KHATULISTIWA SMART : JOURNAL OF ARTIFICIAL INTELLIGENCE
e-ISSN : XXXX-XXXX
p-ISSN : XXXX-XXXX

Khatulistiwa Smart : Journal of Artificial Intelligence
Vol. 1 No. 1 (2024) : 17-25
Available online at : https://journal.literasikhatulistiwa.org/index.php/smart

Sentiment Analysis of Movie Reviews Using Spark on IMDB Review Dataset

Luthfi Nurul Huda[1]
[1] Magister Teknik Informatika Universitas AMIKOM Yogyakarta
luthfinurulhuda@students.amikom.ac.id[1]

Doi
Received: 23 November 2024 *Accepted: 15 Desember 2024* *Published: 31 Desember 2024*

Abstract Analisis sentimen pada ulasan film telah menjadi topik penting dalam penelitian berbasis teks, terutama untuk mendeteksi polaritas sentimen seperti positif, negatif, atau netral. Penelitian ini mengevaluasi kinerja dua algoritma, Support Vector Machine (SVM) dan Logistic Regression (LR), dalam mengklasifikasikan ulasan film dengan menggunakan dataset IMDb yang tersedia untuk umum di Kaggle. Data tersebut dibagi menjadi 80% untuk pelatihan dan 20% untuk pengujian, melalui tahapan preprocessing seperti case folding, tokenisasi, penghilangan stop words, stemming, dan ekstraksi fitur dengan menggunakan Word2Vec. Hasil evaluasi menunjukkan bahwa SVM memiliki akurasi sebesar 76%, mengungguli LR yang mencapai 69%. Keunggulan SVM terletak pada kemampuannya untuk menemukan hyperplane yang optimal dalam ruang berdimensi tinggi, yang sesuai dengan sifat data teks yang jarang dan berbentuk vektor. Sebaliknya, meskipun LR lebih sederhana dan lebih cepat untuk dilatih, model ini menunjukkan kinerja yang lebih rendah karena keterbatasannya dalam menangani hubungan non-linear. Preprocessing terbukti memberikan kontribusi yang signifikan dalam meningkatkan kualitas data input, sementara representasi Word2Vec memberikan fitur-fitur yang berarti untuk mendukung analisis sentimen. Penelitian ini menggarisbawahi pentingnya memilih algoritma yang tepat untuk analisis sentimen berbasis data besar, dengan hasil yang menunjukkan bahwa SVM lebih unggul dalam menangani data teks berskala besar. Penelitian ini berkontribusi dalam memahami efektivitas metode pembelajaran mesin dalam analisis sentimen ulasan film, sekaligus memberikan dasar untuk penelitian di masa depan yang dapat memperluas metode dan set data ke domain lain.

Keywords *fraud detection; credit card; random forest; XGBoost; machine learning*

1. INTRODUCTION

Sentiment analysis is a field of science that has great potential in research and practical applications. Helping to explain that sentiment analysis research, especially text, is increasingly attracting attention both at home and abroad. (Tan et al., 2023). This is an important task for detecting sentiment polarity in text, which is widely applied in e-commerce systems, movie reviews, blogs, and social media (Resyanto et al., 2019), (Mee et al., 2021). In general, sentiment analysis is a process or way to evaluate the extent to which written or spoken language is determined to be a positive, negative, or neutral expression (Mee et al., 2021). Currently, machine-learning is the main method applied to sentiment analysis. As a reference whether sentiment analysis gets good results. There have been many studies related to sentiment analysis, to get conclusions based on data processed or obtained from various text data sources (Pohan et al., 2022).

Big data is a term for data of enormous size and variety. This data also continues to grow very quickly. The three main things that characterize it are the massive amount of data, the speed at which the data is moving, and the wide variety of data. The combination of these three characteristics often makes it challenging to effectively analyze big data (Kurniawan & Mustikasari, 2022). Therefore, a reliable machine learning framework, strategy, and environment are needed so that data analysis can be done more optimally and accurately (Vindua & Zailani, 2023). Not all machine learning tools are capable of handling large data that can be loaded directly into memory on a single computer. One tool that is often used to process data at scale is Apache Hadoop. This tool is designed to support data analytics needs, but it has some important drawbacks. Apache Hadoop generates high overhead when running processes, and relies heavily on storing data and computation results on disk. As a result, Hadoop is less than ideal for use in cases that require iterative processes or low latency (Tumbel et al., 2017).

Apache Spark is a modern framework designed specifically for distributed computing. It is built to optimize tasks that require low latency. With its ability to store data and processing results directly in memory, Apache Spark is an ideal choice for iterative and machine learning applications (Ananda Lubis et al., 2024). Apache Spark comes with an open-source library designed for large-scale data analysis, the Machine Learning Library (MLlib). MLlib provides various features to support machine learning tasks, such as regression, classification, clustering, and rule extraction, making it easier to process data efficiently. The research community has long studied how to apply machine learning effectively. However, the study of machine learning libraries for big data, such as Apache Spark's MLlib, is still limited to date. In addition, not much research has been done to date regarding in-memory data processing such as Apache Spark's MLlib. This research will analyze IMDB Riview movie reviews whose data is quite large. Therefore, researchers will experiment only using the library provided by Apache Spark. Several stages are a challenge in doing sentiment analysis. As done by (Kurniawan & Mustikasari, 2022), The main challenge raised in this research is how to classify fake news in Indonesian with high accuracy and efficiency. Previous research shows that many machine learning algorithms and tools, such as R and Weka, are not optimal for handling large data volumes, mainly due to memory limitations. In addition, the evaluation of big data libraries such as Apache

Spark's MLlib for text classification in Indonesian is still very limited. This study aims to evaluate the performance of MLlib Apache Spark in classifying fake news in Indonesian. The evaluation was conducted using four algorithms, namely Naïve Bayes, Gradient-Boosted Tree (GBT), Support Vector Machine (SVM), and Logistic Regression. The dataset used in this research comes from TurnBackHoax.id, a portal managed by Masyarakat Anti Hoax Indonesia (MAFINDO). The results show that the Naïve Bayes algorithm has the highest accuracy of 83.3% with an F1-score of 0.834, although it takes 16.3 seconds to process the data. Linear Regression recorded the fastest running time, which was 6.46 seconds, with 82% accuracy and F1-score of 0.819. Support Vector Machine gave an accuracy result of 81.8% with a processing time of 8.36 seconds. Meanwhile, Gradient-Boosted Tree yielded 80.2% accuracy and the longest processing time of 5 minutes 15 seconds, due to the iterative nature of the algorithm. Overall, this research shows that MLlib Apache Spark is able to provide quite good results for the classification of fake news in Indonesian, with a combination of adequate speed and accuracy. The next research conducted by (Kurniawan & Mustikasari, 2022) This article discusses the COVID-19 pandemic that has led to various public responses to face-to-face learning policies, ranging from support, objection, to neutral attitudes. However, unstructured public opinion data, large numbers, and the use of informal language are challenges in sentiment analysis. To overcome this, the research uses the Linear Regression method implemented through Apache Spark MLlib and Spark NLP. The results show an accuracy rate of 90.04%, with the distribution of community sentiment as follows: 47.7% neutral sentiment, 31.4% positive sentiment, and 20.9% negative sentiment. This result reflects that the majority of people have a neutral opinion towards the face-to-face learning policy, although there are some who support and reject the policy. These findings provide valuable insights for the government to understand public opinion and as a consideration in designing education policies during the pandemic.

The third research conducted by (Wahyudi, 2018) This study aims to build a personalized movie recommendation system that can overcome data sparsity, cold start, and scalability problems using MLlib Apache Spark. In addition, this study aims to evaluate the effectiveness of the ALS-WR and Cosine Similarity algorithms in generating relevant recommendations based on user preferences. The dataset used comes from movielens.org with three different sizes: 100K, 1M, and 10M. This dataset includes information such as movie title, genre, rating, and user interaction, and is used to train and test the recommendation model. The method used in this research is ALS-WR (Alternating Least Square-Weight Regularization) The process includes data collection, preprocessing (case folding, tokenization, and stopword removal), model training with data division (60% training, 20% validation, and 20% testing), and evaluation using RMSE, precision, and user acceptance test metrics. Data is processed using MLlib Apache Spark to ensure processing efficiency and speed. With the results of RMSE (Root Mean Square Error) shows the prediction error rate. As a result, the 100K dataset produces RMSE 0.96 (validation) and 0.94 (test); the 1M dataset produces RMSE 0.86 (validation) and 0.96 (test); the 10M dataset produces RMSE 0.81 (validation and test).

Based on the background that explains some of the challenges in sentiment analysis using spark, this research will raise how spark uses two algorithms including Support Vector

Machine (SVM) and Linear Regression (LR) to perform sentiment analysis of IMDB Review movie reviews. Based on previous studies such as research by (Wahyudi, 2018) describes the influence of Apache Spark to ensure efficiency and processing speed in large-scale datasets, as well as research conducted by (Kurniawan & Mustikasari, 2022) shows that MLlib Apache Spark is able to provide quite good results for the classification of fake news in Indonesian, with a combination of adequate speed and accuracy. So this research will use spark by comparing two algorithms, Support Vector Machine (SVM) and Linear Regression (LR) to measure the effect of spark in terms of speed in processing machine learning. Secondly, the evaluation results on the two algorithms can be taken into consideration to choose the right algorithm for text classification.

2. RESEARCH METHODS

2.1. Type and Nature of Research

This research is a type of empirical research, which means that this research is based on real data taken from previous research. This data is used to test the hypothesis that stemming and feature extraction using Word2vec can improve the performance of SVM and LR algorithms in analyzing lexicon-based sentiment. This research has experimental characteristics, because there is manipulation of the independent variables, namely the use of stemmers and feature extraction techniques. This manipulation aims to observe the extent of the influence of these two variables on the dependent variable, namely the accuracy and performance of SVM and LR algorithms in sentiment analysis tasks.

On the other hand, this research is also descriptive in nature. The data analysis results obtained were not only tested through variable manipulation, but also compared with previous studies to understand the differences or improvements achieved in the accuracy of SVM and LR in sentiment analysis. Thus, this approach allows researchers to gain in-depth insights into the effectiveness of using stemming and feature extraction, both in the context of their direct influence on SVM and LR and in comparison with previous approaches in similar studies.

2.2. Data Collection Methods

The data sources needed in this research were obtained from two data collection methods as follows:

2.2.1. Literature Review

This research involves data collection by analyzing previous research papers that have relevance to the topics raised, especially sentiment analysis. The data obtained can be in the form of public data available on publication media, public repositories and so on.

2.2.2. Secondary Data Analysis

This involves using data that has been collected by previous researchers. The data obtained can be in the form of archives that are publicly available and allowed to be used, usually found on GITHUB repositories, drivers and so on.

To conduct experiments in this study, researchers used publicly available datasets on Kaggle from the results of literature reviews on several previous papers.

2.3. Data Analysis Method

This research is conducted through several stages of data analysis methods to determine the impact of stemmer and feature extraction on the performance of SVM and LR algorithms in the sentiment analysis process. The first step is to perform data preprocessing using Apache Spark, which includes case folding, cleaning, tokenizing, stopwords removal, and then stemming. This stemming process uses PorterStemmer which is implemented through a library supported by Spark. The next step is data classification using the lexicon method to measure the polarity of the data so that it can provide a class label in the form of positive, negative, or neutral which will be used as the Y value in the SVM and LR algorithms. The third step is feature extraction implemented using Word2Vec with Spark MLlib, which will be used as an independent variable in the SVM and LR algorithms. The last step is to train and evaluate the SVM and LR algorithms using Spark MLlib by calculating accuracy, precision, recall, and F1-Score on several combinations of stemmer and feature extraction to determine the combination that produces the best performance of both.

2.4. Alur Penelitian

To complete this research requires steps that must be completed. The steps or stages carried out in this study can be seen in Figure 1 which is described as follows:

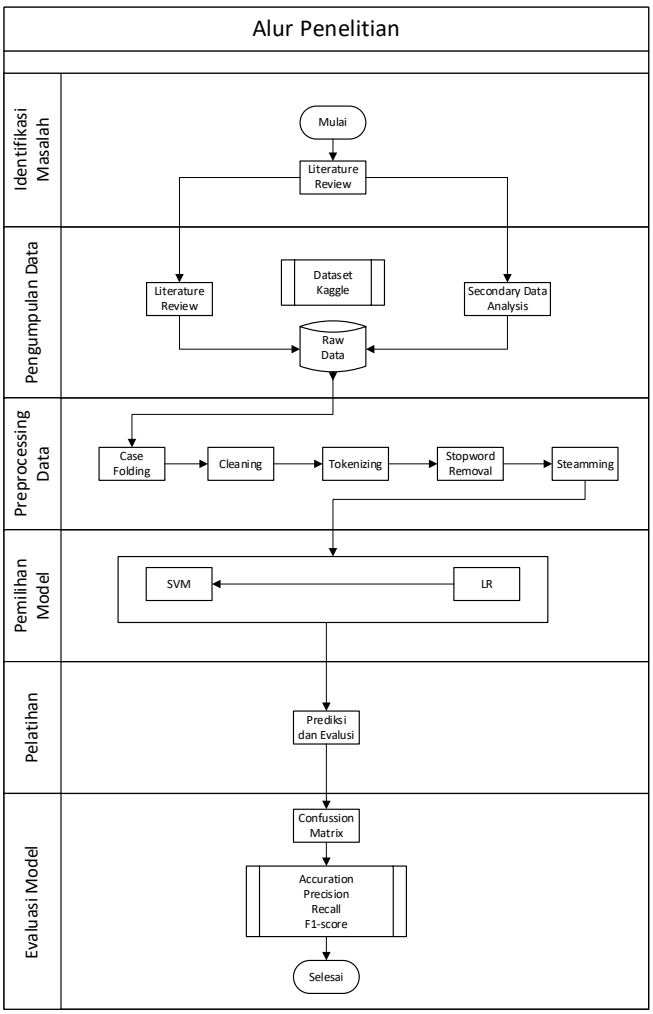


Figure 1. Research Flow

2.4.1. Literature review

Literature review is conducted to identify existing research with similar topics in this case the use of SVM and NB in lexicon-based sentiment analysis. This can help researchers to identify research gaps and formulate questions about the research to be carried out.

2.4.2. Data Collection

As described in the "Data Collection Methods" section, this research uses public data that has been used in previous studies. This was done to save research time because there was no need to do the labeling manually and involve language experts. On the other hand, this can validate the results of the research conducted.

2.4.3. Preprocessing Data

This stage is carried out to process raw (unstructured) data (text) into structured text, with the aim of maximizing the results of sentiment analysis. As explained in the "Data Analysis Methods" section, this process involves case folding, cleaning, tokenizing, stopwords removal and then stemming. The following is an explanation of the steps in preprocessing:

a. Case folding

   This step serves to convert uppercase or capital letters into lowercase or lowercase letters.

b. Cleaning

   Cleaning aims to clean up unimportant things in sentiment analysis such as characters, numbers, punctuation marks, white space and single char.

c. Tokenizing

   Tokenization or tokenizing refers to the separation of text documents (sentences) into small units. Where the unit (word) is called a token. In general, tokenization is the process of breaking sentences into words, in which case each word in the sentence is accommodated in the form of an array.

d. Stopwords Removal

   The function of this process is to remove basic words or words in the stop list that have a high frequency of occurrence, for example connecting words such as "will", "but", "or", "and" and "for".

e. Stemming

   Stemming is a technique to convert words into basic words by removing affixes either suffixes or prefixes attached to words. As in the word "helpful" becomes "help".

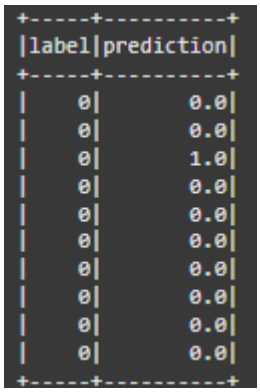
2.4.4. Model Selection

After preprocessing, including case folding, removal of non-alphabetic characters, stop words, and stemming using PorterStemmer, we selected Support Vector Machine (SVM) and Linear Regression (LR) algorithms for sentiment analysis. Text features (X) are generated from numerical representation using Word2Vec, while labels (Y) are determined from rating values: 1 for positive sentiment (Rating ≥ 7) and 0 for negative sentiment (Rating < 7). The dataset was divided into training data (80%) and test data (20%) to train and test the model. Normalization is performed with MinMaxScaler so that the features are in an appropriate range. The model is then evaluated using Confusion Matrix to calculate accuracy, and

can be extended with metrics such as precision, recall, and F1-Score for more comprehensive performance analysis.

3. RESULTS AND DISCUSSION

To analyze the sentiment of movie reviews, the proposed models are Support Vector Machine and Linear Regression and the dataset used is the IMDb Movies dataset which is already available on Kaggle and can be used publicly. The dataset is then divided into 2 parts, namely training data (80%) and testing data (20%). The figure below shows a sample of the results of the data processing that has been done.



label	prediction
0	0.0
0	0.0
0	1.0
0	0.0
0	0.0
0	0.0
0	0.0
0	0.0
0	0.0
0	0.0

Figure 2. Prediksi Label

This label column comes from data that has been processed and used to train the model. The label column itself contains the target value or the final result of the data that the model wants to predict. In this case, the label value is taken from the results of the preprocessing of the dataset, where the label is determined based on the sentiment polarity value. For example:

a. Reviews with Rating ≥ 7 are considered as positive and labeled as 1.
b. Reviews with Rating < 7 are considered as negative and labeled as 0.

These values become the target for the model to learn the patterns in the data. Meanwhile, the prediction column above is generated by the model after the prediction process is performed on the test data. The values in the column show the sentiment predictions made by the SVM and LR models based on the numerical features generated from the Word2Vec process.

All values in the label column are 0, which indicates that this subset of data consists of reviews that are labeled as negative sentiment based on the original data (low rating). Meanwhile, the prediction column reflects the model's classification result of the review sentiment based on the feature data, where a value of 0.0 indicates a negative sentiment prediction and 1.0 indicates a positive sentiment prediction. If the prediction value is equal to the label value, then the model has predicted correctly. For example, in the first row showing 0 | 0.0, this negative review is correctly classified as a negative sentiment. On the other hand, if the prediction value is different from the label value, then the model made a prediction error. For example, in the row 0 | 1.0, the model incorrectly predicted the negative review as a positive sentiment. From the results shown, the model managed to predict most of the reviews correctly (prediction = label), but there was 1 row where the prediction was wrong, namely 0 | 1.0 .

The results of the experiments conducted show that SMV is superior to LR. The performance of the two models will be evaluated based on accuracy, as shown in Table 1:

Table 1. Results and Accuracy

No	Model	Accuracy (%)
1	Support Vector Machine	76
2	Logistic Regression	69

4. CONCLUSIONS

To analyze the sentiment of movie reviews, the proposed models (Support Vector Machine and Logistic Regression) were evaluated on the IMDb dataset after preprocessing and feature extraction. The performance of the models was evaluated based on accuracy, where SVM achieved 76% accuracy, while Linear Regression achieved 69%. The effectiveness of SVM is due to its ability to find the optimal hyperplane for high-dimensional data, which suits the sparse and vectorized nature of text. Logistic Regression, although simpler and faster to train, showed lower performance due to its reliance on linear relationships. Preprocessing steps such as data cleaning, tokenization, stop words removal, and stemming significantly improved the quality of the input data, and Word2Vec provided a meaningful vector representation of the text that allowed both models to classify sentiment well. However, SVM is superior due to its ability to handle high-dimensional and sparse data better.

5. REFERENCES

Ananda Lubis, F., Studi Manajemen, P., Ekonomi Dan Bisnis Islam, F., & Irwan Padli Nasution, M. (2024). Penggunaan Teknologi Big Data untuk Analisis Prediksi Bisnis. *Jurnal Ilmiah Nusantara (JINU)*, *1*(4), 3047–9673. https://doi.org/10.61722/jinu.v1i4.1882

Kurniawan, A. A., & Mustikasari, M. (2022). Evaluasi Kinerja MLLIB APACHE SPARK pada Klasifikasi Berita Palsu dalam Bahasa Indonesia. *Jurnal Teknologi Informasi dan Ilmu Komputer*, *9*(3), 489. https://doi.org/10.25126/jtiik.2022923538

Mee, A., Homapour, E., Chiclana, F., & Engel, O. (2021). Sentiment analysis using TF–IDF weighting of UK MPs' tweets on Brexit[Formula presented]. *Knowledge-Based Systems*, *228*, 107238. https://doi.org/10.1016/j.knosys.2021.107238

Pohan, R., Ratnawati, D., Arwani, I., a, b, c, & d. (2022). Implementasi Algoritma Support Vector Machine dan Model Bag-of-Words dalam Analisis Sentimen mengenai PILKADA 2020 pada Pengguna Twitter. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *6*(10), 4924–4931. http://j-ptiik.ub.ac.id

Resyanto, F., Sibaroni, Y., & Romadhony, A. (2019). Choosing The Most Optimum Text Preprocessing Method for Sentiment Analysis: Case:iPhone Tweets. *Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019*, 2–6. https://doi.org/10.1109/ICIC47613.2019.8985943

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. *Applied Sciences (Switzerland)*, *13*(7). https://doi.org/10.3390/app13074550

Tumbel, C. Z., Sitepu, H., & Hutagalung, M. (2017). Analisis Big Data Berbasis Stream

Processing Menggunakan Apache Spark. *Jurnal Telematika*, *11*(1), 6. https://doi.org/10.61769/telematika.v11i1.145

Vindua, R., & Zailani, A. U. (2023). Analisis Sentimen Pemilu Indonesia Tahun 2024 Dari Media Sosial Twitter Menggunakan Python. *JURIKOM (Jurnal Riset Komputer)*, *10*(2), 479. https://doi.org/10.30865/jurikom.v10i2.5945

Wahyudi, I. S. (2018). Big data analytic untuk pembuatan rekomendasi koleksi film personal menggunakan Mlib. Apache Spark. *Berkala Ilmu Perpustakaan dan Informasi*, *14*(1), 11. https://doi.org/10.22146/bip.32208