International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Impact of Feature Selection for Emotion Detection from Annotated Punjabi Text

Ubeeka Jain*[1], Parminder Singh[2]

Submitted: 27/04/2023 **Revised**: 28/06/2023 **Accepted**: 05/07/2023

Abstract*:* Natural language processing means dealing with data that is understood by humans but not directly understood by machines. Firstly, there is a need to convert that data into a form of numeric data because machines are able to understand and process numeric data. This procedure of data conversion is called feature extraction from text data. In this paper, various feature extraction techniques along with text similarity methods are used to extract numeric values from already processed and cleansed data. These techniques are term frequency-inverse document frequency, along with cosine similarity, jaccard similarity, and euclidean distance methods of text similarity. So, in this way, features are extracted from annotated emotional Punjabi text data for system training and testing for the classification process. In this article, a novel system is proposed for feature selection after feature extraction. The most relevant feature sets are selected by the grasshopper optimization algorithm and provided to the classification model for system training and testing. Comparison of results after feature extraction and after feature selection is done under various statistical performance measures, and these are satisfactory.

Keywords: *Natural Language Processing, Feature Extraction, Feature Selection, Grasshopper Optimization Algorithm, Emotion Detection*

1. Introduction

Every piece of information gathered in real life is substantial. A method is needed to comprehend this data, which is impossible to process manually. The idea of feature extraction enters the picture at this point [1, 2, 3]. Feature extraction is a process for transforming processed data into usable numerical features [4, 5, 6]. Feature extraction is a step in the dimensionality reduction process that splits and compacts raw data into easily handled chunks so that the data processing can be easy and quick [7, 8, 9]. These huge data sets include a large number of variables [10, 13], and handling and processing those variables efficiently is an important aspect [14, 15, 16]. These variables are chosen and combined into valuable and meaningful sets [17, 18]; this process is called the extraction of the best feature sets that are suitable to train a classification model [19, 20, 21]. Rather, these feature sets are reduced but still represent the original data set with great accuracy [22, 23]. So it is clearly observed that the large data sets are efficiently compacted to preserve pertinent information [24, 25, 26]. Thus, using feature extraction, the quantity of repetitive data in the corpus is reduced [27, 28, 29].

Three text similarity techniques are used in the proposed system for feature extraction, together with term frequency and inverse document frequency. These text similarity techniques include Euclidean distance, Cosine Similarity, and Jaccard Similarity. Additionally, the Grasshopper Optimization method [11, 12], a recent approach from the field of swarm intelligence, is used to improve these qualities. This approach has been successfully used to address a variety of optimization issues in a number of contexts. One of the key uses of grasshopper optimization algorithm is feature selection. This algorithm is built on levy flights, which evaluate the aspects of the annotated Punjabi text data that are accepted or rejected. The best feature sets are chosen for the system's training in this way. The proposed approach thus completes the machine learning process. Six fine-grained emotional classes are defined in this study to categorize emotional annotated text data of the Punjabi language using a variety of classification algorithms. Traditional measures are used to evaluate the results, and overall, the system performed well. The study has thus been successfully finished with positive findings.

1.1. Framework and Contribution to the Research

This article describes the comparative study of before feature selection and after feature selection for fine-grained emotion detection in Punjabi annotated text data. In this research article, a procedure is proposed for feature selection so that the enhanced feature sets can classify annotated Punjabi text under six emotional classes in an efficient manner. The contribution of the work covered in this article is discussed below.

[1] *Research Scholar, IKG Punjab Technical University, Kapurthala, Punjab (India),*
[2] *Department of Computer Science & Engineering, Guru Nanak Dev Engineering College, Ludhiana, Punjab (India).*
[1] *Corresponding author. Email: ubeekajain@gmail.com*

- The key contribution of this work is to extract features from cleaned and processed Punjabi annotated text data, which has previously been gathered and validated under this research activity.
- Another important contribution to the direction of this study is feature selection to enhance the received feature sets in order to make the proposed system more accurate and efficient. The most advantageous feature sets for system training and classification have been selected.
- The Grasshopper Optimization Algorithm is employed to locate the most pertinent feature sets for system training. The system responds effectively to the proposed technique.
- A comparative study of feature extraction and enhancement is evaluated by various standard metrics.
- Experimental results are validated by standard metrics and compared with each other to get the best-suited results.

The remaining paper is structured by various sections, as Section 2 explains various feature extraction techniques along with text similarity methods applied in the proposed system. Section 3 brings to light on feature selection by the Grasshopper Optimization Algorithm to enhance the extracted feature sets. Section 4 includes a comparative study of feature extraction and feature selection with experimental results and a discussion of various performance measures. Section 5 brings to light on the conclusion and future scope of the proposed work.

2. Feature Extraction Process

One of the most common and fundamental methods for feature extraction is the term frequency and inverse document frequency. This is employed to pull out features from annotated text data in Punjabi. TF-IDF, which stands for term frequency-inverse document frequency, is a procedure that gives terms specific weights that indicate how significant they are in the document. This makes it easier to determine how frequently a particular word exists in that record and how many other documents it appears in overall. Therefore, the useful algorithm Term Frequency-Inverse Document Frequency employs the frequency of words to assess how important certain words are to a specific document. So, in this study, processed annotated data in Punjabi is first transformed into numerical data. Then TF-IDF is applied on Punjabi annotated text data for feature extraction.

2.1. TF-IDF

TF-IDF is a method, used in the fields of text summarization, machine learning and information retrieval. It measures the significance of string representations (words, phrases, lemmas, etc.) in a record amongst a group of records that is also known as dataset. TF-IDF can be splited into two components: TF (term frequency) and IDF (inverse document frequency). By examining how frequently a specific phrase appears in relation to the document, term frequency analyses documents.

Inverse document frequency indicates how frequently or infrequently it appears in the whole corpus. It is used to determine whether a word is relevant to a corpus of documents. The words those most commonly occur in almost every document like stop words, are not useful for that particular document. IDF can find by the division of total number of documents by the number of documents in which a particular word occurs and then find out the logarithm of the value comes from division. If the word occurs frequently in several documents, then this number is close to 0. Otherwise, the number is close to 1. Both the terms can be formulated by table 1 below.

In the formulation of IDF, n represents total number of document in a corpus and d(i) represent number of documents in which particular word occurs. Multiplication of TF and IDF provides the collective score of a word in a document. The word is more pertinent to that specific document if this score is higher. It can be mathematically formulated below.

Table 1: Feature extraction measures and formulation.

Measure	Formula
Term Frequency	Number of times a term occurs in the document / total number of terms in document
Inverse Document Frequency	[log (n/d(i))]
Term Frequency-Inverse Document Frequency	TF×IDF

To put it briefly, the major shortcoming of this measure is that it takes a long time for large data sets to directly calculate document similarity in the word count space. For this reason, in this study, various text similarity algorithms are used and outlined in the following sections to extract more pertinent features from the Punjabi emotional text data. This is because it is not adequate to extract features just on the basis of TF-IDF due to some drawbacks.

2.2. Text Similarity

Text similarity measures are based on the content, structure, or style of two texts; and provide numerical scores that show how similar text documents are. These tools are helpful for the processes of information extraction and analysis from enormous collections of text data, known as text mining and text analytics. Text similarity methods are applied to find out which texts and documents are most alike. This is also used in document recommendations. The usage of text similarity measure is quite important in the field of natural language processing. There are various text similarity methods, but in this study three most popular methods are used in the development of emotion detection system from annotated Punjabi text data, those are discussed in the following subsections.

2.2.1. Cosine similarity

It can be calculated as if two vectors are almost pointing in the same direction then find out the cosine of the angle between two vectors. It is most appropriate for where replicated text is significant and the specialty of this text similarity method is that, it is capable of handling documents of any size. There is no data size restriction. Cosine similarity measure is applied on Punjabi annotated text corpus for proposed emotion detection system. This is applied to find similarity between text data for better machine learning process.

2.2.2. Jaccard similarity

This represents the ratio of common words to total unique words or we can say the intersection of words to the union of words in both the documents. This measure predicted the data objects like sets. It is determined by dividing the size of the union by the size of the intersection of two sets. It provides results in the range between 0 – 1. 1 means the higher similarity between text files on the other hand 0 means there is no similarity between the given texts. This similarity measure works only on the unique set of words for each sentence and does not consider duplicate text or words in a sentence. For proposed emotion detection system we have applied jaccard similarity measure on Punjabi annotated text corpus in order to find similarity between text files for better machine learning process.

2.2.3. Euclidian Distance

This is the most commonly used type of distance. Vectors of text data are required for the application of this measure. It computes the separation between two points by applying the Pythagoras theorem. The similarity score will decrease and vice versa, depending on how far apart two vectors are from one another. The Euclidean distance measure is applied to the proposed annotated Punjabi emotional text corpus in order to find similarity between texts for a better machine learning process.

3. Feature Selection Process

It involves automatically selecting appropriate feature sets for a machine learning model. Feature selection is necessary to enhance the feature set extracted in the previous section of this paper. Finding the right data for system training and testing in enormous data sets is particularly challenging. In this work, a novel behavior is devised and developed to solve this problem. The best aspects of the data are enhanced using the Grasshopper Optimization Algorithm for system training and testing. This meta-heuristic optimization algorithm is new. It is applied to numerous real-life problems across different fields. To optimize the numerical aspects of an annotated database, grasshopper is designed and developed on the basis of levy flights (LFGOA). This optimization approach enhances both performance and accuracy. While looking for the most relevant collection of characteristics from annotated text data, this LFGOA is kept appropriately stable between exploitation and exploration. Steps for this feature enhancement process are discussed below.

```
1.  Initialize input from feature extraction (take
    output of text similarity measures as input).
2.  Initialize six class labels.
3.  Initialize Levy flights and group order.
4.  Initialize empty arrays for reward and penalty.
5.  Select random population for group food.
6.  Find global food by taking centroid from k-
    means.
7.  Find pairing food.
8.  Apply fitness function
        If value of fitness flag == 0
            Then assign penalty to fitness threshold
        Else
            Assign(100-fitness threshold) to reward
9.  Find summation of rewards.
10. Find summation of penalties.
11. If summation of rewards > summation of
    penalties
            Accept the record
      Else
          Reject the record
12. Find selected indexes as output.
```

4. Experimental Results and Comparative Analysis

Annotated Punjabi text data is already processed, and mathematical features are also extracted, as discussed in the previous sections of this paper. In this section of the paper, the result analysis is done on the feature extraction techniques as explained in the earlier section. The proposed system follows two phases, one is training and another is testing phase. In this study, features are extracted from Punjabi annotated text data, and those mathematical features are used to train the developed model and tested under various scenarios by different classifiers, like SVM, ANN, GNB, etc. The system is evaluated by the performance metric named confusion matrix after feature extraction without feature selection, as is elaborated in following subsections. The results of the feature selection process by the proposed algorithm are discussed in the next subsection below.

4.1. Results by Grasshopper Optimization Algorithm

Grasshopper optimization algorithm is applied to enhance the best suited features of data for system training and testing. This is a recent meta-heuristics optimization algorithm. It is used for many real world problems in various areas. Levy flights based grasshopper is designed and developed to optimize the numeric features of annotated database. Performance and accuracy both are improved by this optimization algorithm. This LFGOA is maintained appropriate stability in between exploitation and exploration while searching the most related set of features from annotated text data. Results of rewards mechanism in terms of selection (reward) and rejection (penality) of most suitable data sets for system training is represented by Figures 1 to 6 below.

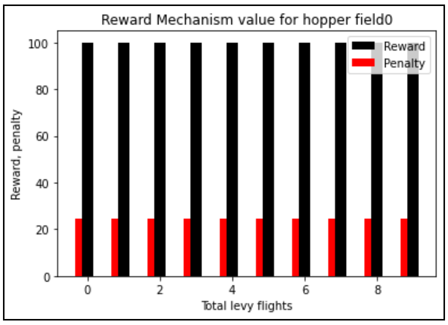

Fig 1. Reward Mechanism for hopper field 1

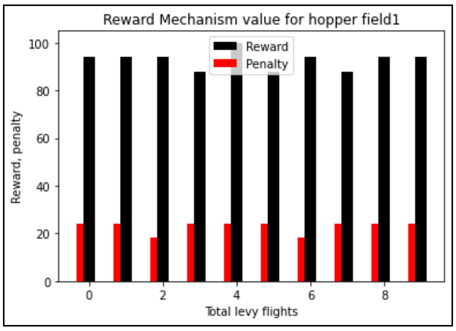

Fig 2. Reward Mechanism for hopper field 2

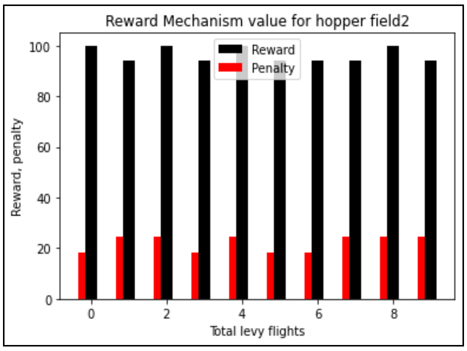

Fig 3. Reward Mechanism for hopper field 3

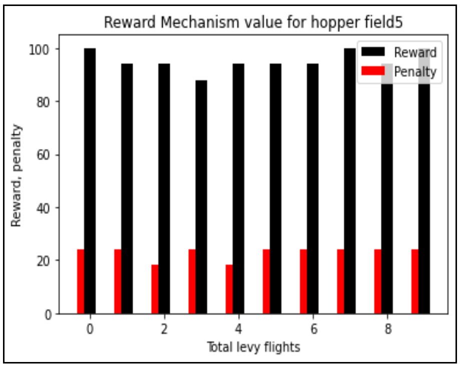

Fig 4. Reward Mechanism for hopper field 4

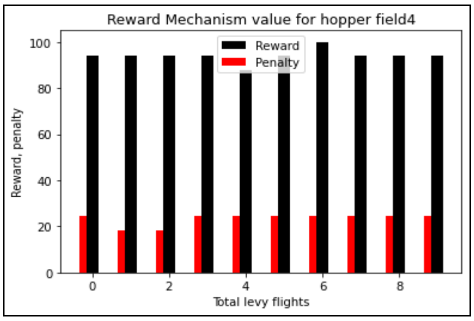

Fig 5. Reward Mechanism for hopper field 5

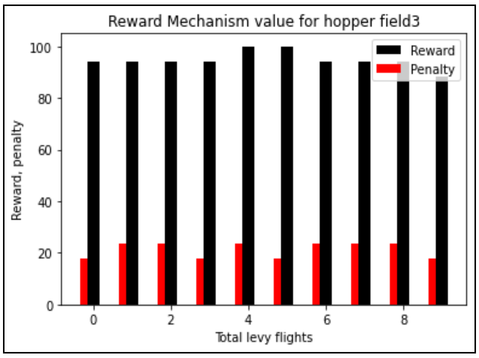

Fig 6. Reward Mechanism for hopper field 6

4.2. Confusion Matrix

In essence, it is a graphical display of the outcomes of the previously mentioned process. This matrix contains an equal number of rows and columns. If the number of classes are unequal in any dataset then it’s accuracy can be easily manipulated. Because a confusion matrix is also known as an error matrix, it can help to determine what kind of errors are in a technique. In this matrix, each row displays the actual class, and each column displays the predicted class. The main idea behind this matrix is to determine how frequently a class's instances are incorrectly classified, or, to put it another way, how frequently a classification model confuses the predicted classes with the target classes. Confusion matrix for all the classes after feature extraction and without application of feature selection algorithm is shown by the Figure 7.

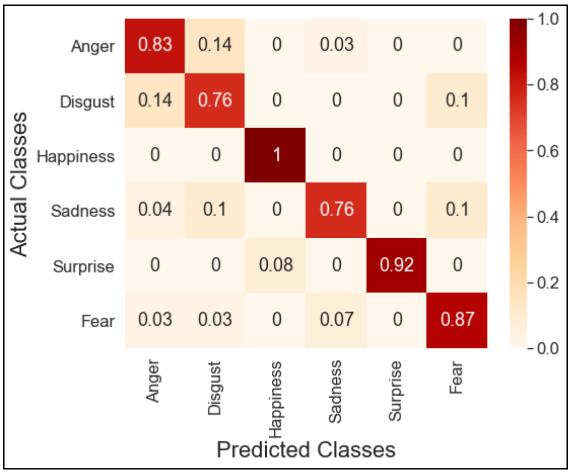

Fig 7. Confusion matrix for all classes before feature selection

From the above confusion matrix, it can be noticed that happiness emotion is most distinguishable and causes fewer confusions with other emotional classes. Surprise and fear classes have sometimes been confused. Anger, sadness, and disgust were also confused, but to a much lesser extent. So, overall the results of this evaluation process are satisfactory. But these results are enhanced by developing a novel behavior with the grasshopper optimization algorithm for feature selection discussed in the section 3 of this paper. So, after applying this feature enhancement algorithm, the most suitable feature sets are selected and provided to the model for training and testing. More accurate results are gathered and shown in Figure 8 of the confusion matrix for all classes after the feature selection algorithm.

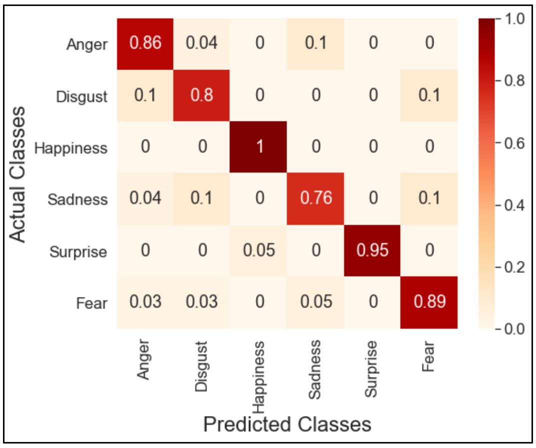

Fig 8. Confusion matrix for all classes after feature selection

From the above confusion matrix, it can be noticed that happiness emotion is also most distinguishable and causes fewer confusions with other emotional classes. Surprise, fear, and anger classes have sometimes confused. Sadness and disgust were also confused, but to a much lesser extent. So, the results of this evaluation process are enhanced, and it provides more accuracy as compared to the prior feature selection process. To wrap up, this is a much healthier approach for accessing the correctness of the classification method.

4.3. Comparative Analysis

In this section of the paper, the results of various experiments on annotated Punjabi text data before and after feature selection by the proposed algorithm for feature enhancement are comparatively measured by various performance metrics.

4.3.1. Accuracy

Accuracy is one of the statistical measures for evaluating classification models. It is represented by the sum of all accurate predictions divided by the sum of all predictions for a dataset. Accuracy in terms of percentage for all the classes are represented by the figure 9 below.

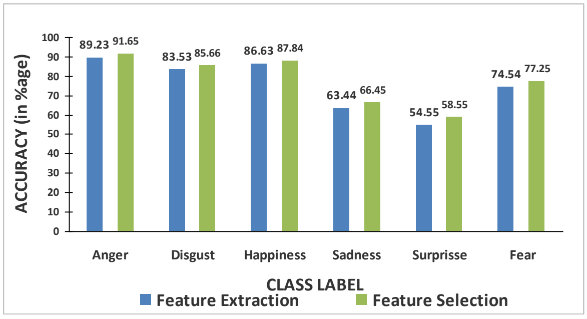

Fig 9. Comparison of accuracy values before and after feature selection

From the above figure, it is observed that a sudden rise in the results happened after the feature selection algorithm. The most relevant feature sets are selected and applied to the model for system training and testing. The accuracy of all the classes is higher as compared to earlier results. The execution time of the system also reduces, and the speed of the system also increases. So, the overall performance of the system improves.

4.3.2. Precision, Recall and F1-Score

All these three are excellent statistical metrics to find the overall performance of the emotion detection system for Punjabi. In this study, all three parameters are calculated and diagrammatically represented by Figures 10, 11, and 12.

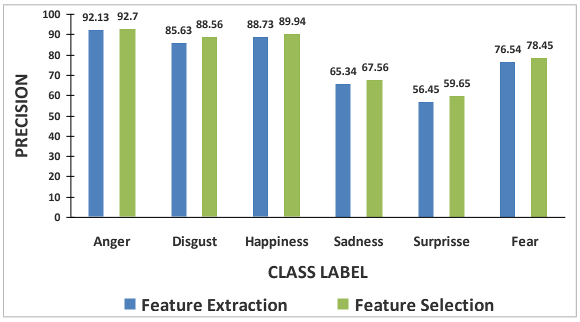

Fig 10. Comparison of precision before and after feature selection

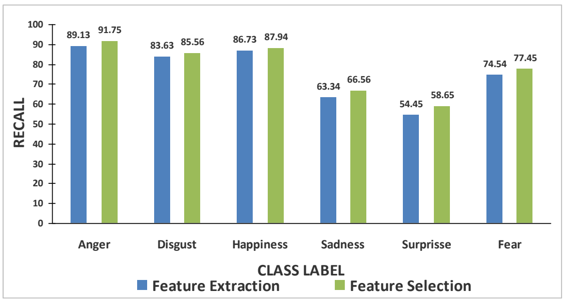

Fig 11. Comparison of recall before and after feature selection

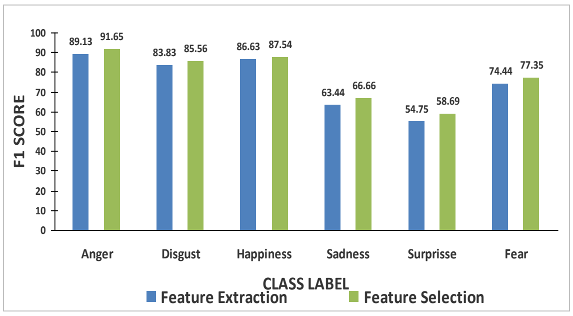

Fig 12. Comparison of f1-score before and after feature selection

From these figures, it is observed that the values of all the measures for all the classes are higher after feature selection. The key idea behind this sudden and improved rise in results is that when the feature selection algorithm is applied after feature extraction, the most relevant set of features are chosen to provide as input to the classification model, which improves the overall system performance. Again, the execution time reduces because of the most relevant feature set selection, and the speed and accuracy increase. The overall system performs well.

5. Conclusion and Future Scope

To improve the process of machine learning, the basic requirement is that the best-suited data sets are passed to this process as input. After pre-processing of text data which is also the part of this research work, mathematical features are derived termed as feature extraction from text data. Extraction of features and selection of features are the key steps of this research work. In this paper, the extraction of feature process and the selection of features process are explained in detail. Feature sets are enhanced by the proposed feature selection grasshopper optimization algorithm. Various feature sets are tested before and after feature selection, and the proposed system is also tested under various statistical performance metrics. It is observed that the system performs well after the application of the feature selection algorithm. In the future, improvements can be made to the database. The number of files for all the emotional categories can be more, and the number of emotional classes can also be more. This system can also be trained and tested by more classification models. More feature extraction techniques and algorithms for feature selection can be explored to get better performance. So, the overall performance of emotion detection system for Punjabi is satisfactory and good.

Acknowledgements

This work was supported by IKG Punjab Technical University, Kapurthala, Punjab (India) and Guru Nanak Dev Engineering College, Ludhiana, Punjab (India). The authors are really thankful to theses organizations for their support in this research work.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] A. Abdi, S. M. Shamsuddin, S. Hasan, J. Piran. Deep Leaning – based Sentiment Classification of Evaluation Text Based on Multi-feature. Information Processing and Management. 2019; 56(4):1245-1259.

[2] S. Kusal, S. Patil, K. Kotecha, R. Aluvalu, V. Varadarajan. AI Based Emotion Detection for Textual Big Data: Techniques and Contribution. Big Data and Cognitive Computing. 2021; 5(43):1-45.

[3] P. Nandwani, R. Verma. A Review on Sentiment Analysis and Emotion Detection from Text.

Social Network Analysis and Mining. 2021; 11(81):1-19.

[4] A. Kaur, V. Gupta. N-gram Based Approach for Opinion Mining of Punjabi Text. Multi-Disciplinary Trends in Artificial Intelligence. 2014; 8(1):81-88.

[5] A. Landowska, M. Szwoch, W. Szwoch, M. R. Wrobel, A. Kolakowska. Emotion Recognition and its Application. Human- Computer Systems Interaction: Backgrounds and Applications. 2014; 3(1):1-13.

[6] M. Krommyda, A. Rigos, K. Bouklas, A. Amditis. An Experimental Analysis of Data Annotation Methodologies for Emotion Detection in Short Text Posted on Social Media. Informatics. 2021; 8(19):2-15

[7] S. Poria, E. Cambria, A. Gelbukh, F. Bisio, A. Hussain. Sentiment Data Flow Analysis by Means of Dynamic Linguistic Patterns. IEEE Computational Intelligence Magazine. 2015; 10(4): 26-36.

[8] Zualkernan, F. Aloul, S. Shapsough, A. Hesham, Y. El-Khorzaty. Emotion Recognition Using Mobile Phones. Computers and Electrical Engineering. 2017; 60(1): 1-13.

[9] S. N. Shivhare, S. K. Saritha. Emotion Detection from Text Documents. International Journal of Data Mining and Knowledge Management Process. 2014; 4(6): 51-57.

[10] B. Karin, P. M. Viviane, S. Aline. Multilingual Emotion Classification Using Supervised Learning: Comparative Experiments. Information Processing and Management. 2017; 53(1): 684-704.

[11] L. Wu, J. Wu, and T. Wang. Enhancing Grasshopper Optimization Algorithm (GOA) with Levy Flight for Engineering. Nature Portfolio. 2023; 13(4): 1-49.

[12] Y. Meraihi, A. B. Gabis, S. Mirjalili, and A. R. Cherif. Grasshopper Optimization Algorithm: Theory, Variants and Applications. IEEE Access. 2021; 9(1): 50001-50024.

[13] A. R. Murthy and A. Kumar. A Review of Different Approaches for Detecting Emotion from Text. Materials Science and Engineering. 2021; 11(10): 1-23.

[14] R. Ahuja, A. Chug, S. Kohali, S. Gupta, and P. Ahuja. The Impact of Features Extraction on the Sentiment Analysis. Procedia Computer Science. 2019; 15(2): 341-348.

[15] J. Akilandeswari and G. Jothi. Sentiment Classification of Tweets with Non-Language Features. Procedia Computer Science. 2018; 14(3): 426-433.

[16] M. Tarhanicova, K. Machova, and P. Sincak. Computers Capable of Distinguishing Emotions in Text. Emergent Trends in Robotics and Intelligent Systems. 2015; 31(6): 57-64.

[17] B. Karin, P. M. Viviane, and S. Aline. Multilingual Emotion Classification Using Supervised Learning: Comparative Experiments. Information Processing and Management. 2017; 53(1): 684-704.

[18] D. Yasmina, M. Hajar, and A. M. Hassan. Using YouTube Comments for Text-Based Emotion Recognition. Procedia Computer Science. 2016; 83(1): 292-299.

[19] E. Yar, I. Delibalta, and L. Baruh. Online Text Classification for Real Life Tweet Analysis. In Proceedings of IEEE Signal Processing and Communication Application Conference (SPCA'16). IEEE, Turkey, 1609-1612.

[20] S. M. Nagarajan and U. D. Gandhi. Classifying Streaming of Twitter Data Based on Sentiment Analysis Using Hybridization. Neural computing and Applications. 2019; 31(5): 1425-1433.

[21] L. Servi and S. B. Elson. 2015. A Mathematical Approach to Gauging Influence by Identifying Shifts in the Emotions of Social Media Users. IEEE Transactions on Computational Social System. 2015; 1(4): 180-190.

[22] S. K. Bharti, R. K. Gupta, P. K. Shukla, M. Bouye, S. K. Hingaa, and A. Mahmoud. Text Based Emotion Recognition Using Deep Learning Approach. Computational Intelligence and Neuroscience. 2022; 22(1): 1-8.

[23] P. Natarajan. Efficient Natural Language Processing Used for Twitter Data Based on Sentiment Analysis. International Journal of Scientific Development and Research. 2019; 4(6): 368-380.

[24] B. S. Sundar, V. Rohith, B. Suman, K. N. Chary. Emotion Detection on Text Using Machine Learning and Deep Learning Techniques. International Journal for Research in Applied Science & Engineering Technology. 2022; 10(6): 2277-2286.

[25] A. A. Maruf, Z. M. Ziyad, M. M. Haque, F. Khanam. Emotion Detection from Text and Sentiment Analysis of Ukraine Russia War Using

Machine Learning Technique. International Journal of Advanced Computer Science and Applications. 2022; 13(12): 868-882.

[26] A. H. Saffar, T. K. Maan, B. Ofoghi. Textual Emotion Detection in Health: Advances and Applications. Journal of Biomedical Informatics. 2023; 137(C).

[27] S. Peng, L. Cao, Y. Zhou, Z. Ouyang, A. Yang, X. Li, W. Jia, S. Yu. A Survey on Deep Learning for Textual Emotion Analysis in Social Networks. Digital Communications and Networks. 2022; 8(5): 745-762.

[28] J. Guo. Deep Leraning Approach to Text Analysis for Human Emotion Detection from Big Data. Journal of Intelligent Systems. 2022; 31(1): 113-126.

[29] Y. Cai, X. Li, J. Li. Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review. Sensors. 2023; 23(5): 1-33.