Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 9 No. 1 2024, pp. 8 – 18

Utilization Of Principal Component Analysis To Improve Emotion Classification Performance In Text Using Artificial Neural Networks

Mahazam Afrad*[1]**, Muljono**[2]
1,2Teknik Informatika, Universitas Dian Nuswantoro
*E-mail : mahazam01@gmail.com**[1]*, muljono@dsn.dinus.ac.id*[2]

Pujiono[3]
Teknik Informatika, Universitas Dian Nuswantoro
E-mail : pujiono@dsn.dinus.ac.id[3]

Abstract - Emotions, being transient and variable, differ across locations, times, and individuals. Automatic emotion identification holds significant importance across various domains, such as education and business. In education, emotional analysis contributes to intelligent electronic learning environments, while in business, it aids in assessing customer satisfaction with products. This study advocates the application of Principal Component Analysis (PCA) to enhance the performance of text emotion classification using the Artificial Neural Network (ANN) method. PCA, a pattern identification method, reduces text dimensions, improving the classification process by determining word similarities. PCA offers the advantage of dimension reduction without compromising information integrity. The classification approach involves two stages: one after PCA dimension reduction and the other without PCA post TF-IDF stage. The study's conclusive findings, incorporating PCA in ANN classification, demonstrated a notable increase in recall for the happy class, reaching 0.92 compared to the pre-PCA score of 0.91. Furthermore, precision in the sadness class improved to 0.90, surpassing the pre-PCA precision of 0.80. This affirms the efficacy of integrating PCA in enhancing the accuracy and performance of emotion classification in text analysis.

Keywords - Artificial Neural network, Principal Component Analysis, Text Emotion

1. INTRODUCTION

Emotions are feelings of a person's state that are not constant and vary from place to place, time to time or person to person [1]. Identifying emotions automatically is very important because it can be applied in various fields, for example in the field of education, emotion analysis can be used for intelligent electronic learning environments, in business emotion analysis is useful for knowing the satisfaction of customers about their products good or bad. Some research on emotion classification includes research conducted by Muljono et al [2] on the classification of Indonesian language emotion text. In this study, the emotions of Indonesian texts were detected and evaluated with four different classification methods, namely Naïve Bayes (NB), J48, K-Nearest Neighbor (KNN) and Support Vector Machine - Sequential Minimal Optimization (SVM-SMO). The experiments used Indonesian text corpus with 1000 sentences which were labeled with 6 emotion labels namely anger, disgust, fear, pleasure, sadness and surprise. The data preprocessing stage consists of tokenization, normalization, stopWord removal, stemming, and Term Frequency - Inverse Document Frequency (TF-IDF) as the extraction feature of the emotion text. Researchers used 10-fold cross validation and split validation for the evaluation of the experiment. Based on the experimental results, SVM-SMO classification provides the best performance. In the evaluation using 10-fold cross validation, the results show that the accuracy of NB, J48, KNN and SVM-SMO are 80.2%, 80.8%, 68.1% and 85.5%. The same results are also shown by evaluating using split validation where the highest accuracy is obtained by SVM-SMO with 86% accuracy.

Emotion sentiment research was also conducted by Tabashum and Chanda in 2019[3], namely about extracting sentiment from text using emotions from the corpus. In this study, researchers introduced an approach to automatically mark words and then analyze them at the sentence and document level. The evaluation of tagged words is explained and the performance based on document tone change has been explained. The applicability of this approach in contextual level analysis is briefly summarized and a tentative emotion tree using word-level emotion tagging is proposed. The dataset used is taken from 1135 hospital incident articles with words labeled with 6 emotion classes namely happy, sad, angry, fear, disgust, and surprise. The result of this study is that the proposed learning method will learn more error-free if the data tree is large enough to estimate the true emotion of a word in a sentence or passage with an accuracy of 82.75%.

In recent years a lot of research identifying emotions using Twitter by analyzing from Twitter tweets to extract user sentiment or emotions about a topic, has grown rapidly. Research on emotion analysis on Twitter social media by M.A. Tocoglu and A. Alpkocak [4]. In this study, researchers took data by surveying 500 universities to write down moments in their lives in each of the 6 emotion categories and obtained a total of 3206 data. The data taken is labeled with 7 categories of emotions. The method used is word feature weighting with TF-IDF and classification with Naive Bayes. The results of the experiments in this study are the average of precision 0.861, recall 0.861 and F-masure 0.86. The results show that the research in this paper is promising to extract emotions from text with a satisfactory level. In 2019 M.A. Tocoglu and A. Alpkocak examined the analysis of emotions in Turkish tweets using Deep Neural networks. In this study [5], researchers found a problem that several studies on emotion analysis using tweets in Turkey none of which used Deep learning on large tweet datasets. As a result, there is a need for performance comparison of different deep learning architectures for tweet emotion analysis in Turkey. The data used is data sourced from Turkish Twitter totaling 195000. Preprocessing the data using a corpus from Turkish NLTK (Natural Language Toolkit) in Python stopWord removal, removing unique characters, links, spaces, and all numeric characters. Experiments also show that the proposed deep learning approach for tweet emotion analysis in Turkey performs better than traditional machine learning approaches.

Research on word embedding has previously been conducted by Jonathan Herzig et al in 2017. Research on English word embedding features by comparing basic Bag-of-Words (BOW) features and Word Embeddings (Word2vec and GloVe). Their experimental results show that combining basic BOW features and word embeddings can improve performance [6]. Word embeddings for tweet emotion classification were also used by Vora et al [7]. Using Random Forest, their model can achieve 91% precision for four emotion classes in English tweets.

Previously presented models about adopting word embedding vectors represent semantic/syntactic information and those models cannot capture emotional relationships between words. Recently, some emotional word embeddings are proposed but require semantic and syntactic information instead. Research conducted by Erdenebileg Batbaatar et al [8]proposed a new neural network architecture, named SENN (Semantic-Emotion Neural network) which can utilize semantic/syntactic and emotional information by adopting pre-trained word representations.

Sentiment analysis on Twitter studied by Saad Shihab Elbagir and Yang Jing [9]in 2019. This research aims to perform detailed tweet sentiment analysis based on ordinal regression using machine learning techniques. The dataset from Twitter is 10000 Twitter posts with 5000 positive tweets and 7000 negative tweets. Data from Twitter is divided into 7000 tweets for training models and 3000 tweets for tests. The results of this study are on evaluation with F1-score Random Forest and decision tree better than others with a value of 0.85. Decision tree accuracy is better than others with 91.81%. For MSR and MAE evaluation, the best decision tree with a small error value is MAE 0.154 and MSE 0.155.

Research on sentiment text was also conducted in India by G. Vinodhini and RM. Chanrasekaran [10] they examined the classification of sentiment text using PCA and Neural network. In the study, researchers about reviews of digital cameras from amazon sites. The dataset taken was then selected into 600 positive sentences and 600 negative sentences. In the first study, the dataset was classified using a neural network with a Backpropagation base and evaluated to determine the results of precision, recall and f-measure. The second research was carried out by adding PCA before being classified with a neural network and also evaluated to determine its value. Then the two studies were compared and the results obtained were the precision, recall and f-measure values using PCA on the Neural network produced higher values, namely on precision 84.7% while without PCA 79.5, Recall using PCA 85.4% and without PCA 82.7% and finally f-measure 85% using PCA - Neural network and 81.5% using only Neural network without PCA.

Based on previous research, in this study the authors chose the PCA and Neural network methods to propose research on the use of Principal Component Analysis (PCA) by reducing the dimensions of the text to improve the performance of text emotion classification using the Artificial Neural Network method.

2. RESEARCH METHOD

The method used in this research is Principal Component Analysis to improve the performance of Artificial Neural Network in the classification of emotions in text. The data source in this study is taken from Kaggle public data: https://www.kaggle.com/Datasets/ishantjuyal/emotions-in-text which contains tweet posts from Twitter with the emotion labels sad, love, fear, anger, surprise and pleasure. In table 1 below is a sample of the dataset which totals 21459.

Table 1. Sample Datasets

Text	***Emotion***
i didnt feel humiliated	*sadness*
i can go from feeling so hopeless to so damned...	*sadness*
im grabbing a minute to post i feel greedy wrong	*anger*
i am ever feeling nostalgic about the fireplac...	*love*
i am feeling grouchy	*anger*
ive been feeling a little burdened lately wasn...d	*sadness*
ive been taking or milligrams or times recomme...	*surprise*
Text	***Emotion***

i feel as confused about life as a teenager or...	*fear*
i have been with petronas for years i feel tha...	*happy*

The 21459 dataset consists of text sentences with emotion labels consisting of sad, love, fear, anger, surprise and pleasure. The number of each emotion label in the text can be seen in Figure 1 below.

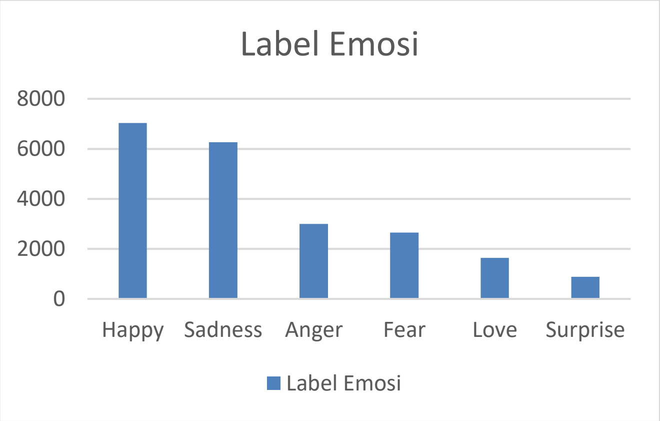

Figure 1. Emotion Label

2.1 Reseacrh Diagram

The research diagram in Figure 2 shows the first stage, namely data pre-processing, raw data or text data is carried out tokenization, StopWord and Stemming stages. The results are then carried out Bag of Word feature extraction and feature weighting using TF-IDF. The next step is to conduct dataset experiments, namely datasets reduced in dimension using PCA and datasets without PCA. The results of the two datasets are then classified using the ANN model to obtain results. The two results are then compared to find out which is better between using PCA on the Dataset or not and alsowe can draw conclusions whether PCA can improve the performance of emotion classification in text using Artificial Neural Network.

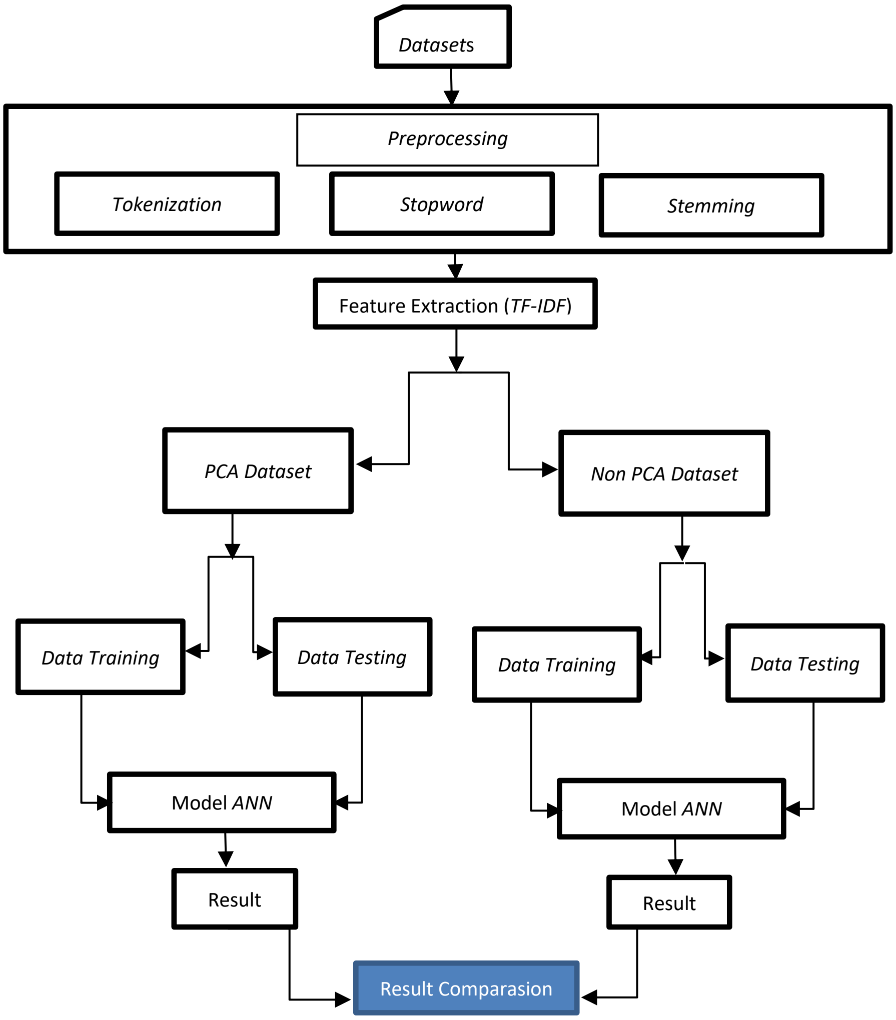

Figure 2. Research Diagram

2.2 Preprocessing

In the early stages of research, namely by forming features in the form of structured data structured data. This is done because the tweet text data comes from Twitter in the form of unstructured text in the form of English sentences with several emoticons and various punctuation marks. punctuation marks. For this reason, some unstructured data processing is required unstructured data at the beginning of processing using some standard text data preprocessing techniques that are often used. standard techniques that are often used. Before the text data is classified, the data is pre-processed by tokenization, stopWord removal and stemming. Tokenization is a process to remove punctuation from sentences in documents. StopWord removal is a process to remove conjunctions that are not included in a topic determination in the document, for example "and", "namely" and "or" [11]. Stemming is a stage to remove affixes on a word until a basic word is obtained in a document.

2.3 Feature Extraction

The pre-processed dataset will then be used for feature extraction using Bag of Word (BOW) and feature weighting using TF-IDF. Classification features are extracted based on the Bag of Words (BOW) model where each document is represented as a vector of weight values calculated from the number of words that appear in the document. Then to get features from the review text data, the TF-IDF method is used, later each term obtained from the tokenization process will be weighted using frequent terms and idf. The results of this TF-IDF Extraction show structured data, where the document shows the nth document index then the label column shows the sentiment class of the analysis or output.

In Bag of Word each document is represented as a vector of weight values calculated from the number of words that appear in the document [12]. The way Bag of Word is used is as follows:

Sentence 1	:'I am high'
Sentence 2	:'Yes I am high'
Sentence 3	:'I am Kidding'

Bag of Words will only take unique words from all the whole words that will be "I am high yes Kidding". To get features from the review text data, the TF-IDF method is used, later each term obtained from the tokenization process will be weighted using term frequency and idf [2]. The results of this TF-IDF Extraction show structured data, where the document shows the nth document index then the label column shows the sentiment class of the analysis or output.TF-IDF can be calculated by the equation below:

$$TFIDF(t) = TF * log\frac{N}{df} \qquad (1)$$

In the TFIDF formula above, it can be seen where t is a term, TF is the total number of terms t that appear in the document, N is the number of documents, and df is the number of documents containing term t. For calculation examples, we can take a sample of 9 words in the Dataset. Previously, the Bag of Word has been made to make the vectorizer can be seen in table 4.3. The next step is to determine N which is the number of documents, here only taking a sample of 9 sentences or documents in the Dataset and for df is the number of documents containing the term t, so for the N /df formula for the term humiliate in the first document, namely feel humiliated, a value of 9 / 1 = 9 can be obtained because the frequency term humiliated in document one appears only once.

2.4 Aplication of PCA model

The application of the model in this study is first with feature selection using Principal Component Analysis (PCA). Principal Component Analysis (PCA) is a method for identifying a pattern in data by determining the similarity of a word. One of the advantages in PCA is that using this method can reduce the number of dimensions that exist in a pattern without reducing the information in the data [13]. In this research, the use of PCA is done after the feature weighting stage using TF-IDF. After PCA modeling, the classification process using ANN is then carried out and evaluated. To find out whether the PCA and ANN models are successful, it will be compared with the dataset without using PCA which is directly applied to classification using ANN. The application of PCA is expected to improve the performance of Artificial Neural Network (ANN).

The calculation of the analysis by the PCA method is based on more than one eigenvalue. The PCA algorithm is usually as follows:

1. Calculate the covariance matrix using equation 2 as below:

   $$Cov(xy) = \frac{\sum xy}{n} - (x)(y) \qquad (2)$$

2. Calculate the eigenvalue using equation 3 as below:

   $$(A - \lambda I) = 0 \qquad (3)$$

3. Calculate the Eigen Vector using equation 4 below:

   $$[A - \lambda I][x] = [0] \qquad (4)$$

4. Determine the new variable (Principal Component) by multiplying the original variable by the Eigen Vector matrix. Meanwhile, what can be explained by the new variable to - I depends on the contribution of pI from each Eigen Value which is calculated based on equation 5 as follows:

   $$p = \frac{\lambda i}{\sum_{j=1}^{d} \lambda j} \times 100\% \qquad (5)$$

2.5 Aplication of ANN Classification

This research was conducted to maximize the accuracy of Artificial Neural Network (ANN). The Artificial Neural Network model itself is a technology based on the study of the brain and nervous system. The ANN model has a specific architectural format, which is inspired by the biological nervous system. The ANN model simulates the electrical activity of the brain and nervous system. A processing element (also known as a neurode or perceptron) is connected to other processing elements. Neurodes are organized in a layer or vector, with the output of one layer serving as input to the next layer and possibly other layers [14]. The application of classification using ANN in this study was carried out in two ways, the first after the dimension reduction stage using PCA and the other without PCA, namely after the TF-IDF stage. The results of ANN classification using PCA and without using PCA will be compared whether the classification performance on emotional text using ANN classification will improve.

2.6 Method Evaluation

Method evaluation in this study using confusion matrix, namely: True Positive TP and True Negative TN, False Positive FP and False Negative FN. TP and TN represent correct classification while FP and FN represent incorrect classification [15]. Confusion Matrix is a method to determine the level of accuracy by calculating the number of correct and incorrect predictions of the true value and predicted value. When validating the order of the existing set of documents will be randomized to avoid grouping documents that come from certain categories.

Table 2. Confusion Matrix

	Positif	**Negatif**	**Precision**
Positif	*True Positive*	*False Positive*	*Precision Positive*

Negatif	*False Negatif*	*True Negatif*	*Precision Negatif*
Recall	*Recall Positive*	*Recall Negatif*	

3. RESULTS AND DISCUSSION

This research uses Principal Component Analysis (PCA) by reducing the dimensions of the text to improve the performance of text emotion classification using the Artificial Neural Network method. The stages that will be carried out include the data preprocessing stage, text data extraction, reducing the dimensions of the text with Principal Component Analysis (PCA) and classified using Artificial Neural Network. This research will compare the accuracy results of text data that has been done PCA which is then calcified using ANN with text data that is directly classified using ANN without reducing the dimensions of the text using PCA.

3.1 Classification Text Emotion using ANN

This research uses a supervised-based neural network method commonly known as Multi-layer Perceptron (MLP) because the dataset has labels to be classified. Class MLPClassifier implements a multi-layer perceptron (MLP) algorithm that is trained using Backpropagation. In this classification stage, a split dataset is carried out with a ratio of 80% training data, totaling 17167 and 20% testing data, totaling 4292. Classification is done by repeating iteration max 100,200,300,400 and 500 using Python with the scikit-learn module of the neural network classification Multi-Layer Perceptron (MLP) algorithm.

Table 3. ANN Iteration 500

	precision	***recall***	***F1-score***
anger	0.88	0.82	0.85
fear	0.82	0.79	0.80
happy	0.88	0.91	0.89
love	0.78	0.75	0.76
sadness	0.89	0.90	0.89
surprise	0.69	0.74	0.71

The results of neural network classification for text emotion datasets with training and testing data divided by 80% for training data and 20% for testing data can be obtained, namely from each iteration it can be seen that changes in accuracy range between 86%. For the results of testing precision, recall and F1-score in each class, it can be seen that for the sadness class, the highest precision value is 89% at iterations 200, 400 and 500, while for recall the highest value is the same two classes, namely happy and sadness with 91% for happy classes at iterations 400 and 500, while sadness is 91% at iterations 100 and 300. For the f-1 score value, the highest value is obtained in the sadness class, which is 90% at iteration 200.

3.2 Optimation PCA for ANN Classification

The results of PCA with ANN resulted in an accuracy of 0.86 with an increase in classification prediction of 0.997. For the results of precision, recall and f-1 measure can be seen in table 4.

Table 4. Classification Results using PCA and ANN

	precision	***recall***	***F1-score***
anger	0.85	0.82	0.83
fear	0.81	0.79	0.80
happy	0.86	0.92	0.89
love	0.78	0.73	0.75
sadness	0.90	0.89	0.89
surprise	0.76	0.71	0.73

In table 4 the highest recall result is 0.92 in the happy class which exceeds the results before using PCA which is 0.91 and precision in the sadness class is 0.90 while before PCA got the highest value of 0.8.

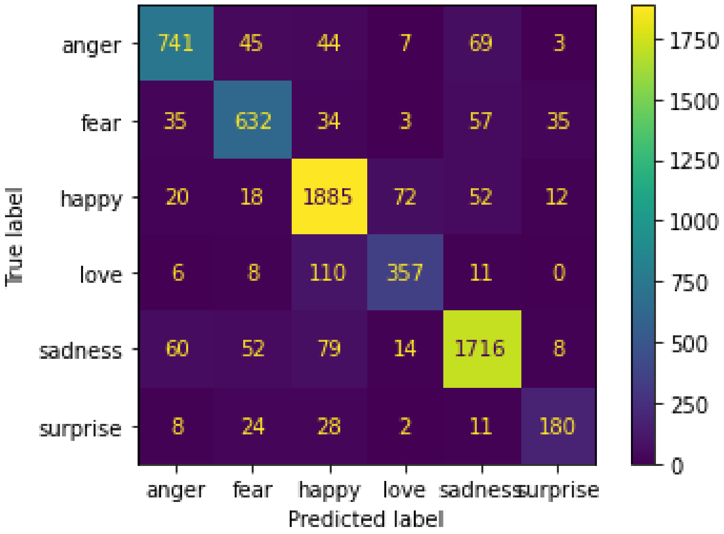

Figure 3. Confusion Matrix PCA-ANN

4. CONCLUSION

Based on the results of research using Principal Component Analysis to Improve Emotion Classification Performance on Text Using Artificial Neural Network as follows:

1. Process research results by comparing ANN classification results using PCA with emotion text classification without using PCA.
2. By using the Principal Component Analysis (PCA) method to reduce the dimensions of the emotional text, the prediction of the class for precision and recall has increased.
3. In the final results of this study by adding PCA which was then classified using ANN showed the highest recall of 0.92 in the happy class which exceeded the results before using PCA which was 0.91 and precision in the sadness class 0.90 while before PCA got the highest value of 0.8.

In this study, the PCA optimization method can indeed improve the performance for classification using ANN but for future research, it can be tried with different datasets to be able to solve different problems, for example with Indonesian language datasets.

REFERENCES

[1] M. S. Saputri, R. Mahendra, and M. Adriani, “Emotion Classification on Indonesian Twitter Dataset,” *Proceedings of the 2018 International Conference on Asian Language Processing, IALP 2018*, pp. 90–95, 2019, doi: 10.1109/IALP.2018.8629262.

[2] Muljono, A. S. Winarsih, and C. Supriyanto, “Evaluation of classification methods for Indonesian text emotion detection,” in *Proceedings - 2016 International Seminar on Application of Technology for Information and Communication, ISEMANTIC 2016*, 2016, pp. 130–133. doi: 10.1109/ISEMANTIC.2016.7873824.

[3] T. Tabashum and S. Chanda, “Sentiment Extraction From Text Using Emotion Tagged Corpus,” in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, 2019, pp. 1–6.

[4] M. A. Tocoglu and A. Alpkocak, “Emotion extraction from turkish text,” *Proceedings - 2014 European Network Intelligence Conference, ENIC 2014*, pp. 130–133, 2014, doi: 10.1109/ENIC.2014.17.

[5] M. A. Tocoglu, O. Ozturkmenoglu, and A. Alpkocak, “Emotion Analysis From Turkish Tweets Using Deep Neural Networks,” *IEEE Access*, vol. 7, pp. 183061–183069, 2019, doi: 10.1109/access.2019.2960113.

[6] J. Herzig, M. Shmueli-Scheuer, and D. Konopnicki, “Emotion detection from text via ensemble classification using word embeddings,” *ICTIR 2017 - Proceedings of the 2017 ACM SIGIR International Conference on the Theory of Information Retrieval*, pp. 269–272, 2017, doi: 10.1145/3121050.3121093.

[7] P. Vora, M. Khara, and K. Kelkar, “Classification of Tweets based on Emotions using Word Embedding and Random Forest Classifiers,” *Int J Comput Appl*, vol. 178, no. 3, pp. 1–7, 2017, doi: 10.5120/ijca2017915773.

[8] E. Batbaatar, M. Li, and K. H. Ryu, “Semantic-Emotion Neural Network for Emotion Recognition From Text,” *IEEE Access*, vol. 7, pp. 111866–111878, 2019, doi: 10.1109/access.2019.2934529.

[9] S. E. Saad and J. Yang, “Twitter Sentiment Analysis Based on Ordinal Regression,” *IEEE Access*, vol. 7, pp. 163677–163685, 2019, doi: 10.1109/ACCESS.2019.2952127.

[10] Institute of Electrical and Electronics Engineers, *International Conference on Information Communication and Embedded Systems : 27-28 February 2014, Chennai, India*.

[11] J. Singh, G. Singh, R. Singh, and P. Singh, “Morphological evaluation and sentiment analysis of Punjabi text using deep learning classification,” *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 5, pp. 508–517, 2021, doi: 10.1016/j.jksuci.2018.04.003.

[12] P. Ahmadi, M. Tabandeh, and I. Gholampour, “Persian text classification based on topic models,” *2016 24th Iranian Conference on Electrical Engineering, ICEE 2016*, pp. 86–91, 2016, doi: 10.1109/IranianCEE.2016.7585495.

[13] S. Narasimhan and S. L. Shah, “Model identification and error covariance matrix estimation from noisy data using PCA,” *IFAC Proceedings Volumes (IFAC-PapersOnline)*, vol. 37, no. 1, pp. 511–516, 2004, doi: 10.1016/s1474-6670(17)38783-9.

[14] H. Cartwright, “Artificial Neural Networks,” *Methods in Molecular Biology*, vol. 1260, pp. 631–645, 2015, doi: 10.1007/978-1-4939-2239_0.

[15] López F.J Ariza, Rodríguez Avi J, and Alba-Fernández M.V “Complete control of an observed confusion matrix,” pp. 1222–1225, 2018.