



Secured Text-Based Emotion Classification Using Machine Learning With NLP

Dr.KavithaSubramani ^{1*}, Nivedha.U², Dr.L.Jabasheela³, Divya.S⁴, Dhuneesha.E⁵

¹ * Professor, Department of CSE, Panimalar Engineering College, Chennai, India.

²UG Scholar, Department of CSE, Panimalar Engineering College, Chennai, India.

³Professor, Department of CSE, Panimalar Engineering College, Chennai, India

⁴UG Scholar, Department of CSE, Panimalar Engineering College, Chennai, India.

⁵UG Scholar, Department of CSE, Panimalar Engineering College, Chennai, India.

*corresponding author email : kavitha.pec2022@gmail.com

Citation: Dr.KavithaSubramani et al. (2024 Secured Text-Based Emotion Classification Using Machine Learning With NLP

Educational Administration: Theory And Practice, 30(5), 901-910

Doi: 10.53555/kuey.v30i5.2986

ARTICLE INFO

ABSTRACT

The collection and evaluation of emotions are for the focus of the sentiment analysis' subfield known as emotion detection. With the ease of obtaining data and the numerous advantages its deliverables provide, many researches are being conducted in the area of text mining and analysis. The proposed approach in text has lately been more well-liked due to its numerous possible applications in marketing, development research, behavioural science, social interaction, automation, etc. In the proposed approach, the text emotion recognition used both speech as well as text to detect emotions. Hence, these methods fall short of creating a useful and flexible system for emotion recognition. A fresh approach was suggested and put into practise for detecting emotions in shortentries. As opposed to conventional methods, which are mostly focused on statistical techniques, this approach attempts to infer and extract the causes of emotions by importing information and theories from other disciplines, such as sociology. The approach of emotion cause extraction is employed as a critical step to enhance the quality of chosen characteristics, and it is based on the idea that a prompting cause event is an essential component of emotion. NLP is used to build the supervised machine learning algorithms, and accuracy metrics are used for comparison. The best model, or the one with the highest accuracy, is then chosen after a comparison of the three, and it is implemented into a webpage. The algorithms used are Linear SVM, Random Forest, Decision Tree Classifier. The three models are then compared and the best one ie the one with highest accuracy is deployed into a webpage. The highest accuracy obtained from the above algorithms is 90%. This work provides a webpage for emotion recognition that takes voice and text input with accurate results.

Keyword: Emotion, sentiment analysis, Supervised Machine learning, Natural language processing, textual emotion recognition

1.Introduction

Billions of users in this world are attracted by social media, which has been expanded quickly throughout the world for the past few years. Through these social networks there is a huge growth in the data generated by the users which has drawn the interest of the researchers to gain any useful information from this data which led to the need for developing NLP(Natural Language Processing) tools and methods. The main subjective data which is nothing but the sentiment analysis is dealt using this NLP. It deals with two different tasks 1.Opinion mining and 2.Emotion recognition. Emotion Recognition has more interest in the research community because it has many benefits in different fields for example, suicide prevention, children and student motivation and performance, reducing cyber-bullying, etc., NLP researchers has applied different methodologies for the textual emotion recognition like Machine Learning(ML), rule based methods, lexical approaches. The most used method is the ML technique because it yields better results than many of the other methods.

The performance and accuracy of a model depends on the size and quality of training dataset. Therefore, the

training data must be carefully chosen to get more accuracy and good performance. This paper demonstrates supervised machine learning technique along with NLP which is used to work with natural human language. Data visualization is done to get better insights about the data. Three algorithms are used for textual emotion recognition. The three models are compared based on accuracy and deployed.

The information about the paper is presented as follows. Section 2 shows the information about related work for emotion recognition. Section 3 presents the proposed work. Section 4 presents the results obtained and discussion about the results. Section 5 and 6 contains performance evaluation and comparative analysis. Section 7 & 8 presents conclusion & future work and references.

2. Related Work

One of the numerous AI-based technologies that has significantly enhanced how people and society can address urgent issues is natural language processing (NLP). They include traditional algorithms-based models like the support vector machines (SVM), hidden markov model (hmm), and decision tree, which were developed in response to the need to improve efficiency and performance by streamlining the emotion annotation process.[1] In light of this, they developed EmoLabel, a tool that identifies the most common emotions through two stages: automatic pre-annotation and human annotation. Emotions are used to categorize the thoughts expressed on social media. Among the various emotions are joy, sadness, fear, disgust, rage, surprise, and trust. In order to engage with people more closely, upcoming artificial intelligence will need to be able to recognise and express emotional states. Although the extraction of emotions from written representations of human conversation has shown encouraging results, the accuracy of acoustic feature-based emotion recognition from audio is still lacking. The suggested methodology is improved by a[2] novel feature extraction method based on Bag of Audio Words (BoAW) and cutting-edge recurrent neural networks. Several scholars have suggested methods for finding anomalies using different text mining approaches. Each remark or tweet is updated in casual human handwriting, and unstructured texts are standardised into a standard format to apply ML algorithms utilising NLPT Natural Language Processing methods. Views[3] expressed on social media are categorised by emotions. These categories include joyful, sad, fear, disgust, wrath, surprise, and trust. This study examines how microblogs may be used to identify anomalies in social media. Text mining and emotion recognition methods from several authors deepen the study. In recent decades, a lot of research has been done on campus or institution security. Assaultants have been deterred from accessing a facility by using licence plate recognition, voice verification, and facial identification separately. A hybrid recognition system may greatly boost security, according to various experts, however hybrid systems aren't frequently covered in the literature. In order to solve this problem,[4] a hybrid driver and vehicle identification module that can identify both the driver and the vehicle was introduced in this study. Face and speech recognition software is used to identify drivers. FaceNet was used to identify faces, and multi-task cumulates convolutional networks were used to crop the faces for facial recognition.[5] Dynamic interactions occurring throughout the data procedure, and they suggest the use of a Dynamic Interactive Multiview Memory Network (DIMMN) model to include relevant information for emotion recognition. Particularly, numerous views are used inside DIMMN to combine information. TFIDF, a metric for word importance in a text, is used in certain research to propose a novel way of emotion detection[6]. Using sentiment analysis, you can tell if the material you're given has a positive, unfavorable, or neutral attitude. Nevertheless, emotion analysis goes farther than that and operates by segmenting the sentiment analysis categories[8]. The research put out makes use of machine learning neural networks to determine the gender and disposition of a speaker. Two modules, notably the blocks for gender identification and emotion detection, must be integrated to complete the assignment. The emotion detection block is built by a convolutional neural network (CNN), trained and tested on a large dataset, while the gender identity block is formed by a simple feed-forward neural network[10]. Individuals utilize websites such as Twitter to textually convey their views and opinions about relevant problems. For greater relationships between individuals and machines, it is crucial to comprehend emotions at a subtler level than just feeling something.[11] Consequently, in this the BERT language model towards emotion detection in Tweets written in Indonesian. Rather than pre-training, which necessitates a large amount of data and resources, employ fine-tuning. The efficiency and efficacy of the suggested model were evaluated using two pre-trained models.

3. Proposed Work

The proposed methodology has the following steps:

- Dataset Collection and Preprocessing
- Data visualization
- Model development
 - Random Forest
 - Linear Support Vector Machine
 - Decision Tree Classifier
- Deployment

Dataset Collection: This is the first phase of textual emotion recognition. The choice of dataset should be done very carefully because the quality and size of dataset is the most important one in supervised machine learning for yielding better results than any other methodologies. This process involves identifying relevant data sources, collecting the data, and preparing it for analysis. Before collecting the data, it is important to identify the purpose and scope of the dataset.

It is important to identify the purpose and scope of the dataset, determining data sources, obtaining necessary permissions, collecting the data, cleaning and preprocessing the data, storing and organizing the data, and analyzing and visualizing the data to gain insights. In the context of text emotion recognition, dataset collection involves collecting a set of text data that is annotated with the corresponding emotion label. For this work, we have chosen “emotion in text” dataset from Kaggle. This contains six emotions namely anger, fear, joy, sad, surprise, love. The shape of dataset is (21549,2).



Fig1.Emotions in dataset

Data Preprocessing: It is the process of cleaning the imbalanced data, unwanted data, duplicate data, missing values. It involves

- Tokenising the data
- Removing stopwords
- Eliminating punctuation
- Encoding the data

Emotions in dataset:

Anger	Fear	Joy	Love	Sadness	Surprise	Total
2993	2652	7029	1641	6265	879	21459

Table 1.Emotions in dataset

Data visualization: Data visualization refers to the graphical representation of data and information. It is a way to communicate complex information in a simple, clear and easy-to-understand manner. By visualizing data, patterns, trends and relationships can be easily identified and interpreted.

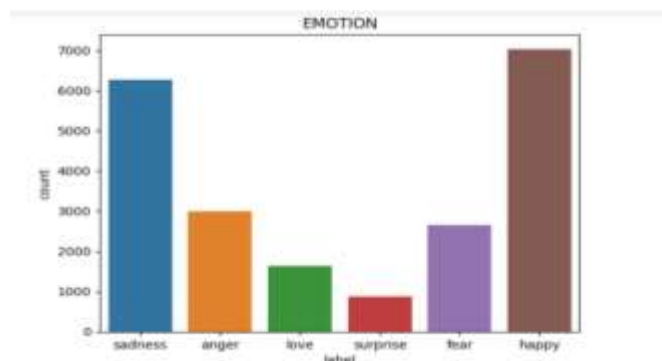


Fig2.Screenshot of Data visualization of different emotions

Using the dataset, the above bargraph was generated. This represents the no: of sentences that represents each emotion. In this paper, we use “matplotlib” and “seaborn” library for data visualization. Using countplot, the above graph was generated. Word clouds are a type of data visualization that can be used to visualize the most commonly used words in a text. Word Clouds were generated for all the emotions in pairwise manner. Overall, data visualization can be a powerful tool for analyzing and understanding patterns in text-based emotion recognition, and can help researchers and practitioners make more informed decisions about how to interpret

and use emotional data.

Model Development

After pre-processing and visualization of data, algorithm implementation takes place. It starts with importing necessary libraries such as pandas, regular expressions, stopwords and WordNetLemmatizer from the Natural Language Toolkit (nltk) for text pre-processing. The preprocessing method is the same for all the three algorithms. Then it loads the dataset using Pandas and preprocesses the text data using the function preprocess_text(), which converts the text to lowercase, removes punctuation and digits, removes stopwords and lemmatizes the words. After pre-processing, the text data is converted to numerical vectors using TfidfVectorizer, which is a method for converting text data into a numerical form that can be used by machine learning algorithms. Then, the data is split into training and testing sets using train_test_split from scikit-learn, and data is trained on the by using the corresponding libraries and functions.

$$\text{Train set} + \text{Test set} = \text{Total Dataset}$$

For example, Decision Tree Classifier uses “from sklearn.tree import DecisionTreeClassifier”. Next, the trained model is used to predict the emotions on the testing data and evaluate the performance using metrics like accuracy, precision, recall, and f1-score.

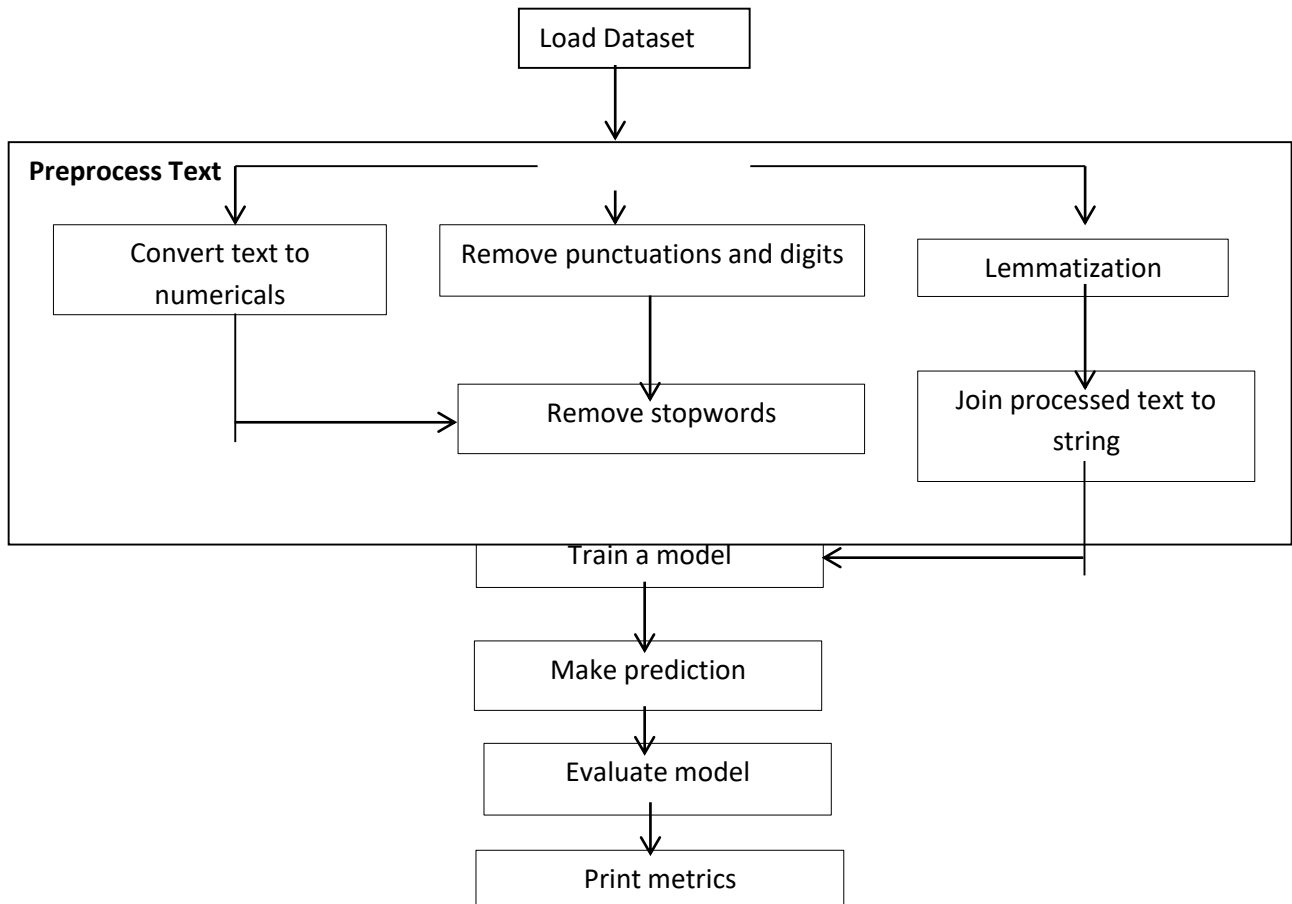


Fig3. Flow of modules

1) Random Forest

Random Forest is trained used the “RandomForestClassifier” module form sklearn.ensemble library. The dataset is split into train and test ie. “X_train, X_test, y_train, y_test”. The size of test dataset is 20% of the total dataset. The value of random state used to train the model using Random Forest is 42. One of the hyperparameters used in random forest is “n_estimators”. The default value is 10 and it is set to 100 in this model. The n_estimators represents the number of decision trees.

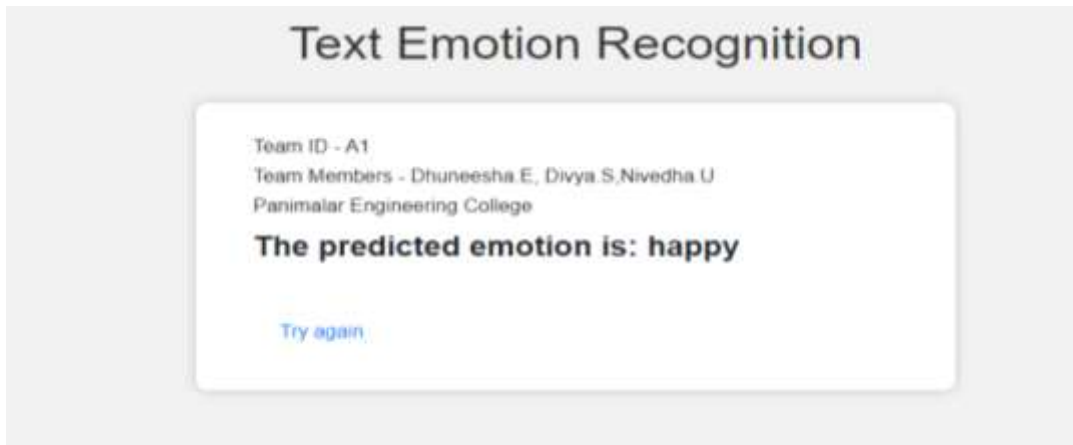


Fig 7. Screenshot of the predicted emotion is displayed

The output screen includes a heading that describes the need of the application, which is detecting emotions in text. There is an input box where users can enter the text they want to analyze. A "Submit" button is provided to initiate the analysis process. After the analysis is completed, the output box displays the detected emotion(s) in the input text. The example shows how the output screen might work for a positive input text that expresses joy. Depending on the specific implementation of the application and the machine learning model behind it, the output could detect and display a variety of different emotions, such as sadness, anger, fear, disgust, or surprise.

5. Evaluation Metrics

i. Random Forest Confusion Matrix

[[536	22	25	4	30	0]
[19	451	10	1	32	18]	
[14	13	1289	32	27	6]	
[3	2	80	230	3	0]	
[18	21	54	13	1164	7]	
[1	19	15	1	5	127]]]	

The TP value for an emotion is the value where the actual value is same as predicted value ie. for anger the TP value is 532. The FN for anger is $[20+29+4+31+1]=85$. The FP for anger is $[18+19+4+23+1]=55$. TN value for anger is $[454+ 8+ 1+ 27+ 23+15+ 1278+ 32+31+6+3+ 85+ 223+ 3+0+19 +57+12+ 1160+ 6+21+18+ 1+ 5+ 122]=3610$

Classification Report

```
In [0]: from sklearn.metrics import classification_report
print("Classification Report of Random Forest\n",classification_report(y_test,y_pred))
```

Classification Report of Random Forest				
	precision	recall	f1-score	support
anger	0.89	0.86	0.88	617
fear	0.85	0.85	0.85	531
happy	0.87	0.93	0.89	1381
love	0.82	0.70	0.75	318
sadness	0.92	0.91	0.92	1277
surprise	0.77	0.73	0.75	168
accuracy			0.88	4292
macro avg	0.85	0.83	0.84	4292
weighted avg	0.88	0.88	0.88	4292

Fig 8. Screenshot of the classification report of Random Forest

**ii. Linear SVM
Confusion Matrix**

$$\begin{bmatrix} 536 & 18 & 19 & 5 & 37 & 2 \\ 17 & 453 & 16 & 0 & 28 & 17 \\ 9 & 11 & 1305 & 27 & 23 & 6 \\ 3 & 2 & 61 & 244 & 8 & 0 \\ 20 & 12 & 33 & 5 & 1203 & 4 \\ 2 & 19 & 13 & 1 & 11 & 122 \end{bmatrix}$$

The TP value for an emotion is the value where the actual value is same as predicted value ie. for anger the TP value is 536. The FN for anger is $[18+19+5+37+2]=81$. The FP for anger is $[17+9+3+20+2]=51$. TN value for anger is $[453+ 16+ 0+ 28+ 17+11+ 1305+ 27+ 23+6+2+ 61+ 244+ 8+0+12 +33+5+ 1203+ 4+19+13+1+11+122]=3624$

Classification Report

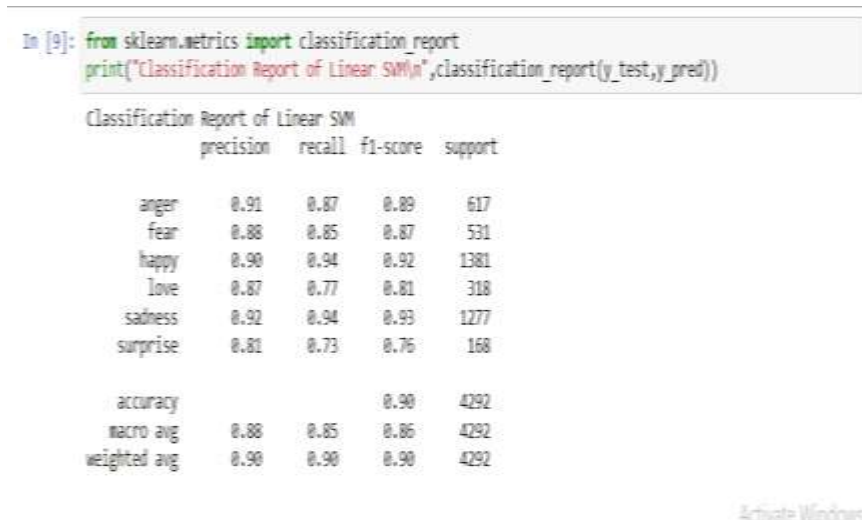


Fig 9. Screenshot of the classification report of Linear SVM

**iii. Decision Tree Classifier
Confusion Matrix**

$$\begin{bmatrix} 535 & 16 & 25 & 7 & 30 & 4 \\ 31 & 442 & 13 & 2 & 21 & 22 \\ 23 & 16 & 1212 & 54 & 66 & 10 \\ 1 & 3 & 69 & 243 & 2 & 0 \\ 30 & 34 & 60 & 18 & 1124 & 11 \\ 3 & 27 & 8 & 1 & 2 & 127 \end{bmatrix}$$

The TP value for an emotion is the value where the actual value is same as predicted value ie. for anger the TP value is 534. The FN for anger is $[17+29+5+27+5]=83$. The FP for anger is $[30+23+1+34+1]=89$. TN value for anger is $[433+ 12+3+ 20+ 333+18+ 1218+ 56+ 57+9+3+ 73+ 236+ 5+0+31 +57+17+ 1126+12+26+12+ 2+2+125] = 3586$

Classification Report

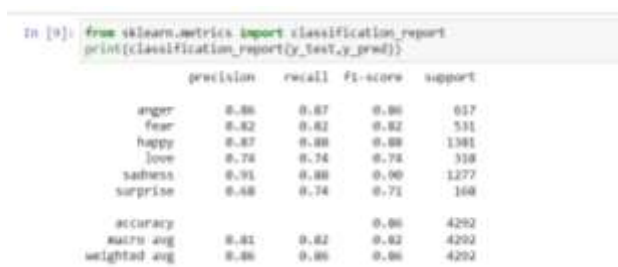


Fig 10. Screenshot of the classification report of Decision Tree Classifier

From the above evaluation metrics, it can be inferred that every value of precision, recall and F1-score has an average value of 80%. Most of it's value is between 80 to 90%. The accuracy of all the three algorithms is more

than 85%.From the confusion matrices, it can be inferred that the diagonal elements have the highest values.These diagonal elements are the values that represent the no: of correct predictions of emotions.In previous research papers on text emotion recognition, the maximum accuracy obtained was 86.75%. The number of emotions used to train the model were also less. In the proposed work, the model is trained to identify 6 types of emotions. Also the highest accuracy obtained from the three models is 90%.

6.ComparativeAnalysis

F1-score and Accuracy

	Anger	Fear	Joy	Love	Sadness	Surprise	Accuracy
Random Forest	0.88	0.85	0.89	0.75	0.92	0.75	88%
Linear SVM	0.91	0.88	0.90	0.87	0.92	0.81	90%
DecisionTree Classifier	0.86	0.82	0.88	0.74	0.90	0.71	86%

Table 2. Accuracy and F1-score of algorithms in proposed work

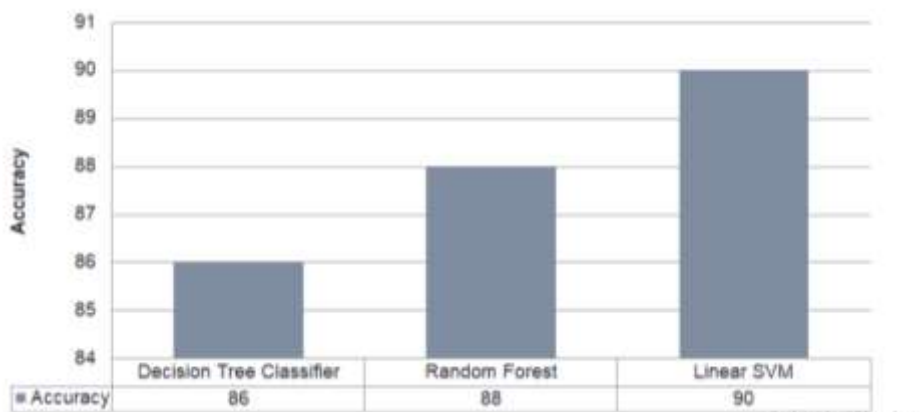


Fig 11.Comparison of algorithms in proposed work

Thus, we have built three models based on three algorithms and each has yielded different accuracies. The first model which was built using Random Forest yielded an accuracy of 88%. The Linear SVM yielded an accuracy of 90% and Decision Tree Classifier model yielded an accuracy of 86%.From this we can say that Linear SVM algorithm yields the highest accuracy. It is then followed by Random Forest. The least accuracy was yielded by Decision Tree algorithm. These models were built using Natural Language Processing technique and Machine Learning algorithms.

S.NO	ALGORITHM	ACCURACY
1	Decision Tree	86%
2	Random Forest	90%
3	Linear SVM	88%

Table 3. Accuracy of algorithms in proposed work

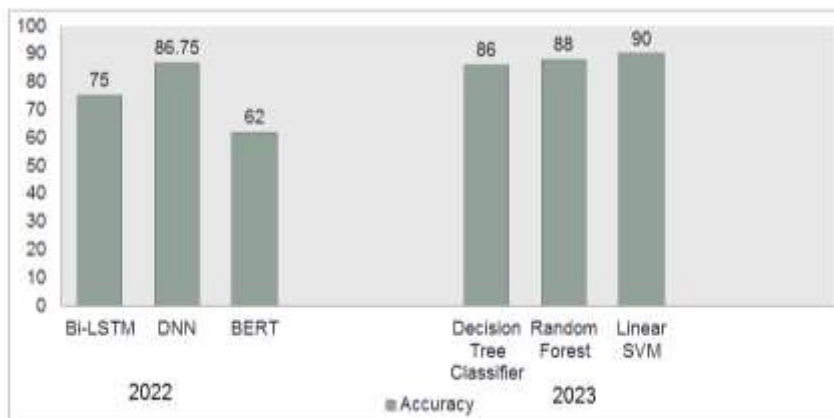


Fig 12.Comparison of accuracies in proposed and related work

Based on the above performance analysis, it is inferred that Linear Support Vector Machine yields highest accuracy. Other two algorithms has also yielded greater accuracy. In 2022 research papers, the algorithms like Bi-LSTM, DNN, BERT yielded 75%, 86.75%,62% accuracy respectively. From this, it is concluded that the proposed work has yielded higher accuracies compared to it's related work and all the three algorithms used in proposed work is greater than 85%.

7.Conclusionand Future Works

Various approaches have been proposed to tackle this problem, including rule-based systems, machine learning methods, and deep learning models. The performance metrics of these models depends largely on the quality and size of the training data, feature engineering, and the choice of the classification algorithm. Nevertheless, recent studies have shown promising results, and emotion recognition based on text has practical applications in many domains, including social media monitoring, mental health diagnosis, and customer service. Furthermore, the development of cross-lingual and cross-cultural emotion recognition models can enable these systems to be used in multilingual and multicultural environments. Finally, the ethical implications of using emotion recognition technology must be carefully considered, and privacy concerns and biases in training data must be addressed to ensure that these systems are developed and used responsibly.

References

1. Lea Canales; Walter Daelemans; Ester Boldrini; Patricio Martínez-Barco, "EmoLabel: Semi-Automatic Methodology for Emotion Annotation of Social Media Text" in IEEE Transactions on Affective Computing (Volume: 13, Issue: 2, 01 April-June 2022)
2. SadilChamishka, IsharaMadhavi, RashmikaNawaratne, DamindaAlahakoon, Daswin De Silva, Naveen Chilamkurti and VishakaNanayakkara , "A voice-based real-time emotion detection technique using recurrent neural network empowered feature modeling" in Multimedia Tools and Applications volume 81, pages 35173-35194, June 2022
3. V. S. Bakkialakshmi and T. Sudalaimuthu , "Anomaly Detection in Social Media Using Text-Mining and Emotion Classification with Emotion Detection" in Part of the Communications in Computer and Information Science book series (CCIS,volume 1697), January 2023
4. Abu Sayeed, Azmain Yakin Srizon, Md. MehediHasan, Jungpil Shin, Md. Al MehediHasan and M. Rasel Mahmud, "A Hybrid Campus Security System Combined of Face, Number- Plate, and Voice Recognition", in Part of the Communications in Computer and Information Science book series (CCIS,volume 1704), January 2023
5. JintaoWen, DazhiJiang, GengTu, ChengLiu, ErikCambria, "Dynamic interactive multiview memory network for emotion recognition in conversation" in Information Fusion Volume 91, March 2023, Pages 123-133
6. VarunSundaram, Saad Ahmed, Shaik Abdul Muqtadeer, R. Ravinder Reddy, "Emotion Analysis in Text using TF-IDF" in 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), March 2021
7. Abd Algani, Y. M., Caro, O. J. M., Bravo, L. M. R., Kaur, C., Al Ansari, M. S., & Bala, B. K. (2023). Leaf disease identification and classification using optimized deep learning. *Measurement: Sensors*, 25, 100643.
8. J. K. S. Al-Safi, A. Bansal, M. Aarif, M. S. Z. Almahairah, G. Manoharan and F. J. Alotoum, "Assessment Based On IoT For Efficient Information Surveillance Regarding Harmful Strikes Upon Financial Collection," 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2023, pp. 1-5, doi: 10.1109/ICCCI56745.2023.10128500.
9. Kaur, C., Kumar, M. S., Anjum, A., Binda, M. B., Mallu, M. R., & Al Ansari, M. S. (2023). Chronic kidney disease prediction using machine learning. *Journal of Advances in Information Technology*, 14(2), 384-391.
10. Eadline, D. (2015). *Hadoop 2 Quick-Start Guide: Learn the Essentials of Big Data Computing in the Apache Hadoop 2 Ecosystem*. Addison-Wesley Professional.
11. M. A. Tripathi, R. Tripathi, F. Effendy, G. Manoharan, M. John Paul and M. Aarif, "An In-Depth Analysis of the Role That ML and Big Data Play in Driving Digital Marketing's Paradigm Shift," 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2023, pp. 1-6, doi: 10.1109/ICCCI56745.2023.10128357.
12. Khan, S. I., Kaur, C., Al Ansari, M. S., Muda, I., Borda, R. F. C., & Bala, B. K. (2023). Implementation of cloud based IoT technology in manufacturing industry for smart control of manufacturing process. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 1-13.
13. M. Lourens, A. Tamizhselvi, B. Goswami, J. Alanya-Beltran, M. Aarif and D. Gangodkar, "Database Management Difficulties in the Internet of Things," 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), Uttar Pradesh, India, 2022, pp. 322-326, doi: 10.1109/IC3I56241.2022.10072614.

14. Abd Algani, Y. M., Caro, O. J. M., Bravo, L. M. R., Kaur, C., Al Ansari, M. S., & Bala, B. K. (2023). Leaf disease identification and classification using optimized deep learning. *Measurement: Sensors*, 25, 100643.
15. Ratna, K. S., Daniel, C., Ram, A., Yadav, B. S. K., & Hemalatha, G. (2021). Analytical investigation of MR damper for vibration control: a review. *Journal of Applied Engineering Sciences*, 11(1), 49-52.
16. Naidu, K. B., Prasad, B. R., Hassen, S. M., Kaur, C., Al Ansari, M. S., Vinod, R., ... & Bala, B. K. (2022). Analysis of Hadoop log file in an environment for dynamic detection of threats using machine learning. *Measurement: Sensors*, 24, 100545.
17. Abd Algani, Y. M., Ritonga, M., Kiran Bala, B., Al Ansari, M. S., Badr, M., & Taloba, A. I. (2022). Machine learning in health condition check-up: An approach using Breiman's random forest algorithm. *Measurement: Sensors*, 23, 100406. <https://doi.org/10.1016/j.measen.2022.100406>
18. Mourad, H. M., Kaur, D., & Aarif, M. (2020). Challenges Faced by Big Data and Its Orientation in the Field of Business Marketing. *International Journal of Mechanical and Production Engineering Research and Development (IJMPERD)*, 10(3), 8091-8102.
19. Zainal, A. G. (2022). Recognition of Copy Move Forgeries in Digital Images using Hybrid Optimization and Convolutional Neural Network Algorithm. *(IJACSA) International Journal of Advanced Computer Science and Applications*, 13(12).
20. Aarif, Mohd, and Ali Alalmal. "Importance of Effective Business Communication for promoting and developing Hospitality Industry in Saudi Arabia." *A case study of Gizan (Jazan)* (2019).
21. Gourank Jain, Satyam Verma, Honey Gupta, Saloni Jindal, MukeshRawat, Kapil Kumar, "Machine Learning Algorithm Based Emotion Detection System", in 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), July 2022
22. Chris Jonathan D and Sujitha Juliet, "Text-based Emotion Recognition using Sentiment Analysis" in 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), May 2022
23. NamrataChaudhari, ReshuAgarwal, MeghnaNarwade, Xinrui Yu, JafarSaniie , "Artificial Intelligence System for Emotion Recognition and Text Analytics" in 2022 IEEE International Conference on Electro Information Technology (eIT), May 2022
24. N. Susithra, K. Rajalakshmi, P. Ashwath, B. Ajay, D. Rohit, S. Stewaugh, "Speech based Emotion Recognition and Gender Identification using FNN and CNN Models", in 2022 3rd International Conference for Emerging Technology (INCET), May 2022
25. KuncahyoSetyoNugroho, FitraAbdurrachmanBachtiar, "Text-Based Emotion Recognition in Indonesian Tweet using BERT" in 2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), February 2022