EMITTER International Journal of Engineering Technology
Vol. 11, No. 1, June 2023, pp. 89~99
DOI : 10.24003/emitter.v11i1.768

A Combination of Lexicon-based and Distributional Representations for Classification of Indonesian Vaccine Acceptance Rates

Katon Suwida[1], Muhammad Yusuf Kardawi[1], Diana Purwitasari[1], Fahril Mabahist[1]

[1]Department of Informatics Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
Corresponding Author: diana@if.its.ac.id

Received February 5, 2023; Revised April 2, 2023; Accepted June 7, 2023

Abstract

When the COVID-19 pandemic hit, the use of vaccines was advertised as the end of the pandemic by the entire world. However, the chances of vaccination depended on the sentiments of society and individuals about the vaccine. People's acceptance of vaccines can change depending on conditions and events. Social media platforms such as Twitter can be used as a source of information to find out the conditions and attitudes of the community toward the program. By implementing a machine learning technique on the COVID-19 vaccine dataset, we hope to impact the classification result with text. This study suggests three distinct machine learning models for classifying texts of the COVID-19 vaccination, namely a model based on the first lexicon using the feature extraction method; second, using the word insertion technique to utilize distribution representation; and third, a combination model of distribution representation and feature extraction based on the lexicon. From the evaluation that has been carried out, we found that a combination of lexicon-based and distributional representation methods succeeded in giving the best results for classifying the level of acceptance of the COVID-19 vaccine in Indonesia with an accuracy score of 71.44% and an F1-score of 71.43%.

Keywords: vaccination, text classification, lexicon-based, distributional representations.

1. INTRODUCTION

Currently, the entire world is campaigning and socializing vaccination as the end of the pandemic, the success of which depends on the public's and individuals' willingness to be vaccinated. Meanwhile, public attitudes toward vaccines can change depending on conditions or current events. Social media platforms such as Twitter are a valuable source of information to determine conditions and the public's attitude toward getting vaccinated. Nearly 3 billion people use at least one social media platform [1]. Recent studies have

demonstrated the benefits of social media analytics for assessing public health and policymakers, such as predicting health conditions [2] and tracking disease outbreaks [3]. In the context of the COVID-19 pandemic, Twitter has been used in various aspects, such as identifying user concerns [4], [5], [6], [7], the spread of misinformation [8], and general sentiment [9], [10].

Several studies have been conducted with various techniques related to the issue and effect of vaccination, such as social media analysis of public attitudes toward the COVID-19 vaccine [11], sentiment analysis of AstraZeneca/Oxford, Pfizer/BioNTech, and Moderna vaccines [12], analysis of the degree of doubt and acceptance of the COVID-19 vaccination [13]. Current research has achieved excellent results, yet some studies focus on specific data sources. Thus, the literature lacks well-validated methods in disparate data sets. This study aims to contribute to the effect of vaccination with text classification by proposing a machine-learning method for the COVID-19 vaccination dataset [14]. In addition, this study will conduct qualitative and quantitative tests regarding the best features and techniques for classifying vaccine acceptance rates for Indonesian people.

Based on these goals, this study offers three different machine-learning models for text classification of the COVID-19 vaccine, namely;

1. A model based on the lexicon using the feature extraction method.
2. Using the word embedding technique to utilize distribution representation.
3. A combination model of distribution representation and feature extraction based on the lexicon is used.

2. RELATED WORKS

Research has been done on a technique for text categorization on COVID-19 vaccination using deep learning and conventional machine learning methods [15]. In this study, the Latent Dirichlet Allocation (LDA) model was used to extract the five subjects whose internet users frequently expressed good and negative opinions about vaccines and vaccination. Then deep learning and traditional methods are applied to identify sentiments based on accuracy, precision, sensitivity, specificity, and F1-score to identify the best method. The deep learning method has outscored another model.

Another study [16] combines word embedding representations and lexicon-based feature frameworks to detect psychological stress for text classification. In this study, word embedding methods using Word2Vec, GloVe, and the FastText technique for lexicon-based feature exploit effective, syntactic, social, and topic-related features. An ensemble model combines two different models. Based on some experience, combining the FastText model and the lexicon-based feature framework resulted in the best-performing model.

Lexicon has been widely used in text classification, but little research has focused on constructing the emotion lexicon. A study [17] has compared the emotion lexicon, and this study made a comparative evaluation of NRC, ESN

(EmoSenticNet), DPM (DepecheMood), and TDPM (Topic Based) emotion lexicon (Depeche). Various emotion lexicons were evaluated using Semeval2007 test data based on ea

ch emotion category and title data set. The experimental results of NRC and DPM have the best evolution compared to ESN and TDPM.

3. ORIGINALITY

Social media, such as Twitter, is often used as a representation of the condition of the community regarding developing issues that concern the community. Many tweets from Twitter are used as research data to assess the level of public acceptance of a public policy. Therefore, it is necessary to process good tweet data to become clean data by removing unnecessary words. That clean data becomes information facilitating research in developing a lexicon-based model and distribution representation for classifying vaccine acceptance rates.

A problem that seeks to differentiate text based on users' expressions of vaccination can be considered as the classification of vaccine acceptance rates from the text. This classification can be accomplished by analyzing specific syntactic and linguistic features using lexicons [18]. The lexicon-based approach will evaluate the frequency and usage level of specific word terms by comparing the words in the text with the lexicon dictionary. Based on the nature of the words used in the sentence, the lexicon can categorize specific overall aspects such as sentiment, emotion, cognition, or topic [19].

Lexicon-based features can assist in delivering broad indications about a text's sentiments, cognitions, or topics. Fortunately, the lexicon cannot comprehend subtler, more complex, contextual cues unique to human language [20]. By enabling the encoding of the semantics and syntax present in words and their representation in vector space as relation offsets, the Word Embedding-based approach offers a solution to this problem. This study employed several pre-trained word embedding methods, including Word2Vec [21] and FastText [22].

In developing the word embedding model, we needed a corpus as a collection of texts that capture language in written form. At this stage, we collected data using non-API scraping techniques on Twitter from January 1, 2020, to October 1, 2022, using keywords such as the following.

The scraped data collected with a total of 30,068 data points are used as a corpus to add to the word domain of the dataset. The data aim to enrich the world in the vaccine domain.

Table 1. Scraping Corpus Keywords on Twitter

Keyword	Number of Tweets
Biofarma	65
Booster	5205
Covid	6070
Indofarma	77
Indovac	412
Moderna	194
Pfizer	339
Sinovac	195
Vaksin	11973
Vaksinasi	2249
Vaksin covid	1818
Vaksin booster	5205
Total	**30068**

4. SYSTEM DESIGN

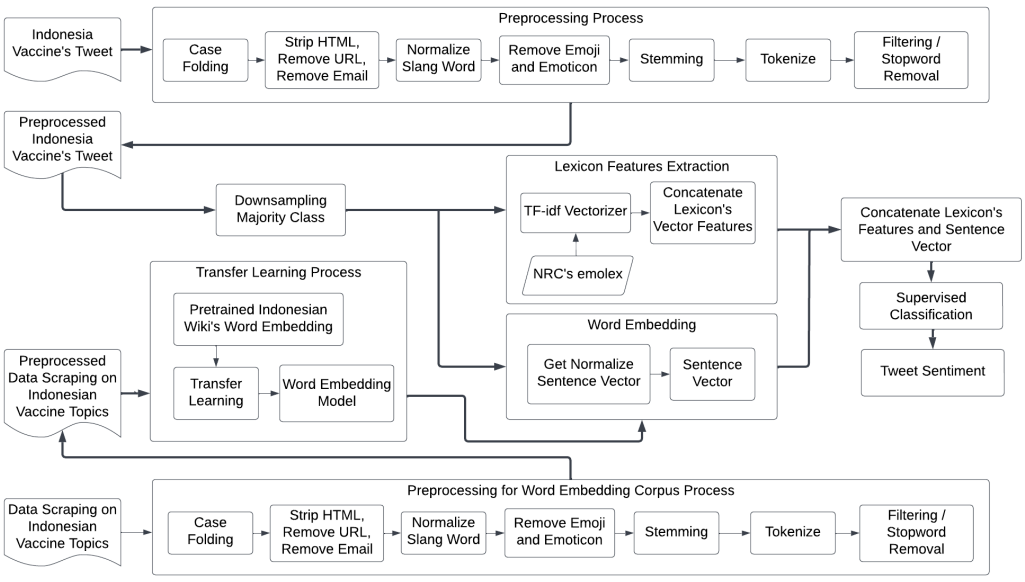

Figure 1. General architecture representation of combination model

As illustrated in Figure 1, developing a text classification system architecture focuses on three main methods: lexicon-based, word embedding, and combining the two methods. The obtained model was classified using four different classification techniques. There are differences in data preprocessing in the word embedding model. Namely, we added a data corpus to build the

Word2Vec [21] dan FastText [22] models. We found that adding a data corpus gave better results in developing the word embedding model.

This study uses datasets from previous studies [14]. In the preprocessing stage, case folding, stripping HTML, removing URLs, removing emails, normalizing slang words, stemming, tokenizing, and filtering/stopword removal is carried out. In addition, a data cleaning process has also been executed to remove irrelevant data such as spam and unrelated data (non-language and non-COVID-19). This study used 9027 data points consisting of 3751 data points in the supporting class, 3299 in the neutral class, and 1977 in the opposing class. To balance the data in each class, we use a down sampling technique to randomly reduce the data in the supporting and neutral classes to produce 5931 data with 1977 data in each class. The data that has undergone the down sampling process is converted into a numerical representation using the word embedding technique. The word embedding model is the result of training using data from the corpus obtained from the scrapping process on the Twitter platform, according to Table 1. Next, we split the data into 80% training data and 20% testing data, with a total of 4745 and 906 data points, respectively.

4.1 Proposed Method

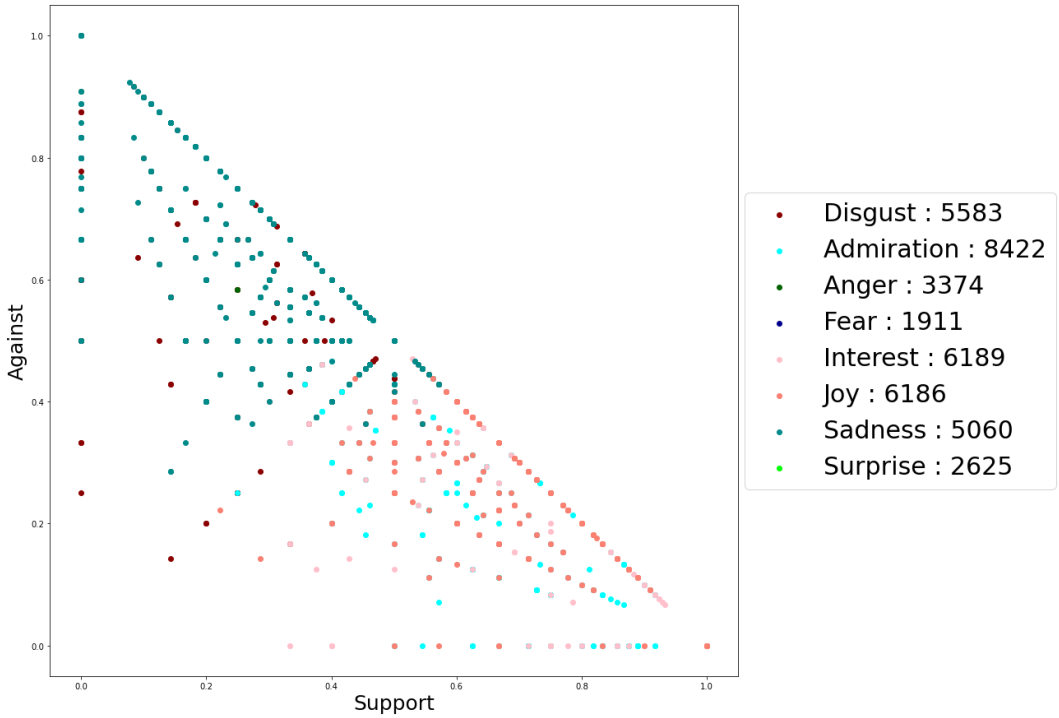

Figure 2. SecticNet distribution of emotion lexicon

At this stage, the aim is to conduct feature selection to help the feature characterization process detect public opinion related to COVID-19 vaccinations. This feature is based on the lexicon and sentiment analysis

methods to extract information from textual data. In this section, we use SecticNet [23] as a Lexicon-based method, which expands the WordNet database with a detailed list of emotions, as follows. Sentiment analysis, also called opinion mining, recognizes, extracts, and processes textual data to get perceptions of the sentiment. These statistics are crucial for providing input on products, services, and other subjects. Sentiment analysis, also known as opinion mining, is the process of removing an opinion or opinions from a document for a particular topic [24].

In developing the lexicon model, we used 36,725 data points consisting of eight types of emotion to determine the text classification of vaccination acceptance in Indonesia. In Figure 2, we show the distribution of the eight types of emotion in the aspect-based sentiment dataset from tweets related to Indonesia's COVID-19 program. It appears that the emotions of awe, interest, joy, and surprise support COVID-19 vaccination, while the emotions of disgust, anger, sadness, and fear do not harm COVID-19.

5. EXPERIMENT AND ANALYSIS

This study carried out three experimental scenarios: classification using the lexicon-based model, the Distributional Representation Model, and the Ensemble Model, a combination of the two. The classification algorithms used in this experiment are Random Forest, XGBoost, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM), which have been tuned to their hyperparameters.

5.1 Lexicon-based Model

The first model employs emotions extracted from the text, including disgust, admiration, anger, fear, interest, joy, and sadness, to classify lexical data from the COVID-19 vaccine dataset. Following the methodology described in Section 4, these emotional features were retrieved. The model aims to assess the effectiveness of lexicon-based features in classifying Indonesia's COVID-19 vaccine acceptance rate. We propose encoding the text into a fixed-length vector using a lexicon-based approach. As shown in Table 2, we then utilize Random Forest, XGBoost, KNN, and SVM algorithms to classify the resulting model based on eight distinct emotional states to evaluate the level of acceptance of the COVID-19 vaccine.

The classification method used has values that are near each other, with the Random Forest technique having the best performance in this lexicon model with the highest F1-score, accuracy, precision, and recall compared to other classification techniques. Random Forest obtained the highest accuracy score at 38.96%, 37.89% for precision, 37.98% for recall, and 37.82% for an F1-score. This result is still far lower than needed to classify vaccine acceptance rates in Indonesia. Lexicon works by classifying all the features in the text with the emotions available in Table 2, generating a new vector, which is then classified using four different methods.

5.2 Distributional Representation Model

The second model takes advantage of distributed representations by using word embedding techniques. A type of computational text representation known as "word embedding" provides words with the same meanings with comparable representations [21]. In the development of this model, there is a conversion from text to vector representation by encoding semantic and syntactic information. We employed different word embedding and then conducted a comparative analysis to obtain the best word embedding technique for classifying text on the level of acceptance of the COVID-19 vaccine in Indonesia. We then retrained the pre-trained Word2Vec [21] and FastText [22] models by scrapping the corpus on Twitter, as shown in Table 1. After obtaining the vector representation, we classified the vectors using Random Forest, XGBoost, KNN, and SVM.

The distributional representation model produced a significantly higher score than the previous lexicon model, which could not comprehend the finer and more detailed qualities and contextual signals inherent to human language [20]. The FastText [22] type of word embedding offered a better score than Word2Vec [21]. Moreover, the XGBoost classification technique provides excellent and stable performance in classifying vector word embedding in classifying vaccine acceptance rates in Indonesia. Each score obtained is stable, with a scoring accuracy of 69.67% in the Word2Vec technique [21]. The highest accuracy score is 70.26% in the FastText technique [22].

5.3 Ensemble Model

Finally, the third model combines lexicon-based features with distributional representations. Figure 1 shows the model's generalized representation. This model combines the value of each vector in the word embedding model with the feature vectors from the lexicon model. The vector information acquired from word embedding is coupled with valuable and linguistic information from lexicon-based features and provided to the machine learning classifier for additional classification procedures. The Ensemble model has provided the best results in classifying text on the level of vaccine acceptance in Indonesia by obtaining scores of accuracy and F1-score that are higher than the other methods.

Based on the experimental results, the Random Forest and XGBoost classification methods offered better results than the KNN and SVM methods. The SVM method consistently ranks lowest in the classification of vaccine acceptance rates in Indonesia for each method used. In addition, based on these results, the word embedding method using FastText [22] shows better results in each classification, with the highest result using the Random Forest method, with an accuracy score of 71.44% and an F1 score of 71.43%.

It is well known that decision trees have a considerable bias when using bare trees and a significant variance when using thorny trees. When employing ensemble approaches, many decision trees are combined to generate higher predictive performance than a single decision tree. Ensemble

decision trees can be performed using a variety of approaches, including boosting and bagging [25]. Random Forest and XGBoost are examples of models in this study that use the ensemble method.

Table 2. Performance Evaluation

Model	Classification	Accuracy (%)	F1-Score (%)
Lexicon	KNN	36.20	35.54
	XGBoost	36.31	35.28
	Random Forest	**38.96**	**37.82**
	SVM	28.36	22.96
Word Embedding with FastText	KNN	63.01	62.60
	XGBoost	**70.26**	**70.26**
	Random Forest	68.57	68.56
	SVM	58.46	58.18
Word Embedding with Word2Vec	KNN	62.93	62.49
	XGBoost	**69.67**	**69.68**
	Random Forest	69.25	69.27
	SVM	58.55	58.26
Ensamble with FastText	KNN	63.26	62.85
	XGBoost	69.33	69.33
	Random Forest	**71.44**	**71.43**
	SVM	58.21	58.0
Ensamble with Word2Vec	KNN	63.18	62.74
	XGBoost	**68.07**	**68.06**
	Random Forest	67.64	67.67
	SVM	58.55	58.34

A group of classifiers with a tree structure can be grouped to form the Random Forest Classifier. Since Random Forest is a more sophisticated variant of bagging, unpredictability is added to it. Instead of splitting each node using the best split among all variables, Random Forest divides each node using the best among a randomly chosen subset of predictors at that node [25].

Gradient boosting machines, shown to push the computational limitations of boosted tree algorithms, are used flexibly and cutting-edge in XGBoost. Boosting is an ensemble technique that adds additional models to

make up for mistakes made by earlier models. Models are added repeatedly until no discernible improvements can be found. Gradient boosting is a technique for developing fresh models that foresee older models' residuals combined to provide the final prediction [26].

6. CONCLUSION

First, this study evaluated the lexicon model while analyzing vaccine acceptance rates. This first experiment used all the features and emotions in the lexicon; the experimental results can be seen in Table 2. The best results were obtained with the random forest method. Next, we experimented with distributional representation with the best score obtained using the FastText technique with an accuracy score of 70.26%. As seen, the best results for classifying the level of acceptance of the COVID-19 vaccine in Indonesia is to combine the models in the first and second experiments by mixing all the features from the lexicon with the detailed information from the distributional representation.

This paper has presented a machine-learning method for classifying vaccine acceptance rates in Indonesia. Our proposed experiment has explored lexicon and two-word embedding techniques. We conclude that the combination of lexicon-based and distributional representation methods gave the best results for classifying the level of acceptance of the COVID-19 vaccine in Indonesia with an accuracy score of 71.44% and an F-Score of 71.43%. It is recommended for future research to apply the combination of lexicon-based and distributional representation methods in various scopes and use better validation methods in different data sets.

Acknowledgments

The author(s) disclosed receipt of the following financial support for the publication of this article. This work is supported by the Department of Informatics Engineering, Institut Teknologi Sepuluh Nopember, Indonesia.

REFERENCES

[1] A. Hussain and A. Sheikh, **Opportunities for Artificial Intelligence-Enabled Social Media Analysis of Public Attitudes Toward Covid-19 Vaccines**, NEJM Catal Innov Care Deliv, pp. 1–7, 2021, doi: 10.1056/CAT.20.0649.

[2] R. M. Merchant et al., **Evaluating the predictability of medical conditions from social media posts**, PLoS One, vol. 14, no. 6, pp. 1–12, 2019, doi: 10.1371/journal.pone.0215476.

[3] L. Samaras, E. García-Barriocanal, and M. A. Sicilia, **Comparing Social Media and Google to Detect and predict severe epidemics**, Sci Rep, vol. 10, no. 1, pp. 1–11, 2020, doi: 10.1038/s41598-020-61686-9.

[4] C. H. Chang, M. Monselise, and C. C. Yang, **What Are People Concerned About During the Pandemic? Detecting Evolving Topics about COVID-19 from Twitter**, J Healthc Inform Res, vol. 5, no. 1, pp. 70–97, 2021, doi: 10.1007/s41666-020-00083-3.

[5] O. Oyebode et al., **Health, psychosocial, and social issues emanating from the COVID-19 pandemic based on social media comments: Text mining and thematic analysis approach**, JMIR Med Inform, vol. 9, no. 4, 2021, doi: 10.2196/22734.

[6] Y. Su, A. Venkat, Y. Yadav, L. B. Puglisi, and S. J. Fodeh, **Twitter-based analysis reveals differential COVID-19 concerns across areas with socioeconomic disparities**, Comput Biol Med, vol. 132, no. March, p. 104336, 2021, doi: 10.1016/j.compbiomed.2021.104336.

[7] H. Jang, E. Rempel, D. Roth, G. Carenini, and N. Z. Janjua, **Tracking COVID-19 discourse on Twitter in north america: Infodemiology study using topic modeling and aspect-based sentiment analysis**, J Med Internet Res, vol. 23, no. 2, 2021, doi: 10.2196/25431.

[8] D. Gerts et al., **'Thought I'd share first': An analysis of COVID-19 conspiracy theories and misinformation spread on Twitter**, JMIR Public Health Surveill, vol. 7, no. 4, p. e26527, 2021.

[9] J. Zhou, S. Yang, C. Xiao, and F. Chen, **Examination of Community Sentiment Dynamics due to COVID-19 Pandemic: A Case Study from a State in Australia**, SN Comput Sci, vol. 2, no. 3, pp. 1–11, 2021, doi: 10.1007/s42979-021-00596-7.

[10] M. Pellert, J. Lasser, H. Metzler, and D. Garcia, **Dashboard of Sentiment in Austrian Social Media During COVID-19**, Front Big Data, vol. 3, October, pp. 1–9, 2020, doi: 10.3389/fdata.2020.00032.

[11] M. Sallam, **Covid-19 vaccine hesitancy worldwide: A concise systematic review of vaccine acceptance rates**, Vaccines (Basel), vol. 9, pp. 1–14, 2021, doi: 10.3390/vaccines9020160.

[12] R. Marcec and R. Likic, **Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines**, Postgrad Med J, pp. 544–550, 2021, doi: 10.1136/postgradmedj-2021-140685.

[13] M. R. Jawad et al., **Advancement of artificial intelligence techniques based lexicon emotion analysis for vaccine of COVID-19**, Periodicals of Engineering and Natural Sciences, vol. 9, no. 4, pp. 580–588, 2021, doi: 10.21533/pen.v9i4.2383.

[14] C. B. P. Putra, D. Purwitasari, and A. B. Raharjo, **Stance Detection on Tweets with Multi-task Aspect-based Sentiment: A Case Study of COVID-19 Vaccination**, International Journal of Intelligent Engineering and Systems, vol. 15, no. 5, pp. 515–526, 2022, doi: 10.22266/ijies2022.1031.45.

[15] M. S. Zulfiker, N. Kabir, A. A. Biswas, S. Zulfiker, and M. S. Uddin, **Analyzing the public sentiment on COVID-19 vaccination in social**

media: Bangladesh context, Array, vol. 15, Sep. 2022, doi: 10.1016/j.array.2022.100204.

[16] S. Muñoz and C. A. Iglesias, **A text classification approach to detect psychological stress combining a lexicon-based feature framework with distributional representations**, Inf Process Manag, vol. 59, no. 5, Sep. 2022, doi: 10.1016/j.ipm.2022.103011.

[17] F. S. Tabak and V. Evrim, **Comparison of emotion lexicons**, in 13th HONET-ICT International Symposium on Smart MicroGrids for Sustainable Energy Sources Enabled by Photonics and IoT Sensors, HONET-ICT 2016, Nov. 2016, pp. 154–158. doi: 10.1109/HONET.2016.7753440.

[18] S. Muñoz and C. A. Iglesias, **A text classification approach to detect psychological stress combining a lexicon-based feature framework with distributional representations**, Inf Process Manag, vol. 59, no. 5, p. 103011, 2022, doi: 10.1016/j.ipm.2022.103011.

[19] C. S. G. Khoo and S. B. Johnkhan, **Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons**, J Inf Sci, vol. 44, no. 4, pp. 491–511, 2018, doi: 10.1177/0165551517703514.

[20] S. Wang, W. Zhou, and C. Jiang, **A survey of word embeddings based on deep learning**, Computing, vol. 102, no. 3, pp. 717–740, 2020, doi: 10.1007/s00607-019-00768-7.

[21] J. D. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, **Distributed Representation of Words and Phrases and their Compositionality**, Advances in Neural Information Processing Systems 26 (NIPS 2013), 2013, doi: 10.18653/v1/d16-1146.

[22] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, **FastText.zip: Compressing text classification models**, pp. 1–13, 2016.

[23] E. Cambria, Q. Liu, S. Decherchi, F. Xing, and K. Kwok, **SenticNet 7: A Commonsense-based Neurosymbolic AI Framework for Explainable Sentiment Analysis**, Proceedings of the Language Resources and Evaluation Conference, no. June, pp. 3829–3839, 2022.

[24] N. R. Prayoga et al., **Unsupervised Twitter Sentiment Analysis on The Revision of Indonesian Code Law and the Anti-Corruption Law using Combination Method of Opinion Word and Agglomerative Hierarchical Clustering**, Emit. Int. J. Eng. Technol., vol. 8, no. 1, pp. 200–220, 2020, doi: 10.24003/emitter.v8i1.477.

[25] N. Bahrawi, **Sentiment Analysis Using Random Forest Algorithm-Online Social Media Based**, J. Inf. Technol. Its Util., vol. 2, no. 2, p. 29, 2019, doi: 10.30818/jitu.2.2.2695.

[26] A. Ogunleye and Q. G. Wang, **XGBoost Model for Chronic Kidney Disease Diagnosis**, IEEE/ACM Trans Comput Biol Bioinform, vol. 17, no. 6, pp. 2131–2140, Nov. 2020, doi: 10.1109/TCBB.2019.2911071.