Proceedings of the International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS 2023)
IEEE Xplore Part Number: CFP22DN7-ART; ISBN: 979-8-3503-0085-7

Comparative Analysis of Emotion Classification using TF-IDF Vector

Mohammed Ali Shaik
School of computer science & Artificial Intelligence
SR university
Warangal, Telangana State, India
niharali@gmail.com

Yamsani Sahithi
School of computer science & Artificial Intelligence
SR university
Warangal, Telangana State, India
yamsanisahithi2003@gmail.com

Mandala Nishitha
School of computer science & Artificial Intelligence
SR university
Warangal, Telangana State, India
reddynishitha31@gmail.com

Rekulapelli Reethika
School of computer science & Artificial Intelligence
SR university
Warangal, Telangana State, India
reethikarekulapelli@gmail.com

Kondabathula Sumanth teja
School of computer science & Artificial Intelligence
SR university
Warangal, Telangana State, India
sumanthkvs1995@gmail.com

Pradhyumn Reddy
School of computer science & Artificial Intelligence
SR university
Warangal, Telangana State, India
chinthalapradhyumnreddy@gmail.com

2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS) | 979-8-3503-0085-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSSAS57918.2023.10331897

***Abstract*— In the current digital era, it is essential to effectively communicate in online settings, and emojis have emerged as a key tool for doing so. In order to improve the precision and usability of emoji selection within digital communication systems, this study offers a novel alternative. The creation of an advanced emotion classification system, supported by a semantic search algorithm, forms the basis of the suggested system. The solution provides a seamless and context-aware experience, in contrast to conventional methods that rely on manual emoji selection or imprecise keyword-based emotion detection. The system surpasses the restrictions of keyword matching by effectively identifying the emotional content within phrases or paragraphs by utilizing cutting-edge natural language processing techniques. Emoji suggestions are given to users by the system along with a Random Forest-based emotion categorization algorithm and real-time emoji-Emotion mapping, redefining how emotions are communicated in digital chats.**

Keywords*— *Machine Learning, Emoji Selection, Classification, Random Forest, Semantic Search, Emotion Classification, Natural Language Processing

I. INTRODUCTION

Nowadays, The transmission of information and feelings via text is now a crucial part of interpersonal communication. The requirement to comprehend and appropriately interpret the attitudes and emotions portrayed in written text is expanding as social media, messaging services, and online interactions become more commonplace. Emojis have become a worldwide language for expressing feelings, giving the online exchanges more richness and depth. The difficulty, however, is in effectively automating the process of categorizing text-based emotions and connecting them with the relevant emojis.

By creating an Emotion Classification System driven by a cutting-edge semantic search algorithm, the proposed paper aims to close this gap. This system is made to quickly and accurately identify the emotions included in user-provided words or paragraphs and couple them with the appropriate emoji. This paper intends to transform interaction and communication experiences across numerous digital platforms by leveraging the capabilities of "Natural Language Processing (NLP)" and "Machine Learning (ML)".

To develop a system that can correctly and automatically identify the emotions represented in the proposed research, the study aims to execute an advanced emotion classification system accompanied by a semantic search algorithm. Poetry is a kind of literature that uses language, rhythm, and imagery to express sentiments and emotions. However, because they are frequently veiled, complicated, and context-dependent, emotions are challenging to analyze and comprehend. Because of this, the study seeks to develop a deep neural network model that can classify manuscripts into several emotion categories by learning from a vast corpus of poetry texts.

II. LITERATURE REVIEW

The primary objective of [1] study is to create and assess machine learning methods for sentiment analysis of Turkish tweets. Emojis can contribute important context for deciphering the text's emotional tone; hence, the research attempts to use them as an extra source of sentiment data [2]. In this study, the classification models is developed for sentiment analysis utilizing two distinct vector visualizations based on BOW and fastText [4]. In NLP, the "bag of words" visualization technique is employed. It entails breaking down a piece of text, like a sentence or document, into a "bag" of single phrases while ignoring grammar and word order and focusing only on the frequency of each word [5].

This study aims to develop a model that can recognize hand-drawn emojis with accuracy [6]. Emojis are visual symbols that are frequently used in digital communication to represent emotions and ideas. Pre-processing would be necessary to make the obtained hand-drawn emoji images consistent and work with the CNN model [7]. Resizing the photos, changing them from color to grayscale, and possibly standardizing the pixel values are all examples of preprocessing techniques. The 400 photos of the eight different emojis have been given to the CNN model. The model has four convolutional layers with 16, 32, 64, and 128 filters each, and a two-size kernel [8].

Reference [9] 's primary goal is to create a model to determine a texter's gender (i.e., how a male or a female produced the text) from brief text messages and emojis. The dataset was initially divided into "training and testing" sets through which the model would be developed, and on the testing set, it would be assessed. They then used text-only classification, emoji-only classification, and "Text and Emoji" based type to predict the gender. A Decision Tree utilizing a Bag of Words produces the most outstanding results [10]. The results, comprising the model's capacity to identify gender from brief sentences and Turkish-language emojis, will be discussed in the paper's conclusion [11].

The major objective of the study is to develop a system that can recognize various emotions in text or content. Happiness or, sadness or, rage or, surprise, and other emotions are only a few examples. Due to its numerous potential uses, interest in multi-label emotion detection is growing. [12]. The goal is to identify every conceivable emotion present in a textual statement. In order to overcome the issues raised above, a “Multi-label Emotion Detection Architecture (MEDA)” is suggested in this study. “Multi-Channel Emotion-Specified Feature Extractor (MC-ESFE)” and “Emotion Correlation Learner (ECorL)” make up the majority of MEDA.[13]

Emojis that are appropriate for the context and emotional expression of the text will be predicted. The fundamental goal is to develop a machine that can accurately predict an emoji that fits the supplied sentence [14]. This study aimed to offer logical emoji suggestions for user input, enhancing the conversation's semantics and making it more apparent to the reader on the contrary side [15]. The suggested method seeks to create a classifier with the highest level of accuracy that can predict the emoji for a given text. Given that it meets the necessary requirements, they applied the Linear SVC classifier. The correlation matrix shows many overlapping emojis in a logical way [16].

Emojis and text are combined early and late to improve opinion mining. Early fusion is a technique used in this study to connect text and emoji data at the start of analysis and use it as a single input for sentiment analysis [17]. This method improves sentiment interpretation by capturing both textual information and emotional clues from emojis. A more thorough sentiment analysis is subsequently produced by combining the results while considering each one's unique peculiarities [18]. In this case, opinion mining techniques make use of distinct methodologies, which includes: lexicon-based, ML-based, and aspect-based sentiment analysis, which aids organizations in adapting their strategy, enhancing their goods and services, and improving customer engagement. The other methods are mainly in Western languages; hence, this study is for Arabic feeling [19].

Emoji use when texting, the significance of emotion categories for sentiment analysis, and the pre-processing of data are all first discussed. The Data Set is then provided. The pre-processing of data is a crucial stage in any machine learning framework, according to this [20]. Before any further processing, raw data or unstructured data must be converted to a structured representation. The algorithm described in this study successfully distinguishes the two participants' messages from various WhatsApp chats. Emojis and sentences are then separated and translated into the target language to make them generic & unified [21].

Emojis and emoticons are pretty helpful for expressing feelings and sentiments in tweets. They can improve the precision and depth of sentiment analysis by being incorporated into tweet sentiment classification written in western languages, and this study focuses on sentiment on Indonesian Twitter [22]. This SVM method can still provide a model to obtain performance even when some of the emojis and emoticons cannot be categorized into specific sentiment groups. The enormous diversity of emoji and emoticon properties is reduced after being converted into sentiment categories, impacting the model's performance [23].

Using emoticons as indications of sentiment to categorize the sentiment of tweets without utilizing labeled training data. To categorize the tweets, this paper employed SVM and a bag of terms [24]. They suggest using emoticons to perform an unsupervised sentiment classification on Twitter data. Lexicon is produced based on emoticons, so classifiers are utilized to categorize tweets from any domain. Results show that semi-supervised (OTAWE) classification without taking Emoticons into account are worse than unsupervised classification-based [9].

Emojis are represented in a numerical vector space so that they convey their subtle emotional undertones and semantic implications. These emoji embeddings can be used to improve tasks involving visual sentiment analysis, such as categorizing the emotion shown in pictures, movies, or other visual content [25]. Emoji connection with their matching visuals in social media still needs to come by. Visual material is analyzed to ascertain the feelings, attitudes, or mood that are conveyed in the pictures or movies. A minimal dimensional embedding consistently outperforms both the more complex and tailored SOTA model and the widely used object-based embedding through assessment of sentiment & emotion recognition [10].

III. Problem Definition

The problem identified for this paper is the incorrect emoji suggestion based on specific words rather than the context of the complete sentence. The existing methods may forecast emojis when users type a chat on social media platforms like Instagram, WhatsApp, Twitter, etc., based just on individual words without taking the complete sentence meaning and intent into account with produced inaccurate or irrelevant results.

A crucial yet challenging task is appropriately understanding and portrayal of emotions in written content. The ability to successfully communicate emotions and feelings is vital for text-based communication platforms, including social media and customer service encounters. Emojis have gained popularity as a means of expressing emotions. However, it is still difficult to automate the process of matching the appropriate emoji to the relevant text, especially when context and nuance are important considerations. In order to accurately detect the emotions expressed in user-input sentences or paragraphs and couple them with the relevant emojis, an emotion classification system that is equipped with a semantic search algorithm is needed. This research solves this fundamental problem.

The semantic search algorithm applied in the study is a system that can retrieve relevant and similar items based on the user query and emotional preference. The semantic search algorithm supports the emotion classification system in two ways:

First, it helps to create a large and diverse corpus of input texts for training and testing the emotion classification model. The semantic search algorithm can crawl and collect information from various online sources, such as websites, blogs, and social media platforms. It can also filter and categorize the data.

Second, it helps to enhance the user experience and satisfaction of the emotion classification system. The semantic search algorithm can allow the user to input a query in natural language. The semantic search algorithm can then use natural language processing techniques to understand the meaning and intent of the query and then rank them according to their semantic similarity and emotional relevance. This way, the semantic search algorithm can provide the user with personalized and contextualized results that match their query and emotional preference.

IV. Proposed Methodology

A. DATASET

The accuracy of the analytic results and the quality of the data are both critically dependent on data preprocessing. Different procedures may be required depending on the type of data being processed and the study's objectives. These procedures increase the effectiveness of data mining and improve the precision of the findings.

In this paper a data set is used which is obtained from Kaggle [12]:tweet_emotions The dataset consists of 3 attributes which are: tweet_id, sentiment, content

	tweet_id	sentiment	content
0	1956967341	empty	@tiffanylue i know i was listenin to bad habi...
1	1956967666	sadness	Layin n bed with a headache ughhhh...waitin o...
2	1956967696	sadness	Funeral ceremony...gloomy friday...
3	1956967789	enthusiasm	wants to hang out with friends SOON!
4	1956968416	neutral	@dannycastillo We want to trade with someone w...

Fig. 1. First five rows of dataset

- tweet_id: It is a unique identification number
- sentiment: It tells the emotion of content
- content: It contains the text to classify the emotion
- The dataset comprises of 40007 rows of data with three columns

The pie chart in figure 2 illustrates several emotional states: “anxiety, neutrality, melancholy, happiness, love, surprise, joy, relief, hate, emptiness, excitement, boredom, and wrath”.

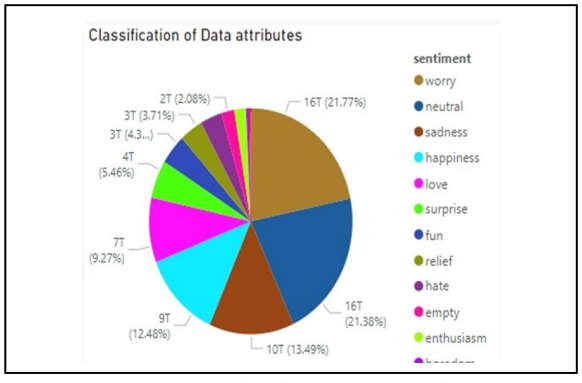

Fig. 2. Classification of data attributes

B. DATA PRE-PROCESSING

Unclean data can be turned into a clean dataset by preprocessing the data.

- Data Collection: The first step entails compiling a broad and representative dataset of text samples that include phrases or paragraphs that express a range of emotions required for system training and evaluation.
- Attribute removal: To acquire the best accuracy and to exclude noisy or irrelevant data, only a small number of characteristics are used to train the model.
- Tokenization: During this stage, the text is separated into individual words or tokens. Tokenization breaks the text into its fundamental parts, making it simpler for the system to analyze and interpret, and is therefore necessary for further analysis.
- Stop Words Removal: Stop words that are regularly used (like "and," "the," and "in") are frequently omitted in order to minimize noise in the data. However, stop words should be eliminated in a way that is sensitive to the context and takes into consideration the unique requirements of the system.
- Encoding Labels: To assist model training in supervised learning, where emotions are labelled, the emotion labels are encoded into numerical values. The training of the emotion classification model requires this step.
- Lowercasing: Making all text lowercase enables case-insensitive matching in later processing phases. By doing this, the system is prevented from treating the same term in several circumstances as a separate entity.
- Remove punctuations: Eliminating punctuation from text standardizes and simplifies the data, boosting its suitability for text-based operations like tokenization and normalization while also lowering noise.

C. DATA SPLITTING

The application transmits authentication data and tokenization data to the tokenization system as part of the tokenization process. The tokenization system validates the data format and authentication details. A token is created using one-way cryptographic methods and stored in a very secure data vault. The program receives the new token for future use. The dataset is typically divided in half, 80:20 each into the “training and test sets”. The model's performance is evaluated on the test set once it has been made trained to classify emotions using the training data. A Data Frame is displayed as a word cloud utilizing the Matplotlib display program and the Python Word Cloud module. Use text data saved in a DataFrame to generate a word cloud visualization with the goal of emphasizing the words that attains the frequently used text. The graphical depiction of word frequencies is shown in figure 3 enlarges words in proportion to frequency of supplied text data.

Fig. 3. Word cloud of content [2,7]

D. ALOGORTHIM

Algorithm: Proposed

Step 1: Define a set of emojis and their corresponding emotion labels, such as 😊 for happy, 😢 for sad, 😡 for angry, etc. You can use a predefined emoji-emotion dictionary or create your own based on your criteria.

Step 2: Capture the user's input text using a camera, microphone, or keyboard. You can use various techniques such as "optical character recognition (OCR)", "speech recognition, or natural language processing (NLP)" to convert the input text into a digital format.

Step 3: Extract the emojis from the input text using regular expressions or other methods. You can use the Unicode standard to identify and match the emojis in the text.

Step 4: Map each emoji to its corresponding emotion label using the emoji-emotion dictionary. You can use a simple lookup table or a more complex algorithm to handle ambiguous or multiple emojis.

Step 5: Display the emoji-emotion mapping results to the user using a "graphical user interface (GUI)" or other methods. You can use various techniques such as color coding, animation, or sound effects to enhance the user experience and feedback.

- Algorithm for "Term Frequency Inverse Document Frequency (TF-IDF)": methodology is a well-liked approach in NLP for performing data retrieval by evaluating the significance of words or terms within a document in the context of a larger group of documents that are often identified as a corpus. It is most frequently used in text analysis, document retrieval, and text mining tasks [19].
- Term Frequency (TF): TF calculates how often a term word or phrase that exists in a certain document to measure the occurrence of a term in a document and evaluates the frequency ratio for every specific term in the given text. It aids in locating the most frequently used words in a given document.[20]

$$\text{TF (Term, Document)} = \frac{\text{Frequenccy of a Term}}{\text{number of terms available}} \quad (1)$$

- Inverse Document Frequency (IDF): IDF calculates the frequency with which a term appears in a corpus of documents. Calculated is the logarithm of the total documents count in the corpus split by the percentage of documents which comprises of term. Higher IDF values are assigned to terms that are more uncommon or distinctive, whereas lower IDF scores are given to terms that are often used throughout several papers.

$$\text{IDF (Term)} = log(\frac{count\ of\ documents\ in\ Corpus}{count\ of\ documents\ with\ Term}) \quad (2)$$

E. MODEL CONSTRUCTION

After dividing the data into testing and training, to tackle the prediction problem, by employing a range of ML models with top six of which are all distinct regression models and the algorithms are trained on the data, and the accuracies of each algorithm are calculated. Finally, a model with good accuracy has been finalized.

- Decision Tree: Building an instructional system which is needed for predicting cost of target variables. This is done by mastering the choice rules created from training data.. The Decision Tree is most frequently used for grouping where each tree is generated through nodes and branches [13]. Each node denotes the classified information with maximum price which is made ready to accept through numerous datasets acquired for performing training of samples across a large record set have an impact on the test set accuracy which is 84.60% [12].
- Random Forest: Data mining and ML techniques employ the "ensemble learning" method known to be random forest for imparting predictions based on several decision trees. With the help of randomly selected data subsets, many decision trees are trained, and their predictions are then aggregated via weighted average or majority voting. Although it needs more compute, it is substantially more accurate and far less prone to over fitting than decision trees.[14]
- KNN: The "k-Nearest Neighbour's (KNN)" model is a simple but effective machine learning technique employed in classification and regression applications. It locates the k nearest data points which are added newly based on input point to predict the class or value of that point utilizing the mean class or value of the k-nearest neighbours. KNN's simplicity and ease of use are its key advantages, which make it a viable choice for rapid prototyping and initial data exploration [11].
- Naive Bayes: Naive Bayes is a statistical machine learning technique that is based on the Bayes theorem. It is highly beneficial for occupations that require categorization. Features are assumed to be conditionally independent, which is a "naive" but frequently accurate assumption. Naive Bayes calculates the likelihood that a data point belongs to a specific class by computing the conditional probability of each feature considering the class and then combining them collectively [16].
- SVM:A popular ML approach is used for imparting classification and regression over applications which uses SVM model which primarily identifies an ideal hyperplane that efficiently divides data into several classes while maximizing the margin between them. Particularly in settings with high-dimensional data or

complicated decision boundaries, this leads to robust and accurate categorization. SVMs are ideal for a variety of applications because they can handle both "linear and non-linear" connections between features using kernel functions [15].

- Logistic Regression: is a type of statistical model which is often known as a logit model that frequently implements the categorization of predictive analytics as the regression process determines probability of occurrence of an event. Since the outcome is the probability attained for the range of the dependent variable is between 0 and 1 [17]. However, it can only be employed in situations comprising binary classification or numerous classes when multiple models are trained and then integrated [18].

F. TESTING

Following model training, in order to evaluate the model performance which is used to perform testing of data and to compare the predicted outcomes to the actual results and measure accuracy by splitting the dataset into 80% testing and 20% training.

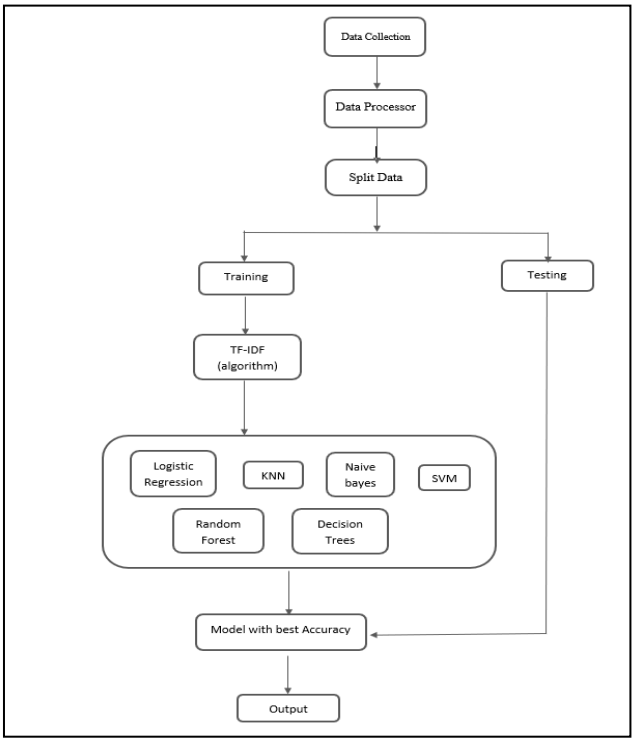

Fig. 4. Flowchart of proposed model

V. RESULTS

Table 1 displays the accuracy percentages for six algorithms using data from the "Kaggle". In this case, Random Forest performs well and 86.015% of the time correctly predicts the outcome. A decision tree has a prediction accuracy of at least 84.6%.

The following table and figure show the outcomes after the proposed approach has been evaluated.

TABLE I. ACCURACY OF "TWEET_EMOTIONS.CSV" DATASET IN DIFFERENT MODELS.

S.No	*Classifiers*	*Accuracy*
1	Decision Tree[12]	84.6075%
2	Random Forest[14]	86.015%
3	KNN[11]	42.12%
4	Naive Bayes[16]	39.21%
5	SVM[15]	68.89%

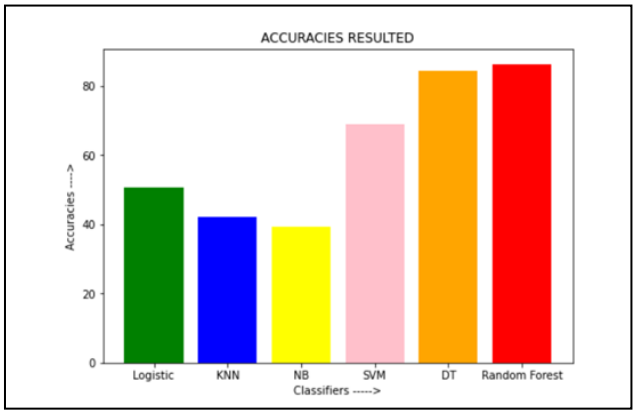

Fig. 5. Accuracy obtained for model construction.

Figure 5 presents a graphical representation of the accuracy percentages for the six used techniques. Random Forest performed admirably, with 86.607% accuracy. A Naive Bayes forecast with the lowest accuracy rate was 39.21%. it is generated after implementing the algorithms in jupiter notebook.

VI. CONCLUSION

After using six different models the accuracy of the classifiers are: 25.95% test accuracy and 84.6075% over all accuracy was attained using the DECISION TREE technique. 32.975% test accuracy and 86.015% over all accuracy was attained using the Random Forest technique. 23.28% test accuracy and 42.12% over all accuracy was attained using the KNN technique.28.112% test accuracy and 39.21% accuracy was attained using the NAIVE BAYES technique.36.6125% test accuracy and 68.89% over all accuracy was attained using the SVM technique.36.75% test accuracy and 50.565% over all accuracy was attained using the LOGISTIC REGRESSION technique. Random forest classifier has produced the maximum accuracy among the other algorithm so; the classifier used in this project will combine the predictions from several trees to increase the precision and decreases over fitting. In data science and AI, it is frequently used for classification and regression problems.

REFERENCES

[1] Rıza Velioğlu; Tuğba Yıldız; Savas Yıldırım "Sentiment Analysis using Learning Approaches over Emojis for Turkish Tweets" 2018 3rd International Conference on Computer Science and Engineering (UBMK) Sarajevo, Bosnia and Herzegovina, pp. 2-5, DOI: 10.1109/UBMK.2018.8566260.

[2] Mehenika Akter; Mohammad Shahadat Hossain; Karl Andersson "Hand-drawn emoji Recognition using convolutional Neural Network" 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE) ,Bhubaneswar, India, pp.1-5, DOI: 10.1109/WIECON-ECE52138.2020.9397933.

[3] İrem Bulut; Muhammed Erdoğan; Beytullah Gönülal; Rüveyda Baş; Özkan Kılıç "Using Short Texts and Emojis to predict the Gender of a Texter in Turkish" 2019 4th International Conference on Computer

Science and Engineering (UBMK) Samsun, Turkey,pp.2-3 ,DOI: 10.1109/UBMK.2019.8907198.

[4] Jiawen Deng; Fuji Ren "Multi-label Emotion Detection via Emotion Specified Feature Extraction and Emotion Correlation Learning". IEEE Transactions on Affective Computing (Volume: 14, Issue: 1, 01 Jan.-March 2023), Tokyo, Japan,pp. 475 – 486, DOI: 10.1109/TAFFC.2020.3034215.

[5] Dheeraj S Nair; N Balagopal "Emoji Prediction from Sentence" 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS) Coimbatore, India, pp.3-4,DOI: 10.1109/ICACCS51430.2021.9441897.

[6] Sadam Al-Azani; El-Sayed M. El-Alfy "Early and late fusion of emojis and text to enhance Opinion Mining" on IEEE Access (Volume: 9),london. UK,pp.4-5,DOI: 10.1109/ACCESS.2021.3108502.

[7] Astha Mohta; Atishay Jain; Aditi Saluja; Sonika Dahiya "Pre-Processing and Emoji Classification of WhatsApp Chats for Sentiment Analysis" 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) Palladam, India, pp. 2-3,DOI: 10.1109/I-SMAC49090.2020.9243443.

[8] Amalia Anjani Arifiyanti; Eka Dyar Wahyuni "Emoji and Emoticon in Tweet Sentiment Classification" 2020 6th Information Technology International Seminar (ITIS) Surabaya, Indonesia , DOI: 10.1109/ITIS50118.2020.9320988

[9] Savitha Hiremath; S H Manjula; Venugopal K R "Unsupervised Sentiment Classification of Twitter Data using Emoticons" 2021 International Conference on Emerging Smart Computing and Informatics (ESCI) Pune, India,pp.7-8, DOI: 10.1109/ESCI50559.2021.9397026

[10] Ziad Al-Halah; Andrew Aitken; Wenzhe Shi; Jose Caballero "Emoji Embedding for Visual Sentiment Analysis" 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW) Seoul, Korea (South), pp.3-5,DOI: 10.1109/ICCVW.2019.00550

[11] b.gnanapriya "emoji based sentiment analysis using knnInternational Journal of Scientific Research and Review Volume 07, Issue 04, April 2019, Annāmalainagar, India.

[12] M. A. Shaik, M. Y. Sree, S. S. Vyshnavi, T. Ganesh, D. Sushmitha and N. Shreya, "Fake News Detection using NLP", 2023 International Conference on Inovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, 2023, pp. 399-405, doi: 10.1109/ICIDCA56705.2023.10100305.

[13] Mohammed Ali Shaik, M. Varshith, S. SriVyshnavi, N. Sanjana and R. Sujith, "Laptop Price Prediction using Machine Learning Algorithms", 2022 International Conference on Emerging Trends in Engineering and Medical Sciences (ICETEMS), Nagpur, India, 2022, pp. 226-231, doi: 10.1109/ICETEMS56252.2022.10093357.

[14] Mohammed Ali Shaik, Dhanraj Verma, P Praveen, K Ranganath and Bonthala Prabhanjan Yadav, (2020), RNN based prediction of spatiotemporal data mining, 2020 IOP Conf. Ser.: Mater. Sci. Eng. 981 022027, doi.org/10.1088/1757-899X/981/2/022027

[15] Mohammed Ali Shaik, "A Survey on Text Classification methods through Machine Learning Methods", International Journal of Control and Automation (IJCA), ISSN:2005-4297,Volume12,Issue-6 (2019), Pp.390-396.