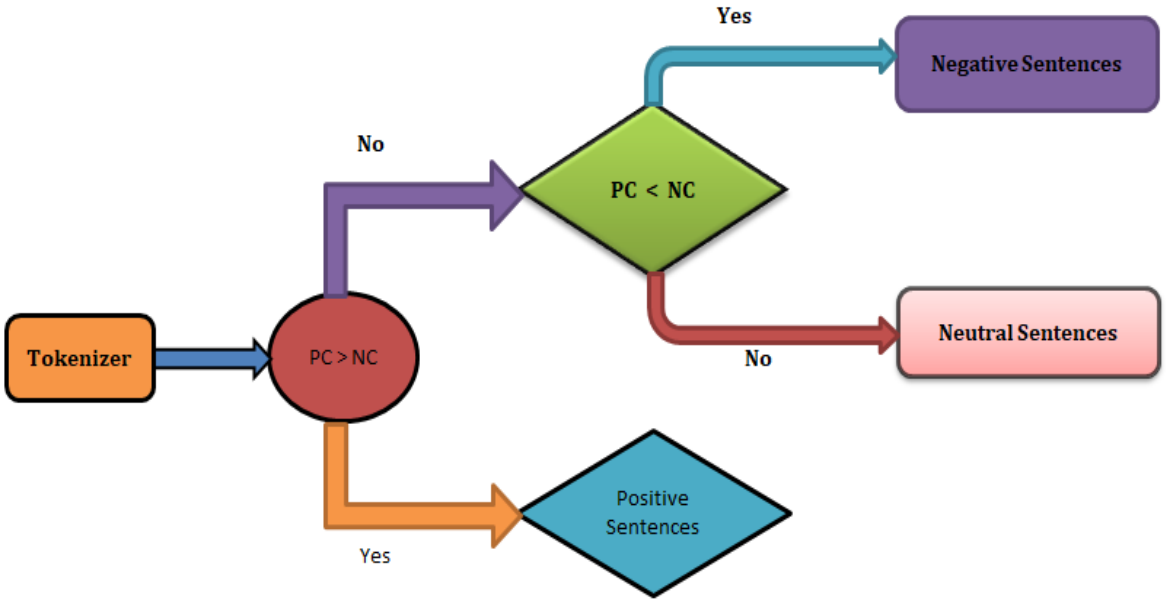

Figure.1: Sentiment Assessment for Kannada Sentences

4. System Configuration

Figure 2 illustrates the architecture for conducting sentiment analysis on Kannada sentences. The process begins with an input text file, which is then tokenized into individual words. These tokens are subsequently compared against two corpora: the positive and negative corpora. In the comparison step, if a token is found in the positive corpus, the positive word count is incremented by one. Conversely, if a token is present in the negative corpus, the negative word count is incremented by one. Following this, a classification decision is made based on the word count comparison.If the negative word count is less than the positive word count, the sentence is classified as positive. Conversely, if the negative word count exceeds the positive word count, the input text is classified as a negative sentence. Finally, when the counts of positive and negative words are equal (NWC = PWC), the sentence is classified as neutral. This approach provides a straightforward method for determining the sentiment of Kannada sentences based on the presence of positive and negative words within the text.

The complete process of computing the sentiment can be given as a linear function of f(x). Suppose to find the sentiment value of a sentence x consisting the word tokens as x0, x1, x2, x3 etc., can be given as per equation 1:

f(x) = t0x0 + t1x1 + t2x2 + t3x3 …. tn-1xn-1 (1)

where,

tag(t) for each token,

where t = +1, if the token is positive

t = -1, if the token is negative

n is the number of tokens in the given sentence.

The above equation can be summarized as per the equation 2

$$f(x) = \sum_{i=0}^{n-1} tixi$$

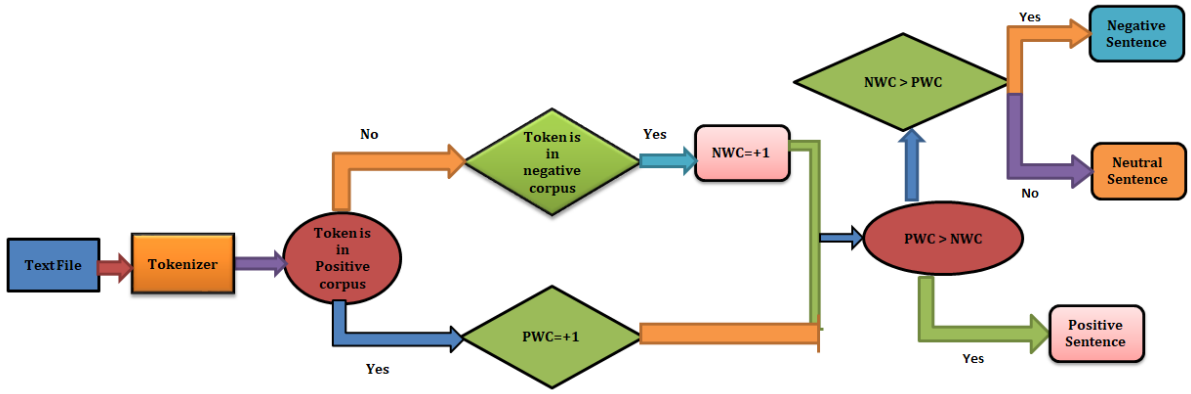

Figure.2: Sentiment Recognition Framework for Kannada Sentences.

Figure 3 depicts the overarching pipeline of the experimental setup, outlining the following key stages: Data Input, Data Pre-processing (which involves Tf-IdfVectorization), Feature Extraction, and culminating in Model Training and Evaluation. Following the training phase, the algorithm can be saved for future use, and at any point, it can be loaded and employed for making predictions by inputting new data instances to the trained Machine Learning model, thus obtaining desired inferences or predictions. This systematic workflow enables efficient data processing, feature extraction, and model utilization for various tasks within the experimental context.

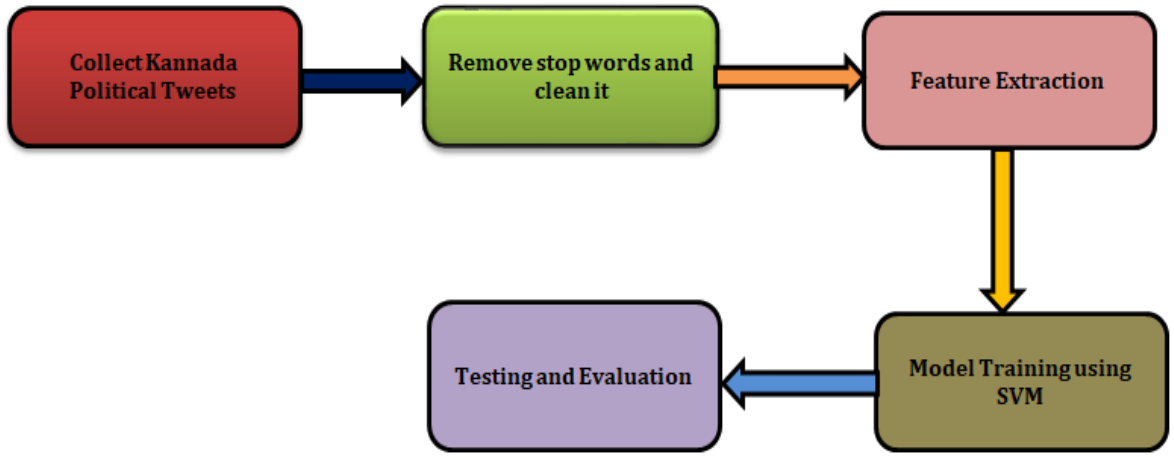

Figure.3: Test Configuration

For the task of sentiment analysis in Kannada, a supervised learning approach is adopted, wherein classification algorithms are employed to categorize sentences and phrases into positive and negative sentiments. The dataset comprises over 100 Kannada tweets, with "pos" and "neg" labels indicating sentiment polarity. Several classification algorithms are considered for this task, including Gaussian Naïve Bayes, Multinomial Naïve Bayes, SVM Classifier, SGD Classifier, K Nearest Neighbours Classifier, and Decision Tree Classifier. In this specific work, the SVM Classifier is chosen as the classification algorithm of choice. The data collection process involves web scraping from Twitter, where popular tweets are selected and manually translated from English sentences into Kannada sentences using Google Translator. The resulting dataset exhibits a structure where each tweet is associated with a sentiment label, facilitating the training and evaluation of the sentiment analysis model.

Table 1. Kannada Tweet and Tag

Kannada Tweet	Tag
ಮತಚಲಾಯಿಸಲಾಗಿದೆ! ಮೊದಲಬಾರಿಗೆದೀರ್ಘಸರತಿಇತ್ತು. ಒಂದುಗಂಟೆಕಾಯುತ್ತಿದ್ದರು. ಉತ್ಸಾಹವನ್ನುನೋಡಲುಒಳ್ಳೆಯದು. ಮತಚಲಾಯಿಸಿಬೆಂಗಳೂರು!	pos
ಪಶ್ಚಿಮಬಂಗಾಳದನಿರಾಶ್ರಿತರಿಗೆ. ನಾನುಹೇಳಲುಬಯಸುತ್ತೇನೆ. ನೀವುಯಾರಿಗೂಹೆದರಬೇಕಾಗಿಲ್ಲ. ನಾವುಇಲ್ಲಿನಿರಾಶ್ರಿತರನ್ನುಭಾರತದಮಗಮತ್ತುಮಗಳಂತೆಪರಿಗಣಿಸುತ್ತೇವೆಮತ್ತುಅವರಿಗೆಪೌರತ್ವನೀಡಲಾಗುವುದು "ಎಂದುಅವರುಹೇಳಿದರು.	pos
ಮುಂಬರುವಲೋಕಸಭಾಚುನಾವಣೆಯಲ್ಲಿರಾಜಕೀಯಪಕ್ಷಗಳಭವಿಷ್ಯವನ್ನುನಿರ್ಧರಿಸುವಲ್ಲಿಮೊದಲಬಾರಿಗೆಮತದಾರರುನಿರ್ಣಾಯಕ ಪಾತ್ರವಹಿಸಲಿದ್ದಾರೆ. ನಇತ್ತೀಚಿನಸಂಚಿಕೆಯಲ್ಲಿಇನ್ನಷ್ಟುಓದಿ	pos
ದೊಡ್ಡಬಿಕ್ಕಟ್ಟಿನಮಧ್ಯೆಬಿಜೆಪಿರಾಜಕೀಯದಲ್ಲಿಮಾತ್ರತೊಡಗಿಸಿಕೊಂಡಿದೆ! ಫೆಡರಲಿಸಂಮತ್ತುದುರುದ್ದೇಶಪೂರಿತಬಂಗಾಳದವೆಚ್ಚದಲ್ಲಿಅಗ್ಗದಬ್ರೌನಿಅಂಕಗಳನ್ನುಗಳಿಸಲುಪ್ರಯತ್ನಿಸುತ್ತದೆ	neg
ಇದುಕೇವಲಪ್ರಾರಂಭಸಾಂಬಿತ್ಸ್ರಾಜ್ನೀವುಮತ್ತುನಿಮ್ಮಪಕ್ಷಬಿಜೆಪಿಭಾರತೀಯಮುಸ್ಲಿಮರಿಗೆಮಾಡಿದಅಪರಾಧಗಳಬೆಲೆಯನ್ನುಶೀಘ್ರ ದಲ್ಲೇಪಾವತಿಸಲಿದೆಇಸ್ಲಾಮೋಫೋಬಿಯಾ_ಇನ್_ಇಂಡಿಯಾ	neg
ಕರ್ನಾಟಕಚುನಾವಣೆಗಳಿಗಾಗಿಇಂಧನಪ್ಯೆಸ್ಟ್ಹೈಕ್ಅನ್ನುತಡೆಹಿಡಿಯಲಾಯಿತು..ಮತ್ತುಎಂದಿನಂತೆಜನರುಬಿಜೆಪಿಯಿಂದಮೂರ್ಖರಾಗಿ ದ್ದರುಈಗಪ್ರತಿದಿನವೂಅದುಹೆಚ್ಚುತ್ತಿದೆ. ಆಮ್ಆದ್ಮಿವೆಚ್ಚದಲ್ಲಿತಮ್ಮಸರ್ಕಾರಭಾರಿಲಾಭಗಳಿಸುತ್ತಿರುವುದರಿಂದಮೋದಿಸರ್ಕಾರಯಾವುದೇಪರಿಹಾರನೀಡುವುದಿಲ್ಲ.	neg

In the sentiment analysis framework, the machine learning classification algorithm Support Vector Machines (SVM) is utilized to achieve accurate sentiment identification. After training the model with the provided corpora (dataset), it is ready for deployment to make real-time predictions.

The process of sentiment analysis is divided into four key steps:

Training the Model: In this initial phase, the SVM model is trained using the provided dataset, which includes annotated examples of Kannada sentences labeled as either positive or negative sentiment. The model learns to recognize patterns and relationships within the data, enabling it to make predictions based on new input.

Deployment for Real-Time Predictions: Once the SVM model is trained and fine-tuned, it is ready for deployment. In this stage, a sentence or text input is provided to the trained model, and it promptly generates an output indicating whether the sentiment expressed in the input is positive or negative. This real-time prediction capability allows for the analysis of sentiment in dynamic and evolving contexts.

These four steps together form the framework for sentiment analysis, utilizing SVM as the classification algorithm to accurately identify sentiment in Kannada sentences and phrases.

Table 2 represents the sample training data used for the sentiment analysis task. The data is organized in the form of phrases and sentences, and each example is assigned a label indicating its sentiment polarity. These labeled examples are then loaded into a data frame for further processing and training of the sentiment analysis model. This structured dataset serves as the foundation for training the machine learning algorithm to recognize sentiment patterns and make accurate predictions on new, unseen text inputs.

Table 2. Visualizing the data

Tweet	Label
ಯಾದಗೀರ್ನಲ್ಲಿಒಟ್ಟಾರೆಮಾನವಅಭಿವೃದ್ಧಿಯಮೇಲೆಸರ್ಕಾರಹೆಚ್ಚಿನಪರಿಣಾಮ ಬೀರುತ್ತದೆ	Pos
ಸರ್ಕಾರವನ್ನುಉರುಳಿಸಲುನಾವುಸ್ಕ್ರಾಟಜಿಯನ್ನುಯೋಜಿಸುತ್ತಿದ್ದೇವೆ	Neg
ಬಿಜೆಪಿಮೊದಲೇಲಾಕ್ಡೌನ್ಪ್ರಾರಂಭಿಸಿರಬೇಕು	Pos
ಈಸರ್ಕಾರರಾಜ್ಯದಬೆಳವಣಿಗೆಗೆಅಡ್ಡಿಯಾಗುತ್ತಿದೆ	Neg
ಬಿಜೆಪಿಸರ್ಕಾರಅಧಿಕಾರಕ್ಕೆಬಂದಾಗಿನಿಂದಗ್ರಾಮವಿದ್ಯುದೀಕರಣಹೆಚ್ಚುತ್ತಿದೆ	Pos

The data is structured with two main components: the actual sentences or phrases and their corresponding labels, where a label of 1 indicates a positive sentiment, and a label of 0 signifies a negative sentiment.To ensure the robustness and reliability of the sentiment analysis model, the dataset is split into two distinct parts: training data and testing data. Typically, a common practice is to allocate 70% of the dataset for training, leaving the remaining 30% for testing. This partitioning allows for the evaluation of the trained classifier's accuracy on unseen data, thereby gauging its generalization performance.

Furthermore, to prevent any potential bias or order-related effects in the data, a randomization step is applied. This randomization is achieved by utilizing the "random state" parameter in the scikit-learn function called "train_test_split()". By setting a fixed value for the random state, the code generates the same sequence of random splits each time it is executed. This consistency ensures that, unless other sources of randomness are involved, the results obtained will be identical upon each run of the code. This repeatability aids in verifying and validating the output, contributing to the stability and reliability of the sentiment analysis model.

The vectorized data, obtained after preprocessing and feature extraction, is then input into a Support Vector Machines (SVM) model for training. SVM is a powerful machine learning algorithm that aims to find a set of hyperplanes capable of effectively separating two different classes within the data. These hyperplanes serve as lines or higher-dimensional boundaries that divide the data space into distinct regions representing the different classes.

Originally, SVM was developed primarily for binary classification tasks, meaning it was designed to distinguish between two classes. However, over time, extensions and adaptations of SVM have been developed to handle multi-class and regression problems as well.

In cases where categorical inputs are present, it's necessary to convert them into binary dummy variables, creating one binary variable for each category. This conversion allows SVM to process categorical data effectively.In simpler terms, the primary goal of an SVM classifier is to determine the most optimal way to separate different classes in the data space based on the provided tuning parameters. These tuning parameters include the choice of kernel function, regularization strength, margin settings, and gamma values, all of which influence the classifier's ability to create an efficient partitioning of the data space, effectively distinguishing between different classes. SVM's strength lies in its ability to handle complex data and find the most discriminative boundaries for classification tasks.

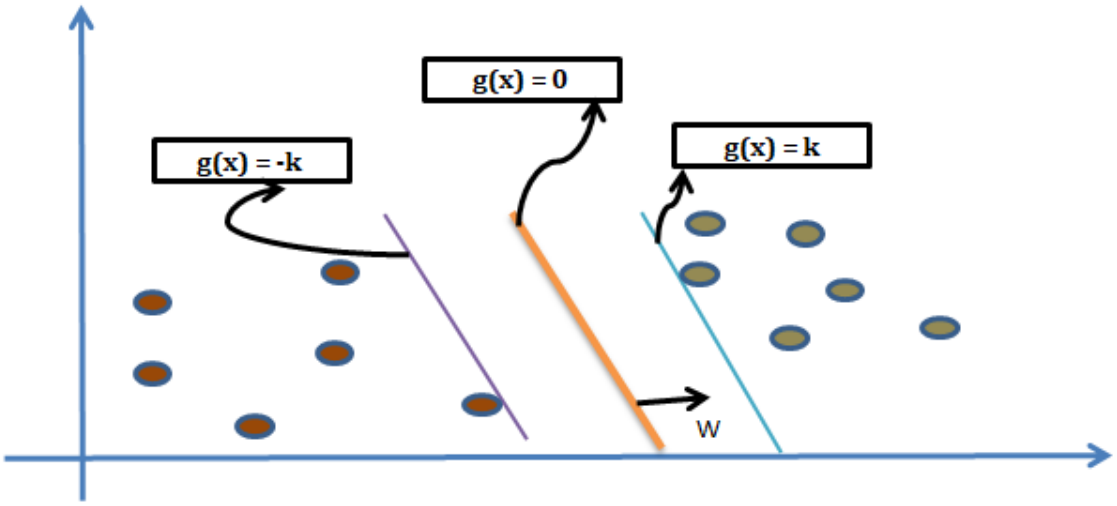

Figure.4: Visual Representation of Data and SVM Decision Boundary

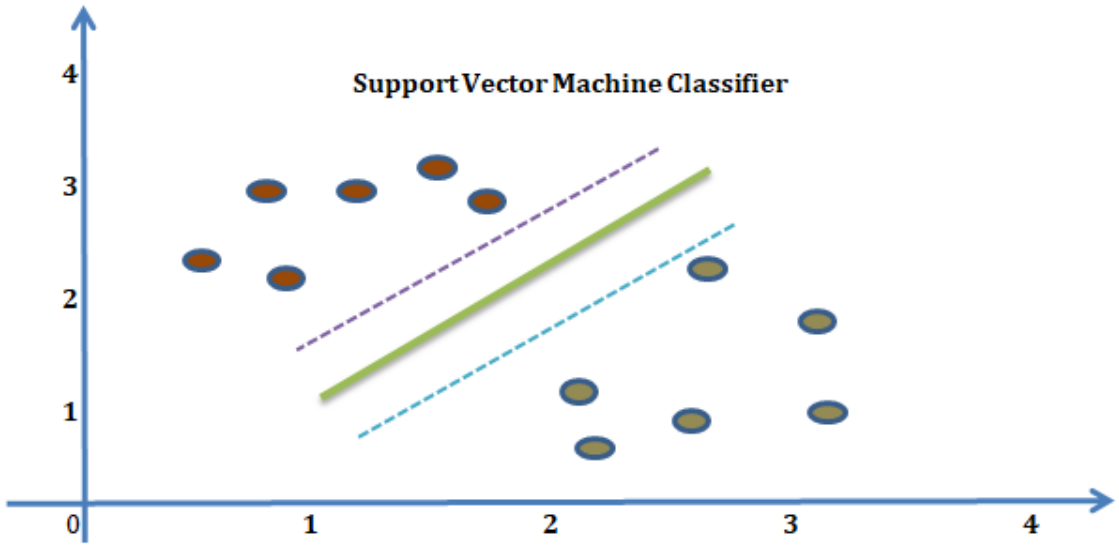

Figure.5: Visual representation of SVM Algorithm's Hyperplanes and Class Division

5. Results and Discussion

The initial corpus, which consists of approximately 5016 lexicons categorized as good and bad words, is prepared for sentiment analysis. These lexicons are used to match against the provided reviews in order to analyze their sentiments. Table 3 provides a sample of the corpus, showcasing examples of good and bad words in the Kannada language. This lexicon serves as a crucial reference for the sentiment analysis process, aiding in the classification of the sentiment expressed in Kannada text. The Kannada reviews are categorized into Positive, Negative, and Neutral sentiments based on the weights assigned to words in the corpora or lexicon tables. Figure 6 illustrates a sample classification, providing an example of how the sentiment analysis model assigns sentiments to Kannada reviews by considering the weights of words present in the lexicon. This classification process helps discern the sentiment expressed in the reviews, allowing for a more nuanced understanding of the text's emotional tone.

Table 3: The sample corpus looks like the following.

Good words sample - Kannada	
ಚೆನ್ನಾಗಿದೆ	ಸಕಾರಾತ್ಮಕ
ಚೆನ್ನಾಗಿದೆ	ಕೈಗೆಟುಕುವ
ಆಡೆತಡೆಇಲ್ಲದೆ	ಅಗ್ಗವಾದ
ವಿಪುಲವಾಗಿವೆ	ಚುರುಕುತನ
ಹೇರಳವಾಗಿ	ಪ್ರಿಯವಾದ
ನಿಖರವಾಗಿ	ಒಪ್ಪಿಗೆಯ
ಕುಶಾಗ್ರಮತಿ	ಸಂತೋಷವಾಗಿ
ಹೊಂದಿಕೊಳ್ಳಬಲ್ಲ	ವಿಸ್ಮಯಗೊಳಿಸು
ಲಾಭ	ಆಶ್ಚರ್ಯಚಕಿತನಾದನು
ಅನುಕೂಲಕರ	ಬೆರಗು
ಮಹತ್ವಾಕಾಂಕ್ಷೆಯಿಂದ	ಅದ್ಭುತ
ಸುಧಾರಿಸುವಲ್ಲಿ	ಮೆಚ್ಚುಗೆ
ಸ್ನೇಹಪರ	ಆಸ್ವಾದಿಸುತ್ತಾನೆ
ಮನರಂಜಿಸುವ	ಹೆಮ್ಮೆ
ಉಲ್ಲಾಸವಾಗುವಂತೆ	ನಿಬ್ಬೆರಗುಗೊಳಿಸು
ಗಮನಾರ್ಹ	ಅತ್ಯುತ್ತಮ
ಪ್ರಶಂಸಿಸುತ್ತೇವೆ	ಬ್ಲಾಕ್‌ಬಸ್ಟರ್
Badwords sample - Kannada	
ಚೆನ್ನಾಗಿಲ್ಲ	ಅಸಂಗತವಾಗಿ
ಅಸಹಜ	ಅಸಂಬದ್ಧತೆ
ರದ್ದುಪಡಿಸುವಂತೆ	ನಿಂದನೆ
ಅಸಹನೀಯ	ತಳವಿಲ್ಲದ
ಹಠಾತ್	ಪ್ರಪಾತ
ಥಟ್ಟನೆ	ಆಕಸ್ಮಿಕ
ತಲೆಮರೆಸಿಕೊಂಡು	ಕಲಬೆರಕೆ
ಕಹಿಗೊಳಿಸು	ದ್ವಂದ್ವಾರ್ಥತೆಯನ್ನು
ನೋವು	ಅಸ್ಪಷ್ಟ
ಉಲ್ಬಣಗೊಳಿಸಬಹುದು	ಅಸ್ಥಿರತೆ
ಸಂಕಟಕೊಡು	ಚಂಚಲ
ಅನ್ಯಾಯಕ್ಕೊಳಗಾದ	ಹೊಂಚುದಾಳಿಯಿಂದ
ಗಾಬರಿಗೊಂಡ	ಸರಿಯಿಲ್ಲದ

ಓದಲುಯಾತನಾಮಯ	ವೈಷಮ್ಯ
ಕಡುದುಃಖ	ಉದಾಸೀನತೆ
ಅಪಾಯಕಾರಿ	ನಿರಾಸಕ್ತಿ
ಗಾಬರಿಯಾಗುವಂತೆ	ನೋಡಲಾಗದಂತಹ

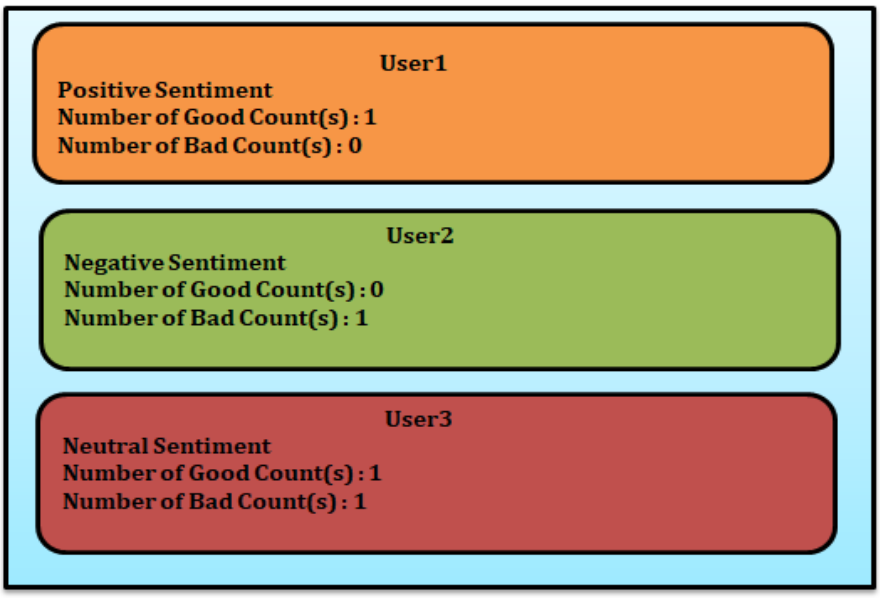

Figure.6: Results of Categorization

Figure 7 displays the polarity classification results for all the provided reviews. These results are projected onto a separate file named "KannadaReviews.txt." This file likely contains the sentiment labels (Positive, Negative, or Neutral) assigned to each review, offering a convenient way to examine and reference the sentiment classifications for the entire set of Kannada reviews.

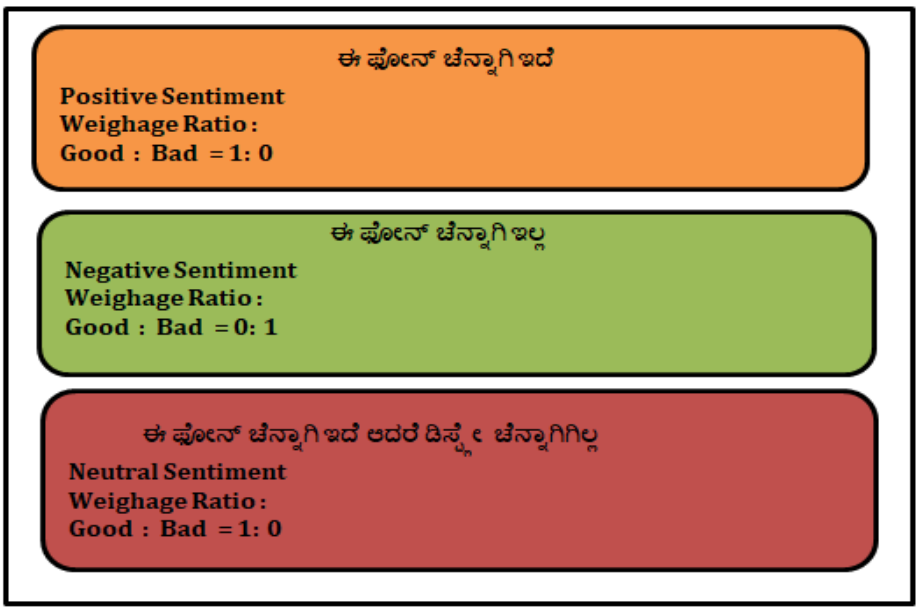

Figure.7: Prototype Result of the Sorting

Figure 8 displays a comprehensive analysis of 2500 mobile product reviews following their classification. As illustrated in the bar graph, the majority of sentences in the dataset are positive reviews, with negative reviews following closely behind, and the fewest belonging to neutral reviews for mobile devices. These findings are derived from the corpus of data under consideration. Additionally, it's evident that the outcomes are directly influenced by the size of the available corpus, which, in turn, has a direct impact on the quality of the analysis conducted. This suggests the potential for further development of a larger and more dynamic corpus.

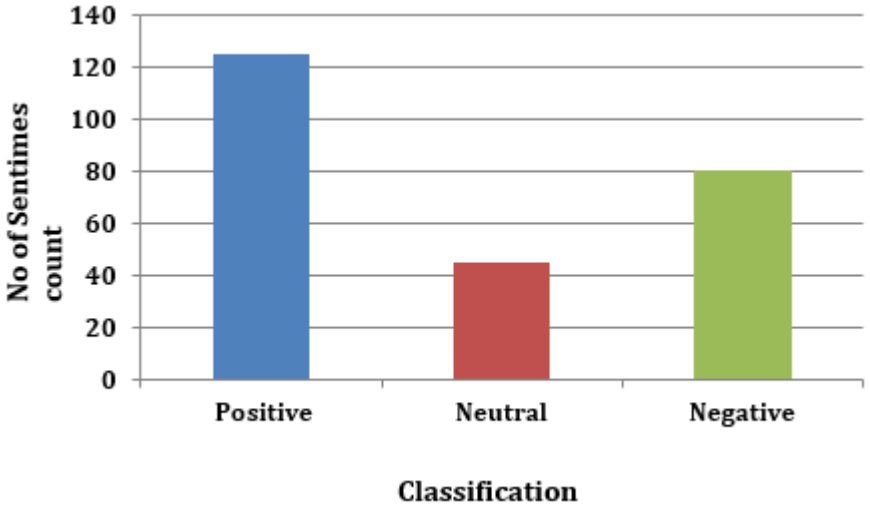

Figure.8: Comprehensive Classification Analysis

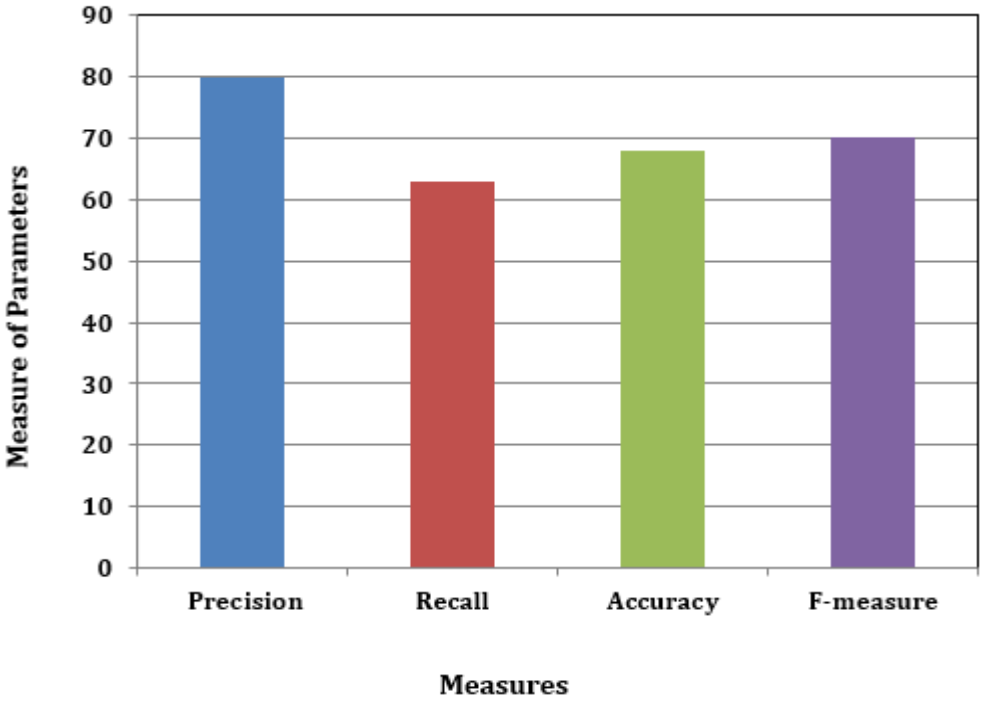

Figure.9: Classifier Assessment Metrics

The above results highlight the essential metrics derived from the model, compared to manual results. Figure 9 provides an overview of the performance metrics of the classifiers obtained from the models. Accuracy and Precision metrics offer valuable insights into the parsed data. The Recall metric sheds light on the challenges posed by higher-level morphological natural language processing (NLP), such as identifying negative sentences and sarcasm, which the model struggled to classify accurately.

The SVM model achieved a performance score that exceeded the expected range of 75-78%. This success can be attributed to the meticulous selection of hyperparameters and the effective preprocessing of the dataset. SVM, as a machine learning algorithm, is particularly well-suited for scenarios involving small and well-defined datasets that exhibit clarity without ambiguity. Leveraging these strengths, the approach adopted for this sentiment analysis task yielded highly favorable results. In addition to the standard evaluation metrics, Figure.10. showcases the extended metrics used for assessing the performance of the Machine Learning model. These extended metrics provide a comprehensive understanding of how well the model performs and its effectiveness in accurately classifying sentiment in Kannada sentences and phrases.

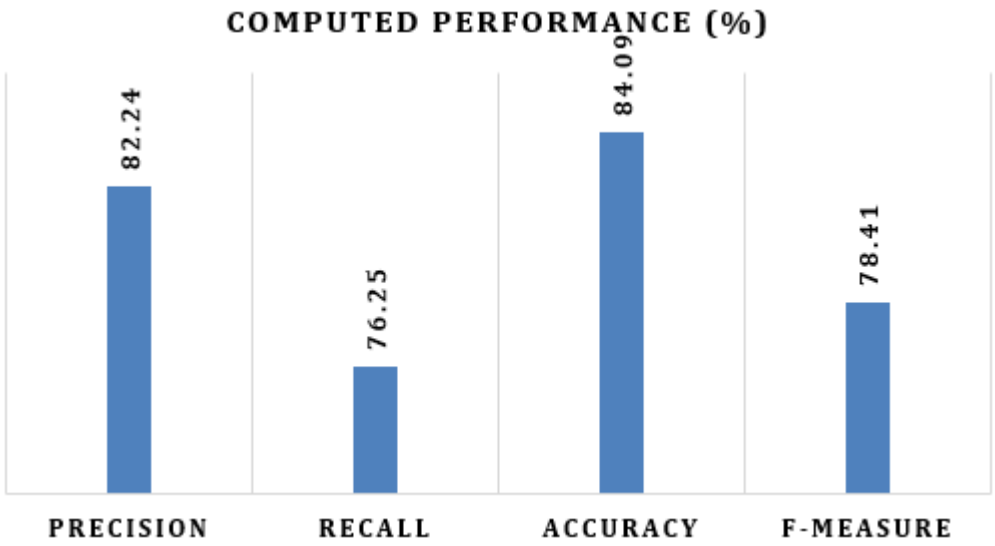

Figure.10: Performance Metrix

When comparing the performance of the chosen algorithm to implementations referenced in the literature survey, it was observed that the algorithm did not achieve the desired accuracy and precision score of 80%. Figure 11 provides visual examples that illustrate the outcomes during the testing phase of the selected algorithm. These examples likely demonstrate instances where the algorithm may have misclassified or encountered challenges in accurately determining sentiment in Kannada sentences. Evaluating these cases can help identify areas for potential improvement and fine-tuning in future iterations of the sentiment analysis model.

```
In [9]: fun('ಅವರು ಜನರನ್ನು ಲೂಟಿ ಮಾಡುತ್ತಿದ್ದಾರೆ')
Input Text:  ['ಅವರು ಜನರನ್ನು ಲೂಟಿ ಮಾಡುತ್ತಿದ್ದಾರೆ']
Sentiment: Negative

In [10]: fun('ನಾವು ರಾಜ್ಯದ ಜನರ ಪರವಾಗಿ ನಿಲ್ಲುತ್ತೇವೆ')
Input Text:  ['ನಾವು ರಾಜ್ಯದ ಜನರ ಪರವಾಗಿ ನಿಲ್ಲುತ್ತೇವೆ']
Sentiment: Positive

In [11]: fun('ನಾವು ಸರ್ಕಾರವನ್ನು ಉರುಳಿಸುತ್ತೇವೆ')
Input Text:  ['ನಾವು ಸರ್ಕಾರವನ್ನು ಉರುಳಿಸುತ್ತೇವೆ']
Sentiment: Negative

In [12]: fun('ನಾವು ಯಾವುದೇ ಭ್ರಷ್ಟ ಆಚರಣೆಗಳು ಮತ್ತು ಹಗರಣಗಳನ್ನು ಅನುಸರಿಸುವುದಿಲ್ಲ')
Input Text:  ['ನಾವು ಯಾವುದೇ ಭ್ರಷ್ಟ ಆಚರಣೆಗಳು ಮತ್ತು ಹಗರಣಗಳನ್ನು ಅನುಸರಿಸುವುದಿಲ್ಲ']
Sentiment: Negative

In [13]: fun('ರಾಜ್ಯದ ಅಭಿವೃದ್ಧಿ ಸರ್ಕಾರದ ಮುಖ್ಯ ಕಾರ್ಯಸೂಚಿಯಾಗಿದೆ')
Input Text:  ['ರಾಜ್ಯದ ಅಭಿವೃದ್ಧಿ ಸರ್ಕಾರದ ಮುಖ್ಯ ಕಾರ್ಯಸೂಚಿಯಾಗಿದೆ']
Sentiment: Positive
```

Figure. 11. Results after testing the model

6. Conclusion

The retail industry holds significant potential, especially when businesses expand their operations to regional markets. In this context, the work on Kannada opinion classification discussed in this article offers tremendous benefits. The method outlined in this article provides a comprehensive classification of Kannada reviews into positive, negative, or neutral polarities. This work intends to further enhance its capabilities by generating a more extensive corpus and classifying sentiments based on entire documents, rather than just at the sentence level. This hierarchical approach involves computing sentiment from the word to sentence, sentence to paragraph, and paragraph to document, allowing for a more holistic understanding of sentiment within larger textual contexts. Achieving an accuracy rate of 80% in non-English text analysis, especially in Kannada, is a commendable accomplishment. The work is indeed praiseworthy given the inherent challenges posed by the language, such as grammar ambiguity. One of the primary challenges faced during this project was obtaining data that is free from ambiguity in the Kannada language, which is a complex and nuanced language. To further enhance the accuracy of the sentiment analysis model, one effective approach is to expand the dataset with unambiguous Kannada text. Increasing the volume of high-quality data can help the model better understand and classify sentiments in Kannada text.

Furthermore, advancements in mathematical computation are being explored, with a focus on assigning weightage to specific tokens. Some words carry more significant weight in a sentence compared to others, and this can have a substantial impact on the overall sentiment polarity. To accommodate this, there is a need for an upgraded corpus model capable of storing words with varying levels of weightage. This enhanced functionality aims to determine the degree of negative or positive polarity in a sentence by computing results based on diverse weight distributions, ultimately leading to more accurate predictions.

Additionally, a hybrid methodology that combines both Machine Learning and Non-Machine Learning techniques could be explored. This hybrid approach leverages the strengths and advantages of both computing paradigms, potentially leading to even more accurate sentiment analysis results in Kannada. By incorporating a variety of techniques and methodologies, it becomes possible to address the unique challenges presented by languages like Kannada and improve the overall performance of sentiment analysis models.

References

[1] Shankar R, Suma Swamy: Corpora based Classification to perform Sentiment Analysis of Mobile Product Reviews in Kannada Language, International Journal of Engineering and Technology 8(5), 5186-5191 (2020).

[2] Rohini V, Merin Thomas, Dr. Latha CA: Domain-based Sentiment Analysis in Regional Language-Kannada. International Journal of Engineering Research and Technology 8(5), 5186-5191 (2018).

[3] K. M. Anil Kumar, et al: Analysis of Users Sentiments from Kannada Web Documents, Eleventh International Multi-Conference on Information Processing 54, 247-256 (2015).

[4] YashaswiniHegde, S.K Padma: Sentiment Analysis using Random Forest Ensemble for Mobile Product Reviews in Kannada, 2017 IEEE 7th International Advance Computing Conference, 777-782 (2017).

[5] Impana P, Jagadish S Kallimani: Cross-Lingual Sentiment Analysis for Indian Regional Languages, International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques 17, 867-872 (2017).

[6] Jayashree R, Srikantha Murthy K: An analysis of sentence level text classification for the Kannada language, International Conference of Soft Computing and Pattern Recognition, 147-151 (2011).

[7] S. Parameshwarappa, V. N Narayana, G.N Bhrarthi, "A Novel Approach to Build Web Corpus", International Conference on Computer Communication and Informatics (ICCCI-2012).

[8] Jayashree R, Sreekanta Murthy K, "An Analysis of Sentence level Text Classification for the Kannada Language" International Conference of Soft Computing and Pattern Recognition (SoCPaR)- 2011.

[9] Deepamala N, Ramnath Kumar P, "Polarity Detection of a Kannada Document" IEEE-2015.

[10] Shankar R, Suma Swamy, A Survey on Sentimental Analysis in Different Indian Dialects, International Journal of Advanced Research in Computer and Communication Engineering, Vol5, 1072-1076. 10.17148/IJARCCE.2016.54262.

[11] Piyush Arora, Sentiment Analysis for Hindi Language, MS Thesis IIIT-H 2013.

[12] Sandeep Chandran,Bhadran V K, Santhosh George, Manoj Kumar P, "Document Level Sentiment Extraction for Malayalam", International Conference On Recent Advances In Engineering, Science & Technology (Icon 2015)
[13] Anu Sharma, Sentiment Analyzer using Punjabi Language, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 9, September 2014
[14] Muralikrishna H, Ananthakrishna T, Kumarasharma, HMM Based Isolated Kannada Digit Recognition System using MFCC, International Conference on Advances in Computing Communications and Informatics (ICACCI)-2013
[15] KalyanamaliniSahoo, V EshwarchandraVidyasagar, Kannada WordNet - A Lexical Database, TENCON 2003