International Journal of Natural Language Processing (IJNLP)
Volume 2, Issue 1, January-June 2024, pp. 12-20, Article ID: IJNLP_02_01_002
Available online at https://iaeme.com/Home/issue/IJNLP?Volume=2&Issue=1
Journal ID: 1667-1255

SENTIMENT ANALYSIS ON EMPLOYEE LAYOFFS BASED ON HYBRID FEATURE EXTRACTION AND LONG SHORT TERM MEMORY NETWORK

Ranjit Kumar S
Software Engineer, Rangsons Aerospace Private Limited, Bangalore, India

ABSTRACT

In recent decades, sentiment analysis has become crucial for understanding the opinions and emotions expressed in different forms of communication, namely speech, text, etc. Particularly, in the scenario of employee layoffs, sentiment analysis assists organizations in understanding the overall emotional impact of the layoffs on the workforce. In the initial phase of this research, twitter data is acquired using rapid-miner software. Subsequently, data pre-processing is performed by employing regular expressions. The removal of punctuation, stopwords, and conversion of text data to lowercase reduces noise in the collected dataset that enables the classification model to focus more on relevant information. After pre-processing the twitter data, feature vector extraction is accomplished by applying Term Frequency-Inverse Document Frequency (TF-IDF) and spacy.word2vec techniques. In this context, hybrid feature extraction improves the classification model's ability in capturing rich textual information that results in more accurate sentiment classification. In the final phase, a Long Short Term Memory (LSTM) network is applied to classify the categories of user sentiments, namely positive, neutral, and negative. Compared to other traditional classification models, the LSTM network is computationally effective, and achieved a higher accuracy of 96.30%, recall of 97.12%, and precision of 96.88% on the collected dataset.

Keywords: Long Short-Term Memory Network, Rapid-Miner, Sentiment Analysis, Term Frequency-Inverse Document Frequency, Word2vec.

Cite this Article: Ranjit Kumar S, Sentiment Analysis on Employee Layoffs Based on Hybrid Feature Extraction and Long Short Term Memory Network, International Journal of Natural Language Processing (IJNLP), 2(1), 2024, pp. 12-20.
https://iaeme.com/Home/issue/IJNLP?Volume=2&Issue=1

1. INTRODUCTION

After a layoff, analyzing the sentiments/emotions of employees is vital for understanding the impact on mental health that can affect organizational culture and productivity [1]. Furthermore, it helps to provide counselling services for mitigating negative effects like resentment, stress, etc. Sentiment analysis gives an idea about future layoff strategies and human resource practices that makes employees more compassionate and effective [2-3]. In the present scenario, twitter is an emerging platform to share social messages compared to other social media platforms like WhatsApp, Telegram, Instagram, LinkedIn, Facebook, etc. [4-7]. Presently, several machine and deep learning-based models have been developed for analysing the sentiments of users [8]. However, most of the traditional machine and deep learning-based models focus only on textual feature information to construct vector representations of twitter data [9-10]. By analyzing existing literatures, it has been found that traditional models fail to extract relevant, informative feature information necessary for achieving superior classification performance [11-12]. For highlighting this particular concern, a hybrid feature extraction is performed in this research article, and combined with a novel deep learning model for sentiment analysis.

The contributions of this research article are defined below;

- Performed data pre-processing by eliminating punctuations, stop words, non-English characters, tabs, spaces, etc. in the collected tweets. In the context of sentiment analysis, pre-processing enhances the overall accuracy of the classification model by organizing and cleaning data effectively.
- Next, hybrid feature extraction is carried out by combining TF-IDF and spacy. word2vec techniques. These two techniques extracts highly rich and informative feature information from the text to achieve precise sentiment classification.
- Used an LSTM network for classifying the categories of user sentiments such as positive, neutral, and negative. In comparison to other classification models, an LSTM network is less prone to the overfitting problem while handling large datasets. Furthermore, an LSTM network is highly flexible by means of input and output configurations.

This article is organized as follows. A literature survey of existing articles is presented in section 2. The methodologies undertaken are detailed in section 3. The empirical analysis and the conclusion of this research article are presented in sections 4 and 5.

2. RELATED WORKS

A few recent literatures on the topic of 'sentiment analysis' are surveyed in this section. Baqach and Battou [13] presented a novel deep learning based framework for sentiment analysis, which incorporates four models such as support vector machine, attention mechanism, LSTM, and Convolutional Neural Network (CNN). In this literature study, the presented deep learning based framework's performance was analysed on three labelled datasets, and validated using three different evaluation measures. The empirical outcomes revealed that this framework achieved superior performance in comparison to other baseline models, particularly in the context of sentiment analysis. Vidyashree and Rajendra [14] introduced an improved sentiment analysis model for identifying the polarity of tweets, namely negative, neutral, and positive. In this study, an improved deep learning model named the Stochastic Gradient Neural Network (SGNN) was introduced for categorizing the user's sentiments based on their tweets. The SGNN model obtained better classification performance related to other traditional models.

Bania [15] implemented an automated sentiment classification model named CovDLCNet, which was based on the concept of LSTM network. Firstly, a one graph theoretic approach was employed to eliminate similar and duplicate tweets. Then, the polarity score of every tweet was computed using the CovDLCNet model, which categorized the tweets into three types, namely positive, neutral, and negative. Extensive simulation analysis revealed that the CovDLCNet model obtained a higher classification accuracy score compared to other baseline models. Parveen et al. [16] initially pre-processed the collected tweets, and then, feature extraction was performed using a log term frequency-based modified inverse class frequency technique. Subsequently, feature selection and classification were accomplished by implementing a hybrid mutation-based white shark optimization algorithm and a Gated Attention Recurrent Network (GARN). The evaluation measures (f1-measure, recall, precision, and accuracy) demonstrated the efficacy of the GARN model.

Correspondingly, Aslan et al. [17] initially extracted feature vectors from the collected tweets utilizing fast-text Skip-gram and CNN models. Subsequently, optimal feature information was selected by employing an arithmetic optimization algorithm, which was finally passed into the K-nearest neighbor model for classifying tweets as neutral, negative, and positive. In addition, Vatambeti et al. [18] integrated the elephant herd optimization algorithm with a bi-directional LSTM network for twitter sentiment analysis regarding online food services. As discussed in the literature section, deep learning models are very effective for the application of sentiment analysis. To further enhance classification performance, we have performed hybrid feature extraction and combined it with a deep learning model. Hybrid feature extraction is important in sentiment analysis for capturing informative feature information from the text that improves the model’s accuracy and robustness.

3. METHODOLOGY

In the context of sentiment analysis, the proposed framework comprises four steps, such as twitter data collection, data pre-processing, feature vector extraction, and classification. The sentiment analysis related to employee layoffs provides numerous benefits, namely refining communication strategies, understanding public perception, and managing corporate reputation. Initially, real time twitter data are collected using rapid-miner software, and further, the collected data are pre-processed by performing regular expressions. From the pre-processed data, feature vectors are extracted by employing a hybrid feature extraction (TF-IDF and Word2Vec). Lastly, an LSTM network is used for classifying user sentiments into positive, negative, and neutral categories. A brief explanation about the proposed framework is presented as follows, and it is graphically illustrated in figure 1.

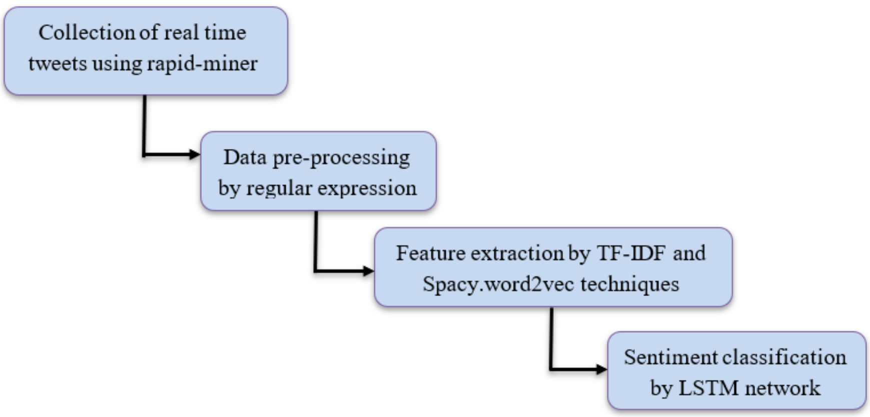

Figure 1. Processes involved in the proposed framework

3.1. Data collection and pre-processing

As mentioned in earlier sections, rapid-miner software is utilized to acquire twitter data. Approximately, 10,500 tweets are collected based on the query 'employee layoffs' and an 'extract sentiment' operator is used to generate labels for the collected tweets. The data is exported from rapid-miner in CSV file format. In the data pre-processing phase, the following processes are performed after collecting tweets related to the topic of ''employee layoffs'

- Elimination of non-English characters.
- Elimination of unnecessary lines, tabs, and spaces in the collected tweets.
- Elimination of '@' symbols and hashtag symbols (#pressure, #security, etc.).
- Elimination of numbers, punctuations, and special characters from the collected tweets, because these elements do not contribute to improve the classification performance.

3.2. Hybrid feature extraction

From the pre-processed twitter data, hybrid feature vectors are extracted by integrating TF-IDF and Spacy.word2vec techniques [19-20]. Firstly, TF-IDF is a common vectorization technique, which effectively extracts discriminative feature vectors from pre-processed tweets. It calculates how frequently a term appears in a tweet, and this technique is mathematically expressed in equations (1) and (2). The TF-IDF technique superiorly highlights salient features and focuses more on meaningful words that improves the accuracy of a classification model, and maintains effectiveness across diverse and large datasets.

$$TF = \frac{Number\ of\ times\ a\ term\ appers\ in\ a\ tweet}{Total\ terms\ in\ a\ tweet} \tag{1}$$

$$IDF = log\frac{Total\ \ tweets}{Number\ of\ tweets\ with\ \ term} \tag{2}$$

Secondly, the Spacy.word2vec technique is employed for learning the relationships between words. The word2vec technique extracts semantically rich and robust feature vectors for sentiment analysis. It has the ability in understanding and encoding contextual meanings from words that improves the predictive power of a classification model leading to accurate interpretations of text data. The extracted feature vectors are finally passed as input into a classification model (an LSTM network) to classify the categories of sentiments, namely positive, negative, and neutral.

3.3. Sentiment classification

The extracted feature vectors are processed by the LSTM network, which captures the long-range information and temporal dependencies required for sentiment analysis [21]. The LSTM network comprises three gates, namely input gate, forget gate, and output gate. Each LSTM unit receives three inputs such as, the unit state c_{t-1}, feature vectors x_t from the TF-IDF and Spacy.word2vec techniques, and the output from the prior time step h_{t-1}. Here, c denotes the cell state and t represents the time step. On the other hand, the LSTM network produces two outputs at the current time step such as, the cell state c_t and the current output h_t. In the LSTM network, the gate concept $g(x)$ is introduced for connecting hidden layers [22]. The term $g(x)$ is mathematically illustrated in equation (3), where; $W, \sigma, and\ b$ denote the weight function, sigmoid activation function and bias, respectively.

$$g(x) = \sigma(Wx + b) \tag{3}$$

In the LSTM network, the input gate i_t avoids unnecessary changes to the memory by controlling the input that enters the cell state c_t. Furthermore, it determines how much information from the current input is added to the cell state c_t. The mathematical expression of the input gate i_t is presented in equation (4). Here, $[h_{t-1}, x_t]$ denotes the long-term connection, b_i indicates the bias of the input gate i_t, and W_i represents the weight matrix of the input gate i_t.

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i) \tag{4}$$

Additionally, the forget gate f_t selectively retains or eliminates information from the cell state c_t. By eliminating unneeded data or retaining informative data, the forget gate f_t reduces the overfitting risk in the LSTM network. The mathematical expression for the forget gate f_t is denoted in equation (5). Here, W_f *and* b_f represent the weight matrix and bias of the forget gate f_t. On the other hand, the output gate manages both the current cell state c_t and the prior cell state c_{t-1}, and also determines the current output value h_t. The mathematical expression for the output gate o_t is represented in equation (6). Here, W_o *and* b_o state the weight matrix and bias of the output gate o_t, and $tanh$ denotes the tangent activation function [23]. The tangent activation function and sigmoid activation function are mathematically expressed in equations (7-10).

$$f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f) \tag{5}$$

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o), \text{ where } h_t = o_t \times tanh\,(c_t) \tag{6}$$

$$tanh(z) = y = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{7}$$

$$tanh(z) = 1 - y^2 \tag{8}$$

$$\sigma(z) = y = \frac{1}{1 + e^{-z}} \tag{9}$$

$$\sigma'(z) = y(1 - y) \tag{10}$$

Furthermore, the status of the input unit c_t' is defined on the basis of the last output and input gates, and this statement is mathematically stated in equation (11). In the current input unit c_t', the cell state c_t is multiplied by the input gate i_t, and in the element state c_{t-1}, the cell state c_t is multiplied by the forget gate f_t. These new products are integrated and expressed in equation (12). The current memory c_t' and the long-term memory c_{t-1} are incorporated for generating a cell state c_t, which is expressed in equation (12). Here, W_c *and* b_c represent the weight matrix and bias of the cell state c_t.

$$c_t' = tanh\,(W_c \times [h_{t-1}, x_t] + b_c) \tag{11}$$

$$c_t = f_t \times c_{t-1} + i_t \times c_t' \tag{12}$$

The parameters considered in the LSTM network are determined as follows, loss function as categorical cross-entropy, optimizer as Adam, epochs is 100, total classes is 3, batch size is 8, and learning rate is 0.001. The empirical results of the proposed framework are given in section 4.

4. RESULTS

The proposed framework (hybrid feature extraction with LSTM network) is implemented using the Python programming language. The following libraries are utilized for executing this framework such as Numpy, Pandas, Sklearn, SpaCy, TensorFlow, and Keras. This framework is simulated on a system with 16GB of memory, an Intel i5 12th generation processor, a 1TB hard disk, and an NVIDIA GeForce RTX 3050 graphics card.

The effectiveness of the proposed framework (hybrid feature extraction with LSTM network) is evaluated by means of accuracy, recall, and precision. The mathematical formulas of accuracy, recall, and precision are presented in equations (13-15).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \tag{13}$$

$$Recall = \frac{TP}{TP+FN} \times 100 \tag{14}$$

$$Precision = \frac{TP}{TP+FP} \times 100 \tag{15}$$

Here, a False Negative (FN) represents a scenario where the LSTM network incorrectly predicts an actual positive sample as negative, and a False Positive (FP) denotes a scenario where the LSTM network incorrectly predicts an actual negative sample as positive. On the other hand, a True positive (TP) represents a scenario where the LSTM network correctly predicts an positive sample as positive, and a True Negative (TN) states that the LSTM network correctly predicts a negative sample as negative.

As stated in table 1, the proposed framework's performance is compared with four existing classification models, namely decision tree, random forest, Artificial Neural Network (ANN) and Recurrent Neural Network (RNN). Furthermore, we have validated the results using three different feature combinations (TF-IDF, Word2vec, and hybrid features). By analyzing table 1, the proposed framework achieved an accuracy of 96.30%, a recall of 97.12%, and a precision of 96.88%, which are higher in comparison to other classification models. The overall performance analysis is visually presented in figures 2 and 3.

Table 1. Performance analysis of the proposed framework with different classifiers

Feature extraction	Classifiers	Accuracy (%)	Recall (%)	Precision (%)
TF-IDF	Decision tree	88.25	87.98	89.66
	Random forest	88.27	88.34	90.37
	ANN	89.30	90.30	90.70
	RNN	90.76	91.14	91.76
	LSTM	88.20	89.08	90.82
Word2vec	Decision tree	91.16	91.10	91.84
	Random forest	92.12	90.85	92.48
	ANN	92.75	92.82	93.72
	RNN	91.80	92.44	92.80
	LSTM	91.38	91.80	92.77
Hybrid features (TF-IDF and Word2vec)	Decision tree	94.20	95.03	94.20
	Random forest	94.71	96.75	94.60
	ANN	95.29	95.80	95.86
	RNN	95.80	96.74	96.25
	LSTM	**96.30**	**97.12**	**96.88**

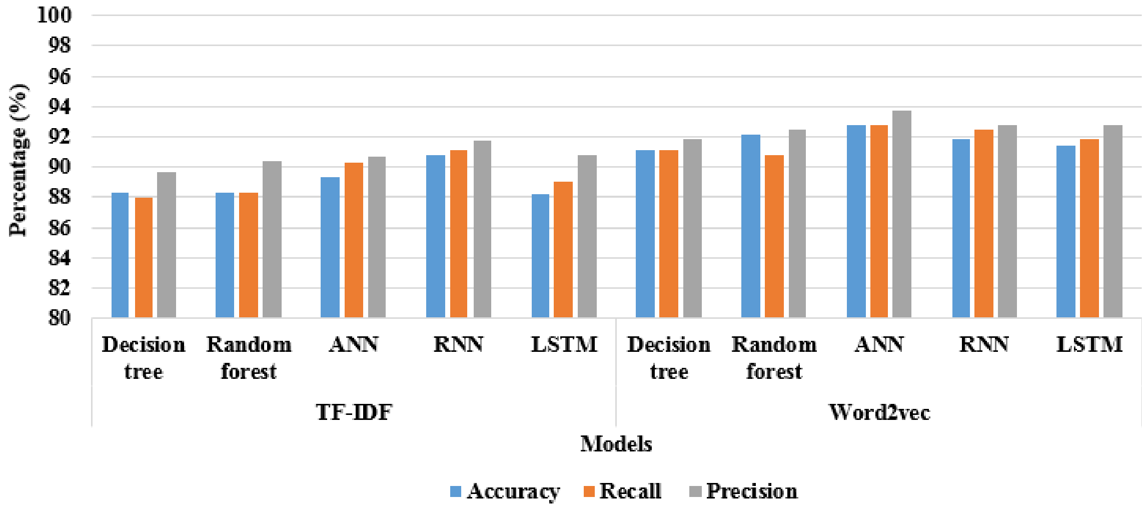

Figure 2. Graphical analysis of the LSTM network and other classification models using individual features

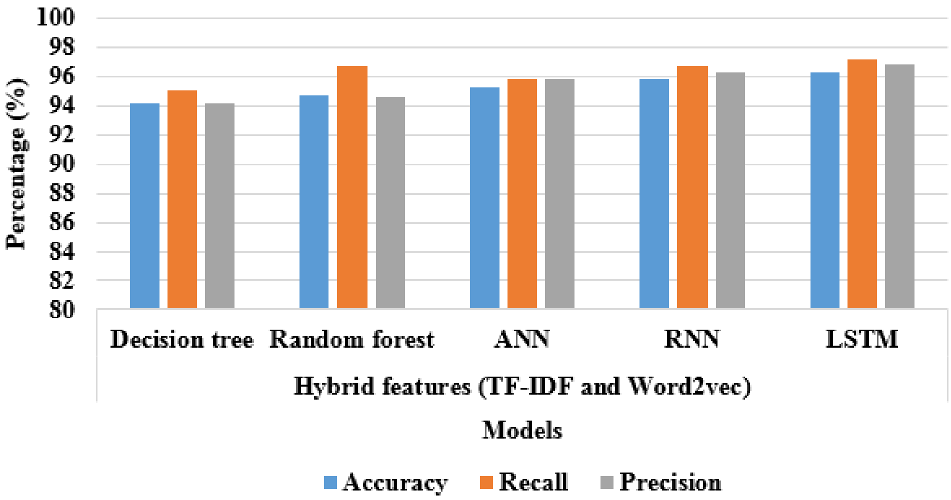

Figure 3. Graphical analysis of the LSTM network and other classification models using hybrid features

5. CONCLUSION

In the present scenario, sentiment analysis is an emerging tool utilized for interpreting and classifying emotions expressed in text data. The machine learning models are vital in the field of sentiment analysis, which is effective in interpreting, analysing, and classifying subjective information in text data. In this context, the extensive experimental investigation revealed that the proposed framework achieved superior performance in comparison to the existing traditional classification models. On the collected dataset, the proposed framework (hybrid feature extraction with LSTM network) obtained an accuracy of 96.30%, a recall of 97.12%, and a precision of 96.88%. In future work, an ensemble-based deep learning model can be implemented for further enhancing the performance of sentiment analysis. Here, analyzing the sentiments of employees regarding their layoff helps in finding potential risks like internal conflicts, legal issues, and reputational damage. Furthermore, early detection of negative sentiments allows organizations to overcome these risks proactively, before the escalation.

REFERENCES

[1] Fatouros, G., Soldatos, J., Kouroumali, K., Makridis, G. and Kyriazis, D., 2023. Transforming sentiment analysis in the financial domain with ChatGPT. Machine Learning with Applications, 14, p.100508.

[2] Kaur, G. and Sharma, A., 2023. A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis. Journal of big data, 10(1), p.5.

[3] Wang, D., Guo, X., Tian, Y., Liu, J., He, L. and Luo, X., 2023. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. Pattern Recognition, 136, p.109259.

[4] Braig, N., Benz, A., Voth, S., Breitenbach, J. and Buettner, R., 2023. Machine learning techniques for sentiment analysis of COVID-19-related twitter data. IEEE Access, 11, pp.14778-14803.

[5] Manias, G., Mavrogiorgou, A., Kiourtis, A., Symvoulidis, C. and Kyriazis, D., 2023. Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data. Neural Computing and Applications, 35(29), pp.21415-21431.

[6] Wadhwani, G.K., Varshney, P.K., Gupta, A. and Kumar, S., 2023. Sentiment analysis and comprehensive evaluation of supervised machine learning models using Twitter data on Russia-Ukraine war. SN Computer Science, 4(4), p.346.

[7] Prasanna, M.S.M., Shaila, S.G. and Vadivel, A., 2023. Polarity classification on twitter data for classifying sarcasm using clause pattern for sentiment analysis. Multimedia Tools and Applications, 82(21), pp.32789-32825.

[8] Abiola, O., Abayomi-Alli, A., Tale, O.A., Misra, S. and Abayomi-Alli, O., 2023. Sentiment analysis of COVID-19 tweets from selected hashtags in Nigeria using VADER and Text Blob analyser. Journal of Electrical Systems and Information Technology, 10(1), p.5.

[9] Diantoro, K., Soderi, A., Rohman, A. and Sitorus, A.T., 2023. Sentiment Analysis of Public Opinion on the 2024 Presidential Election in Indonesia Using Twitter Data with the K-NN Method. Digitus: Journal of Computer Science Applications, 1(1), pp.1-10.

[10] Cheng, T., Han, B. and Liu, Y., 2023. Exploring public sentiment and vaccination uptake of COVID-19 vaccines in England: a spatiotemporal and sociodemographic analysis of Twitter data. Frontiers in Public Health, 11, p.1193750.

[11] Widodo, D.A., Iksan, N. and Sunarko, B., 2023. Sentiment analysis of Twitter media for public reaction identification on COVID-19 monitoring system using hybrid feature extraction method. International Journal of Intelligent Systems and Applications in Engineering, 11(1), pp.92-99.

[12] Park, J. and Seo, Y.S., 2023. Twitter sentiment analysis-based adjustment of cryptocurrency action recommendation model for profit maximization. IEEE Access.

[13] Baqach, A. and Battou, A., 2023. CLAS: A new deep learning approach for sentiment analysis from Twitter data. Multimedia Tools and Applications, 82(30), pp.47457-47475.

[14] Vidyashree, K.P. and Rajendra, A.B., 2023. An improvised sentiment analysis model on twitter data using stochastic gradient descent (SGD) optimization algorithm in stochastic gate neural network (SGNN). SN computer science, 4(2), p.190.

[15] Bania, R.K., 2024. CovDLCNet: LSTM based deep learning network for multiclass sentiment analysis on COVID-19 public tweets. Multimedia Tools and Applications, pp.1-33.

[16] Parveen, N., Chakrabarti, P., Hung, B.T. and Shaik, A., 2023. Twitter sentiment analysis using hybrid gated attention recurrent network. Journal of Big Data, 10(1), p.50.

[17] Aslan, S., Kızıloluk, S. and Sert, E., 2023. TSA-CNN-AOA: Twitter sentiment analysis using CNN optimized via arithmetic optimization algorithm. Neural Computing and Applications, 35(14), pp.10311-10328.

[18] Vatambeti, R., Mantena, S.V., Kiran, K.V.D., Manohar, M. and Manjunath, C., 2024. Twitter sentiment analysis on online food services based on elephant herd optimization with hybrid deep learning technique. Cluster Computing, 27(1), pp.655-671.

[19] Khan, Z.A. and Rekha, V., 2023. Fake News Detection Using TF-IDF Weighted with Word2Vec: An Ensemble Approach. International Journal of Intelligent Systems and Applications in Engineering, 11(3), pp.1065-1076.

[20] Susanto, A.D., Pradita, S.A., Stryadhi, C., Setiawan, K.E. and Hasani, M.F., 2023, October. Text Vectorization Techniques for Trending Topic Clustering on Twitter: A Comparative Evaluation of TF-IDF, Doc2Vec, and Sentence-BERT. In 2023 5th International Conference on Cybernetics and Intelligent System (ICORIS) (pp. 1-7). IEEE.

[21] Edara, D.C., Vanukuri, L.P., Sistla, V. and Kolli, V.K.K., 2023. Sentiment analysis and text categorization of cancer medical records with LSTM. Journal of Ambient Intelligence and Humanized Computing, 14(5), pp.5309-5325.

[22] Yadav, V., Verma, P. and Katiyar, V., 2023. Long short term memory (LSTM) model for sentiment analysis in social data for e-commerce products reviews in Hindi languages. International Journal of Information Technology, 15(2), pp.759-772.

[23] Mohbey, K.K., Meena, G., Kumar, S. and Lokesh, K., 2023. A CNN-LSTM-based hybrid deep learning approach for sentiment analysis on Monkeypox tweets. New Generation Computing, pp.1-19.

Citation: Ranjit Kumar S, Sentiment Analysis on Employee Layoffs Based on Hybrid Feature Extraction and Long Short Term Memory Network, International Journal of Natural Language Processing (IJNLP), 2(1), 2024, pp. 12-20

Abstract: https://iaeme.com/Home/article_id/IJNLP_02_01_002

Article Link:
https://iaeme.com/MasterAdmin/Journal_uploads/IJNLP/VOLUME_2_ISSUE_1/IJNLP_02_01_002.pdf

✉ **editor@iaeme.com**