JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1109-1116
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6299

Sentiment Analysis using Random Forest and Word2Vec for Indonesian Language Movie Reviews

Fahriza Ichsani Rafif*, Mahendra Dwifebri Purbolaksono, Widi Astuti

Faculty of Informatics, Informatics, Telkom University, Bandung, Indonesia
Email: [1,*]fahrizaichsani@student.telkomuniversity.ac.id, [2]mahendradp@telkomuniversity.ac.id, [3]widiwdu@telkomuniversity.ac.id
Correspondence Author Email: fahrizaichsani@student.telkomuniversity.ac.id

Abstract−The film industry in recent years has become one of the industries that people are most interested in. The convenience of watching movies through streaming services is one of the reasons why watching movies is so popular. This ease of access resulted in a large selection of available movies and encouraged the public to look for movie reviews to find out whether the movies was good or bad. Freedom of expression on the internet has resulted in many movie reviews being spread. Therefore, sentiment analysis was conducted to see the positive or negative of these reviews. The method used in this research is Random Forest and Word2Vec skip-gram as feature extraction. The Random Forest classification was chosen because Randomforest is a highly flexible and highly accurate method, while Word2Vec Skip-Gram is used as a feature extraction because it is an efficient model that studies a large number of word vectors in an irregular text. The best model obtained from this experiment is a model built with stemming, Word2Vec with 300 dimensions, and a max_depth value of 23, achieving an f1-score of 83.59%.

Keywords: Sentiment Analysis; Random Forest; Word2Vec; Movie Review

1. INTRODUCTION

The film industry continuously produces works, leading to the presence of numerous online platforms that provide a space for movie enthusiasts to provide evaluations and comments on the movie. Online media, particularly movie review websites such as IMDb, Rotten Tomatoes, and Metacritic, have a significant influence on the industry. These movie review websites not only serve as platforms for experts to critique movie but also allow everyone to provide their evaluations and comments. These reviews are useful for movie enthusiasts to determine whether a film is worth purchasing or not [1]. Due to the abundance of data from these reviews, many studies have been conducted, including sentiment analysis.

Opinions are an important aspect of our lives as they can influence behavior and decisions. In the movie industry, movie production companies always seek to know the opinions of consumers or the public about their movies. Movie enthusiasts also want to know other peoples opinions on movie they intend to buy or watch, which is why sentiment analysis has become a supporting field of study [2]. Sentiment analysis is a field that analyzes opinions and sentiments towards a service or product in the form of text [3]. This field analyzes peoples reviews of a movie and categorizes them as positive or negative reviews.

The classification performance of sentiment analysis depends on the features present in the text, which can sometimes pose challenges that may reduce the classification performance. Therefore, feature extraction and selection play a crucial role in addressing such issues [4]. This research is expected to contribute and assist future researchers in selecting methods for data that utilize the Indonesian language, as well as helping and facilitating the public in choosing the films they want to watch. In this study, the feature used is Word2Vec. In a research [5] conducted by Serkan Ballı 2019, Word2Vec has been shown to achieve very high accuracy of 99.64%. The method used in this research is Random Forest, which has proven to be one of the most successful methods among others.

The Random Forest approach will be employed as the chosen method for this study. This method has the advantage of producing high accuracy and efficiently handling large datasets [6]. In a study [7] conducted by Isaiah Steinke 2022, Random Forest achieved an accuracy of 85.27% using a dataset consisting of movie reviews as well.

Research [8] conducted by Saeed Mian Qaisar 2020 discusses sentiment analysis of IMDb movie reviews. This experiment employed Long Short-Term Memory (LSTM), and the dataset consisted of a total of 50.000 samples obtained from the IMDb website. The dataset was split into training and testing data, each containing 25.000 samples. Within these 25.000 samples, there were 12.500 positive and 12.500 negative samples. The outcome of this testing yielded an accuracy of 89.9%. This research is good enough to be used as a reference for researchers who also use movie reviews datasets, but this study used the LSTM method while the researchers used a Random Forest.

Research [9] conducted by A.Jihad 2021 focuses on sentiment analysis of movie reviews using the Random Forest algorithm and Word2Vec feature extraction. The dataset, collected from the IMDb movie review website, comprises 50.000 data samples evenly distributed with positive and negative labels (25.000 each). This study achieves the highest accuracy of 75.76% by employing skip-gram Word2Vec with a dimension of 300 and applying Adaptive Boosting to the base model. This research is very useful for researchers to use because it uses a movie reviews dataset and also uses the same method, namely Random Forest and Word2Vec.

Research [10] conducted by W.Widayat 2021 performs sentiment analysis on movie reviews using the LSTM deep learning method and Word2Vec. The dataset consists of 25.000 movie reviews, with each review

Submitted: 09/06/2023; Accepted: 14/07/2023; Published: 23/07/2023

Volume 7, Nomor 3, Juli 2023, Page 1109-1116
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6299

containing 233 words. Both CBOW and skip-gram methods are applied to Word2Vec to generate vector representations for each word in the corpus. The word vector dimensions used are 50, 60, 100, 150, 200, and 500. The best result is obtained with a dimension of 100, achieving an accuracy of 88.17%, while the lowest result is obtained with a dimension of 500, yielding 85.86%. This study uses the movie review and Word2Vec datasets which are useful as references for researchers to use, due to the similarity in the use of the dataset which is a movie review and uses the Word2Vec method as well. but this study uses a different method, namely LSTM.

Research [11] conducted by F.W.Kurniawan 2020 discusses sentiment analysis of Indonesian tweets using Word2Vec. This study employs Support Vector Machine (SVM) as the classification method and Word2Vec for feature extraction. The dataset, collected from Twitter, consists of positive and negative sentiment tweets, which undergo preprocessing to remove noise. Word2Vec is evaluated using two architectures: CBOW and skip-gram. The results of this research include a precision of 64.4%, recall of 58%, and an f1-score of 61.1%. This research is used as a reference for datasets that use Indonesia language. This study also uses the same feature extraction, namely Word2Vec, but uses a different method, namely Support Vector Machine (SVM).

Research [12] conducted by H. Juwintho 2020 focuses on sentiment analysis of Indonesian tweets. This study utilizes deep learning with the Deep Convolutional Neural Network algorithm and Word2Vec. The dataset consists of 999 Indonesian tweets collected from Twitter. The evaluation results yield the best accuracy of 76.40%. This research also serves as the same reference as previous research, namely to become a reference for datasets that use Indonesia language. This study also uses a different method, namely the Deep Convolutional Neural Network, but uses the same feature extraction, namely Word2Vec.

2. RESEARCH METHODOLOGY

The system to be developed is a sentiment analysis of Indonesian movie reviews using the Random Forest and Word2Vec methods. The flowchart of the system to be built can be seen in figure 1:

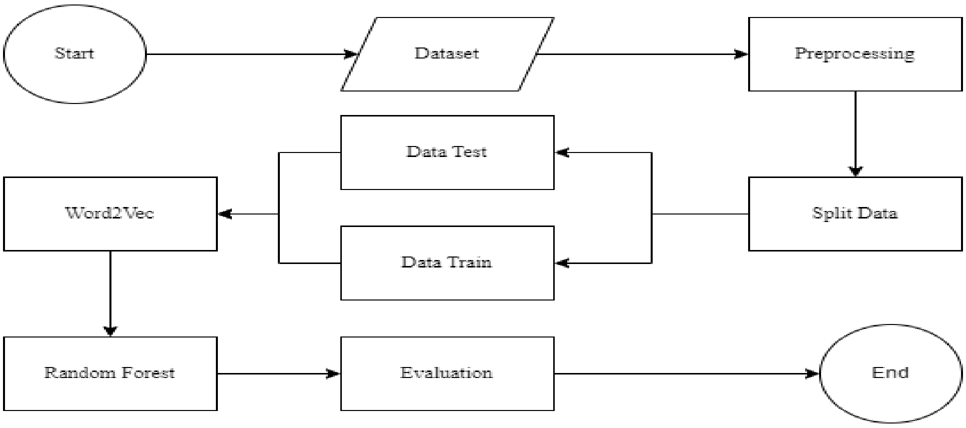

Figure 1. System Design

2.1 Dataset

The dataset to be used will be obtained from the website movfreak.blogspot.com, which is an Indonesian movie review website. The collected data consists of 5.529 reviews that have been labeled. The labeled data resulted in 2.884 positive data and 2.645 negative data. After the data preprocessing, it was discovered that the dataset contains duplicates, necessitating the removal of duplicate data. This process yielded a total of 4.208 data with 2.875 positive and 1.333 negative.

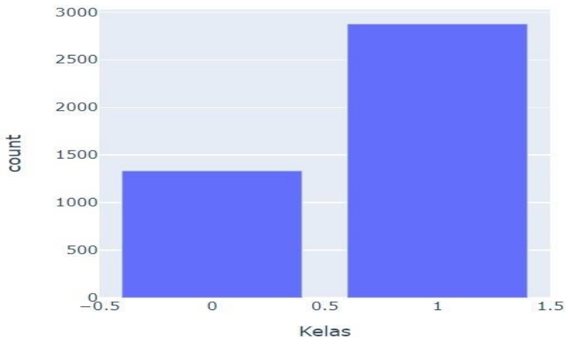

Figure 2. Dataset Distribution

Submitted: 09/06/2023; Accepted: 14/07/2023; Published: 23/07/2023

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1109-1116
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6299

2.2 Preprocessing

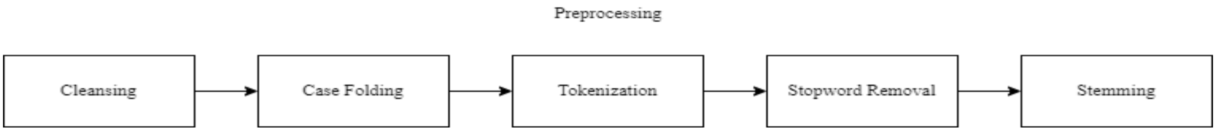

Figure 3. Preprocessing

After the data has gone through the labeling process with 2.875 positive and 1.333 negative, preprocessing will be carried out on the data. Data preprocessing is an important process for machine learning. As one of the important processes to be used in machine learning, preprocessing requires good understanding and assistance on problematic data and therefore will make the data we have more efficient and reliable [13]. The preprocessing stage in this experiment as shown in figure 3 will be divided into 5 stages, namely cleansing, case folding, tokenization, stopword removal, stemming. This stage is carried out to eliminate incomplete and inconsistent data. Data must be processed so that the data mining process in sentiment analysis functions properly.

2.2.1 Cleansing

Cleansing is a process that removes unnecessary characters in sentences that make sentences inefficient [14]. Example is presented in table 1:

Table 1. Cleansing

Text	Cleansing
“Poster nya bener bener bikin ekspetasi tinggi setinggi langit. Tapi filmnya meh. Gw tuh berharap mereka liatin megalodon sepenuhnya yg besar sebesar kapal titanic mungkin. Tapi pas nonton filmnya kita malah diliatin bayi ikan hiu.”	“Poster nya bener bener bikin ekspetasi tinggi setinggi langit Tapi filmnya meh Gw tuh berharap mereka liatin megalodon sepenuhnya yg besar sebesar kapal titanic mungkin Tapi pas nonton filmnya kita malah diliatin bayi ikan hiu”

2.2.2 Case Folding

Case folding involves transforming all uppercase letters within the data into lowercase letters, resulting in improved structure and increased accuracy of the data [15]. Example is presented in table 2:

Table 2. Case Folding

Text	Case Folding
“Poster nya bener bener bikin ekspetasi tinggi setinggi langit Tapi filmnya meh Gw tuh berharap mereka liatin megalodon sepenuhnya yg besar sebesar kapal titanic mungkin Tapi pas nonton filmnya kita malah diliatin bayi ikan hiu”	“poster nya bener bener bikin ekspetasi tinggi setinggi langit tapi filmnya meh gw tuh berharap mereka liatin megalodon sepenuhnya yg besar sebesar kapal titanic mungkin tapi pas nonton filmnya kita malah diliatin bayi ikan hiu”

2.2.3. Tokenization

Tokenization will transform a sentence into tokens based [15]. Example is presented in table 3:

Table 3. Tokenization

Text	Tokenization
“poster nya bener bener bikin ekspetasi tinggi setinggi langit tapi filmnya meh gw tuh berharap mereka liatin megalodon sepenuhnya yg besar sebesar kapal titanic mungkin tapi pas nonton filmnya kita malah diliatin bayi ikan hiu”	“[poster], [nya], [bener], [bener], [bikin], [ekspetasi], [tinggi], [setinggi], [langit], [tapi], [filmnya], [meh], [gw], [tuh], [berharap], [mereka], [liatin], [megalodon], [sepenuhnya], [yg], [besar], [sebesar], [kapal], [titanic], [mungkin], [tapi], [pas], [nonton], [filmnya], [kita], [malah], [diliatin], [bayi], [ikan], [hiu]”

2.2.4 Stopword Removal

Stopword removal is used to delete or modify words or sentences that have little or no impact or influence [15]. Example is presented in table 4:

Table 4. Stopword Removal

Text	Stopword removal
“[poster], [nya], [bener], [bener], [bikin], [ekspetasi], [tinggi], [setinggi], [langit], [tapi], [filmnya], [meh], [gw],	“[poster], [nya], [bener], [bener], [bikin], [ekspetasi], [langit], [filmnya], [meh], [gw],

Submitted: 09/06/2023; Accepted: 14/07/2023; Published: 23/07/2023

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1109-1116
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6299

Text	Stopword removal
[tuh], [berharap], [mereka], [liatin], [megalodon], [sepenuhnya], [yg], [besar], [sebesar], [kapal], [titanic], [mungkin], [tapi], [pas], [nonton], [filmnya], [kita], [malah], [diliatin], [bayi], [ikan], [hiu]"	[tuh], [berharap], [liatin], [megalodon], [sepenuhnya], [yg], [kapal], [titanic], [pas], [nonton], [filmnya], [diliatin], [bayi], [ikan], [hiu]"

2.2.5 Stemming

Stemming is a stage that converts a word into its base form [15]. Example is presented in table 5:

Table 5. Stemming

Text	Stemming
"[yg], [mengganggu], [gue], [keputusan], [keputusan], [bodoh], [karakternya], [nginep], [rumah], [tua], [yg], [udah], [hancur], [ga], [rumah], [warga], [si], [ario], [bayu], [si], [asmara], [abigail], [minimal], [adegan], [yg], [effort], [utk], [cari], [penginapan], [yg], [layak], [rumah], [warga], [minimal], [diskusi], [mah], [ga], [langsung], [nginep], [aja], [rumah], [bobrok], [yg], [ga], [layak], [gitu], [ga], [ngeri], [makhluk], [halus], [penjahat], [aneh], [bgt], [si], [marissa], [ngaku], [rahayu], [keputusan], [konyol], [flashback], [udah], [melucuti], [separuh], [nyawa], [film]"	"yg ganggu gue keputusan keputusan bodoh karakter nginep rumah tua yg udah hancur ga rumah warga si ario bayu si asmara abigail minim adegan yg effort utk cari nginap yg layak rumah warga minim diskusi mah ga langsung nginep aja rumah bobrok yg ga layak gitu ga ngeri makhluk halus penjahat aneh bgt si marissa ngaku rahayu keputusan konyol flashback udah lucut separuh nyawa film"

2.3 Split Data

After completing the preprocessing stage, The subsequent action is to carry out data splitting on the dataset. During this phase, the data will be split, allocating 80% for training data and 20% for testing data. The results after the data split is presented in table 6:

Table 6. The Distribution of The Split Data

Split Data	Data Train	Data Test
Total Data	3366	842
Positive	2273	602
Negative	1093	240

2.4 Word2Vec

After completing data splitting, the next step is feature extraction using Word2Vec. Word2Vec is one of many word embeddings. Word embedding is a natural language processor learning technique which describes a word as a vector. Word2Vec describes a word as a vector based on several features it has, including the use of dimensions in vectors [16]. Word2Vec has 2 types of models, namely Skip-Gram models and Continuous Bag of Words (CBOW) models. the two models have different functions. Skip-Gram is an efficient model that studies a large number of word vectors in an irregular text, while the Continuous Bag of Words (CBOW) models have a function to predict the target word based on the context of the surrounding words [17]. Among the two available methods, skip-gram is chosen for this research. The formula for this method is as follows:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\ne 0}\log p(w_{t+j}|w_t) \quad (1)$$

Explanation:
w_t = word center
w_{t+j} = word after word center
c = size of the training context

2.5 Random Forest Classification

After feature extraction using Word2Vec is done, the next step is the classification process. The method used is Random Forest. Random forest is a highly flexible and highly accurate method. It has the capability to deal with both regression and classification tasks, including tasks involving multiple classes. It exhibits a relatively high speed in both training and generating predictions. Additionally, Random Forest is algorithm that can effectively handle sizable datasets containing noise and high dimensionality while also providing insights into the significance of each variable during the classification process. It is robust against statistical assumptions and does not necessitate extensive preprocessing. Additionally, Random Forest reduced variability when compared to a single decision tree. Moreover, Random Forest is more robust compared to boosting methods. Importantly, Random Forest highlights several crucial capabilities, including the ability to address overfitting, apply differential class

Volume 7, Nomor 3, Juli 2023, Page 1109-1116
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6299

weighting, detect outliers, and handle missing values effectively. [18]. An illustration of the workflow of Random Forest is presented in figure 4:

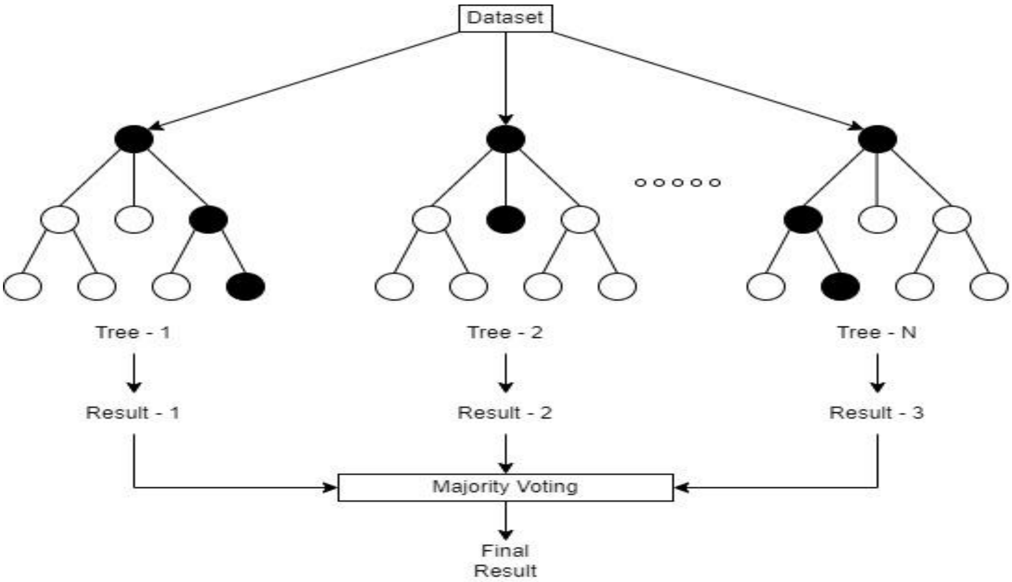

Figure 4. Random Forest Model

2.6 Evaluation

After the data has completed the preprocessing phase and data split, then it goes through the feature extraction and classification processes, the system will be evaluated using a confusion matrix. Confusion matrix is a stage that useful for measuring the systems effectiveness by examining the prediction results of the built classification system [19]. Table of confusion matrix information is presented in table 7:

Table 7. Confusion Matrix

Confusion Matrix		**Predicted Class**	
		Positive	**Negative**
Actual Values	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Description
True Positive = correctly predicted as positive
True Negative = correctly predicted as negative
False Positive = incorrectly predicted as positive
False Negative = incorrectly predicted as negative

In this experiment the confusion matrix will calculate the f1-score, precision, and recall. The following are the respective formulas for f1-score, precision, recall:

The f1-score can be described as the balanced average of precision and recall, calculated as the harmonic mean. It offers a fair assessment of both precision and recall simultaneously and is particularly useful when you want to find an optimal trade-off between the two metrics [20]. Below is the equation representing the f1-score:

$$F1 - Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (2)$$

Precision quantifies the proportion of accurately predicted positive instances to the overall instances predicted as positive [20]. The folowing is the formula for the precision:

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

Recall measures the proportion of accurately predicted positive cases compared to the overall number of positive cases. The following is the formula for the recall:

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

3. RESULT AND DISCUSSION

This study utilizes movie review information sourced from movfreak.blogspot.com website, which has gone through the preprocessing stage, resulting in a total of 4208 datasets. These datasets will be divided into train data

Submitted: 09/06/2023; Accepted: 14/07/2023; Published: 23/07/2023

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1109-1116
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6299

and test data with 80:20 ratio, generating 3366 train data and 842 test data. Subsequently, feature extraction will be conducted using the Word2Vec. The chosen Word2Vec model for this study is the skip-gram model. Once the data has been split and features have been extracted, a Random Forest algorithm will be employed for classification purposes. After all the processes have been done several scenarios will be conducted. The first scenario involves testing with and without stemming to observe the impact of stemming on the built system. The second scenario involves testing the Word2Vec feature extraction by selecting different dimensions to evaluate the performance of feature extraction in the system. The third scenario focuses on optimizing the Random Forest using the max_depth parameter. The scenario information presented in table 8:

Table 8. Scenario

Scenario	Experiment
Scenario 1	Testing with and without stemming to observe the impact of stemming on the built system
Scenario 2	Testing the Word2Vec feature extraction by selecting different dimension to evaluate the performance of feature extraction in the system
Scenario 3	Optimizing the Random Forest using the max_depth parameter

3.1 Comparison Stemming and Without Stemming

In the first scenario, there will be a test conducted to observe the consequences of utilizing stemming compared to not utilizing stemming. Before getting the results from testing this scenario, the data has been analyzed utilizing Word2Vec with 300 dimensions and employing the Random Forest classification method to process the data. The results of this scenario is presented in table 9:

Table 9. Comparison Result

Preprocessing	Precision	Recall	F1 Score
With Stemming	**72.96%**	**96.57%**	**83.12%**
Without Stemming	72.36%	96.40%	82.67%

According to the evaluation results, the inclusion of stemming has a minimal effect on the performance of the classification model. In this case, with stemming, the model achieved a precision of 72.96%, recall of 96.57%, and f1-score of 83.12%. On the other hand, without stemming, the model achieved a precision of 72.36%, recall of 96.40%, and f1-score of 82.67%. Although there is a small difference between the two approaches, the difference is not significant. Precision gauges the correctness of positive predictions, recall evaluates the models capacity to capture positive instances, and the f1-score offers a trade-off between precision and recall. In both cases, with or without stemming, the model demonstrated good performance with high recall values, indicating the ability to correctly identify the majority of positive instances. However, precision and f1-score were slightly lower, indicating some cases of false positive or incorrect positive predictions. In this case, the use of stemming can help reduce the variation of similar words in the text, allowing the model to identify more general patterns and improve recall. However, stemming can also remove specific suffix information, which affects precision. Although the performance difference between the two approaches is not significant, the decision to use stemming enough to affect the results of the f1-score to be higher, namely 83.12%. This proves research [9] that the use of stemming results in a higher f1-score compared to without stemming.

3.2 The Effect of Dimension in Word2Vec

In scenario 2, a test will be carried out to see the effect of the Word2Vec dimension on the system. The dimensions to be used are dimensions 100 and 300. The evaluation of these dimensions will be conducted by examining the data that has been subjected to stemming during preprocessing, along with the utilization of the Random Forest classification method. The outcomes of the second testing scenario are presented in table 10:

Table 10. Dimension Comparison in Word2Vec

Dimension Word2Vec	Precision	Recall	F1 Score
300 Dimension	**72.96%**	**96.57%**	**83.12%**
100 Dimension	72.17%	96.40%	82.55%

Based on the test results in the table above, the 300 dimensions Word2Vec model, it achieved a precision of 72.96%, recall of 96.57%, and f1-score of 83.12%. On the other hand, the 100 dimensions Word2Vec model achieved a precision of 72.17%, recall of 96.40%, and f1-score of 82.55%. Comparing the two dimensions, there is a slight difference in the performance metrics. The 300 dimensions model shows a slightly higher precision, recall, and f1-score compared to the 100 dimensions model. This suggests that the 300 dimensions representation captures more nuanced semantic information and helps improve the models ability to correctly classify instances. It should be emphasized that the disparity in performance between the two aspects is not substantial. Both models

Submitted: 09/06/2023; Accepted: 14/07/2023; Published: 23/07/2023

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1109-1116
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6299

demonstrate good performance with high recall values, indicating their ability to identify positive instances accurately. The slight variation in precision and f1-score may be attributed to the trade-off between precision and recall in the classification process. In summary, the evaluation results indicate that the Word2Vec model with stemming and 300 dimensions exhibits slightly improved performance in precision, recall, and f1-score compared to the 100 dimensions model. This proves research [9] that the use of Word2Vec 300 dimension model results in a higher f1-score compared to Word2Vec 100 dimension models.

3.3 Random Forest Optimalization Using max_depth Parameter

In scenario 3, Random Forest optimalization will be carried out using the max_depth parameter. In the context of Random Forest algorithm, the max_depth parameter refers to the maximum depth of each individual decision tree in the forest. By setting the max_depth parameter, system can limit the depth of the trees in order to prevent overfitting and improve generalization. The range of values to be used for max_depth in this experiment is from 1 to 50. This scenario will use the results from the previous scenario, namely Word2Vec with dimension 300 which has gone through the stemming process, this is because Word2Vec with dimension 300 which has gone through the stemming process gets the most f1-score high among others, namely 83.12% and this third scenario will try to increase the f1-score. The test results for scenario 3 are visually represented in figure 5:

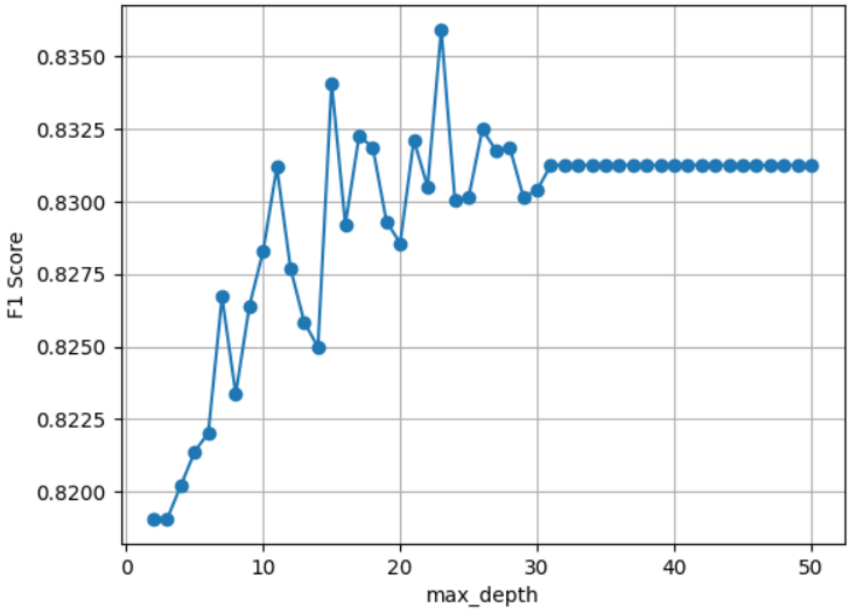

Figure 5. Performance Graph of max_depth

According to the information presented in Figure 5, it is noticeable that the f1-score demonstrates an upward trend as the max_depth value rises. This positive trend continues until it reaches its peak at max_depth = 23, with an f1-score of 83.59%. After that point, the f1-score stabilizes as max_depth is further increased up to 50. It can be concluded that increasing the value of max_depth generally results in improved model performance. However, after reaching max_depth = 23, further increases do not yield significant improvements in the f1-score. This indicates that the model is already sufficiently complex at max_depth = 23 to learn the patterns from the available data.

4. CONCLUSION

Based on the conducted testing on Sentiment Analysis using Random Forest and Word2Vec on Indonesian Movie Reviews, the following conclusions can be drawn. The built system will undergo three scenarios, starting with evaluating the impact of stemming on data processed using Word2Vec with a dimension of 300 and using a Random Forest classification, comparing it with data processed without stemming. The second scenario involves examining how the dimension of Word2Vec impacts data that has undergone stemming during the preprocessing phase. The dimensions to be used are dimensions 100 and 300. The third scenario is random forest optimalization using the max_depth parameter for data that uses Word2Vec with a dimension of 300 and has gone through a stemming process. Based on the obtained results, stemming has an effect on the performance of the developed system. In scenario 1, stemming is shown to produce higher performance compared to without stemming. The dimension in Word2Vec also influences the systems performance. Scenario 2 demonstrates that the 300 dimensions Word2Vec outperforms the 100 dimensions. The max_depth parameter also affects the systems performance. Scenario 3 shows that higher values of max_depth generally improve the model performance, and in this study, the highest result was achieved with max_depth = 23. In conclusion, the best results are obtained by

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 3, Juli 2023, Page 1109-1116
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v7i3.6299

the model built using stemming, a 300 dimensions Word2Vec, and max_depth = 23, achieving an f1-score of 83.59%.

REFERENCE

[1] A. Andreyestha and A. Subekti, "Analisa Sentiment Pada Ulasan Film Dengan Optimasi Ensemble Learning," J. Inform., vol. 7, no. 1, pp. 15–23, 2020, doi: 10.31311/ji.v7i1.6171.
[2] S. Bhatia, M. Sharma, and K. K. Bhatia, "Sentiment Analysis and Mining of Opinions," Stud. Big Data, vol. 30, no. May, pp. 503–523, 2018, doi: 10.1007/978-3-319-60435-0_20.
[3] S. Mukherjee, "Sentiment Analysis," ML.NET Reveal., pp. 113–127, 2021, doi: 10.1007/978-1-4842-6543-7_7.
[4] J. Khan, A. Alam, and Y. Lee, "Intelligent Hybrid Feature Selection for Textual Sentiment Classification," IEEE Access, vol. 9, pp. 140590–140608, 2021, doi: 10.1109/ACCESS.2021.3118982.
[5] S. Ballı and O. Karasoy, "Development of content-based SMS classification application by using Word2Vec-based feature extraction," IET Softw., vol. 13, no. 4, pp. 295–304, 2019, doi: 10.1049/iet-sen.2018.5046.
[6] A. Ramadhan, B. Susetyo, and Indahwati, "Penerapan Metode Klasifikasi Random Forest Dalam Mengidentifikasi Faktor Penting Penilaian Mutu Pendidikan," J. Pendidik. dan Kebud., vol. 4, no. 2, pp. 169–182, 2019, doi: 10.24832/jpnk.v4i2.1327.
[7] I. Steinke, J. Wier, L. Simon, and R. Seetan, "Sentiment Analysis of Online Movie Reviews using Machine Learning," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 9, pp. 618–624, 2022, doi: 10.14569/IJACSA.2022.0130973.
[8] S. M. Qaisar, "Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory," 2020 2nd Int. Conf. Comput. Inf. Sci. ICCIS 2020, no. November 2020, 2020, doi: 10.1109/ICCIS49240.2020.9257657.
[9] M. A. A. Jihad, Adiwijaya, and W. Astuti, "Analisis sentimen terhadap ulasan film menggunakan algoritma random forest," e-Proceeding Eng., vol. 8, no. 5, pp. 10153–10165, 2021.
[10] W. Widayat, "Analisis Sentimen Movie Review menggunakan Word2Vec dan metode LSTM Deep Learning," J. Media Inform. Budidarma, vol. 5, no. 3, p. 1018, 2021, doi: 10.30865/mib.v5i3.3111.
[11] F. W. KURNIAWAN, "Analisis Sentimen Twitter Bahasa Indonesia dengan Word2Vec," Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 2, no. 2, pp. 4704–4713, 2020, [Online]. Available: https://openlibrary.telkomuniversity.ac.id/home/catalog/id/159923/slug/analisis-sentimen-twitter-bahasa-indonesia-dengan-word2vec.html%0A/home/catalog/id/159923/slug/analisis-sentimen-twitter-bahasa-indonesia-dengan-word2vec.html
[12] H. Juwintho, E. Setiawan, and J. Santoso, "Sentiment Analysis Twitter Bahasa Indonesia Berbasis WORD2VEC Menggunakan Deep Convolutional Neural Network," J. Teknol. Inf. dan Ilmu Komput., vol. 7, no. 1, pp. 181–188, 2020, doi: 10.25126/jtiik.202071758.
[13] E. A. Felix and S. P. Lee, "Systematic literature review of preprocessing techniques for imbalanced data," IET Softw., vol. 13, no. 6, pp. 479–496, 2019, doi: 10.1049/iet-sen.2018.5193.
[14] D. J. Putri and M. Dwifebri, "Text Classification of Indonesian Translated Hadith Using XGBoost Model and Chi-Square Feature Selection," vol. 4, no. 4, pp. 1732–1738, 2023, doi: 10.47065/bits.v4i4.2944.
[15] I. Prayoga and M. D. P, "Sentiment Analysis on Indonesian Movie Review Using KNN Method With the Implementation of Chi-Square Feature Selection," vol. 7, pp. 369–375, 2023, doi: 10.30865/mib.v7i1.5522.
[16] S. Al-Saqqa and A. Awajan, "The Use of Word2vec Model in Sentiment Analysis: A Survey," ACM Int. Conf. Proceeding Ser., no. December, pp. 39–43, 2019, doi: 10.1145/3388218.3388229.
[17] R. P. Nawangsari, R. Kusumaningrum, and A. Wibowo, "Word2vec for Indonesian sentiment analysis towards hotel reviews: An evaluation study," Procedia Comput. Sci., vol. 157, pp. 360–366, 2019, doi: 10.1016/j.procs.2019.08.178.
[18] T. Zhu, "Analysis on the applicability of the random forest," J. Phys. Conf. Ser., vol. 1607, no. 1, 2020, doi: 10.1088/1742-6596/1607/1/012123.
[19] F. Rahmad, Y. Suryanto, and K. Ramli, "Performance Comparison of Anti-Spam Technology Using Confusion Matrix Classification," IOP Conf. Ser. Mater. Sci. Eng., vol. 879, no. 1, 2020, doi: 10.1088/1757-899X/879/1/012076.
[20] F. Khairani, A. Kurnia, M. N. Aidi, and S. Pramana, "Predictions of Indonesia Economic Phenomena Based on Online News Using Random Forest," SinkrOn, vol. 7, no. 2, pp. 532–540, 2022, doi: 10.33395/sinkron.v7i2.11401.

Submitted: **09/06/2023**; Accepted: **14/07/2023**; Published: **23/07/2023**