A Virtual Journaling Understanding and Tracking Daily Emotions

S. Senthilnathan
Assistant professor,
Velalar College of Engineering and Technology
Nitheshprawin.N, Sanjai.R, Shanmugapriya.V
Velalar College of Engineering and Technology

Abstract— Journaling is a popular tool for introspection and personal development. It can assist people in gaining insight into their emotions, behaviors, and thoughts. Virtual journaling applications have grown in popularity as a convenient way to track and analyze daily experiences since the advent of technology. The development of a virtual journaling application for understanding and tracking daily emotions is proposed in this paper. The proposed application will analyze the text entered by the user and extract emotional information using natural language processing techniques. The application will display emotional data visualizations, allowing users to identify patterns in their emotional experiences. Users will be able to set goals and track their progress in the direction of emotional well-being. The application will also make personalized recommendations for activities or exercises that will benefit users. Manage their emotions. The proposed application has the potential to enhance emotional awareness and promote emotional wellbeing.

I. INTRODUCTION

The application will leverage natural language processing techniques to analyze text entered by the user and extract emotional information. Users will be able to visualize their emotional data and identify patterns in their experiences, set emotional wellbeing goals, and track progress towards them. The application will also offer personalized suggestions for managing emotions. This virtual journaling application has the potential to enhance emotional awareness and promote emotional wellbeing, providing a convenient tool for individuals to reflect on their daily experiences and gain insights into their emotional states.

Named Entity Recognition (NER) plays a key role in information extraction, allowing for the identification of “entities” (e.g., Person, Location). NER is widely used in machine translation, question answering information retrieval, and automatic summarization [2]. Originally, Person, Organization, Location were the first three entities considered for semantically classifying words. In the following years, new entities were defined to meet domain-specific needs (e.g., in the medical or legal sectors) [3], [4]. Various criteria must be taken into consideration before selecting and using a NER software such as its performance, cost, documentation, license, and etcOur methodology will be presented in order to give the possibility to reproduce the experiments. Using our reproductible methodology, our results show that StanfordNLP perfoms between 15% and 30% better on selected corpora than the other software tested. However, we were not able to retrieve the same results as the ones we can find in the literature for every tested software. The difference observed can be up to 66%. Evidence that existing state-of-the-art studies lack of information for experiment reproducibility purposes and differences in results is discuss ensures a fair and meaningful evaluation and comparison of NER software .The interpretation of emotions and further classification from the given text has attracted a variety of studies based onSA. The digital platforms can provide a rich source of texts and other related content that can be further utilized by analyze techniques to derive the potential sentiment of the writers. While the manual approach may not be feasible for a massiveamount of data, the automated process requires a series of pre-processing steps to computationally operate such data; this is acrucial step because of the presence of potentially uninforma-tive or erroneous fragment(s) within the given text. One of therecent studies presented a combination of various preprocess-ing methods such as replacing contractions, replacing negations with antonyms, removing stopwords, punctuation, numbers, emoticons, etc., as well as tokenization, word stem-ming to handle negations, to name a few; several experi-ments were also conducted with the individual, as well as combinational, preprocessing step(s). The selection of, as well as the order of performing, such preprocessing steps require significant attention.Named Entity Recognition (NER) plays a key role in information extraction, allowing for the identification of

"entities" (e.g., Person, Location). NER is widely used in machine translation, question answering information retrieval, and automatic summarization Sentiment analysis is used to extract and quantify the attitudes,opinions, and emotions expressed in a piece of text, such as product reviews, social media posts, customer feedback, and news articles. Emotions like joy, anger, sadness, or fear.

1 Motivation

The application will leverage natural language processing techniques to analyze text entered by the user and extract emotional information. Users will be able to visualize their emotional data and identify patterns in their experiences, set emotional wellbeing goals, and track progress towards them. The application will also offer personalized suggestions for managing emotions. This virtual journaling application has the potential to enhance emotional awareness and promote emotional wellbeing, providing a convenient tool for individuals to reflect on their daily experiences and gain insights into their emotional states.One important point about the reported evaluation studies is that they often lead to difference in results, sometimes in a substantial way, when evaluating a same software. Let us mention, for example, the study of in which 7 NER software are compared (cf., TABLE I). The authors conclude that NLTK and OpenNLP have similar results, and most importantly better than StanfordNLP. However, one may wonder why the results of StanfordNLP are that low considering that the official StanfordNLP website2 announces much better results (note: both studies having used the same corpus for evaluation purposes). Such divergence of results is emphasized in Fig. 2, where the F1-score obtained by 4 out of the 6 evaluation studies presented in TABLE II-B have been reported, which all evaluate a same software (StanfordNLP in this case) based on a common set of corpora (CoNLL 2003, Ritter, MSM2013). The result is unmistakable, especially considering the Ritter corpus, as the F1-score of [16] is more than double the one of. It should also be noted that existing evaluation studies often fail to provide all the necessary information to allow the reproducibility of experiments, resulting in the impossibility to obtain similar results for comparison purposes. For example, looking at the 6 evaluation studies reported in TABLE I, only 24% of the evaluated NER software (i.e., 6 out of 25) provide information about the type of classifier used, while only 16% detail the version of the software used for the experiment. This lack of information, and the impossibility to replicate the evaluation, thus make it very difficult to replicate the experiments, and most importantly to be able to judge the quality and completeness of the results. To overcome this problem, we proceeded in a two-stage fashion. First, First, we contacted the authors of the reported studies in order to request for the missing information; so far, only Pinto et al. provided new information, which is highlighted in bold in TABLE I. Second, we propose and present a clear and replicable comparison study in the next section.

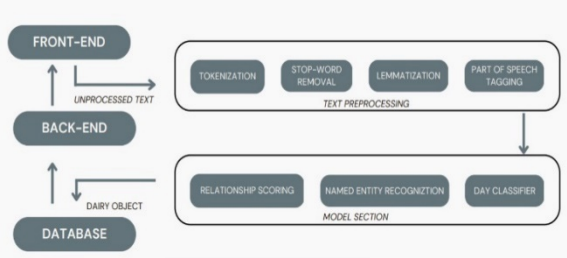
II. RELATED WORK

The identification of the inherent emotion from the given textfragment has been a subject of interest for many researchers. Among the wide variety of its applicability, SA has been explored using different datasets, series of preprocessing steps, and various computational methods; also, some specificfeatures have also been targeted to improve the sentiment classification.The use of social media has spread around the world; the diversity of users and their inputs to such social platforms generates enormous data. Therefore, the usage of automatedtext classification is highly demanded to identify sentiment present within the text. It can also be noticed that such data are likely to have diversified ways of expressions and hence, the preprocessing steps play a vital role in identifying the sen- timents; the study has also shown the impact of preprocessing steps on the accuracy of ML algorithms.For instance, thestudy compared most widely used preprocessing techniques using four ML classifiers on two datasets, namely, SS- Twitter and SemEval. The study conducted by combining sev- eral techniques showed that the results varied depending on the classifier and the combination of techniques used. The combination of replacing Uniform Resource Locators (URL)s and user mentions, replacing contractions, removing num- bers, replacing repetitions of punctuation, and lemmatization was considered as a preferable combination The study [18] presented a comparison between prepro-cessing techniques on data collected from Twitter for SA. Results showed that basic cleaning techniques involving stop-words removal, removing punctuation, URLs, hashtags, etc. along with stemming, improved the performance. However, using a dictionary to detect and correct misspelled words did not improve

the performance but reduced the elaboration-time for cleaning the raw-text. The other methods such as replacing negations, replacing emoticons, removing stop- words, also improved the performance. Some of these meth- ods reduced noise while others increased the relevance of concepts. Subsequently, various preprocessing steps were applied to convert the users' comments, i.e., sentences into numerical vectors; the preprocessing included various operations such as normalization, lowercasing, removal of accent, extra blank spaces, hyphens, punctuation marks, as well as watermarks, followed by stopwords removal and tokenization. Comparisons indicated that SVM attained higher accuracy with Word2vec as compared to SVM with term frequency–inverse document frequency (TF–IDF) in study. Also, a semi-supervised sentiment hashtag embed-ding (SHE) model was proposed to preserve semantic as well as sentiment distribution of the hashtags.

III.PROPOSED APPROACH

In this article, the proposed approach considers a specific order of preprocessing steps and utilizes ANN model for sentiment classification; a graphical representation of the overall approach is as shown in Fig. 1. A detailed explanation of the steps carried out is as follows; while the procedure is initiated with a collection of multi-domain datasets, here, we restrict our experiments up to three binary datasets.



Fig(proposed approach)

User Interface: Design a user-friendly interface for the application that allows users to easily input and track their daily emotions.

Emotion Categories: Create a list of emotion categories that users can choose from when inputting their emotions. This list could include basic emotions such as happy, sad, angry, etc., as well as more nuanced emotions such as anxious, frustrated, content, etc.Daily Entries: Allow users to input a daily entry of their emotions. This could be done through a text entry box or by selecting the appropriate emotion category from a list.

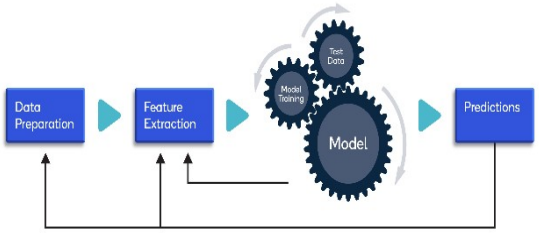
Over a period of time.

Insights and Analysis: Provide insights and analysis of the user's emotional patterns over time. This could involve using machine learning algorithms to identify trends and patterns in the user's emotional data.Additional Features: Consider adding additional features such as a daily gratitude prompt, inspirational quotes, or self-care tips to help users manage their emotions more effectively.Privacy and Security: Ensure that the application is designed with privacy and security in mind. This could involve implementing strong encryption for user data, restricting access to user data, and implementing other security measures as needed. Overall, the goal of this approach is to provide users with a tool for understanding and managing their emotions more effectively. By tracking their emotions over time, users can gain valuable insights into their emotional patterns and make changes to improve their overall well-being.

1.1 Preprocessing

Data preprocessing is a data mining technique that incorpo-rates a range of operations to clean and transform raw data into an understandable format. The online text contains a lot of noise and uninformative parts such as HTML tags, adverb-tenements, stop words that do not contribute to

the sentiment of the text and add noise, and hence, should be removed. Text preprocessing improves the performance of the clas- sifier and accelerates the classification process. In this study, we have performed preprocessing in an iden- tified order (PPR) which is different from the order given in the study, i.e., EPR. For simplicity of understanding, the considered set of operations along with their order are dem- onstrated in Figs. 2 and 3 for EPR and PPR, respectively. Based on the individual preprocessing methods, it was identified that high accuracy in all classifiers was achieved by replacing contractions, removing numbers, replacing repetitions of punctuation, lemmatizing, stemming, and handling negations for the SS-Twitter dataset and by replacing contractions, removing numbers, replacing repetitions of punctuation, and removing stopwords for SemEval dataset . Here, we select the techniques that are stated to have significant impact in and prepare a preprocessing pipeline of EPR. Thus, we include replacement of URLs and user mentions, contrac- tion replacement, numbers removal, repetitions of punctua- tion replacement, and lemmatization in EPR. On the other hand, the proposed preprocessing, PPR, begins with replace- ment of URLs, user mentions, and contractions, followed by removal of numbers and replacement of repeating punctua- tion, conversion to lowercase, tokenization, stopwords removal, stemming, and replacement of negations with alto- nyms as illustrated in Fig. 3; here, lemmatization is claimed to be mutually exclusive to stemming and hence, should not be used together. The importance of each preprocessing step,as well as its order, is discussed in the following sections.this removal, repetitions of punctua- tion replacement, and lemmatization in EPR. On the other hand, the proposed preprocessing, PPR, begins with replace- ment of URLs, user mentions, and contractions, followed by removal of numbers and replacement of repeating punctua- tion, conversion to lowercase Tokenization, stop words removal.



1.1.1 Sentimental Analysis

Sentiment analysis will add some more categories to obtain more efficient results. It has greater focus on polarity. This type follows a 5-star rating like system and classifies opinions as: •Very positive •Positive •Neutral •Negative •Very negative For instance, see the classification of a survey response given below: (iii) Emotion detection Emotion detection is another method of sentiment analysis, used to detect emotions like excited, happiness, irritation, anger, sadness, and the like. Lexicons -lists of words and the emotions they convey are most commonly used in emotion detection. Advanced systems apply complex machine learning algorithms in detecting the emotions. Consider the examples below: (IV) Aspect-based Sentiment Analysis This method of sentiment analysis gives better insights on opinions of the writer and gives preference to different features of the product or service mentioned in the given opinion. Consider a product review, which contains the reviewer's opinions about different properties or views of a product such as the price, efficiency, integrations to other devices or services, mobile version, etc. The following examples show the result of aspect-based analysis. (v) Multilingual sentiment analysis multilingual sentiment analysis is a complex process. It demands preprocessing of the input.

1.1.2 Tokenization

Tokenization is the process of fragmenting the text into tokens such as words, numbers, punctuation marks, or any other special symbol by locating the ending point of a word and beginning of the next word called word boundaries. This cannot be used for languages wherein words do not have clear boundaries such as Chinese and Thai. Many of the tokens generated,

such as articles.Tokenization and “classic” encryption effectively protect data if implemented properly, and a computer security system may use both. While similar in certain regards, tokenization and classic encryption differ in a few key aspects. Both are cryptographic data security methods and they essentially have the same function, however they do so with differing processes and have different effects on the data they are protecting .Tokenization is a non-mathematical approach that replaces sensitive data with non-sensitive substitutes without altering the type or length of data. This is an important distinction from encryption because changes in data length and type can render information unreadable in intermediate systems such as databases. Tokenized data can still be processed by legacy systems which makes tokenization more flexible than classic encryption. In many situations, the encryption process is a constant consumer of processing power, hence such a system needs significant expenditures in specialized hardware and software. Another difference is that tokens require significantly less computational resources to process. With tokenization, specific data is kept fully or partially visible for processing and analytics while sensitive information is kept hidden. This allows tokenized data to be processed more quickly and reduces the strain on system resources. This can be a key advantage in systems that rely on high performance. In comparison to encryption, tokenization technologies reduce time, expense, and administrative effort while enabling teamwork and communication this can be a key advantage systems that rely on high performance. There are different ways to tokenize text, but the most common approach is to split the text into words based on whitespace and punctuation. This approach is known as word-level tokenization. However, there are other approaches such as character-level tokenization, sub-word tokenization, and byte-pair encoding.

3.1.3 Stopwords Removal

Stopwords removal is the process of removing the words, such as articles and pronouns, occurring with high fre- quency across the text, and are irrelevant to the task of SA [5]. It is not necessary to analyze them as they not contain any useful information for SA [5]; however, the stopwords removal must be carried out carefully to ensure

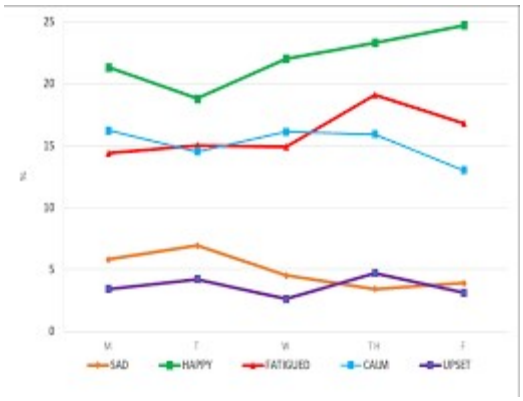
3.1.4 Lemmatization

Lemmatization is the process of reducing a word to its base or root form, known as the lemma. In the context of sentiment analysis, lemmatization can be used to normalize words and improve the accuracy of the analysis.For example, consider the words "happy" and "happier." By applying lemmatization, both words can be reduced to the base form "happy," which can help the sentiment analysis model to recognize that they convey similar emotions. Stemming transforms derived words to their base form. For example, the word “play” can have variations like “playing”, “played”, “plays”, etc. These variations are clubbed to their base form “play” after applying stemming. Thus, stemming helps to club many different variations of a token in a single entity [69]. For our experimentations, we have used Porter Stemmer [70] for word stemming.Lemmatization can be performed using various techniques such as rule-based approaches or machine learning models. Some popular libraries that can be used for lemmatization in Python include NLTK, spaCy, and TextBlob. Overall, incorporating lemmatization into sentimental analysis can help to improve the accuracy and consistency of the results by reducing variations in word forms and normalizing the language used in the analysis. The morphological analysis would need the extraction of the correct lemma of every word. To simplify it, let's just say that lemmatization is a linguistic term refers to the act of grouping together words that have the same root or lemma but have different inflections or derivatives of meaning so they can be analyzed as one item. The process of lemmatization seeks to get rid of inflectional suffixes and prefixes for the purpose of bringing out the word’s dictionary form.

IVRESULT ANALYSIS

During this step, we check for the word “not” in each sentence and check if the word next to it has antonyms or not. During PPR, we have performed this step after stopwords removal. This is because articles such as “a”, “an”, “the” do not have an antonymA virtual journaling application for understanding and tracking daily emotions typically involves the collection and analysis of data from users' self-reported emotions over time. The collected can be analyzed using various statistical and machine learning techniques to gain insights into users' emotional patterns and to develop personalized recommendations for improving emotional well-being. One common approach to

analyzing the data is to perform sentiment analysis, which involves extracting emotional indicators such as positive, negative, or neutral sentiments from the text data entered by users. The sentiment analysis can be combined with data on other emotional factors such as intensity, frequency, and duration of emotions, to provide a more comprehensive understanding of users' emotional states. Once the data is collected and analyzed, virtual journaling applications can provide various insights and recommendations to users based on their emotional patterns. For example, the application can provide



V.CONCLUDING REMARKS AND FUTURE DIRECTIONS

This paper proposes the use of preprocessing techniques such as contraction replacement, replacement of punctua- tion and numbers with space, conversion to lowercase to maintain uniformity in the text, tokenization, stop words removal, and word stemming, and replacing negations with antonyms in an identified order to improve the performance of the classifier. Feature extraction (FE) is carried out using the nlp model followed by feature selection (FS) using IG to select relevant and important features as it gives better performance than the other commonly used FS methods. In this study the proposed virtual journaling application for understanding and tracking daily emotions has the potential to provide a convenient tool for individuals to reflect on their daily experiences, gain insight into their emotional states, and promote emotional wellbeing. By leveraging natural language processing techniques to analyze the text entered by the user and extract emotional

Information, the for managing emotions and help users identify patterns in their emotional experiences. With the ability to set goals and track progress towards emotional wellbeing, users can actively work towards improving their emotional health. Overall, this application has the potential to enhance emotional awareness and promote emotional wellbeing, providing a valuable tool for individuals to improve their overall quality of life.

REFERENCES

[1] C. Dhaoui, C. M. Webster, and L. P. Tan, "Social media sentiment analysis: Lexicon versus machine learning," J. Consum. Marketing, vol. 34, no. 6, pp. 480–488, 2017.
[2] D. Mungra, A. Agrawal, and A. Thakkar, "A voting-based senti- ment classification model," in Intelligent Communication, Control and Devices, Berlin, Germany: Springer, 2020, pp. 551–558.
[3] Y. Liu, J.-W. Bi, and Z.-P. Fan, "Multi-class sentiment classifi- cation: The experimental comparisons of feature selection and machine learning algorithms," Expert Syst. Appl., vol. 80, pp. 323–339, 2017.
[4] P. Melville, W. Gryc, and R. D. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification," in Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2009, pp. 1275–1284.
[5] S. Symeonidis, D. Effrosynidis, and A. Arampatzis, "A comparative evaluation of pre-processing techniques and their interactions for twit- ter sentiment analysis," Expert Syst. Appl., vol. 110, pp. 298–310, 2018.
[6] V. Narayanan, I. Arora, and A. Bhatia, "Fast and accurate senti- ment classification using an enhanced naive bayes model," in Intelligent Data Engineering and Automated Learning, Berlin, Ger- many: Springer, 2013, pp. 194–201.
[7] A. Kumar and G. Garg, "The Multifaceted Concept of Context in Sentiment Analysis," in Cognitive Informatics and Soft Computing, Berlin, Germany: Springer, 2020, pp. 413–421.
[8] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment clas- sification using machine learning techniques,"

2002, arrive: 0205070.
[9] L.-S. Chen, C.-H. Liu, and H.-J. Chiu, "A neural network based approach for sentiment classification in the blogosphere," J. Infor-metrics, vol. 5, no. 2, pp. 313–322, 2011.
[10] G. Forman, "An extensive empirical study of feature selection metrics for text classification," J. Mach. Learn. Res., vol. 3, no. Mar, pp. 1289–1305, 2003.
[11] A. Thakkar and R. Lohiya, "Role of swarm and evolutionary algo- rithms for intrusion detection system: A survey," Swarm Evol. Comput. vol. 53, 2020, Art. No. 100631.
[12] A. Thakkar and R. Lohiya, "Attack classification using feature selection techniques: A comparative study," J. Ambient Intell. Humanized Comput. vol. 12, no. 1, pp. 1249–1266, 2021.
[13] A. Thakkar and R. Lohiya, "A review on machine learning and deep learning perspectives of IDS for IoT: Recent updates, secu- rity issues, and challenges," Arch. Comput. Methods Eng., vol. 28,no. 4, pp. 3211–3243, 2021.
[14] K. Chaudhari and A. Thakkar, "Survey on handwriting-based personality trait identification," Expert Syst. Appl., vol. 124, pp. 282–308, 2019.
[15] R. Sharma, H. Rajvaidya, P. Pareek, and A. Thakkar, "A compara- tive study of machine learning techniques for emotion recognition," in Emerging Research in Computing, Information, Communication and Applications, Berlin, Germany: Springer, 2019, pp. 459–464.
[16] K. Chaudhari and A. Thakkar, "A comprehensive survey on travel recommender systems," Arch. Comput. Methods Eng., vol. 27, pp. 1–27, 2019.
[17] S. Renjith, A. Sreekumar, and M. Jathavedan, "An extensive study on the evolution of context-aware personalized travel recommender sys-tems," Informat. Process. Manage. vol. 57, no. 1, 2020, Art. No. 102078.
[18] A. Thakkar, N. Jivani, J. Padasumbiya, and C. I. Patel, "A new hybrid method for face recognition," in Proc. Nirma Univ. Int. Conf. Eng., 2013, pp. 1–9.
[19] C.Nagarajan and M.Madheswaran - 'Experimental verification and stability state space analysis of CLL-T Series Parallel Resonant Converter' - Journal of ELECTRICAL ENGINEERING, Vol.63 (6), pp.365-372, Dec.2012.
[20] C.Nagarajan and M.Madheswaran - 'Performance Analysis of LCL-T Resonant Converter with Fuzzy/PID Using State Space Analysis'- Springer, Electrical Engineering, Vol.93 (3), pp.167-178, September 2011.
[21] C.Nagarajan and M.Madheswaran - 'Stability Analysis of Series Parallel Resonant Converter with Fuzzy Logic Controller Using State Space Techniques'- Taylor & Francis, Electric Power Components and Systems, Vol.39 (8), pp.780-793, May 2011.
[22] Nagarajan and M.Madheswaran - 'Experimental Study and steady state stability analysis of CLL-T Series Parallel Resonant Converter with Fuzzy controller using State Space Analysis'- Iranian Journal of Electrical & Electronic Engineering, Vol.8 (3), pp.259-267, September 2012.
[23] K. Chaudhari and A. Thakkar, "A comprehensive survey on travel recommender systems," Arch. Comput. Methods Eng., vol. 27, pp. 1–27, 2019.
[24] S. Renjith, A. Sreekumar, and M. Jathavedan, "An extensive study on the evolution of context-aware personalized travel recommender sys-tems," Informat. Process. Manage. vol. 57, no. 1, 2020, Art. No. 102078.
[25] A. Thakkar, N. Jivani, J. Padasumbiya, and C. I. Patel, "A new hybrid method for face recognition," in Proc. Nirma Univ. Int. Conf. Eng., 2013, pp. 1–9.
[26] P. Pareek and A. Thakkar, "A survey on video-based human action recognition: Recent updates, datasets, challenges, andapplications," Artif. Intell. Rev., vol. 54, no. 3, pp. 2259–2322, 2021.
[27] G.Neelakrishnan, K.Anandhakumar, A.Prathap, S.Prakash "Performance Estimation of cascaded h-bridge MLI for HEV using SVPWM" Suraj Punj Journal for Multidisciplinary Research, 2021, Volume 11, Issue 4, pp:750-756
[28] R. Patel, C. I. Patel, and A. Thakkar, "Aggregate features approach for texture analysis," in Proc. Nirma Univ. Int. Conf. Eng., 2012, pp. 1–5.
[29] A. Thakkar, D. Mungra, and A. Agrawal, "Sentiment analysis: An empirical comparison between various training algorithms for artificial neural network," Int. J. Innov. Comput. Appl., vol. 11,no. 1, pp. 9–29, 2020