Impact of Feature engineering for Improved Sentiment Analysis in Amazon Product Reviews Using K-Nearest Neighbor

Nitami Lestari Putri[1*], *Budi* Warsito[2], and *Bayu* Surarso[3]

[1]Magister of Information System, School of Postgraduate Studies, Diponegoro University, Semarang 50275, Indonesia
[2]Department of Statistics, Faculty of Science and Mathematics, Diponegoro University, Semarang 50275, Indonesia
[3]Department of Mathematics, Faculty of Science and Mathematics, Diponegoro University, Semarang 50275, Indonesia

Abstract. Online reviews are an important factor that encourages consumers to make purchases through e-commerce. However, it is challenging to objectively assess the sentiments expressed by actual consumers due to the prevalence of fraudulent reviews. This study focuses on sentiment analysis and seeks to uncover the best feature combinations based on review and reviewer centric approach. The results of the study show that the combination of feature Rating, VerifiedPurchase, ReviewLengths, and (CV+TF-IDF) = 1400 words with the application of KNN classification provides the best accuracy rate of 83%. The results of this study can assist consumers in making purchasing decisions and seller in increasing the value of their products and services based on the feedback provided by customers.

1 Introduction

Electronic commerce, often known as e-commerce, has a significant impact on the success of failure of businesses today. E-commerce platforms have many benefits, including the ability to process more orders quickly, respond to customer or marketplace needs, and offer a variety of payment options [1]. The Amazon, which Jeffrey P. Bezos Created in 1994, is one of the most well-known online retailers in the world today. The Amazon River, the biggest river in the world, is where the term Amazon originates [2]. Other retail products like electronics and cosmetics have been added to Amazon's product line. Amazon has achieved success and is currently ranked highly among other well-known internet retailers worldwide [3].

Customers in online transaction often read product or service reviews posted online by prior users to obtain granular product suggestions and make purchase choices [4]. Online reviews are helpful to prospective buyers and sellers as well. Vendors might create their extra marketing plans using the feedback given by customers [5]. However, several issues also

[*] Corresponding author: nitamiputri38@gmail.com

developed. Such as there are fake reviews that can cause financial losses for e-commerce businesses and can mislead consumers to make the wrong decision [4]. Fake reviews can also make it difficult to know the true sentiment on reviews shared by real buyers. Based on the existing problems, an analysis of the review text provided by e-commerce users using text mining techniques is needed. One of the popular text mining techniques is the sentiment analysis. Sentiment analysis is a technique used in online retail to determine how a customer feels about certain item or company [6].

There are two approaches that can be used in extracting sentiment from online reviews. The first is a machine learning approach to automatically classify reviews that need training data. The second is a lexicon-based approach that requires a dictionary that contains a predefined lexicon and information about the polarity of words related to sentiment [7, 8]. Most previous studies have applied both approaches to sentiment analysis, with various methods used in each approach. This study focuses on a lexicon-based approach for determining each word's polarity value in text reviews to identify the right sentiment for each review. Previous studies have applied various methods for lexicon-based sentiment analysis. The Textblob library is used in this study instead of other available methods because it is a faster and smaller library and enables accurate text sentiment analysis [9].

Feature selection in text mining can be a consideration to be able to perform text analysis accurately. In several of the earlier research, factors that are labelled as review centric and reviewer centric were considered to identify fraudulent review dataset. There are, however, still few studies that take review-centric and reviewer-centric aspect into account to determine the appropriate sentiment. To effectively determine sentiments in reviews, this study suggests using characteristics that may be divided into review and reviewer-centric aspects. This study evaluated several feature combinations to determine how they influenced the ability to discern review sentiment. The Count Vectorizer (CV) and TF-IDF were combined as a feature extraction technique and the most common word frequencies were chosen to make comparisons. To determine which model is the most accurate, the results of the feature combination analysis are compared using the K-Nearest Neighbor (KNN) approach. The KNN algorithm was chosen because more successful with huge volumes of data and it can be used as a classification for various domains [10]. To assist companies and customers in making wise judgments, this study aims to ascertain the impact of several factors on determining the appropriate sentiment in Amazon product reviews.

2 Literature Review

Previous studies on feature combinations based on review and reviewer-centric features, particularly for the identification of fraudulent reviews in e-commerce, have been conducted. There are still few studies in the field of sentiment analysis that classify features based on review-centric and reviewer-centric aspects to effectively identify sentiment in reviews. To identify fraudulent reviews, some earlier studies including work by Birim et al., used review and reviewer-centric feature characteristics. The goal is to choose the optimum feature combination to reliably detect fraudulent reviews. The score of sentiment, topic distribution based on LDA, cluster distribution based on AHC, and sparse matrix using CountVectorizer have been chosen as the review-centric characteristics. To detect fraudulent reviews, to reviewer-centric features ReviewLengths and VerifiedPurchase are also utilised. [6]. Another study conducted by Daiv et al., considered a combination of rating features, verified purchase, and review length to identify fake review based on review and reviewer-centric criteria [5]. Martinez et al., use a review-centric to identify characteristics of fraudulent and honest reviews in the context of the hotel industry and consider the polarity of the review's sentiments. Considering both positive and negative reviews as well as fraudulent and honest reviews, were chosen [11].

Sentiment analysis may be conducted using a variety of machine learning methods. By contrasting 5 machine learning methods, Akter et al., were able to identify the sentiments expressed in product reviews on “Daraz”, a Bangladeshi e-commerce website. The result show that use of the KNN algorithm along with the TF-IDF vectorization produces the best accuracy of the other algorithms [12]. Another study conducted by Jazuli et al., used KNN algorithm to automated data labelling and execute sentiment analysis. Therefore, tertiary schools can employ KNN to categorize Twitter user evaluations and evaluate and appraise higher education services [13]. Qorib et al., performed a sentiment analysis to examine COVID-19 vaccine hesitancy by contrasting a combination 3 sentiment methods, 5 machine learning algorithms, and 3 combinations of vectorization methods. The result show that the Textblob+ TF-IDF +LinearSVC combination has the best performance compared to the other models [14].

3 Theoretical Backgrounds

3.1 Textblob Library

Textblob is popular lexicon-based sentiment analysis approach for python that offers streamlined text processing. The simple Textblob API simplifies several common text processing and NLP tasks, such as language translation, tokenization, classification, sentiment analysis, and others [9]. In Textblob, to calculate the polarity value using equation (1), where a positive class is denoted by a polarity score of +1, and a negative class by a score of -1. Textblob applies the concept of a simple arithmetic mean to calculate polarity ($\bar{P}$) where the number of polarities of words ($\sum_{i=1}^{n} X_i$) is divided into the number of times, the value (n) is the word that appears in the lexical data [15].

$$\bar{P}(x) = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1}$$

3.2 Review Centric and Review Centric Feature

Review centric and reviewer centric approaches are commonly used to derived relevant features from fake reviews. However, it is possible that this approach can be used in other text analysis such as sentiment analysis. According to Martinez et al., The qualities that a bag of words offers form the base of review-centric techniques. As a result, the effectiveness of the classifier in identifying reviews depends on the choice of pertinent phrases. Meanwhile, the reviewer-centric concentrate on the features gleaned from the reviewer’s profile traits and behaviour patterns [11]. The review centric approach aims to capture textual and characteristic information from each review and the reviewer centric aims to find the behavioural characteristics of the reviewers from reviews they write [6].

3.3 Feature Extraction

3.3.1 CountVectorizer

Class features and text feature extraction techniques are numerically calculated using CountVectorizer (CV) [16]. The length of vector created from the CV is equal to the length of the dictionary word, which causes the vector to be aligned [17]. Table 1 is an example of a CV matrix which produces a 3x5 matrix containing sentences such as “magazine is good”, ”magazine is average”, and ”magazine is nice”. Based on the sentence, there are 3 documents and 5 different features such as magazine, is, good, average, and nice. Each 1 in the matrix denotes the existence of a feature, and each 0 denotes the lack of a feature from a certain text.

Table 1. CountVectorizer matrix.

	Feature1	**Feature2**	**Feature3**	**Feature4**	**Feature5**
Doc1	1	1	1	0	0
Doc2	1	1	0	1	0
Doc3	1	1	0	0	1

3.3.2 TF-IDF

The Term Frequency-Inverse Document Frequency (TF-IDF) is composed of two measures : Term Frequency (TF) and Inverse Document Frequency (IDF). The TF-IDF can assess the significance of the words for the dataset's documents. The TF-IDF is shown in equation (2), where the frequency of term is represented by tf and determined from count (c), the term (t) in the document (d) and is shown as $(tf(t,d) = C_{td})$. The frequency of occurance of term is converted to a binary feature, where 1 signifying the word's existence and 0 signifying its absence from the document [18].

$$tfidf = tf(t,d) \times idf(t,d) \qquad (2)$$

Equation (3) displays the IDF for the word w in a text document (t). T denotes the corpus's overall document count, while $df(t)$ denotes the total numer of the documents in term (t) [18].

$$idf(t,d) = 1 + \log \frac{T}{(1+df(t))} \qquad (3)$$

3.4 K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a classification technique that doesn't make any assumptions about the distribution of the underlying data, instead relying on feature similarity. KNN classifies data points according to the majority vote of their neighbours, with data points going into the category that K closest neighbours agree on the most. KNN determines decision limits for more than two classes when classifying new points. A hyperparameter that adjusts this classifier's complexity is the number of K neighbours [19].

Cosine similarity is a popular metric in text analysis. Cosine similarity has the benefit of not being dependent on the size of the feature vector when text is processed as a set of words [20]. Equation (4), in which the θ is angle formed by vectors A and B, is representation of cosine similarity. A and B stand for the feature vectors, while A_i^2 and B_i^2 respectively denote their lengths [21].

$$Similarity = \cos\theta = \frac{\sum_{i=1}^{p} A_i \times B_i}{\sqrt{\sum_{i=1}^{p} A_i^2} \times \sqrt{\sum_{i=1}^{p} B_i^2}} \qquad (4)$$

4 Research Method

The following part provides a description of the research technique employed in this study based on the research framework illustrated in Figure 2:

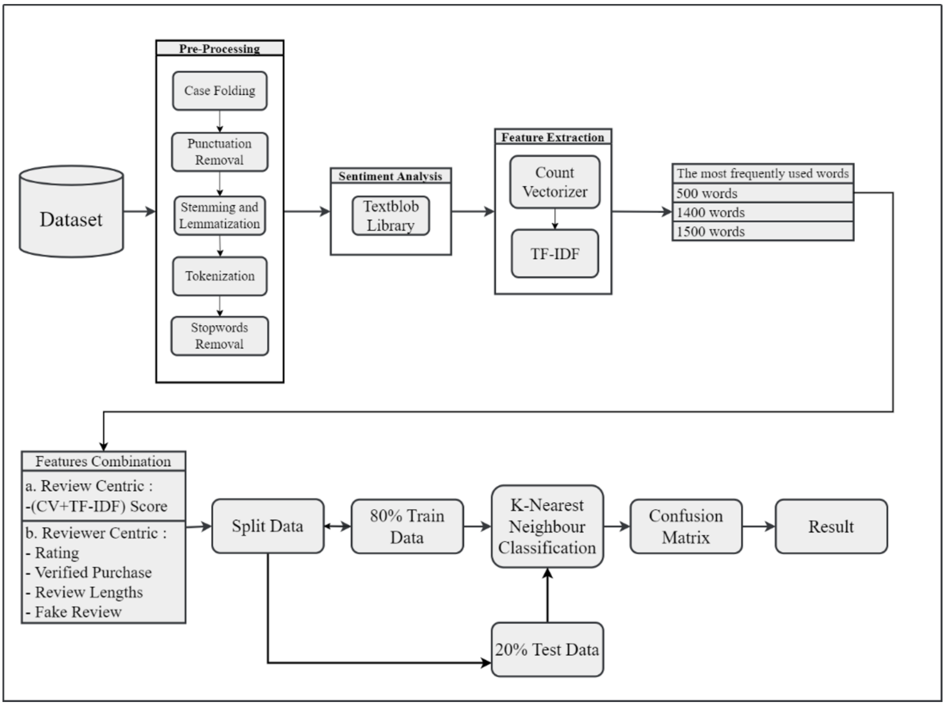

Fig.2. Research Framework

4.1 Dataset Description

The fake product reviews dataset from Amazon was used for this study. In this dataset, consists of 21,000 reviews. Each review has features such as document ID, label indicating fake and real reviews, a Rating, a VerifiedPurchase, the ProductCategory, the ProductID, the ProductTitle, the ReviewTitle, and the ReviewText. This dataset may be found at the following URL: https://github/ aayush210789/Deception-Detection-on-Amazon-reviews-dataset it was collected from Aayush Saxena's Github repository.

4.2 Pre-Processing

The data obtained is then pre-processed, which aims to reconstruct the text into a form that is easier to digest for machine learning algorithms. In this study, there are five phases of pre-processing. These steps include case folding, punctuation removal, stemming and lemmatization, tokenization, and stopword removal.

4.2.1 Case Folding

The case folding technique involves transforming all of the letters in the dataset to uppercase or all lowercase in order to simplify data collecting processing and decrease memory use [22].

4.2.2 Punctuation Removal

All punctuation from the sentence is removed during this process. The process of text analysis will not be impacted by any of these punctuation marks, which have no discernible meaning [6].

4.2.3 Stemming and Lemmatization

Combining the two vocabulary normalization techniques of stemming and lemmatization improved the recommended model's performance in this study [14]. The word is transformed into its basic form throughout the stemming process. Prefixes, suffixes, and confixes are intended to be removed during the stemming process. Then proceed with the lemmatization stage. Lemmatization is the process of condensing properly inflected words and making sure that these fundamental terms are in the right dictionary language. Lemmatization is based on original tense and vocabulary [14].

4.2.4 Tokenization

The technique of tokenizing involves separating a text into tokens, or distinct words. As an example of the following sentence "I'm looking for a cheap wardrobe" is changed to ["I'm", "looking", "for", "a", "cheap", "wardrobe"] after the tokenization process is carried out [14].

4.2.5 Stopwords Removal

Stopwords are very common words with a primary structural function that are used repeatedly in text. These words serve merely as syntax, bearing minimal significance and having little bearing on textual analysis. They do not identify the topic. To make it easier to find significant terms, you can delete these words from the text [23].

4.3 Lexicon-Based Approach Using Textblob Library

For sentiment analysis, this study uses a lexicon-based methodology. One lexicon-based technique for sentiment analysis is the Textblob library. The polarity value of each word in text reviews is determined in this study using the Textblob library. The positive, negative, and neutral sentiment labels are then established after determining the polarity value.

4.4 Feature Extraction

This study uses CountVectorizer (CV) to calculate the frequency of words that appear and TF-IDF to calculate the weight value of each word in the review. In the CV process, the 500, 1400, and 1500 words that appear most often in the review are selected to form a vector. After the text data has been vectorized and the frequency of each word has been determined, the weight value of each word is determined using TF-IDF.

4.5 Feature Combination

The feature combination and product reviews sentiments are determined using review centric and reviewer centric techniques. The following characteristics, which are divided into review and reviewer-centric categories to detect sentiments in the Amazon product review dataset, are described below:

4.5.1 Review Centric

In this study, the results of feature extraction using CV and TF-IDF are categorized as review centric feature. This is because the results of feature extraction show the frequency or importance of a word in a review and captured textual and characteristic information from each review.

4.5.2 Reviewer Centric

In this study, rating, verified purchase, review lengths, and fake review are categorized as reviewer centric. The following is a description of the four features as follows [5, 6]:

a. Rating
The rating feature contains product rating values given by e-commerce users consisting of ratings 1 to 5. This feature represents user satisfaction or dissatisfaction with the products they buy. This feature can be used to validate that the reviews written and the ratings provided by users are directed in one direction only and do not contradict.

b. Verified Purchase
The verified purchase feature in the Amazon product review dataset is a feature that verifies reviewers who buy products on Amazon have purchased the product and did not receive a product with a large discount. This feature can be used to find out which buyers bought the product and used it and know the sentiment from the reviews they wrote.

c. Review Lengths
Machine learning models may be trained using the review lengths feature. In sentiment analysis research, the length of the review can be used to determine the characteristics of the reviewer from the length and shortness of the reviews written.

d. Fake Review
The fake review feature on the Amazon product review dataset indicates a review based on a fake/real review. The fake review feature can be used to provide information that fake/real reviews have good/bad sentiments in product reviews.

In this study, features chosen based on review and reviewer centric criteria are employed. A variety of models are created based on a mix of attributes and the frequency at which a word appears to make comparisons. Three models in total are offered for comparison. The three suggested models are shown below and may be seen in Table 2.

Table 2. Proposed models.

Feature Matrices	The Frequency of Word
(CV+TF-IDF) Score	500
Rating, VerifiedPurchase, ReviewLengths, (CV+TF-IDF) Score	1400
Rating, VerifiedPurchase, ReviewLengths, FakeReview, (CV+TF-IDF) Score	1500

4.6 K-Nearest Neighbor Classification

The data is divided into 20% of test data and 80% of train data. After the data is divided, the KNN algorithm is used for the classification with 5-fold cross validation. In this study, the similarity between instances is calculated using the cosine similarity between feature vectors. To test and determine the best accuracy value, K is set from k=1 to k=10.

5 Result and Discussion

The Jupyter-Lab environment employs Python 3.9.2 for all data pre-processing, sentiment analysis, and KNN classification. Following the pre-processing of the text data, the following text analysis procedures were finished and described:

5.1 Sentiment Analysis

When the polarity score has been established using a Textblob, a sentiment label is provided. The thresholds for positive, negative, and neutral sentiment labels are determined when Positive sentiment is defined as having a polarity score is >0, negative sentiments as having a polarity score is <0, and neutral sentiment as having a polarity score is =0. The polarity scores and sentiment labels for Amazon product reviews, shown in Table 3, is detailed below:

Table 3. Polarity score and sentiment label.

Reviews	Polarity Score	Sentiment Label
['camera', '2', 'week', 'pictur', 'great', 'easi', 'use', 'love']	0.65	Positive
['unfortun', 'didnt', 'work', 'made', 'sick', 'throw', 'onc', 'two', 'time', 'tri']	-0.71	Negative
['wa', 'look', 'inexpen', 'desk', 'calcolatur', 'work', 'doe', 'everyth', 'need', 'onli', 'issu', 'tilt', 'slightli', 'one', 'side', 'hit', 'ani', 'key', 'rock', 'littl', 'bit', 'big', 'deal']	0	Neutral

Table 4 shows the total number of sentiments. A total of 21000 reviews, 17,275 are positive reviews, 2,154 are negative reviews, and 1,571 are neutral reviews.

Table 4. The total of sentiment labels.

Sentiment Labels	Total
Positive	17,275
Negative	2,154
Neutral	1,571
Total	21,000

5.2 Feature Extraction

In this study, CV and TF-IDF were used for feature extraction. Word frequency distributions are built for each review in the CV sparse matrix. Each review's sparse matrix for each suggested model is built using the top 500, 1400, and 1500 words' frequency distributions. The weight of each word is then determined using TF-IDF. Using CV, a premade sparse matrix made from text data in the form of a vector is transformed.

5.3 K-Nearest Neighbor Classification

5-fold cross validation is used during the classification process to assess how well each suggested model is working. Testing is then performed by adjusting the KNN's parameter from k=1 to k=10. Table 5 displays the KNN's accuracy results.

Table 5. Results of each proposed model's performance.

K=	**(CV+TF-IDF)= 500 words**	**Rating, VerifiedPurchase, ReviewLengths, (CV+TF-IDF)= 1400 words**	**Rating, VerifiedPurchase, ReviewLengths, FakeReview, (CV+TF-IDF)= 1500 words**
1	0.739	0.762	0.766
2	0.660	0.710	0.705
3	0.797	0.808	0.806
4	0.789	0.804	0.795
5	0.808	0.821	0.817
6	0.813	0.822	0.816
7	0.820	0.826	0.821
8	0.821	0.827	0.825
9	**0.824**	0.828	0.825
10	0.823	**0.830**	**0.826**

Table 5 shows that the combination of features from rating, verified purchase, review lengths, and (CV+TF-IDF) = 1400 words obtained the best score at k = 10 with an accuracy value of 83%. This proves that the combination of these features is effective in classifying review sentiment polarity accurately. Additionally, with a different model, (CV+TF-IDF) = 500 words obtains an accuracy of 82,4% at k = 9, and a combination of features from rating, verified purchase, review lengths, and fake reviews, (CV+TF-IDF) = 1500 words obtains an accuracy of 82,6% at k = 10. The accuracy obtained in this study shows that the combination of features used can capture important aspects of customer sentiment expressed in Amazon product reviews.

6 Conclusion

This study proposes several feature combination models that aim to compare which feature combinations can accurately identify the sentiments stated in the review. The results show that the combination of features such as Rating, VerifiedPurchase, ReviewLengths, and (CV+TF-IDF)=1400 words with the application of KNN classification provides a significant accuracy rate of 83% at k=10. This indicates how the proposed model is successful at correctly classifying review sentiment on Amazon product reviews. These results demonstrate the potential to utilize selected features and KNN classification in sentiment analysis tasks and can assist consumers and sellers in understanding customer sentiment. By understanding the sentiments expressed by customers, both consumers and sellers can make the right decisions regarding purchases and increase the value of products sold in e-commerce.

References

1. V.C. Rodrigues, L.M. Policarpo, D.E.D. Silveira, R.D.R.Righi, C.A.D. Costa, J.L.V. Barbosa, R.S. Antunes, ELERAP, **56** (2022)

2. J.R. Wells, G. Danskin, G. Ellsworth, Harvard Bussiness School : Amazon, 191-199 (2018)
3. C. Tangmanee and C. Jongtavornvitaya, IJEBEG, **14**, 225-245 (2022)
4. S.N. Alsubari, S.N. Desmukh, M.H. Aladhaileh, F.W. Alsaade, T.H.H. Aldhyani, ABB, **2021**, 1-11 (2021)
5. K. Daiv, M. Lachake, P. Jagtap, S. Dhariwal, P.V. Gutte, IRJET, **7**, 2107-2112 (2020)
6. S.O. Birim, I. Kazancoglu, S.K. Mangla, A. Kahraman, S. Kumar, Y. Kazancoglu, J. Bus. Res, **149**, 884-900 (2022)
7. V. Bonta, N. Kumaresh, N. Janardhan, AJCST, **8**, 1-6 (2019)
8. J. Kim and C. Lim, J.AEI, **49** (2021)
9. W. Aljeedani, F. Rustam, M.W. Mkaouer, A. Ghallab, V. Rupapara, P.B. Washington, E. Lee, I. Ashraf, J. Kno. Sys, **255** (2022)
10. S. Atmadja, G. Gumilar, AJRI, **1**, 14-19 (2019)
11. M.R. Martinez-Torres, S.L. Toral, Tourism Management, **75**, 393-403 (2019)
12. M.S. Akter, M. Begum, R. Mustafa, *Bengali Sentiment Analysis of E-Commerce Product Reviews Using K-Nearest Neighbors*, 2021 International Conference on Information and Communication Technology for Sustainable Development, (ICICT4SD), 27-28 February, Dhaka (2021)
13. A. Jazuli, Widowati, R. Kusumaningrum, *Auto Labeling to Increase Aspect-Based Sentiment Analysis Using K-Nearest Neighbors Method*, The 7th International Conference on Energy, Environment, Epidemiology and Information System, (ICENIS), 9-10 August, Semarang, Indonesia (2022)
14. M. Qorib, T. Oladunni, M. Dennis, E. Ososanya, P. Cotae, J.ESWA, **212** (2023)
15. G.A.D.O. Junior, R.T.D.S. Jr, R.D.O. Albuquerquer, L.J.G. Villalba, J. Com. Com, **174**, 154-171 (2021)
16. J.S. Yang, C.Y. Zhao, H.T. Yu, H.Y. Chen, *Use GBDT to Predict The Stock Market*, in Proceedings 2019 International Conference on Identification, Information and Knowledge in The Internet of Things, (IIKI2019), 25-27 October, Jinan, China (2019)
17. W. Zheng, J. Gao, X. Wu, F. Liu, Y. Xun, G. Liu, X. Chen, J.JSS, **168** (2020)
18. S. Kaur, P. Kumar, P. Kumaraguru, Soft. Comp, **24**, 9049-9069 (2020)
19. J. Kiani, C. Camp, S. Pezeshk, J. Comp. Struc, **2018**, 108-122 (2019)
20. P. Cunningham and S.J. Delany, ACM. Comp. Surv, **54**, 1-25 (2021)
21. X. Han, G. Yang, S. Qu, G. Zhang, M. Chi, *A Weighted Algorithm Based on Physical Distance And Cosine Similarity for Indoor Localization*, 2019 14th IEEE Conference on Industrial Electronics and Applications, (ICIEA), 19-21 June, Xi'an, China (2019)
22. T. Mustaqim, K. Umam, M.A. Muslim, *Twitter Text Mining for Sentiment Analysis on Government's Response to Forest Fires With Vader Lexicon Polarity Detection and K-Nearest Neighbor Algorithm*, The 6th International Conference on Mathematics, Science, and Education. (ICMSE), 9-10 October 2019, Semarang, Indonesia (2019)
23. N. Alami, M. Meknassi, N. En-Nahnahi, Y.E. Adlouni, O. Ammor, J. ESWA, **172** (2021)