Jurnal Teknik Informatika (JUTIF)
Vol. 5, No. 3, June 2024, pp. 669-674

DOI: https://doi.org/10.52436/1.jutif.2024.5.3.1366
p-ISSN: 2723-3863
e-ISSN: 2723-3871

COMPARISON PERFORMANCE OF WORD2VEC, GLOVE, FASTTEXT USING SUPPORT VECTOR MACHINE METHOD FOR SENTIMENT ANALYSIS

Margaretha Anjani[1], Helena Nurramdhani[2]

[1]Informatics, Faculty of Computer Science, Universitas Pembangunan Nasional "Veteran" Jakarta, Indonesia
[2]Information System, Faculty of Computer Science, Universitas Pembangunan Nasional "Veteran" Jakarta, Indonesia
Email: [1]margarethanjani@upnvj.ac.id, [2]helenairmanda@upnvj.ac.id

(Article received: September 02, 2023; Revision: October 07, 2023; published: May 18, 2024)

Abstract

Spotify is a digital audio service that provides music and podcasts. Reviews received by the application can affect users who will download the application. The unstructured characteristic of review text is a challenge in text processing. To produce a valid sentiment analysis, word embedding is required. The data set that is owned is divided by a ratio of 80:20 for training data and testing data. The method used for feature expansion is Word2Vec, GloVe, and FastText and the method used in classification is Support Vector Machine (SVM). The three word embedding methods were chosen because they can capture semantic, syntactic, and contextual meanings around words when compared to traditional engineering features such as Bag of Word. The best performance evaluation results show that the GloVe model produces the best performance compared to other word embeddings with an accuracy value of 85%, a precision value of 90%, a recall value of 79%, and an f1-score of 85%.

Keywords: *fasttext, glove, sentiment analysis, support vector machine, word2vec.*

1. INTRODUCTION

This current life surrounded by various technological successes which are increasingly rapidly increasing, of course, increasing human needs so as to produce a variety of the latest media. The latest media is a means of telecommunication technology with digitization and can be accessed widely which can be reached anywhere and anytime [1]. With a wide selection of audio media that can be used today it is often confusing for audio media users to find out which audio media is of high quality.

Spotify is one of the online audio media with a lot of digital enthusiasts that presents music and podcasts with a complete and new list that can be listened to both offline and online [2]. With conditions of intense competition between the Spotify application and several other online audio media, it has become a major concern in improving and perfecting an application. In connection with the presence of Google Play reviews, it can launch the views and opinions of the user's assessment of the application. Reviews are sentences or text that contain comments or ratings on the application. However, some users often find it difficult to understand actual information because the information provided by users is in text form. Where the text must be able to tell the views of the reviewer. Because of this, a lot of research has been done on machine learning to develop software that can assist with analysis.

Sentiment analysis is included in the textual data processing section which carries out analysis based on evaluations, sentiments, opinions, and behavior, as well as the emotions of a person who is a source for obtaining information about sentiments that can have positive or negative values contained in opinions [3]. Analysis, sentiment produces benefits and influences on research and applications rapidly [4]. Based on research conducted by [5] there are various variations that can be used when carrying out an analysis, one example of which is by applying the word embedding method in the analysis. The word embedding method also allows matching words that are similar based on measuring the semantic distance between the vectors embedded in the word [6]. Word embedding describes every word in the document in vector form. Where the vector represents the projection of the word in the vector space. The position of the learned word comes from the text or the words around it. Word embedding allows you to retrieve the semantic and syntactical meanings of words.

The analysis carried out is a performance comparison of the word-embedding method which will be applied to the Support Vector Machine algorithm for reviews of the use of the Spotify application on the Google Play Store. The word embedding methods used are the Word2Vec, GloVe, and FastText methods. After that, the researcher draws conclusions to prove which word embedding method has the highest levels of accuracy, precision, recall, and f1-score.

2. RESEARCH METHODS

The literature review continues throughout the data destruction process leading up to the classification and serves as a guide for conducting the research. Sources of literary studies in the form of e-books, national magazines, and some international magazines [7]. Research method can be seen in Figure 1.

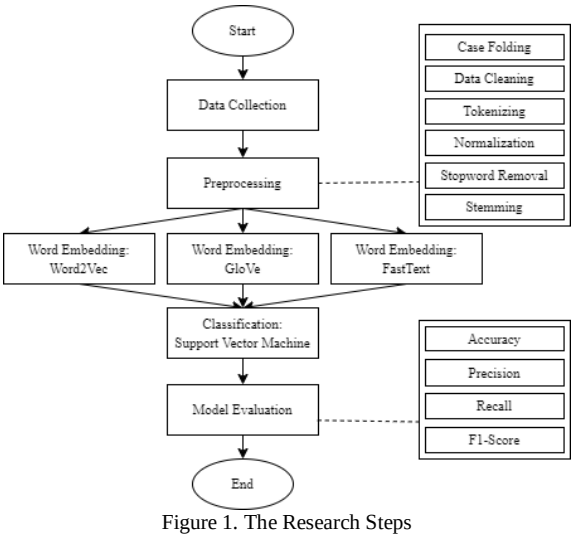

Figure 1. The Research Steps

Based on Figure 1, there are 5 (five) steps to be carried out in this reasearch, including data collection and labelling, preprocessing data, feature extraction with word embedding, split data and classification, and model evaluation. The discussion of each step is as follows:

2.1. Data Collection and Labelling

Data collection uses scraping techniques. Where web scraping is an activity of collecting data automatically from the internet by writing a program that is stored in a file with a certain format and parsing the data to extract the necessary information [8]. Labeling is done on reviews with a rating of 1, 2 and 3 will be labeled negative or a value of 0, while reviews with a rating of 4 and 5 will be labeled positive or a value of 1 [9].

2.2. Preprocessing

Text preprocessing is the process of converting unstructured data into structured data. The purpose of text preprocessing is that the data to be used has smaller dimensions, noise will be reduced, and the data will become more structured so that the data can be further processed [10]. A collection of several documents that have unstructured format can also be called corpus [11]. Text mining has several steps including text extraction using a certain technique, text preprocessing or commonly known as text processing, and text weighting or indexing, as well as text [12]. According to [13] the following are the steps of text preprocessing:

2.2.1. Case Folding

Case folding is a step to equalize all letters into lowercase letters so as not to make mistakes during the tokenizing step.

2.2.2. Data Cleaning

Data cleaning is a steps to reduce noise by eliminating or deleting all numbers, links, punctuation, and symbols as well as other than letters of the alphabet.

2.2.3. Tokenizing

Tokenizing is a step to cut or separate sentences into words for word.

2.2.4. Normalization

Normalization is a step for changing abbreviated words into standard words, the researcher makes a dictionary containing abbreviated words and standard words.

2.2.5. Stopword Removal

Stopword Removal is a step to take important words from the results of the previous tokenizing step. This step uses two techniques, including stop lists (removing unnecessary words) and word lists (saving important words).

2.2.6. Stemming

Stemming is a step to return words that have basic words by removing affixes to these words.

2.3. Feature Extraction Word Embedding

Word embedding is a method of inserting words, then converting words into continuous vector based on a predetermined length, so that they are not limited by a larger vocabulary. The word embedding model allows us to associate similar words by measuring the semantic distance between vectors embedded in these words [7].

2.3.1. Word2Vec

Word2Vec is a method that studies each word in a corpus to generate individual word vectors that depend on the local statistical word context. The vector representation has the property of a relationship with the related word during the training process [14]. The result of the vector grouping of similar words is a word dictionary which contains a collection of word similarity [15].

2.3.2. GloVe

GloVe or what can be called Global Vector is an algorithm to get a vector representation of words [16]. GloVe is a method that studies every word in a corpus

to generate vectors for individual words that depend on the context of global statistical words. Glove has a great similarity between vocabulary and other vocabulary, so this value can be used to expand the features of sentiment analysis [17].

2.3.3. FastText

FastText is a method that learns or applies n-gram characters to produce a numeric representation. Where FastText has the advantage of relatively faster processing time and is able to process words that do not appear in the dictionary (vocabulary), which in Word2Vec can cause errors [18].

2.4. Classification Support Vector Machine

Support Vector Machine (SVM) is a technique for making predictions in classification and regression [19]. SVM is a classification method using the basic concept which is a harmonious combination derived from computational theory in the previous decades. SVM itself requires positive and negative training sets to make the best decision when dividing positive data with negative data in an n-dimensional space or what is commonly called a hyperplane [20]. According by [21], Example of a hyperplane in class can be seen in Figure 2.

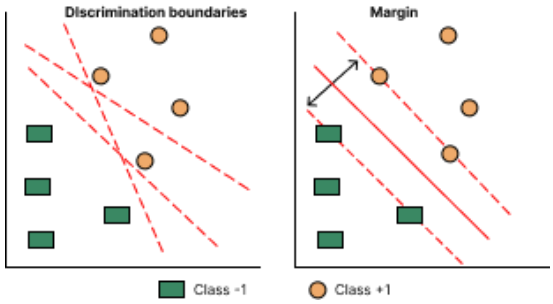

Figure 2. Hyperplane Separator

2.5. Confussion Matrix

The confusion matrix will show the total sample in the right class and the total sample in the wrong class. The True Positive (TP) and True Negative (TN) values will refer to the total sample with the class that is estimated to be correct. Whereas the False Positive (FP) and True Negative (FN) valueswill refer to the total sample with a class that will be estimated to be wrong [22]. Combinations of actual values and prediction values can be seen in Table 1.

Table 1. Confussion Matrix

		Prediction Class	
		Positive	*Negative*
Actual Class	*Positive*	*True Positive* (TP)	*False Positive* (FP)
	Negative	*True Negative* (FN)	*True Negative* (TN)

Table 1 describes the confusion matrix class which contains 4 elements in it with the provisions for setting the element values as follows:

a. True Positive (TP) is positive data that classified with positive data, so it is called true or true as positive data.
b. True Negative (TN) is negative data that classified with negative data, so it is called true or true as negative data.
c. False Positive (FP) is negative data that classified with positive data, so it is called false or wrong as positive data.
d. False Negative (FN) is positive data that classified with negative data, so it is called false or wrong as negative data.

The confusion matrix is used to calculate performance metrics, where calculate the performance of the model that has been generated. The following is an understanding of performance metrics:

1. Accuracy

Accuracy represents accuracy of the model when it classifies correctly or in other words it produces the ratio of correct predictions to all existing data. Accuracy compares the total True Positive, and True Negative to the total data.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

2. Precision

Precision represents accuracy of the data on the predicted results given by the model. Precision compares total True Positives to total True Positives and False Positives.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

3. Recall

Recall represents the success of the model in recovering information or in other words the ratio that is predicted to be positive compared to the overall positive data results. Precision compares total True Positives to total True Positives and False Negatives.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

4. F1-Score

F1-Score represents the average division of the weighted precision and recall.

$$F1 - Score = \frac{2\times P\times R}{P+R} \quad (4)$$

3. RESULTS AND DISCUSSIONS

3.1. Data Collection and Labelling

The data collection step is carried out using a scraping technique on the Google Play Store. The scraping technique is carried out using the Python programming language which uses the google-play-scraper library. The scraping technique was carried out on October 18 2022 and April 17 2023 with

review data from the Spotify application as many as 1512 of the latest reviews and in Indonesian which were saved using the csv format. The dataset was taken from 09 October 2022 to 17 October 2022 and 01 April 2023 to 16 April 2023. And stored in the form of csv data format which has columns totaling 3 variables, namely Comment, Rating, and Date. The result of scraping the Spotify application review data can be seen in Figure 3.

	Comment	Rating	Date
0	Sportify is good	5.0	2022-10-17 20:23:03
1	Lagu yg di sukai gabisa diurut? Di playlist jg...	1.0	2022-10-17 20:23:03
2	Bagus	5.0	2022-10-17 20:21:54
3	Ribet	2.0	2022-10-17 20:21:02
4	Oke punya	5.0	2022-10-17 20:19:52

Figure 3. Scraping Data Results

The collected Spotify application review data will proceed to the labeling stage. Data labeling into two positive reviews and negative reviews. The data labeling is adjusted to the rating on the application review on the Google Play Store, where ratings 1, 2, and 3 will be labeled as negative sentiment or with a value of 0, while ratings 4 and 5 will be labeled as positive sentiment or with a value of 1. In The labeling process is not only adjusted to the rating on the review, but also verified by Indonesian users who validate the assessment of a review data whether it is appropriate to include positive or negative sentences based on a subjective point of view by Indonesian users. The final results of the total labeling of the spotify application reviews can be seen in Table 2 below.

Table 2. Total Labelling

Sentiment	Negative	Positive
Total	756	756

Table 2 is the total results of labeling the review data used, there were 756 reviews labeled negative sentiment and 756 reviews labeled positive sentiment. With a total review data of 1512 reviews which were labeled manually by matching ratings and reviews.

3.2. Preprocessing

The preprocessing step is carried out in several steps to be able to remove noise and words that have no meaning. The steps to be carried out include case folding, data cleaning, tokenizing, normalization, stopword removal, and stemming as shown in Table 3.

Table 3. Preprocessing

Preprocessing	Sebelum	Sesudah
Case Folding	Kok ga bisa login sih? Padahal jaringan udah bagus pake data seluler ga bisa pake wifi juga ga bisa.. :(	kok ga bisa login sih? padahal jaringan udah bagus pake data seluler ga bisa pake wifi juga ga bisa.. :(
Data Cleaning	kok ga bisa login sih? padahal jaringan udah bagus pake data seluler ga bisa pake wifi juga ga bisa.. :(	kok ga bisa login sih padahal jaringan udah bagus pake data seluler ga bisa pake wifi juga ga bisa
Tokenizing	kok ga bisa login sih padahal jaringan udah bagus pake data seluler ga bisa pake wifi juga ga bisa	[kok, ga, bisa, login, sih, padahal, jaringan, udah, bagus, pake, data, seluler, ga, bisa, pake, wifi, juga, ga, bisa]
Normalization	[kok, ga, bisa, login, sih, padahal, jaringan, udah, bagus, pake, data, seluler, ga, bisa, pake, wifi, juga, ga, bisa]	[kok, tidak, bisa, masuk, sih, padahal, jaringan, sudah, bagus, pakai, data, seluler, tidak, bisa, pakai, internet, juga, tidak, bisa]
Stopword Removal	[kok, tidak, bisa, masuk, sih, padahal, jaringan, sudah, bagus, pakai, data, seluler, tidak, bisa, pakai, internet, juga, tidak, bisa]	[masuk, jaringan, bagus, pakai, data, seluler, pakai, internet]
Stemming	[masuk, jaringan, bagus, pakai, data, seluler, pakai, internet]	[masuk, jaring, bagus, pakai, data, seluler, pakai, internet]

3.3. Feature Extraction

The Word2Vec, GloVe, and FastText methods as word embedding at the feature extraction step use the genism library which provides Word2Vec, GloVe, and FastText calculations to form vectors containing similarity weights. The implementation uses a window with a value of 3 as the range in similarity and determines the size of the vector with a length of 1000. Then the average or mean calculation of all word vectors in the sentence will be applied to represent the embedding vector set can be seen in Table 4.

Table 4. Vector Result of Word2Vec, Glove, and FastText Models

Kata	Word2Vec	GloVe	FastText
keren	-0.000088	0.042555	-0.001348
bagus	-0.000053	-0.003743	-0.001396
baru	-0.000092	0.129566	-0.001374
putar	-0.000105	-0.015709	-0.001581
lagu	-0.000089	-0.015919	-0.001844
banget	-0.000096	-0.099364	-0.001394

3.4. Classification

The classification process uses 80% training data and 20% testing data which are distributed randomly and fairly. The classifier uses the classification_report library from sklearn. After the feature extraction modeling process using training data, the model will be tested using data testing to determine the performance of the model by classifying the test data.

3.5. Model Evaluation

Based on results of performance evaluation of three word embedding models, there is a comparison

between the word embedding model trials between Word2Vec, GloVe, and FastText. The values that will be compared from the evaluation results are accuracy, precision, recall, and f-1 score. The results of the model evaluation calculations are written in Table 5.

Table 5. Comparision Evaluation Results

No	Data	*Word Embedding*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
1	*Data Testing*	*Word2Vec*	65%	74%	44%	56%
		GloVe	85%	90%	79%	85%
		FastText	71%	76%	62%	68%
2	*Data Training*	*Word2Vec*	64%	72%	46%	57%
		GloVe	89%	92%	85%	88%
		FastText	78%	81%	73%	77%

Table 5. shows the obtained results of the comparison of the three feature extraction model experiments using the evaluation results including accuracy values, precision values, recall values and f1-score values. Of the three feature extraction model experiments, the GloVe model produced the best evaluation results on data testing and training data, then the FastText model on data testing and training data, and the last Word2Vec model on data testing and training data. The comparative description of the three word embedding models using Data Testing can be seen in Figure 4 and also the comparative using Data Training can be seen in Figure 5.

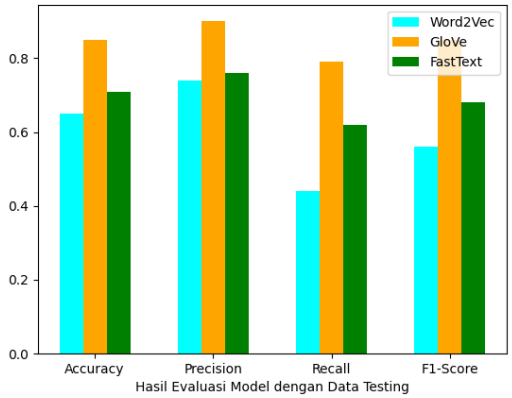

Figure 4. Graph Model Evaluation using Data Testing

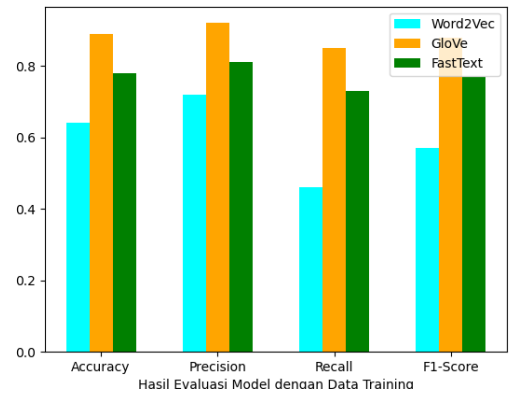

Figure 5. Graph Model Evaluation using Data Training

Based on Figure 4 and Figure 5, the comparative description of the three word embedding models that have been implemented, it can be understood that in this study, the GloVe word embedding model using the SVM algorithm has the best performance, then in the second position is the FastText word embedding model using the SVM algorithm, and the last position Word2Vec word embedding model using the SVM algorithm for processing spotify application review data. This can be proven by the increase in each evaluation value, where the GloVe model is in the first rank, the FastText model is in the second rank, and the Word2Vec model is in the third rank.

The Glove model has the highest performance evaluation score due to the efficient use of statistics, where GloVe combines global statistics to get word vectors, unlike Word2Vec and FastText which rely on local statistics to get word vectors. The second position FastText models which have the second highest performance evaluation value can be due to being an extension of the Word2Vec model which treats each word as a composition of n-gram characters. Meanwhile, the Word2Vec model is the oldest model that treats every word in the corpus to produce a vector for each word.

4. CONCLUSION

The performance evaluation results are seen from the level of accuracy, precision, recall, and f1-score applying the confusion matrix calculation, the Word2Vec method produces a accuracy of 65%, a precision of 74%, a recall of 44%, and a f1-score of 56%. The GloVe method produces a accuracy of 85%, a precision of 90%, a recall of 79%, and a f1-score of 85%. While the FastText method has an accuracy of 71%, a precision of 76%, a recall of 62%, and a f1-score of 68%. Based on the evaluation results above, the use of the GloVe method is better than the FastText and Word2Vec methods for processing the Spotify application review dataset using the SVM algorithm in this study. Based on the evaluation results between the application of the word embedding method, comparisons can be made. The GloVe method produces the highest evaluation value than the Word2Vec and FastText methods for processing spotify application review data with the SVM algorithm in this study, due to the use of efficient statistics, where GloVe combines global statistics to obtain word vectors, unlike Word2Vec and FastText which rely on local statistics to get the word vector.

REFERENCES

[1] D. McQuail, *Teori Komunikasi Massa McQuail 1, 6E* (6th ed.). Salemba Humanika, 2011.

[2] Spotify, *About Us, Spotify*, 2022 https://www.spotify.com/us/about-us/contact/ (accessed Sep 16, 2022).

[3] F. V. Sari and A. Wibowo, "Analisis Sentimen Pelanggan Toko Online Jd.Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi"

Jurnal SIMETRIS, vol. 10, no. 2, pp. 2252–4983, 2019.

[4] G. A. Buntoro, "Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter," *Integer Journal*, vol. 2, no. 1, pp. 32–41, 2017. https://t.co/jrvaMsgBdH

[5] A. Nurdin, B. A. S. Aji, A. Bustamin, and Z. Abidin, "Perbandingan Kinerja Word Embedding Word2vec, Glove, Dan Fasttext Pada Klasifikasi Teks", *Jurnal TEKNOKOMPAK*, vol. 14, no. 2, pp. 74–79, 2020.

[6] M. D. Rhman, A. Dhunaidy, and f. Mahananto, "Penerapan Weighted Word Embedding pada Pengklasifikasian Teks Berbasis Recurrent Neural Network untuk Layanan Pengaduan Perusahaan Transportasi," *JURNAL SAINS DAN SENI ITS*, vol. 10, no. 1, pp. 2337–3520, 2021.

[7] S. Fransiska, Rianto, and A. I. Gufroni, "Sentiment Analysis Provider by.U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method," *Scientific Journal of Informatics*, vol. 7, no. 2, pp. 2407–7658, 2020. http://journal.unnes.ac.id/nju/index.php/sji

[8] R. Mitchell, *Web Scraping with Python: Collecting More Data from the Modern Web* (A. MacDonald, Ed.; Second Edition). O'Reilly Media, Inc, 2018.

[9] H. Nguyen, A. Veluchamy, M. L. Diop and R. Iqbal, "Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches," *SMU Data Science Review*, vol. 1, no. 4, 2018. https://scholar.smu.edu/datasciencereviewAvailableat:https://scholar.smu.edu/datasciencereview/vol1/iss4/7http://digitalrepository.smu.edu.

[10] D. A. Fauziah, A. Maududie, and I. Nuritha, "Klasifikasi Berita Politik Menggunakan Algoritma K-nearst Neighbor (Classification of Political News Content using K-Nearest Neighbor)," *BERKALA SAINSTEK*, vol. 6, no. 2, pp. 106–114, 2018.

[11] B. Titania, *PENERAPAN METODE TEXT MINING DAN SOCIAL NETWORK ANALYSIS PADA JEJARING SOSIAL TWITTER*, 2020.

[12] A. M. Pravina, I. Cholissodin, and P. P. Adikara, "Analisis Sentimen Tentang Opini Maskapai Penerbangan pada Dokumen Twitter Menggunakan Algoritme Support Vector Machine (SVM)," *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, vol. 3, no. 3, pp. 2789–2797, 2019. http://j-ptiik.ub.ac.id

[13] E. Sonalitha, S. R. Asriningtias, and A. Zubair, *Text Mining* (Pertama). Graha Ilmu, 2021.

[14] A. S. Girsang, *Word Embedding dengan Word2vec*, 2020. https://mti.binus.ac.id/2020/11/17/word-embedding-dengan-word2vec/#:~:text=Word%20embeddings%20adalah%20proses%20konversi%20kata%20yang%20berupa,merepresentasikan%20sebuah%20titik%20pada%20space%20dengan%20dimensi%20tertentu (accessed Sep 30, 2022)

[15] H. F. Naufal and E. B. Setiawan, "Ekspansi Fitur Pada Analisis Sentimen Twitter Dengan Pendekatan Metode Word2Vec," *E-Proceeding of Engineering*, vol. 8, no. 5, pp. 10339, 2021.

[16] J. Pennington, R. Socher, and C. D. Manning, *GloVe: Global Vectors for Word Representation*, 2014. https://nlp.stanford.edu/projects/glove/ (accessed Sep 20, 2022).

[17] M. D. D. Sreya, and E. B. Setiawan, "Penggunaan Metode GloVe untuk Ekspansi Fitur pada Analisis Sentimen Twitter dengan Naïve Bayes dan Support Vector Machine," *E-Proceeding of Engineering*, vol. 9, no. 3, 2022.

[18] A. S. Girsang, *Word Embedding dengan FastText*, 2021. https://mti.binus.ac.id/2021/12/31/word-embedding-dengan-fasttext/#:~:text=Word%20embedding%20menangkap%20informasi%20semantik%20dan%20kata%20sintaksis%2C,oleh%20Facebook%20yang%20dapat%20digunakan%20untuk%20word%20embedding (accessed Sep 20, 2022).

[19] B. Santosa, *Data Mining : Terbaik Pemanfaatan Data untuk Keperluan Bisnis* (Ed. 1, Cet. 1). Graha Ilmu, 2007.

[20] I. Cholissodin, Sutrisno, A. A. Soebroto, U. Hasanah, and Y. I. Febiola, *AI, Machine Learning & Deep Learning*, 2019. https://www.researchgate.net/publication/348003841

[21] A. Pratama, *Klasifikasi Kondisi Detak Jantung Berdasarkan Hasil Pemeriksaan Elektronikardiografi (EKG) Menggunakan Binary Decision Tree-Support Vector Machin (BDT-SVM)*, 2016.

[22] K. N. Utami and E. B. Setiawan, "Ekspansi Fitur dengan FastText pada Klasisikadi Topik dengan Metode Naïve Bayes-Support Vector Machine (NBSVM) di Twitter," *E-Proceeding of Engineering*, vol. 9, no. 3, pp. 1872, 2022. https://t.co/C1SAKZKniG.