Siber Nusantara Publisher
ARTIFICIAL RESEARCH:
Artificial Intelligence and Robotic Journal
E-ISSN: xxxx-xxxx
P-ISSN: xxxx-xxxx
https://journal.e-siber.org/ARTIFICIAL siberpublisher.info@gmail.com +62 822 1012 3600

DOI: https://doi.org/10.38035/artificial.v1i1

The Role of Data Pre-Processing Techniques and Classification Algorithms on the Accuracy of Sentiment Analysis in Social Medi: A Literature Review

Salsabila Dwi Fitri[1], Yorasakhi Ananta[2]
[1]Jambi University, Jambi, Indonesia, dwifitrisalsabila@gmail.com
[2]Andalas University, Padang, West Sumatra, Indonesia, yorasakhiananta27@gmail.com

Corresponding Author: dwifitrisalsabila@gmail.com

Abstract: The development of digital technology and the explosion of data on social media have increased the need for accurate sentiment analysis to understand public opinion. This article aims to systematically review the role of data pre-processing techniques and classification algorithms in improving the accuracy of sentiment analysis in social media. Through the Systematic Literature Review (SLR) approach, more than 30 scientific articles from trusted sources were reviewed between 2018 and 2024. The results of the study show that effective pre-processing such as tokenization, stemming, and stop word removal significantly improve the quality of input data, while algorithms such as SVM, Random Forest, and deep learning provide the best performance in sentiment classification. This article is expected to be a conceptual reference for further research and the development of a more precise sentiment analysis system.

Keywords: Sentiment Analysis, Social Media, Data Preprocessing, Classification Algorithm, Model Accuracy

INTRODUCTION

Social media has become the main platform for people to express opinions, criticisms, and praises of various issues, products, or services. According to Statista (2023), global social media users have reached 4.89 billion people, which shows great potential in exploring public opinion data. One of the main approaches used to evaluate public perception on social media is sentiment analysis, which is the process of identifying and categorizing opinions as positive, negative, or neutral.

However, the biggest challenge in sentiment analysis lies in the quality of raw data that is unstructured, contains slang, emojis, spelling errors, and various other forms of noise. To overcome this, data preprocessing techniques are crucial so that the information obtained is clean and can be processed efficiently by the classification algorithm (Yenter & Verma, 2017).

On the other hand, the selection of the classification algorithm also plays a vital role in determining the success of sentiment analysis. Research by Kowsari et al. (2019) showed that deep learning algorithms such as LSTM outperform traditional approaches in capturing complex sentence context.

Several studies have compared various combinations of pre-processing techniques and classification algorithms. For example, a study by Ali (2023) showed that the combination of lemmatization with Random Forest resulted in an accuracy of 87% in sentiment analysis of product reviews.

In another study, Mishra and Kumar (2020) emphasized the importance of using a pre-processing pipeline tailored to specific languages and domains to improve the effectiveness of classification models.

With the increasing complexity of social data and the advancement of machine learning algorithms, remapping and synthesis of previous research are needed to formulate the best approach in sentiment analysis.

This article discusses the role of data pre-processing techniques and classification algorithms on the accuracy of sentiment analysis on social media. This study is a literature review in the fields of computer science and data science.

Based on this background, the purpose of writing this article is to build a hypothesis for further research, namely to formulate: 1) The role of data pre-processing techniques on the accuracy of sentiment analysis; and 2) The role of classification algorithms on the accuracy of sentiment analysis.

METHODS

The method of writing this literature review article uses the Library Research and Systematic Literature Review (SLR) approaches. The analysis was carried out qualitatively, with the main sources coming from online applications such as Google Scholar, Mendeley, and other relevant online academic applications.

This approach allows researchers to conduct a systematic search of academic literature relevant to data pre-processing techniques, classification algorithms, and accuracy in sentiment analysis on social media.

In the context of qualitative analysis, literature review is used consistently with the methodological assumptions underlying the exploratory approach. One of the reasons for using qualitative analysis is because this research is exploratory in nature, namely aiming to explore in depth the role of techniques and methods in improving the accuracy of sentiment analysis (Ali, H., & Limakrisna, 2013).

Article searches were conducted on publications from 2018 to 2024, with main keywords such as: "sentiment analysis", "text preprocessing", "classification algorithm", "social media text mining", and "machine learning for sentiment".

Inclusion criteria include articles published in reputable journals, available in English or Indonesian, and presenting an evaluation of model accuracy in the context of sentiment analysis.

Data analysis was carried out by summarizing the main findings of each article, grouping the approaches used, and assessing the effectiveness of each combination of techniques on the accuracy of sentiment classification results.

RESULT AND DISCUSSION

Definition and Significance of Sentiment Analysis

Sentiment analysis is an approach in computer science that is used to automatically identify and classify opinions or emotions in text (Liu, 2012). In the context of social media, sentiment analysis becomes important because platforms such as Twitter, Facebook, and Instagram generate large amounts of data that reflect public opinion. This technology is widely used in business to understand consumer perceptions of certain products or services (Ali, 2023). Several studies have stated that the accuracy of sentiment analysis is highly dependent on the quality of the data and the algorithm used (Zhang et al., 2018). Therefore, this study discusses two main factors that affect the accuracy of sentiment analysis, namely data pre-processing techniques and classification algorithms. The main focus is to understand how the combination of the two can improve performance in opinion classification.

Data Pre-Processing Techniques in Sentiment Analysis

Data pre-processing is the initial stage in the sentiment analysis pipeline that aims to clean and prepare text data for further processing (Bird, Klein, & Loper, 2009). This process includes removing punctuation, numbers, symbols, uppercase to lowercase conversion, stemming, and stopword removal (Kouloumpis et al., 2011). A study by Ali (2023) showed that a combination of tokenization and lemmatization can increase classification accuracy by 12%. This stage also serves to reduce noise in the data, especially in the context of social media which contains a lot of slang, abbreviations, and emoticons. By performing data normalization and cleaning thoroughly, the classification algorithm can work more efficiently. Therefore, the quality of pre-processing is crucial in achieving accurate analysis results.

Dimensions of Effective Pre-Processing Techniques

Some dimensions of pre-processing techniques include tokenization, normalization, stemming, and noise removal (Gonçalves et al., 2013). Each of these dimensions plays a role in optimizing the input for the classification algorithm. Tokenization breaks down text into smaller word units, while stemming reduces words to their basic form. Research by Khan et al. (2020) showed that stopword removal can increase model speed and accuracy by up to 10%. The combination of these techniques needs to be adjusted to the characteristics of the language and data domain used. Thus, there is no one-size-fits-all approach for sentiment analysis cases.

Challenges in Social Media Data Preprocessing

Data from social media tends to be unstructured and contains a lot of noise such as emojis, abbreviations, and slang. According to research by Eisenstein (2013), preprocessing that is not in accordance with the local context can cause a decrease in accuracy. In the Indonesian context, the use of mixed languages (code-switching) is also a challenge (Sari & Wicaksono, 2022). Techniques such as spelling correction and slang normalization need to be further developed. A study by Pratama (2021) showed that without slang normalization, algorithm accuracy decreased by up to 15%. Therefore, it is important to develop preprocessing techniques that are contextual and adaptive to linguistic trends in social media.

The Role of Classification Algorithms in Sentiment Analysis

Classification algorithms are the core of sentiment analysis systems that determine the sentiment category of text input. Commonly used algorithms include Naïve Bayes, Support Vector Machine (SVM), and Random Forest (Pang & Lee, 2008). A study by Ali (2023) shows that the SVM algorithm excels in accuracy, especially for data that has gone through an optimal pre-processing stage. In addition, deep learning algorithms such as LSTM and BERT have also begun to be widely used in recent studies. However, computational complexity is a challenge for real-time implementation. The selection of algorithms must consider the trade-off between accuracy and processing efficiency.

Dimensions of Classification Algorithm Performance Evaluation

Evaluation of classification algorithms is carried out using metrics such as accuracy, precision, recall, and F1-score (Manning et al., 2008). According to a study by Nurhayati et al. (2022), the F1-score metric better represents the model's performance in handling imbalanced data. Random Forest shows high accuracy but tends to overfit on small datasets. On the other hand, Naïve Bayes has high speed but lower accuracy. Another study by Dewi and Prabowo

(2023) emphasized the importance of cross-validation to avoid bias. Comprehensive evaluation is needed so that the classification results are reliable.

Combination of Pre-Processing Techniques and Classification Algorithms

The use of appropriate pre-processing techniques can significantly improve the performance of classification algorithms. For example, the combination of stopword removal and SVM resulted in 85% accuracy on the Indonesian Twitter dataset (Yuliana & Lestari, 2021). In another study, the use of Indonesian stemming together with the Random Forest algorithm achieved 83% accuracy (Ramadhan & Siregar, 2020). This proves that there is a strong relationship between the quality of pre-processing and the effectiveness of the algorithm. Therefore, adjusting the preprocessing strategy must be done together with the selection of the algorithm. With this approach, the accuracy of sentiment analysis can be maximized.

Comparative Study: Classical Algorithms vs. Deep Learning

A study by Rahmawati et al. (2021) compared the SVM and LSTM algorithms for sentiment analysis of e-commerce product reviews. The results showed that LSTM outperformed SVM in terms of accuracy (92% vs. 85%) after intensive pre-processing. However, LSTM requires more computing resources and training time. The advantage of LSTM lies in its ability to capture the context and order of words in a sentence. However, in resource-limited scenarios, classical algorithms such as SVM or Naïve Bayes are still effective choices. Therefore, the choice of method still needs to be adjusted to the objectives and capacity of the system.

Influence of Pre-Processing on Multilingual Datasets

Sentiment analysis on social media in a multilingual country like Indonesia faces its own challenges. Mixed languages (code-switching) often appear in one text, thus requiring adaptive pre-processing techniques (Riyanto & Saputra, 2021). The study showed that the use of local dictionary-based lemmatization increased classification accuracy by 10%. In addition, the integration of automatic translation can help handle foreign words that are not recognized by the model. Pre-processing that combines language identification and transliteration is more effective on multilingual data. Therefore, the flexibility and locality of pre-processing techniques are important in the context of Indonesian sentiment analysis.

Effectiveness of Combination of Ensemble Classification Algorithms

Ensemble models such as Random Forest and XGBoost show high performance in sentiment analysis with big data (Hastie et al., 2009). A study by Wulandari and Sari (2023) showed that the combination of ensemble with TF-IDF-based pre-processing was able to achieve 91% accuracy. The ensemble technique works by combining decisions from several decision tree models, making it more resistant to overfitting. However, the drawback is the higher computation time compared to a single algorithm. For real-time sentiment monitoring cases, this is an important consideration. Even so, for the needs of long-term predictions and high accuracy, ensembles are a very effective choice.

Comparison of Efficiency Between Algorithms

Each algorithm has advantages and limitations in terms of efficiency. SVM and Naïve Bayes excel in fast training times, suitable for systems with limited resources (Ali, 2023). On the other hand, deep learning such as BERT and LSTM require GPUs and large datasets to provide optimal performance. A study by Prasetyo and Rakhmawati (2022) showed that Naïve Bayes produced 78% accuracy in 30 seconds of training time, while BERT reached 93% but required more than 2 hours of training time. This efficiency must be adjusted to the objectives of the project. Therefore, the selection of algorithms is not only based on accuracy, but also operational efficiency.

Contribution of Feature Engineering in Sentiment Analysis

In addition to pre-processing and classification techniques, feature engineering plays an important role in improving accuracy. Techniques such as TF-IDF, word embeddings, and n-grams have been shown to be effective in forming text representations (Jurafsky & Martin, 2021). A study by Anggraini et al. (2021) showed that the use of n-grams increased the accuracy of SVM from 76% to 84%. Word embeddings such as Word2Vec and GloVe also strengthen the understanding of the context of words in the text. Integrating feature engineering with the analysis pipeline can provide richer and more accurate results. Therefore, feature engineering is an integral element in the development of sentiment analysis systems.

Case Study: Public Sentiment towards Government Policy

Sentiment analysis is also widely applied to understand public opinion towards government policies. Research by Lestari and Nugroho (2022) used SVM to classify opinions about PPKM policies from Twitter. The results showed that 62% of sentiments were negative with a classification accuracy of 87% after thorough pre-processing. The techniques used included stemming, stopword removal, and symbol filtering. This study confirms that sentiment analysis can be an important tool in data-driven policy making. Thus, optimization of pre-processing techniques and classification algorithms has a direct impact on socio-political understanding.

The Role of Sentiment Analysis in E-Commerce

In e-commerce, sentiment analysis is used to assess customer satisfaction based on product reviews. Research by Susanto and Rahayu (2023) shows that the classification of positive and negative sentiments can help sellers improve services. They used a combination of TF-IDF, stopword removal, and the XGBoost algorithm, resulting in 90% accuracy. These results support the importance of a solid preprocessing strategy to improve analytical results. In addition, sentiment data is also used in recommendation systems and product improvement. Therefore, the use of sentiment analysis in this sector is greatly influenced by the quality of pre-processing and classification.

Integration of Sentiment Analysis in Decision Systems

Many organizations are starting to integrate sentiment analysis results into decision-making systems. By visualizing analytical results, managers can understand public opinion trends in real time (Ali, 2023). Research by Damanik and Pertiwi (2021) shows that integrating a sentiment analysis dashboard increases the speed of response to issues by 35%. This technology utilizes the output of the classification model to provide strategic insights. However, the accuracy of this information is highly dependent on the pre-processing stages and algorithms used. Therefore, system integration must be supported by a solid and standardized data pipeline.

Theoretical Implications and Recommendations for Further Research

Based on the literature study discussed, it can be seen that the relationship between pre-processing techniques and classification algorithms greatly affects the accuracy of sentiment analysis. The theoretical contribution of this study is to identify the synergy between preprocessing and machine learning in public opinion processing. Recommendations for further research are to explore the integration of hybrid models that combine deep learning techniques with linguistic-based preprocessing. In addition, it is important to evaluate the performance of the model in various social, cultural, and linguistic contexts. With this approach, the results of sentiment analysis can be more representative and applicable. This study opens up space for the development of local frameworks in the Indonesian context.

CONCLUSION

This article has systematically reviewed various studies that discuss the influence of data preprocessing techniques and classification algorithms on the accuracy of sentiment analysis in social media. The findings show that preprocessing techniques such as tokenization, stemming, lemmatization, and stopword removal have a significant impact on sentiment classification results. On the other hand, the selection of algorithms such as Naïve Bayes, SVM, Random Forest, to deep learning models such as BERT also determines the quality of the prediction output. The combination of appropriate preprocessing and appropriate algorithms can produce higher accuracy, as proven in many of the reviewed studies. This study also underlines the importance of local context, especially in dealing with multilingual and slang data in Indonesian social media. Therefore, the development of a sentiment analysis pipeline that is tailored to the characteristics of local data is an important recommendation for future research and practical applications.

REFERENCE

Ali, H. (2023). *Data Science dan Penerapan Big Data di Indonesia*. Jakarta: Gramedia Pustaka Utama.

Ali, H., & Limakrisna, N. (2013). *Metodologi Penelitian: Aplikasi dalam Pemasaran*. Jakarta: Mitra Wacana Media.

Anggraini, S., Nurhadi, D., & Maulana, A. (2021). Penerapan N-gram dan SVM dalam Analisis Sentimen Produk. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 8(2), 110-117.

Damanik, R., & Pertiwi, L. (2021). Implementasi Dashboard Analisis Sentimen untuk Pengambilan Keputusan Instansi Pemerintah. *Jurnal Sistem Informasi*, 14(3), 225-234.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing* (3rd ed.). Pearson.

Lestari, D., & Nugroho, A. (2022). Analisis Sentimen Kebijakan Pemerintah di Media Sosial Menggunakan Support Vector Machine. *Jurnal Ilmu Komputer dan Informasi*, 15(1), 17–24.

Prasetyo, E., & Rakhmawati, R. (2022). Perbandingan Akurasi Naïve Bayes dan BERT dalam Klasifikasi Sentimen Twitter. *Jurnal Teknologi dan Sistem Komputer*, 10(3), 239–248.

Riyanto, D., & Saputra, R. (2021). Pengaruh Pra-Pemrosesan Multibahasa pada Akurasi Analisis Sentimen. *Jurnal Penelitian Ilmu Komputer Indonesia*, 6(1), 1–9.

Susanto, B., & Rahayu, F. (2023). Pemanfaatan Analisis Sentimen untuk Menilai Kepuasan Pelanggan pada Marketplace. *Jurnal Teknologi Informasi dan Komputer*, 11(2), 87–93.

Wulandari, R., & Sari, H. (2023). Optimalisasi TF-IDF dan XGBoost untuk Klasifikasi Sentimen Ulasan Produk. *Jurnal Ilmu Komputer dan Rekayasa Perangkat Lunak*, 9(2), 78–86.