Int. J. Nonlinear Anal. Appl.
Volume 12, Special Issue, Winter and Spring 2021, 595-604
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.5378

Text mining based sentiment analysis using a novel deep learning approach

Enas Fadhil Abdullah[a], Suad A. Alasadi[b,*], Alyaa Abdulhussein Al-Joda[c]

[a] *Faculty of Education for Girls, University of Kufa, Al- Najaf, Iraq*
[b] *College of Information Technology, University of Babylon, Babil, Iraq*
[c] *Engineering Technical College of Al-Najaf,Al-Furat Al-Awsat Technical University(ATU), Al-Najaf, Iraq*

(Communicated by Madjid Eshaghi Gordji)

Abstract

Leveraging text mining for sentiment analysis, and integrating text mining and deep learning are the main purposes of this paper. The presented study includes three main steps. At the first step, pre-processing such as tokenization, text cleaning, stop word, stemming, and text normalization has been utilized. Secondly, feature from review and tweets using Bag of Words (BOW) method and Term Frequency _Inverse Document Frequency is extracted. Finally, deep learning by dense neural networks is used for classification. This research throws light on understanding the basic concepts of sentiment analysis and then showcases a model which performs deep learning for classification for a movie review and airline_ sentiment data set. The performance measure in terms of precision, recall, F1-measure and accuracy were calculated. Based on the results, the proposed method achieved an accuracy of 95.38% and 93.84% for a movie review and Airline_ sentiment, respectively.

Keywords: Sentiment analysis, Deep learning, DNN, Text mining.

1. Introduction

Data Mining (DM) is one of the most active computer science fields. It's a promising young area. Data mining has been of the highest significance to the information industry and community in recent years since it has been extensively available and the immediate need to turn this data into usable information and expertise [13, 17]. DM focuses on extracting secret information from different data storage facilities, data marts and repositories [16]. Text mining is a data mining technology dedicated to textual data management. It usually refers to the extraction from unstructured information text interesting and non-trivial trends or awareness [12]. Text mining is essentially a multidisciplinary field. This includes knowledge collection, statistics, algebra, computer learning, language processing

*Corresponding author

Email addresses: inasf.alturky@uokufa.edu.iq (Enas Fadhil Abdullah), suad.alsady@uobabylon.edu.iq (Suad A. Alasadi), dt.alyaa@atu.edu.iq (Alyaa Abdulhussein Al-Joda)

Received: March 2021 *Accepted:* June 2021

and the development of natural languages. The most normal form of information storage is the type of textual records. According to a new survey, 80% of business information is contained in text documents [16]. Data mining's general objective is to retrieve data from databases automatically. Text Mining is the same global challenge but applied in particular to unstructured textual data.
The vast use of text data means that information can be derived from it more intensively. However, the process of text mining for unstructured data is considerably more complicated. Data mining uses a range of text processing methodologies, one of which is natural language processing (NLP).
The analyses of linguistic evidence, most usually in the form of texts such as papers or published materials, by means of analytical techniques, are natural language processing (NLP). The objective of natural linguistic processing is essentially to create a description of the text by using linguistic insights to apply structure to the unstructured natural language. The syntactical nature of this structure can be that it captures the grammatical relation between the components –or more semantic– of the text [26]. The implementation of texts for pre-processing includes the Syntactic Analysis, Tokenizer, Semantic, Stop Word Removal, and Stemming module [18].
Text data can also be evaluated at various representation stages. Text results, for example, can easily be processed as a mouthpiece or viewed as a string of words. However, it is important to replicate text knowledge semantically in most implementations so that more meaningful research and analysis can be carried out. For example, it might be possible to discover more fascinating trends than to view text as a word bag by representing text data on named persons, organizations and locations and their relationships [1]. In this paper leveraging text mining techniques for sentiment analysis using sentiment analysis datasets are movie review and airline_sentiement. Using deep learning for classification, collecting texts mining and deep learning in this paper.

2. Overview of Sentiment Analysis

Sentiment analysis is one of the most common areas of text mining. This research will identify documents on the basis of the feelings. Only text attributes can be chosen and used for modeling purposes based upon the dataset and data attributes not all data have been added to the study. The aim is to view the optimistic, negative and neutral word weight. the chosen attributes. Both tweets were scanned and the score was given to the outcome of the sentiment analysis. The results are based on positive and negative terms based on the positive and negative files [5]. The feeling can usually be divided into two categories (positive or negative) or three groups (positive, negative, and neutral classes).
Sentiment analysis can technically be subdivided into four method types [8]: Machine learning is an algorithm that makes the model for supervised classification based on training data. The lexicon method, by using the semantic direction of words or sentences in a text, includes the measurement of the polarity of emotion, it looks for opinions on the basis of rules and classifies them according to the number of positive and negative terms. This method is expressed by classes, then divided into ranks and sentiment analysis has some classification algorithms, such as Naive Bayes, and Support Vector Machine (SVM). Both algorithms are generally known as standard algorithms in terms of modeling paradigm. But lately, with the appearance of a modern way of using DL techniques, a new age of text classification activities can be carried out by sentimental analysis [14]. Text mining can be considered to go beyond access to information for additional users to interpret and ingest information and make decision-making even easier [27].
Opinion mining, sentiment analysis and subjectivity analysis are similar fields with similar objectives for computer techniques to be developed and implemented to process opinion text collections or reviews. Additional analysis objectives include the development of heuristics or methods to identify,

grade or summarize feelings in relation to certain things, events or subjects [24].
Simply, sentiment analysis approaches are based on two kinds of techniques. The first is to base these technologies on lexicon- or corpus-based techniques on decision-making treaties including k-Nearest Neighbors (k-NN), Conditional Random Field (CRF), Hidden Markov (HMM), Single Dimensional Classification (SDC) and Sequential Minimal Optimization (SMO), relating to emotion classification methodologies [22]. Second, the solution to machine learning is in three classes: I monitored; (ii) half-monitored, and (iii) uncontrolled.

3. Deep Learning Model

G.E. Hinton first suggested deep learning in 2006 and is part of the Deep Neural Network process [9]. Neural network is a community of interconnected neurons focused on a mathematical model for the retrieval and transmission of information in recent decades and in particular in the area of systems recognition, modeling and control applications [10]. The neural network has a human brain presence which has many neurons making up a powerful network [2]. Deep learning networks can provide training in both supervised and unattended areas [25].
The designs of the depth learning network involve several different networks including the DNN (Dense Neural Learning), the CNNs and the RNN (Recurrent Neural Networks). In text processing, vectors, word representation estimates, sentence classification, sentence modelling and feature presentation, the neural networks are highly beneficial [28].
The definition, construction, and evaluation of a deep learning model can be divided into several features:

- Layers Number
- Layers Type
- Units (neurons) number in each Layer
- Single layer Activation Functions
- The size of input and output

3.1. Deep Learning Layers

For deep learning models there are several kinds of layers. CNNs classify images or recognize objects, with recurring layers employed by RNns common to sorting natural languages and understanding of text. CNNs classify images or detect objects. We're going to use layers Dense and Dropout. The most prominent and typical form of layer is a thick layer - it is just a normal neural network layer that connects each neuron to the neurons of the last and next layers. Each dense layer is activated using the input and weight of the synapses to determine the output of its neurons.
Dropout layers are only layers of regularization which alter some inputs randomly to 0. This reduces the likelihood of neural network overfitting.

3.2. Activation Functions

There are also several kinds of layer activation functions. All connects the input and weights of the neuron to each other and distinguishes the network. Very common features are the ReLU, the Sigmoid feature and the Linear feature. A few different roles are going to be mixed.

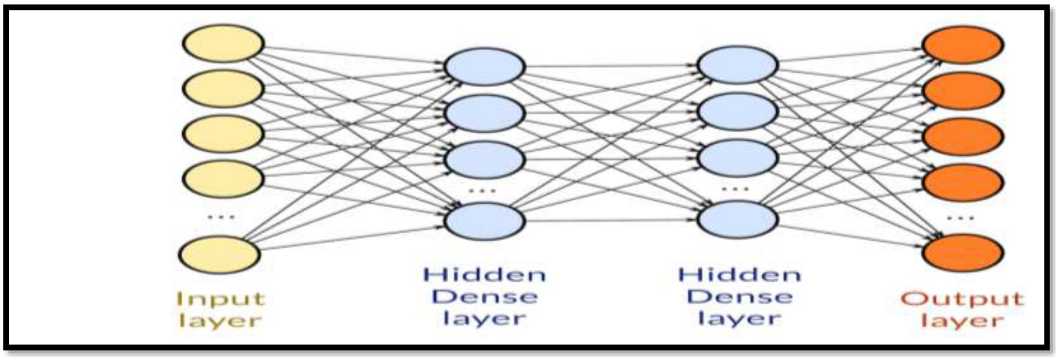

Figure 1: Deep learning layers

3.3. Input and Output layers

Models have an input layer and a display layer as well as opaque layers. Figure 1 shows input and output layers:

In the input layer, the number of neurons is the same as in our results. E.g., the train df and the test df data frames have 67 features - so there would be 67 neurons in our input layer. This is where our details entered.
The number of neurons depends on your objective for the output layer. And we estimate the price – a single value, only one neuron would be used. Models of classification must have output neuron class-number. You have to choose to compile the model:

- **The Loss Function** - The closest the model is to the target, the lower the error. To keep track of success, diverse problems involve various loss functions.
- **The Optimizer** - The algorithm that optimizes the output of the loss function.
- **Metrics** - Metrics for model assessment. For instance, it would make sense for a Mean Squared error loss function to use the Mean Absolute Error to measure and to use many other metrics.

4. Performance Measures

The purpose of a classification assignment is to classify the objects most applicable to a particular customer in the sense of the product recommendation. The two most well recognized classification measures are precision and recall; they also serve to measure the consistency of the knowledge recovery tasks as a whole [19]. Precision and reminder are the most common measurements for the recall of details, but the measurements of accuracy allow users to predict to what the items they prefer are calculated as fractions of $hitsu$, the number of correctly suggested user items (u). The precision metric (P) covers the cumulative number of impacts of the recommended pieces ($|recsetu|$) as shown in Equation (4.1) [20]:

$$Precision_u = \frac{|hitsu|}{|recsetu|} \tag{4.1}$$

The $Recall(R)$ calculates the proportion of hits with the potential optimum number of hits due to the test scale as clarified in Equation ($|testsetu|$) (4.2).

Table 1: Confusion matrix

Actual class	Predict Class	
	+	-
+	TP	FN
-	FP	TN

airline_se	text
neutral	@VirginAmerica What @dhepburn said.
positive	@VirginAmerica plus you've added commercials to the experience... tacky.
neutral	@VirginAmerica I didn't today... Must mean I need to take another trip!
negative	@VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces & they have little recourse
negative	@VirginAmerica and it's a really big bad thing about it

Figure 2: Sample of airline dataset

$$Recall_u = \frac{|hitsu|}{|testsetu|} \tag{4.2}$$

Harmonic mean of precision and recall as shown in Equation (4.3) [3, 4].

$$F\_score = \frac{2(Precision.Recall)}{Precision + Recall} \tag{4.3}$$

Classification accuracy: it is the ratio of properly categorized instances to a total number of instances [11, 23, 6, 15].

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{4.4}$$

Table 1 interpreted the symbols $(TP, FN, FP, and\ TN)$ as follow [29, 7]:

1. True Positive (TP): If the case is positive, it is considered positive.
2. False Negative (FN): If the case is positive, but it is considered as negative.
3. True Negative (TN): If the case is negative and it is considered as negative.
4. False Positive (FP): If the case is negative, but it is considered as positive.

5. Description of Data Set

The first is Moview reviews (Mr), which is a compilation of 25,000 films reviewed from the IMDB archive, marked by feelings (positive/negative). This is a framework implementation on two sets. The name of the two subdirectories in that folder, 'pos' and 'neg,' refers to our auto-classification classification (sentence) of component files [21]. Secondly, airline_ sentiment dataset with 14641 record that each record represented tweet (text) and three classes are negative (9178 record), neutral (3099 record) and positive (2363 record). Twitter is a framework for microblogging which enables you to send and receive short messages called tweets. Tweets with a length of up to 140 characters and links to related websites and services. Figure 2 shows sample of data set.

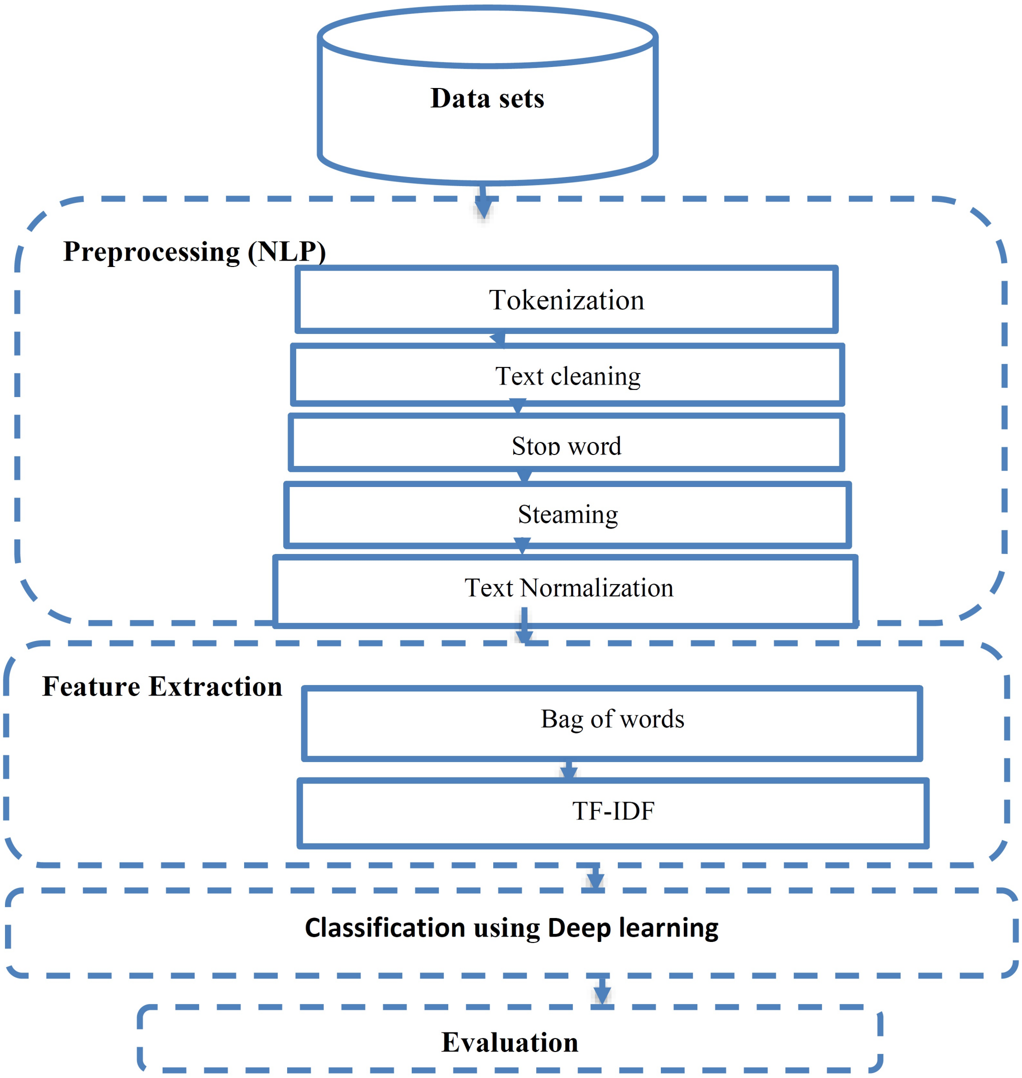

Figure 3: the block diagram of the proposed method

6. Proposed System

The proposed system includes three main phases as shown in Figure 3.

6.1. Preprocessing

The preprocessing phase has been applied on two data sets through five stages:

A. Tokenization can be defined as the procedure of dividing the text into these smaller pieces, to clarify more converting texts into words within a matrix by search line by line for space and put the words in a single matrix after converting texts into words. Tokenization is very important in this respect of review analysis.

B. Text cleaning: is the process of cleaning texts from symbols like $(@, \&, numbers, andetc.)$ after a tokenization process.

C. Stop word: It is too frequent to delete terms from the corpus because they usually are generic words like "a," "an," "the" or "of" that aren't discriminatory throughout a document. Those words are called stop words as well.

D. Stemming: is technique traces the word to its origin according to a technique online snow ball for the purpose of calculating features such as word computer and computing the origin word is compute.

E. Text Normalization: is process transformation text with capital letter to small letter such as "COMPUTER "to "computer", "THE" to "the", and etc.

6.2. Feature Extraction

This phase has been applied on data sets through two stages:

A. Bag of Word: BoW is a representation of text that describes the occurrence of words within a document such as text "hello", the term is repeated five times as represented term frequency is important step for next stage.

B. Term Frequency –Inverse Document Frequency: TF-IDF is calculate as show in Equation (6.1):

$$W(t,d,D) = tf_{t,d} * \log \frac{N}{df} \tag{6.1}$$

Where:

- $W(t,d,D)$: High term frequency weight (in the particular text) and low document frequency of the term in the whole document range.
- $tf_{t,d}$: is the quantity of times that term t occurs in document d.
- N: document total number.
- df: documents containing the term number.

The output of step is weight of repeated words in document, for example, there are two documents, and to know the weight of the word "hello", as the word in the first document is repeated three times, and in the second document it is repeated once, run of Equation TF-IDF as follows:

$$Weightofhello = 3 * \log \frac{2}{4} \tag{6.2}$$

6.3. Classification by Deep learning

Deep learning by neural network or called Dense Neural Network (DNN), this stage can be summarized as follows:

A. Training process: sequential model (linear stack of layers) summarization of stage as follow:

- Embedding layer (input layer): mapping word values (The values we got from the stage of features extracting by TF-IDF) to vector of real number with parameter (5000 layers, output= 50 number of vectors).

Table 2: performance measures

Data set	Precision	Recall	F1	Accuracy
Movie review	97.5	94.5	95.98	95.38
Airline	94.37	93.22	93.8	93.84

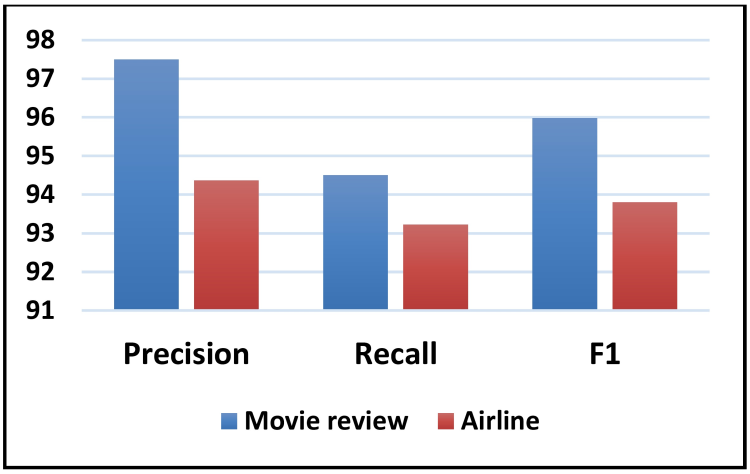

Figure 4: Precision, recall, and F1 measure

- Flatten (hidden layer): vectors 2D to shape 1D.
- Dense (output layer): calculated activation function.
- Outline of the learning (training) algorithm: with parameters batch size is 100, k-fold is 10, and epoch is 10.

B. Update weight using adam optimizer algorithm based on training data, and then finally step of proposed system is calculated precision, recall, F1, and accuracy.

7. Results and Discussions

The methodology of this paper based on sentiment analysis by text mining, classification using Dense Neural Network (DNN), and calculated performance measures (precision, recall, F1, and accuracy) for two data sets movie review and airline_ sentiment. Tables 2 show details performance measures of two datasets.

Figure 4 and Figure 5 represented precision, recall, and F1 and accuracy of classification for two datasets, respectively.

8. Conclusion

Text mining based on deep learning can be useful in the reviews and tweets classification task. In this paper, the use of sentiment analysis based on text mining has been investigated. The results demonstrated the outperformance of the proposed method especially for review classification.

References

[1] C.C. Aggarwal and C.X. Zhai, *Mining Text Data*, Springer-Verlag New York, 2012.

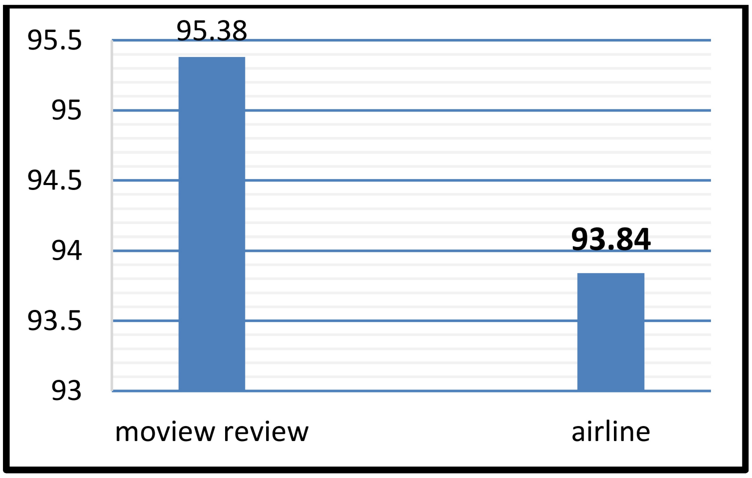

Figure 5: Accuracy of classification

[2] N. Al-A'araji, E. Al-Shamery and A. Abdulhussein, *ARNN for enhancing drift detection of data stream based on modified page hinckley model*, J. Engin. Appl. Sci. 13(10) (2018) 8281-8291.

[3] E. Al-Turkey, G. Al-Sultany, and H. Almamory, *Enhancing content-based recommender system by using enriched user profile*, J. Engin. Appl. Sci. 12(10) (2017) 8858–8863.

[4] B. Arkok and A.M. Zeki, *Classification of Qur'anic topics based on imbalanced classification*, Indonesian J. Elect. Engin. Comput. Sci. 22(2) (2021) 70–79.

[5] M.A. Burhanuddin, R. Ismail, N. Izzaimah, A.A. Mohammed and N. Zainol, *Analysis of mobile service providers performance using naive bayes data mining technique*, Int. J. Elect. Comput. Engin.8(6) (2018) 5153–5161.

[6] R. Burke, M.P. Omahony and N.J. Hurley, *Robust Collaborative Recommendation, Recommender Systems Handbook*, In: F. Ricci, L. Rokach, B. Shapira, and P. Kantor (eds) Recommender Systems Handbook, Springer, Boston, 2015.

[7] H. Chehili, S.E. Aliouane, A. Bendahmane and M.A. Hamidechi, *DeepEnz: Prediction of enzyme classification by deep learning*, Indonesian J. Elect. Engin. Comput. Sci. 22(2) (2021) 500–507.

[8] A. Collomb, C. Costea, D. Joyeux, O. Hasan and L. Brunie, *A study and comparison of sentiment analysis methods for reputation evaluation*, Rapport de recherche RR-LIRIS2014-002, (2014).

[9] M. Day and C. Lee, *Deep learning for financial sentiment analysis on finance news providers*, IEEE/ACM Int. Conf. Adv. Social Networks Anal. Min. (2016) 1127–1134.

[10] M.S. Gaya, M.U. Zango, L.A. Yusuf, M. Mustapha, B. Muhammad, A. Sani, A. Tijjani, N.A. Wahab and M.T.M. Khairi, *Estimation of turbidity in water treatment plant using hammerstein-wiener and neural network technique*, Indonesian J. Elect. Engin. Comput. Sci. 5(3) (2017) 666–672.

[11] D. Gupta, A. Malviya and S. Singh, *Performance analysis of classification tree learning algorithms*, Int. J. Comput. Appl. 55 (2012).

[12] J. Han, M. Kamber and J. Pei, *Data mining: Concepts and Techniques*, San Fransisco: Morgan Kaufmann, 2006.

[13] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, 2ed., Elsevier, 2011.

[14] B. Heredia, T.M. Khoshgoftaar, J. Prusa and M. Crawford, *Cross-domain sentiment analysis: an empirical investigation*, IEEE 17th Int. Conf. Inf. Reuse Integ. (2016) 160–165.

[15] A.F. Hidayatullah, *The influence of stemming on Indonesian Tweet sentiment analysis*, Proc. Int. Conf. Elect. Engin. Comput. Sci. Inf. Palembang, Indonesia, 2(1) (2015) 127–132.

[16] S. Hussain, N.A. Dahan, F.M. Ba-Alwib and N. Ribata, *Educational data mining and analysis of students' academic performance using WEKA*, Indonesian J. Elect. Engin. Comput. Sci. 9(2) (2018) 447–459.

[17] A.D. Indriyanti, D.R. Prehanto, and T.Z. Vitadiar, *K-means method for clustering learning classes*, Indonesian J. Elect. Engin. Comput. Sci. 22(2) 2021 227–233.

[18] A. Jain, G. Kulkarni and V. Shah, *Natural language processing*, IJCSE 6(1) (2018).

[19] D. Jannach, M. Zanker, A. Felfernig and G. Friedrich, *Recommender Systems: An Introduction*, Cambridge University Press, 2010.

[20] M.H. Jopri, M.R.A.b. Ghani, A.R. Abdullah, T. Sutikno, M. Manap and J. Too, *Naïve bayes and linear discriminate analysis based diagnostic analytic of harmonic source identification*, Indonesian J. Elect. Engin. Comput. Sci. 20(3) (2020) 1626–1633.

[21] Keras.io, *IMDB movie review sentiment classification dataset*, Aug 2021.
[22] J. Singh, G. Singh and R. Singh, *A review of sentiment analysis techniques for opinionated web text*, CSI Trans. ICT 4(2) (2016) 241–247.
[23] S. Srivastava, *Weka: a tool for data preprocessing, classification, ensemble, clustering and association rule mining*, Int. J. Comput. Appl. 88 (2014).
[24] Q. Tul Ain, M. Ali, A. Riaz, A. Noureen, M. Kamran, B. Hayat, and A. Rehman, *Sentiment analysis using deep learning techniques: a review*, Int. J. Adv. Comput. Sci. Appl. 8(6) (2017).
[25] P. Vateekul and T. Koomsubha, *A study of sentiment analysis using deep learning techniques on Thai Twitter data*, 13th Int. Joint Conf. Comput. Sci. Software Engin. (2016) 1–6.
[26] K. Verspoor and K.B. Cohen, *Natural Language Processing*, Encyclopedia of Systems Biology, Springer, New York, 2013.
[27] S. Wakade, C. Shekar, K.J. Liszka, and C.-C. Chan, *Text mining for sentiment analysis of Twitter data*, Proc. Int. Conf. Inf. Knowledge Engin. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.
[28] Y. Zhang, M.J. Er, N. Wang, M. Pratama and R. Venkatesan, *Sentiment classification using comprehensive attention recurrent models*, Int. Joint Conf. Neural Networks (IJCNN) (2016) 1562–1569.
[29] L. Zhou and H. Wang, *Loan default prediction on large imbalanced data using random forests*, TELKOMNIKA Indonesian J. Elect. Engin. 10(6) (2012) 1519–1525.