IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 8, August 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.625

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Sentimental Analysis for Kannada Language

Aishwarya OS, Ashwini KL, Deeksha CM, Lakshmi S, TP Channabasavaraju

UG Students, Department of C.S.E., SSIT, Sri Siddhartha Academy of Higher Education, Tumkur, Karnataka, India

Assistant Professor, Department of C.S.E., SSIT, Sri Siddharth Academy of Higher Education, Tumkur, Karnataka, India

ABSTRACT: Natural Language Processing has a variety of uses, including sentiment analysis. Reading, comprehending, and extracting meaningful information from unstructured material are the objectives of natural language processing. Almost 3.2 billion people are active internet users right now. All business has gone online as a result of advances in technology, including e- commerce, movie ticket buying, education, and other areas. In order to advance a firm, it is crucial to understand user opinions on social media. Sentiment Analysis can be used to examine these users' opinions. Sentiment analysis in the Kannada language has received a fair amount of earlier research, with an accuracy rate of 72%. The methodologies utilised for sentiment analysis of texts in Indian regional languages utilising natural language processing are discussed in this work. focused on classifying the four emotions of anger, fear, joy, and sadness using machine learning algorithms like Linear SVC, Logistic Regression, SGD, K-Nearest Neighbors, Multinomial Naive Bayes, and Random Forest Classifier. The Linear Support Vector Classifier, with an accuracy of 87.25%, is the algorithm that performs the best overall.

KEYWORDS: Sentiment Analysis, Social Media, Kannada Language, Machine Learning, Emotions, Accuracy, Algorithms.

I. INTRODUCTION

The use of technology has been increasing rapidly during the previous few years. Because of the quicker communication methods made possible by social media platforms like Twitter, Facebook, and WhatsApp, among others, digital technology has revolutionized how people live. Over 3.2 billion people use the internet regularly at this time. All industries, including ecommerce, movie ticketing, education, and others, have gone online as technology has developed. To advance a firm to new heights, it is crucial to understand user opinions on social media. Sentiment Analysis can be used to examine the opinions of these users. India is one of the most linguistically diverse country in the world with 22 national languages. Only five percentage of Indian population can communicate effectively in English, while restof the people are comfortable with their regional languages. Due to lack of resource availability, Sentiment analysis in Kannada language has not been explored extensively. Models, approaches, and strategies used in natural language processing explain how the language we use affects how we think and the outcomes we get. The term "sentiment analysis" describes the systematic identification, extraction, quantification, and study of affective states and subjective information using natural language processing, text analysis, computational linguistics, and biometrics. Sentiment analysis is a task in natural language processing and information extraction that looks for the writer's emotions as they are expressed in reviews, inquiries, and requests, whether they are positive or negative. Sentiment analysis broadly seeks to ascertain a speaker's or writer's attitude towards a subject or the general polarity of a document. His or her judgement, judgement, affective state, or purposeful emotional communication can all be categorized as attitude. Sentiment analysis has become increasingly necessary in recent years due to the exponential rise in internet usage and public opinion exchange. The proposed system focuses on classifying the emotions such as Anger, fear, joy, and sadness for textual based sentiment analysis. It may be sentence-based, classifying each and every sentence in the text that expresses sentiment. Sentiment analysis can be phrase-based, where the phrases in a sentence are categorized according to the polarity based on some patterns of their recurrence. Linear Support Vector Classifier with an accuracy of 87.25%, Stochastic Gradient Descent with 85.25%, Logistic Regression with 84.25%, Random Forest Classifier with 85.75%, Multinomial Naive Bayes with 85.50% and K-Nearest Neighbors with an 74.50%. The Linear Support Vector Classifier, which has an overall accuracy

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

of 87.25%, outperforms all other algorithms, while K-Nearest Neighbors, which has an accuracy of 74.50%, has got the least performance among all.

II. LITERATURE SURVEY

Here are summaries of papers on sentimental analysis, including details of the year, authors, title, and a brief introduction to their work:

A.Pasumpon Pandian [1] proposed performance evaluation and comparison using deep learning techniques in sentiment analysis (2021). With the use of several ensemble models and by emphasizing the various features needed for sentiment analysis, the author of this study combined two techniques. The performance analysis of surface and deep ensembles, the advantages of deep learning, and the characterization of current approaches are all included in this work. The data is broken down into six data sets, three of which contain 43 tweets totaling 728 words. These studies demonstrate that the combination of information from many sources, such as impact word vectors, generic features, and surface characteristics, will lead to an improvement in sentiment analysis tasks.

Joshi and Pathak [2] proposed Sentiment Analysis on code-mixed Dravidian Languages, A Non- Linguistic Approach (2021). To perform sentiment analysis on this kind of language, many people utilize a combination of two or more languages, known as code mixed language. In this study, the authors extracted features from the data set using TF-IDF and the Keras Tokenizer API. Machine learning (MNB, DTC) and neural networks (ANN, CNN) approaches are used on these extracted characteristics, and the results are submitted to the HASOC 2021 shared problem to enhance transfer learning (BERT). They used 40230, 69675, and 15800 unique words from the Dravidian languages of Malayalam, Telugu, and Kannada, respectively, . Precision recall and F1 score of 0.78 were obtained for Malayalam language MNB utilizing combined word n-gram. With an F1 score of 0.65 for Tamil, ANN did better. In terms of Kannada language, CNN and ANN scored 0.63 and 0.64 on the F1 scale, respectively. The drawback of this strategy is that, in contrast to character-gram feature and combined word feature, machine learning approach fails horribly to code mixed sentences.

Rakshita et. al.[3] Proposed Sentiment analysis of Indian regional languages on social media (2021). This model primarily uses five regional languages, including Telugu, Tamil, Malayalam, Hindi, and Kannada. First, tweets are scraped from Twitter using Twitter's API. Next, text blob is used. Finally, customer reviews are given various sentiment scores and classified as positive, negative, or neutral by using test classification model. Sentiments are determined using the input data's semantic relations and the frequency of each pre-defined word. This results in precise output. Emojis and images might be taken into consideration for this dataset for any additional information, but the downside of this study is that it only uses text data from Twitter.

Basiri and Nemati [4] proposed a novel fusion based deep learning model for sentiment analysis of COVID-19 tweets (2021). For the purpose of sentiment analysis of Tweets relatedto the Corona virus from eight different countries, the authors of this study implement a novel method based on the combination of four Deep Learning models (NBSVM, DistillBERT, BiGRU, and FastText) and one traditional supervised machine learning model (CNN). They used data from Twitter and Google trends as two sources for this investigation. The model is trained using the Stanford Sentiment140 Data set. Positive tweets in this study indicate that the individual is upbeat about the illness, but negative tweets indicate that they are concerned or unhappy about potential negative effects on their lives. The drawback of this study is that it ignores how worldwide Covid- 19 news and statistics affect the general mood of other nations, and that the keywords chosen for information seeking and tweet filtering are independent of the nation for which the dataset is being extracted. The benefit of this work is that the fusion models that are suggested may accurately analyze the coronavirus-related dataset and can be utilized to identify feelings with promising performance.

Bera and Ghose [5] proposed Sentiment Analysis of Multilingual Tweets Based on Natural Language Processing (NLP) (2021). The authors of this study first conducted a thoroughanalysis of natural language processing (NLP) using simple neural networks (SNN), convolutional neural networks (CNN), and long short-term memory (LSTM) neural networks. They then created an amalgamated model by layering a CNN on top of the LSTM. A total of 4,000 samples of reviews in the languages of Bengali, Hindi, and English are used to create the outputs for the models mentioned above.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The Sentiment Analysis dataset's accuracy of classification for the proportion of positive and negative statements is 84.1%, with an average time complexity.

Ranjitha and Bhanu [6] proposed an improved sentiment analysis for Dravidian language Kannada using decision tree algorithm with efficient data dictionary (2021). The authors of this work suggested a method that computationally recognizes and classifies opinions conveyed in a piece of Kannada-language writing in order to ascertain if the writer has a favorable, negative, or neutral attitude toward a certain topic or product. The decision tree technique for kannada sentiment analysis is used to achieve this. Additionally, the training's data set is culled from a variety of websites, including Prajwani, One India News, and Wedunia. The results were obtained with 85% accuracy, 0.78 precision, and 0.79 recall. This study's drawback is that some terms in the Kannada language cause confusing messages to be produced by machine translation, which leads to incorrect results.

Madan and Ghose [7] proposed Sentiment Analysis for Twitter Data in the Hindi Language (2021). In this work, sentiment analysis for Hindi-language Twitter data was conducted. To connect to the Twitter app and extract the tweets, they used the Tweepy API. The resource utilized to create a sentiment classifier is an improved Hindi SentiWordNet. To categorize the various sentiments of tweets, they have used various classification approaches, including LR, MNB, SVM, Decision Tree, and Nearest Neighbor. They have completed sentiment classification and, at long last, sentiment analysis using both the LBA and HBA approaches. Only Hindi tweets are taken into account, which is a flaw in the system. Using LR and MI, this system has a 93.2% accuracy rate.

Kulkarni et. al [8] Proposed a Marathi tweet-based sentiment analysis data set (2021). In this paper, the authors show the baseline classification results utilizing CNN, LSTM, UMFiT, and BERT base deep learning models as well as the first significant publicly available Marathi Sentiment Analysis Dataset L3CubeMahaSent. There are 16,000 unique tweets in the data collection, which are divided into three categories. Using CNN yielded an accuracy of 83.24%, LSTM yielded an accuracy of 82.89%, UMFiT produced an accuracy of 80.80%, and BERT produced an accuracy of 84.13%.

Rajani Shree and Shambhavi [9] proposed POS tagger model for Kannada text with CRF++ and deep learning approaches (2020). In this article, authors demonstrate two strategies for training parts of speech tagging on Kannada words. Supervised machine learning is the first strategy. CRF ++0.50. The second strategy combines a deep learning technique with word embedding. The data set that was utilized in this implementation. Downloaded from TDIL are 1200 sentences with tags in Kannada. 76% accuracy when using CRF++ 0.50. is acquired. Additionally, 71% accuracy is achieved using the deep learning technique. The study's drawbacks include the necessity for a large amount of words to train the model (for example,different verb tenses where a single verb tag for all action words would increase accuracy) and the challenge in identifying proper and common nouns for the model.

Sharma and Ghose [10] proposed sentimental analysis of twitter data with respect to General Elections in India (2020). The authors of this study used the Twitter API to gather tweets. The tweet is analyzed using the R Language, which is also utilized for pre-processing. Two applicants were given consideration for this study. AYLIEN, a text analysis extension of Rapid Miner, is used in this study to do named entity extraction. They also employed the NRC dictionary-based approach and the lexicon-based approach for sentiment analysis. The noise in the tweets and the lack of sentiment labels made the algorithm more difficult. The outcome of ten was in complete agreement with the actual election outcomes discovered in May 2019.

III. METHODOLOGY

The steps involved in kannada language sentiment analysis are shown in Fig. 5.1 above. The sentences will be taken from manually prepared dataset. After taking a custom input sentence, a sentence is tokenized, or broken up into smaller words. Data cleaning is taking out unnecessary information from reviews that doesn't provide any value, like punctuation, commas, and other elements. Stop words in sentences include the terms the, which, and other words that have no semantic value. The technique of stemming involves tracing a word backto its origin. The practice of classifying groupings of texts into distinct categories is known as classification.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

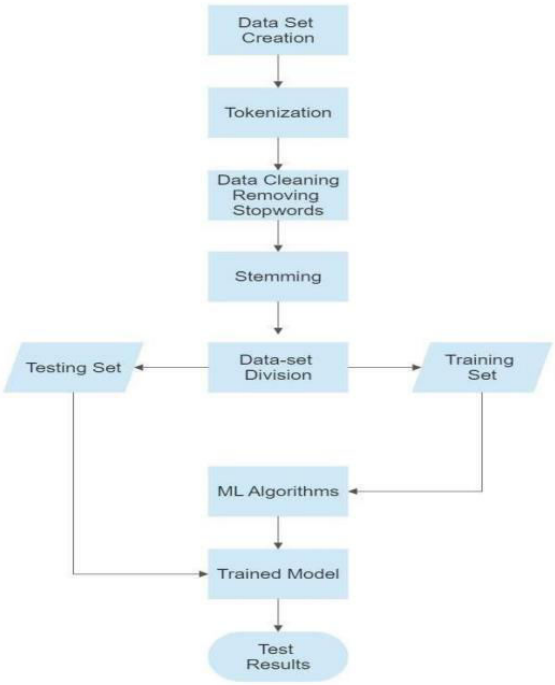

Figure 1: Steps involved in Sentiment Analysis of Indian Regional Language (Kannada)

1. Dataset Creation

A machine learning dataset is a collection of data that is used to train the model. The machine learning algorithm learns how to make predictions using a dataset as an example. Text, image, audio, video, and numeric data are among the common sorts of data. In order for the algorithm to comprehend what the desired output is, the data is typically first labelled or annotated.Data set consists of 604 joy, 520 anger, 520 sad and 356 fear sentences as shown in fig. 5.1

2. Tokenization

As shown in the fig. 5.1 Tokenization involves cutting the raw text into manageable pieces. Tokenization divides the original text into tokens, which are words and sentences. These tokens aid in context comprehension or model development for NLP. By examining the word order in the text, tokenization aids in comprehending the text's meaning. For example, the text “It is raining” can be tokenized into ‘It’, ‘is’, ‘raining’. Tokenization can be done using a variety of tools and frameworks. Some of the libraries that can be utilized to do the work include NLTK, Gensim, and Keras.

3. Data Cleaning

In NLP, data cleaning is a crucial step. The dataset is like a collection of words that the computer cannot understand without cleansing the data as shown in fig.5.1. In this process, duplicate, incorrect, and peripheral data elements are found, and the undesired material is modified, replaced, or deleted. In natural language processing (NLP), data cleaning entails removing numerous punctuation symbols, such as the comma ' , ', colon ' : ', exclamation mark' ! ', hyphen ' - ', question mark ' ? ', apostrophe ' " ', brackets' { } , [] , () ', semicolon ' ; ', ellipsis (***), and (...).

4. Removing Stopwords

From the fig.5.1, Stopwords are any words or phrases that add no meaning to a statement in any language. The real meaning of the statement will not change if these stopwords are removed. The data size will drop as a result of eliminating these stopwords, and the model's training time will also shorten while performance and accuracy increase. The NLTK library is one of the oldest and most widely used Python libraries for natural language processing. In the corpus module, NLTK aids in locating the list of stop words and promotes their removal. The text must be broken up

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

into words in order to remove stop words; if the word is found in the list of stop words provided by NLTK, it is removed. It gives you the option to add or remove stop words from the list of stop words already present in NLTK.

5. Stemming

In Natural Language Processing (NLP), the process of stripping a word down to its root or suffix- and prefix-affixed word stem is known as stemming as shown in fig. 5.1. In contrast, a stemming algorithm normalizes language by condensing various word forms to their standard form. Through the removal of affixes, the base form of the words is extracted using this method.It is analogous to trimming a tree's branches back to the trunk. For instance, the stem of the words "eating," "eats," and "eaten" is "eat". Stemming is used by search engines to index the words. Because of this, a search engine can only record the stems of a word rather than all of itsvariations. 5.6 Dataset Division In data science or machine learning, data splitting comes into the picture when the given data is divided into two or more subsets so that a model can get trained, tested and evaluated as shown in fig. 5.1. In practice or in real-life projects, data splitting is an important aspect, and it becomes a must when models are based on the data as it ensures the making of machine learning models. Usually, we create two or three parts of the main dataset. If two splits are there, it means one will be utilised for training and another one will be used for testing.

6. Training Set

The training data is the biggest (in -size) subset of the original dataset, which is used to train or fit the machine learning model. Firstly, the training data is fed to the ML algorithms, which lets them learn how to make predictions for the given task.

7. Testing Set

Once we train the model with the training dataset, it's time to test the model with the test dataset. This dataset evaluates the performance of the model and ensures that the model can generalize well with the new or unseen dataset. The test dataset is another subset of original data, which is independent of the training dataset. However, it has some similar types of features and class probability distribution and uses it as a benchmark for model evaluation once the model training is completed. Test data is a well-organized dataset that contains data for each type of scenario for a given problem that the model would be facing when used in the real world. Usually, the test dataset is approximately 20-25% of the total original data for an ML project.

Machine Learning Algorithms

Preprocessing will take place before classification, as was stated in prior sections. The classification process, which divides tweets into four categories—joy, anger, fear, and sad—is a crucial part of sentiment analysis. We employ categorization algorithms in order to categorize sentences. SGD Classifier, Linear SVC,K-Neighbors Classifier, Multinomial NB, and Random Forest Classifier are some of the classification methods we employed in our model. Using these algorithms, sentences are to be classified, and effectiveness is to be evaluated.

1. Linear Support Vector Classifier

The fundamental goal of linear SVC is to categorize the given data and provide the best fit hyperplane for the data. Some features can be fed to the classifier after obtaining the hyperplane to see what class it would predict. First, draw a straight line that is unbiased to all classes and has the distance between the +support vector and line equal to the -support vector. Second, the narrow margin values indicate that the support vectors are overly sensitive. Support vector variance won't be robust if the dataset or classifier change.

2. Random Forest Classifier

Supervisory learning algorithms like Random Forests are used. Both classification and regression are accomplished with it. The algorithm is simple and very adaptable. The trees that make up a forest. A forest will grow larger as more trees are present. On arbitrary picked data samples, it creates a few decision trees, receives predictions from each tree, and then uses voting to determine which is the best option. For example, picture classification and recommendation engines are only two examples of how Random Forest is used. Additionally, Random Forest is utilized to categorize dependable loan applicants, uncover fraudulent behavior, and anticipate diseases. A dataset's key attributes are chosen via the Boruta Algorithm.

3.k-NN

The k-nearest neighbors (k-NN) algorithm is a simple machine learning method used for classification and regression. In classification, it assigns a data point to a class based on the majority class of its k closest neighbors. In regression, it predicts a value based on the average of the values of its k nearest neighbors. The algorithm is non-parametric and instance-based,

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

meaning it doesn't explicitly learn a model but instead relies on the distance metric (like Euclidean distance) to make predictions.

4. Logistic Regression

Logistic regression is a statistical method used for binary classification, where the outcome is a categorical variable with two possible values (e.g., yes/no, true/false). It models the probability that a given input belongs to a particular class by applying a logistic function (also known as the sigmoid function) to a linear combination of the input features. This results in an output between 0 and 1, which can be interpreted as the probability of the input belonging to the positive class. Logistic regression is widely used due to its simplicity, interpretability, and effectiveness in many practical scenarios.

5. Multinomial NB

Multinomial Naive Bayes (Multinomial NB) is a variant of the Naive Bayes algorithm specifically designed for classification tasks involving discrete data, such as word counts in text classification. It assumes that the features (e.g., word occurrences) follow a multinomial distribution, making it particularly effective for tasks like spam detection, sentiment analysis, and document categorization.

In Multinomial NB, the algorithm calculates the probability of each class given the feature values and assigns the input to the class with the highest probability. It works under the assumption of conditional independence, meaning it assumes that the presence of a particular feature in a class is independent of the presence of any other feature, given the class. Despite this simplifying assumption, Multinomial NB performs well in many practical scenarios, especially with large datasets.

6. Stochastic Gradient Descent

The SGD (Stochastic Gradient Descent) classifier is a linear model that optimizes using stochastic gradient descent, making it highly efficient for large-scale and high-dimensional datasets. It supports models like SVM and logistic regression, updating model parameters incrementally with each training example. This approach speeds up training but can introduce noise, potentially leading to less stable convergence. Regularization techniques such as L1, L2, or Elastic Net can be applied to prevent overfitting, making SGD a popular choice for tasks like text classification and image recognition.

IV. RESULTS AND DISCUSSIONS

Final Output

```
Enter A sentence
ವೈಭವ್ ನ ನಡೆಗಳು ಅಥವಾ ವರ್ತನೆಯಿಂದ ನಾನು ಸಿಟ್ಟಾಗಿದ್ದೇನೆ ಮತ್ತು ಕಿರಿಕಿರಿಗೊಂಡಿದ್ದೇನೆ
After Cleaning and Stopwords Removal
ವೈಭವ್ ನ ನಡೆಗಳು ವರ್ತನೆಯಿಂದ ಸಿಟ್ಟಾಗಿದ್ದೇನೆ ಕಿರಿಕಿರಿಗೊಂಡಿದ್ದೇನೆ

After Stemming
ವೈಭವ್ ನ ನಡೆ ವರ್ತನೆ ಸಿಟ್ಟಾಗಿ ಕಿರಿಕಿರಿಗೊಂಡಿ

Output of each Algorithms
['Anger', 'Anger', 'Anger', 'Joy', 'Anger', 'Anger']
The Sentence Is classified as a  Anger Sentence
```

Fig. 6.8 Final output of an anger sentence

```
Enter A sentence
ನಿರಾಕರಣೆಯ ಭೀತಿ ಆತ್ಮವಿಶ್ವಾಸವನ್ನು ಕಾಪಾಡಿಕೊಳ್ಳಲು ಕಷ್ಟವಾಗಬಹುದು
After Cleaning and Stopwords Removal
ನಿರಾಕರಣೆಯ ಭೀತಿ ಆತ್ಮವಿಶ್ವಾಸವನ್ನು ಕಾಪಾಡಿಕೊಳ್ಳಲು ಕಷ್ಟವಾಗಬಹುದು

After Stemming
ನಿರಾಕರಣೆಯ ಭೀತಿ ಆತ್ಮವಿಶ್ವಾಸ ಕಾಪಾಡಿಕೊಳ್ಳಲು ಕಷ್ಟವಾಗಬಹುದು

Output of each Algorithms
['Fear', 'Fear', 'Fear', 'Fear', 'Fear', 'Fear']
The Sentence Is classified as a  Fear Sentence
```

Fig. 6.9 Final output of a fear sentence

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The accuracy comparison of all the methods utilized in our model is shown in figure 6.12 above. With 2000 Kannada sentences, our model was trained. Our data has been divided into training and testing. 0.2 of the total data set was testing data. We concluded from the comparison that Linear SVC outperformed all other classifiers. The accuracy rating was 87.25%. SGD classifier accuracy was 85.25%, while that of Logistic Regression was 84.25%. The accuracy of the Random Forest classifier was 85.75%. The accuracy for multinomial naive bayes was 85.50%. K Neighbors Classifier did poorly since it had the lowest accuracy 74.50%.

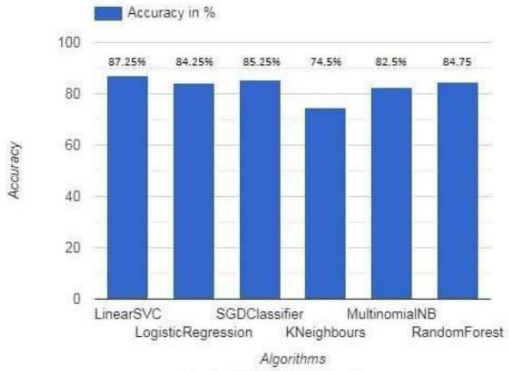

Fig. 6.12 Results comparison

V. CONCLUSIONS

In this project,a classification algorithm of machine learning is proposed to do sentimental analysis for kannada language.. Internet usage has significantly increased in recent years. From shopping to schooling, everything has shifted online. The use of social media for communication is widespread. Therefore, social media generates a lot of data each day. Sentiment analysis is therefore crucial in identifying company insights and achieving significant financial returns. For the English language, there are many sophisticated models for sentiment analysis. The Kannada language has relatively little sentiment analysis, though. We have made an effort to provide a model that is effective for categorizing sentences in kannada using several classification techniques. As part of our project, we gathered 2000 Kannada sentences and manually classified them as joy, angry, fear, or sad sentences. We then used these sentences to train our model. The preparation of data before classification is crucial. It improves model performance and reduces the dataset to a higher level. Preprocessing techniques include tokenization, data cleaning, stop word removal, and stemming. Feature extraction comes after data has been preprocessed. For feature extraction, we employed the TF-IDF approach. We've employed a variety of classification techniques, including Linear SVC, Logistic Regression, SGD, K-Nearest Neighbors, Multinomial Naive Bayes, and Random Forest Classifier. The best performing algorithm overall, the Linear Support Vector Classifier, with an accuracy of 87.25%. Because the effectiveness of the model depends on the data.

REFERENCES

1. Pandian, A. Pasumpon. "Performance evaluation and comparison using deep learning techniques in sentiment analysis." Journal of Soft Computing Paradigm (JSCP) 3.02 (2021): 123- 134.
2. Joshi, Prasad A., and Varsha M. Pathak. "Sentiment Analysis on Code-mixed Dravidian Languages, A Non-linguistic Approach." (2021).
3. Rakshitha, Kakuthota, et al. "Sentimental analysis of Indian regional languages on social media." *Global Transitions Proceedings* 2.2 (2021): 414-420.
4. Basiri, Mohammad Ehsan, et al. "A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets." Knowledge-Based Systems 228 (2021): 107242.
5. Bera, Abhijit, Mrinal Kanti Ghose, and Dibyendu Kumar Pal. "Sentiment analysis of multilingual tweets based on Natural Language Processing (NLP)." International Journal of System Dynamics Applications (IJSDA) 10.4 (2021): 1-12.
6. Ranjitha, P., and K. N. Bhanu. "Improved sentiment analysis for dravidian language-kannada using dicision tree algorithm with efficient data dictionary." IOP Conference Series: Materials Science and Engineering. Vol. 1123.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

No. 1. IOP Publishing, 2021. (2021).

7. Madan, Anjum, and Udayan Ghose. "Sentiment Analysis for Twitter Data in the Hindi Language." 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2021.
8. Kulkarni, Atharva, et al. "L3cubemahasent: A marathi tweet-based sentiment analysis dataset." arXiv preprint arXiv:2103.11408 (2021).
9. Rajani Shree, M., and B. R. Shambhavi. "POS tagger model for Kannada text with CRF++ and deep learning approaches." Journal of Discrete Mathematical Sciences and Cryptography 23.2 (2020): 485-493.Sharma, Ankita, and Udayan Ghose. "Sentimental analysis of twitter data with respect to general elections in India." Procedia Computer Science 173 (2020): 325-334.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details