Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 10, Nomor 1, Februari 2025, pp 58-67
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

Analysis of Performance Labelling Sentiment Between K-Means Indobert And Inset Lexicon-Based

Rama Dona Ariyatma[1], Bagus Priambodo[2]
[1,2]Informatics Engineering, Universitas Mercu Buana, Jakarta, Indonesia
Email: 41520120064@mercubuana.ac.id[1], bagus.priambodo@mercubuana.ac.id[2]

Abstract

Sentiment analysis, a natural language processing technique, plays a key role in identifying opinions or sentiments from textual data. Accurate sentiment labelling within a dataset significantly impacts the performance of sentiment analysis models. However, manual labelling can be time-consuming. Many researchers utilize lexicon-based methods for sentiment labelling, but lexicons are often limited in reflecting topic-specific nuances, potentially leading to inaccurate sentiment representation. This inaccuracy can negatively affect classification models. Inset Lexicon (Indonesia Sentiment Lexicon) provides a pre-weighted list of sentiment words for sentiment analysis in Indonesian. This study aims to explore the use of K-means clustering as an automatic sentiment labelling technique and compare it to the performance of Inset Lexicon. For K-means clustering, IndoBERT is employed as the embedding model. The objective of this research is to evaluate the accuracy of automatic sentiment labelling by comparing it with actual data to assess the performance of both methods. The experiment accuracy shows that K-means with IndoBert achieves 74.79%, higher than Inset Lexicon that achieves only 59.82%.

Keywords: *Sentiment, K-means Cluster, Labeling, Inset Lexicon, IndoBERT*

1. INTRODUCTION

Sentiment analysis is a natural language processing technique aimed at identifying sentiments or opinions expressed in a text. This method is widely applied in various fields, from social media monitoring to product review analysis. One of the key challenges in sentiment analysis is the accuracy of sentiment labelling. If the sentiment labels are incorrect, the performance of the sentiment analysis model built from such a dataset can be significantly impacted.Accurate label selection requires substantial time and effort, particularly when using lexicon-based techniques. One commonly used method is the Inset Lexicon (Indonesian Sentiment Lexicon), a list of words labelled with sentiment in the Indonesian language, which assigns sentiment labels to data based on word weights. This lexicon was developed by Koto [1] and has been extensively used by researchers for sentiment labelling. For instance, Fatani and Irawan [2] applied Inset Lexicon in sentiment analysis for evaluating the LRT Jabodetabek service, while Shaleha and colleagues [3] used it for analyzing sentiments related to the 2024 elections in Indonesia. Similarly, Anam and co-researchers [4] employed an Inset Lexicon to label data in their sentiment analysis of online learning. Desi Musfiroh and her team [5] also utilized this lexicon for sentiment labelling in their research. Many other studies have used Inset Lexicon as a foundation for sentiment labelling before proceeding with classification or model building. Despite the advantages of speeding up the labelling process, using Inset Lexicon can result in datasets with inaccurate labels. This occurs because lexicon-based

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 10, Nomor 1, Februari 2025, pp 58-67
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

methods may not always fit every topic or context, potentially leading to biased sentiment scores, which in turn affect the model's performance.

Several studies have explored the use of clustering techniques to address these challenges. For example, Fenina [6] conducted a sentiment analysis using K-means clustering on 1,407 data points, which yielded two clusters with proportions of 91% for cluster 1 and 9% for cluster 2. Mofiz Mojib Haider and his team used K-means clustering for automatic text summarization, achieving the highest BLEU score of 0.894 [7]. Another study by Jayasree Ravi and Sushil Kulkarni applied BERT embeddings to convert text into vectors, attaining the highest accuracy using K-means clustering [8]. Similarly, Kirk Alvin employed K-means clustering to analyze question banks and categorize them based on difficulty levels [9]. In their study, Alvin Subakti and colleagues [10] demonstrated that BERT-based text representation outperformed TF-IDF when used with K-means clustering, recommending BERT for text data representation.

Therefore, this study will explore the performance of sentiment analysis using Inset Lexicon and compare it with the K-means clustering method to determine positive, negative, and neutral sentiments based on actual data. This research also aims to assess the accuracy of sentiment labels generated through Inset Lexicon and clustering methods, while proposing strategies to improve the quality of sentiment labeling. Consequently, the objective of this study is to understand how to automate sentiment labeling using Inset Lexicon and K-means clustering and evaluate the results using manually labeled data.

2. RESEARCH METHODOLOGY

In conducting the research, the workflow or stages are as follows: starting with data collection, data pre-processing, automatic data labelling, and evaluation of the results with manual labelling. These stages are illustrated in Figure 1.

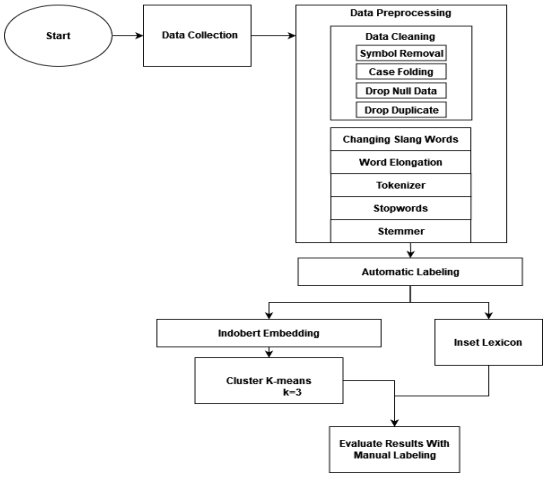

Figure 1. Research Methods

2.1. Data Collection

The data used in this study consists of secondary data. First, the Inset Lexicon-based dataset was downloaded from GitHub

(https://github.com/fajri91/InSet/). Additionally, the actual sentiment data was obtained from IndoNLU or directly from this link: https://github.com/IndoNLP/indonlu/tree/master/dataset [11]. The dataset used is smSA, which has been annotated by an Indonesian language expert with three sentiment labels: positive, negative, and neutral. For this research, We utilized a training dataset comprising 11,000 data points. The SmSA sentiment dataset is presented in Table 1.

Tabel 1. Dataset SmSA

Sentiment	Twitter
positive	betapa bahagia nya diri ini saat unboxing paket dan barang nya bagus ! menetapkan beli lagi !

2.2. Preprocessing

Before conducting the research or applying automatic labelling, text processing must be performed to ensure that the data used is clean and ready for analysis. The following are the stages of preprocessing:

a. Symbol Removal: All symbols and special characters, such as irrelevant punctuation marks, are removed to simplify the data.
b. Case Folding: All text is converted to lowercase to ensure consistency and avoid differences in meaning due to capitalization.
c. Null Data Removal: Rows or columns containing null values are removed to ensure the integrity and completeness of the data.
d. Duplicate Data Removal: Duplicate rows containing identical information are deleted to prevent redundant data from skewing the analysis results.
e. Slang Word Conversion: Informal or slang words in the text are converted to their standard forms to maintain consistency with formal language. The process of converting slang words into formal words uses the IndoNLP library, with the slang word data sourced from Salsabila[12].
f. Word Elongation: Overly elongated words (e.g., "*kenapaaaa*") are shortened to their root form ("*kenapa*") to maintain data uniformity.
g. Tokenization: divides a string of characters using spaces and may include removing certain characters[13].
h. Stopwords Removal: Common words that do not carry significant informational value (such as "*dan*," "*di*," "*yang*") are removed from the text.
i. Stemming: Stemming is a rule-based method used to trim the suffixes of words[14].

The results of the data preprocessing can be seen in Table 2, which highlights the changes in the data after the cleaning process.

Tabel 2. Result preprocessing

Before	After
betapa bahagia nya diri ini saat unboxing paket dan barang nya bagus ! menetapkan beli lagi !	betapa,bahagia,nya,unboxing,paket,barang,nya ,bagus,tetap,beli

2.3. IndoBERT embedding

IndoBERT is a language model based on the BERT architecture, specifically designed to process Indonesian text. The model is trained using a large dataset called Indo4B, which consists of approximately 4 billion words sourced from various text types, including social media, online news, and other articles. IndoBERT comes in several variants, such as IndoBERT-lite and IndoBERT-large, each with a different number of parameters and model capacities, depending on the complexity of the model [11].

In this research, the model used is IndoBERT-base, with 124.5 million parameters, comprising 12 layers along with 12 attention heads, an embedding dimension of 768, and a feed-forward network dimension of 3072. This model is an example of monolingual contextual pre-trained embedding, meaning it is trained exclusively on Indonesian text, making it more effective at understanding the context and meaning of words in the language. More specifically, the feature extraction model used is "indobenchmark/indobert-base-p1," with a maximum token limit of 512, which will later be applied for K-means clustering.

2.4. K-means Clustering

A straightforward iterative clustering method that utilizes distance as a metric to group data is known as K-means clustering. Given a set number of clusters K in a dataset, the algorithm calculates the average distances, determines the initial centroids and each cluster is represented by one centroid [15]. Let O = {O1, O2, ..., On} represent a set of n data samples to be divided into K clusters, C = {Ci, i = 1, ..., k}. The objective of K-means clustering is to minimize the sum of squared errors (SSE) over all k clusters, as defined below:

$$J(c) = \sum_{i=1}^{k} \sum_{O_1 \in C_1} (O_1 - Z_1)^2 \quad (1)$$

Where Ci, Zi, Ol, and k denote the i-th cluster, the centroid of the i-th cluster, the data points assigned to the i-th cluster, and the total number of clusters, respectively[16]. In this case, a value of k=3 was used, as the goal was to determine whether it was possible to obtain clusters representing positive, negative, and neutral sentiments.

2.5. *Indonesia sentimen* (Inset Lexicon)

A common example of this technique is the InSet Lexicon, developed for the Indonesian language, which can categorize sentiments as positive, negative, or neutral. Generally, This approach calculates a score for each word in the text using the lexicon. The text is then classified as positive if the score is above 0, negative if it is below 0, and neutral if the score is 0[17].

2.6. Evaluation

Evaluation will be carried out using external evaluation, especially on data that has been manually labelled by Indonesian language experts, namely from the SMSA dataset. Therefore, the reliability of the data cannot be doubted. A confusion matrix will be used as the primary evaluation tool, along with evaluation measures

including accuracy, precision, recall, and F1-score. The confusion matrix is a powerful and simple metric commonly used to assess the accuracy of a classification model by comparing actual and predicted results[18]. In confusion matrix evaluation, there are four statistical measures: true positive (TP), true negative (TN), false positive (FP), and false negative (FN)[19]. In this study, metrics accuracy is used to evaluate the performance of the proposal method. The accuracy formula is explained in (2).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

2.7. Google colab

Google Colab is widely recognized as an essential tool in the current era of data science. It is popular and widely used due to its free, open-access nature, provided by Google to anyone with a Gmail account[20]. This research utilizes Google Colab as the primary platform for conducting the experiments.

2.8. Cosine Similarity

A similarity metric that determines the cosine of the angle between two vectors located in a multidimensional area. This measure is based on the orientation of the vectors rather than their magnitude. For two attribute vectors, A and B, cosine similarity is computed using the dot product and magnitudes[21], which is formulated as follows:

$$Cosine\ Sim(A,B) = \frac{A \cdot B}{|A||B|} \quad (3)$$

(A) represents the centroid point, and (B) serves as the reference point for sentiment word opinions. For the opinion reference, the data is derived from the NLTK opinion corpus, selecting positive and negative words. Since this lexicon is in English, it is first translated into Indonesian, followed by the removal of duplicate words.

3. RESULT AND DISCUSSION

3.1. Implementing Lexicon Based Inset

In the use of the InSet Lexicon, specifically the word weighting process previously explained, the results of this process can be seen in Table 3.

Table 3. Result Inset Lexicon Labeling

Sentence	Lexicon Pos	Lexicon Neg	Lexicon Final
betapa(4),bahagia(4)(1),nya,unboxing,paket(4),barang,nya,bagus(2)(4),tetap(3),beli(2)(-3)	19	-8	11

Table 3 shows that the sentence is positive as the final_lexicon has a weight of 11, which is above zero. Conversely, the final_lexicon has a negative (-) value, the sentence is considered negative.

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 10, Nomor 1, Februari 2025, pp 58-67
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

3.2. Implementing K-means Clustering

The results of the clustering process will be visualized. Before that, we apply TSNE to reduce the dimensionality of the data for a two-dimensional visual representation. The visual results can be seen in Figure 2.

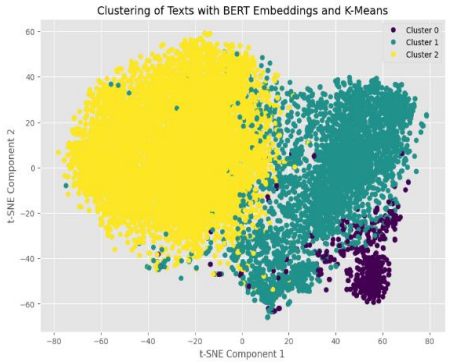

Figure 2. The Visualization of clustering results using k-means

The number of clusters obtained can be seen in Figure 3, visualized using a bar chart.

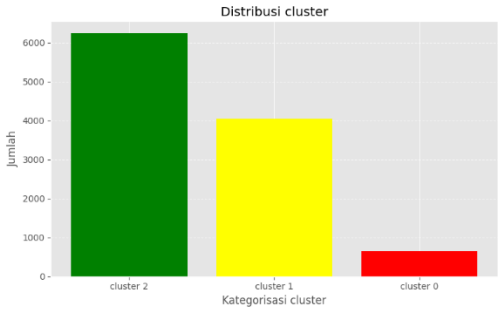

Figure 3. The distribution of clustering result

Cluster 2 contains 6,236 data points, cluster 1 contains 4,049, and cluster 0 contains 647. After clustering, cosine similarity was applied to determine whether a cluster was positive, negative, or neutral. The distance of cosine similarity between the cluster and the actual data is presented in Table 4.

Table 4. The distance of Cosine Similarity

Cluster	Positive Similarity	Negative Similarity	Final Lexicon
0	0.8062	0.8021	Positive
1	0.4355	0.4382	Negative
2	0.3039	0.2928	Positive

The results of each cluster are evaluated using cosine similarity, with neutral_threshold = 0.001 as the limit for determining whether a cluster is neutral and looking for the one with the highest similarity. The findings show that cluster 2 tends to be more positive, while cluster 1 is more negative.

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 10, Nomor 1, Februari 2025, pp 58-67
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

To confirm the characteristics of each cluster, a word cloud visualization was used to display the words included in each cluster. The word cloud for cluster 0 can be seen in Figure 4, for cluster 1 in Figure 5, and for cluster 2 in Figure 6.

Figure 4. Wordcloud of Cluster 0

Figure 5. Wordcloud of Cluster 1

Figure 6. Wordcloud of Cluster 2

After analyzing the word cloud visualizations and the number of words in each cluster, it was confirmed that cluster 2 is positive, further analysis we obtained that cluster 2 topics are about food restaurants in Bandung. Cluster 1 deals with political topics, leaning more towards negativity. Meanwhile, cluster 0 appears balanced with both positive and negative words. However, the researcher used cosine similarity as the primary reference for determining the clusters, categorizing cluster 0 as positive, despite the slight difference.

3.3. Confusion Matrix Evaluation

The automatic labelling results will be evaluated. The results of automatic labelling using the lexicon method with heatmap visualization can be seen in Figure 7.

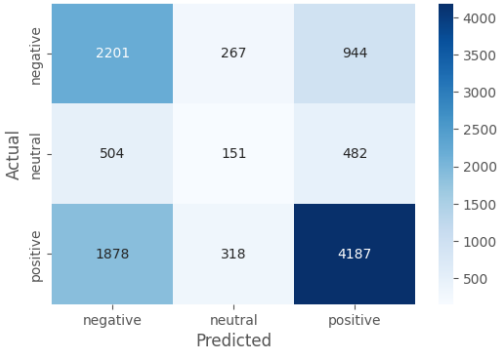

Figure 7. Visual Heatmap confusion matrix Inset lexicon

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 10, Nomor 1, Februari 2025, pp 58-67
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

The accuracy of automatic sentiment labelling obtained from the Inset Lexicon is 59.82%. In addition, automatic labelling using K-means can be visualized through a heatmap with the lexicon in Figure 8.

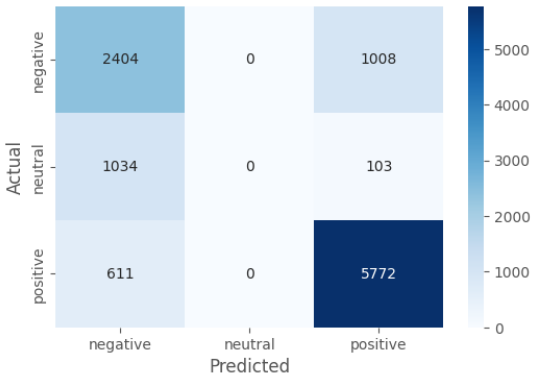

Figure 8. Visual Heatmap confusion matrix k-means clustering

The accuracy obtained using K-means clustering with IndoBERT embedding in sentiment labelling is 74.79%.

3.4. Discussion

Previous research by Jayasree Ravi provided insights into the efficiency of clustering techniques in separating Twitter data, which inspired the use of labelling techniques through clustering based on text embeddings. One of the main points is the use of BERT embeddings, which demonstrated the highest accuracy. Based on this research, our main focus is to evaluate the accuracy performance of sentiment labelling using inset-lexicon and K-means clustering, to offer an alternative strategy for automatic sentiment labelling that can reduce time, cost, and effort. The research results show that using K-means clustering as a labelling technique yields better results compared to inset-lexicon. Figure 9 illustrates the comparison between the two methods, where the accuracy of K-means clustering with IndoBERT embedding proves to be superior to inset-lexicon.

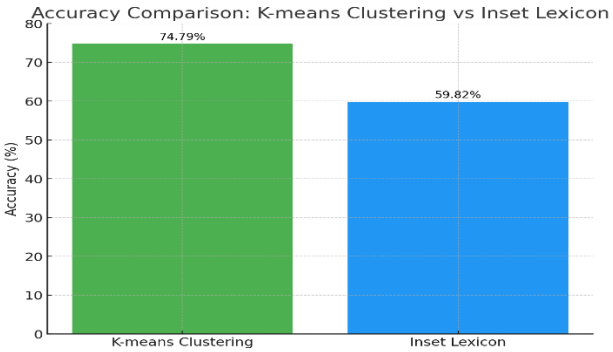

Figure 9. Comparison of Accuracy Labeling Sentiment K-means Clustering and Inset Lexicon

4. CONCLUSION

Our interest in this research is that the use of K-means clustering with IndoBERT as an embedding method can be applied to automatically detect sentiment opinions and even identify topics in each cluster, such as cluster 2 related to food, cluster 1 to politics, and cluster 0 also to politics but with a smaller scale of opinion. The results of this study, which aimed to explore sentiment labelling using K-means clustering compared to the Inset Lexicon, show that K-means clustering outperforms with an accuracy of 74.79%, while the Inset Lexicon achieved 59.82%. This research demonstrates that clustering can indeed be used to label sentiment opinions. For future research, it is suggested to explore other clustering techniques or to compare different embedding methods to see if accuracy can be further improved.

REFERENCES

[1] F. Koto And G. Y. Rahmaningtyas, "Inset Lexicon: Evaluation Of A Word List For Indonesian Sentiment Analysis In Microblogs," In *2017 International Conference On Asian Language Processing (Ialp)*, Ieee, Dec. 2017, Pp. 391–394. Doi: 10.1109/Ialp.2017.8300625.

[2] I. I. Fatani And H. Irawan, "Twitter, Instagram, Youtube Speak: Understanding Sentiments On Lrt Jabodebek Services Via Inset Lexicon, Indobert And Bertopic Approaches," *Journal Of Electrical Systems*, Vol. 20, No. 4s, Pp. 1028–1035, Apr. 2024, Doi: 10.52783/Jes.2147.

[3] S. Shaleha, A. Saputri, And H. M. Wicaksana, "Sentiment Analysis With Supervised Topic Modelling On Twitter Data Related To Indonesian Election 2024," In *2023 International Conference On Computer, Control, Informatics And Its Applications (Ic3ina)*, Ieee, Oct. 2023, Pp. 37–42. Doi: 10.1109/Ic3ina60834.2023.10285800.

[4] M. K. Anam, T. A. Fitri, A. Agustin, L. Lusiana, M. B. Firdaus, And A. T. Nurhuda, "Sentiment Analysis For Online Learning Using The Lexicon-Based Method And The Support Vector Machine Algorithm," *Ilkom Jurnal Ilmiah*, Vol. 15, No. 2, Pp. 290–302, Aug. 2023, Doi: 10.33096/Ilkom.V15i2.1590.290-302.

[5] D. Musfiroh, U. Khaira, P. E. P. Utomo, And T. Suratno, "Analisis Sentimen Terhadap Perkuliahan Daring Di Indonesia Dari Twitter Dataset Menggunakan Inset Lexicon," *Malcom: Indonesian Journal Of Machine Learning And Computer Science*, Vol. 1, No. 1, Pp. 24–33, Mar. 2021, Doi: 10.57152/Malcom.V1i1.20.

[6] R. Nainggolan, F. Adline, T. Tobing, And E. J. G. Harianja, "Analysis Sentiment In Bukalapak Comments With K-Means Clustering Method," *International Journal Of New Media Technology)*, Vol. 9, No. 2, P. 87, 2022.

[7] M. M. Haider, Md. A. Hossin, H. R. Mahi, And H. Arif, "Automatic Text Summarization Using Gensim Word2vec And K-Means Clustering Algorithm," In *2020 Ieee Region 10 Symposium (Tensymp)*, Ieee, 2020, Pp. 283–286. Doi: 10.1109/Tensymp50017.2020.9230670.

[8] J. Ravi And S. Kulkarni, "Text Embedding Techniques For Efficient Clustering Of Twitter Data," *Evol Intell*, Vol. 16, No. 5, Pp. 1667–1677, Oct. 2023, Doi: 10.1007/S12065-023-00825-3.

[9] K. A. S. Awat And M. A. Ballera, "Applying K-Means Clustering On Questionnaires Item Bank To Improve Students' Academic Performance," In *2018 Ieee 10th International Conference On Humanoid, Nanotechnology, Information*

Technology,Communication And Control, Environment And Management (Hnicem), Ieee, Nov. 2018, Pp. 1–6. Doi: 10.1109/Hnicem.2018.8666409.

[10] A. Subakti, H. Murfi, And N. Hariadi, "The Performance Of Bert As Data Representation Of Text Clustering," *J Big Data*, Vol. 9, No. 1, Dec. 2022, Doi: 10.1186/S40537-022-00564-9.

[11] B. Wilie *Et Al.*, "Indonlu: Benchmark And Resources For Evaluating Indonesian Natural Language Understanding," Sep. 2020.

[12] N. Aliyah Salsabila, Y. Ardhito Winatmoko, A. Akbar Septiandri, And A. Jamal, "Colloquial Indonesian Lexicon," In *2018 International Conference On Asian Language Processing (Ialp)*, Ieee, Nov. 2018, Pp. 226–229. Doi: 10.1109/Ialp.2018.8629151.

[13] F. Anisa Nirmala, M. Jazman, N. Evrilyan Rozanda, And F. Nur Salisah, "Cyberbullying Sentiment Analysis Of Instagram Comments Using Naïve Bayes Classifier And K-Nearest Neighbor Algorithm Methods," Vol. 5, No. 5, Pp. 1213–1219, 2024, Doi: 10.52436/1.Jutif.2024.5.5.1997.

[14] H. A. Shehu *Et Al.*, "Deep Sentiment Analysis: A Case Study On Stemmed Turkish Twitter Data," *Ieee Access*, Vol. 9, Pp. 56836–56854, 2021, Doi: 10.1109/Access.2021.3071393.

[15] C. Yuan And H. Yang, "Research On K-Value Selection Method Of K-Means Clustering Algorithm," *J (Basel)*, Vol. 2, No. 2, Pp. 226–235, Jun. 2019, Doi: 10.3390/J2020016.

[16] H. Xie *Et Al.*, "Improving K-Means Clustering With Enhanced Firefly Algorithms," *Appl Soft Comput*, Vol. 84, P. 105763, Nov. 2019, Doi: 10.1016/J.Asoc.2019.105763.

[17] D. E. Sondakh, S. W. Taju, M. G. Tene, And A. E. T. Pangaila, "Sistem Analisis Sentimen Ulasan Aplikasi Belanja Online Menggunakan Metode Ensemble Learning Sentiment Analysis System For Online Shopping Application Reviews Using Ensemble Learning Method," *Cogito Smart Journal |*, Vol. 9, No. 2, 2023.

[18] H. S. Priyanka And R. Ashok Kumar, "Sentiment Analysis Using Machine Learning Based Ensemble Model For Food Reviews," *International Journal Of Innovative Research In Applied Sciences And Engineering*, Vol. 4, No. 3, Pp. 690–694, Sep. 2020, Doi: 10.29027/Ijirase.V4.I3.2020.690-694.

[19] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, And R. Budiarto, "Evaluating Trust Prediction And Confusion Matrix Measures For Web Services Ranking," *Ieee Access*, Vol. 8, Pp. 90847–90861, 2020, Doi: 10.1109/Access.2020.2994222.

[20] P. Kanani And Dr. M. Padole, "Deep Learning To Detect Skin Cancer Using Google Colab," *Int J Eng Adv Technol*, Vol. 8, No. 6, Pp. 2176–2183, Aug. 2019, Doi: 10.35940/Ijeat.F8587.088619.

[21] P. Sitikhu, K. Pahi, P. Thapa, And S. Shakya, "A Comparison Of Semantic Similarity Methods For Maximum Human Interpretability," Oct. 2019, [Online]. Available: Http://Arxiv.Org/Abs/1910.09129