Journal Sensi:Strategic Of Education in Information System
Accreditation rank 5
SK Dirjen DIKTI, RisTek No. 10/C/C3/DT.05.00/2025; E-ISSN: 2655-5298 P-ISSN: 2461-1409

Research Article **Open Access (CC–BY-SA)**

Comparison Classification for Indonesian Twitter Hate Speech and Abusive Detection

Ibnu Mas'ud[1], Tubagus Toifur[2], Aolia Ikhwanudin[3], Muhamad Yusuf[4], AgiantoSyamhalim[5], Anas Nasrulloh[6]
[1] UIN Sultan Maulana Hasanuddin Banten,[2,3,4,5,6] Institut Teknologi Tangerang Selatan
[1] Study Program of Informatic, Science and Technology Faculty, UIN Sultan Maulana Hasanuddin Banten
[2,4] Study Program of Information Technoloy, Computer Science Faculty, Institut Teknologi Tangerang Selatan
[3] Study Program of Informatic, Computer Science Faculty, Institut Teknologi Tangerang Selatan
[5,6] Study Program of Information System, Computer Science Faculty, Institut Teknologi Tangerang Selatan
e-mail: [1]ibnumasud@uinbanten.ac.id, [2]tubagus@itts.ac.id, [3]aolia@itts.ac.id, [4]yusuf@itts.ac.id, [5]agianto@itts.ac.id, [6]anas@itts.ac.id

Abstract

Hate speech and offensive language on social media, particularly Twitter in Indonesia, have become a serious problem that can threaten the social and psychological stability of users. This study aims to analyze and detect such harmful content using a multi-label classification approach, which is more representative in capturing the complexity of real-world language. The research methodology involves collecting data through the Twitter API, which is then subjected to an intensive preprocessing stage, including data cleaning and text normalization using a slang dictionary. We apply machine learning algorithms such as Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest Decision Tree (RFDT). To handle the multi-label characteristics, Binary Relevance (BR), Label Power-set (LP), and Classifier Chains (CC) transformation techniques are used. The results show that the RFDT algorithm with LP transformation provides the best performance with an accuracy rate of 81.2%. This finding confirms that text normalization and the selection of appropriate label transformation techniques are crucial in improving detection accuracy. The results of this study are expected to provide a foundation for the development of a smarter automated content moderation system for Indonesian-language social media.

Keywords — *Hate speech, abusive language, multi-label classification, Indonesian Twitter, Random Forest.*

1. INTRODUCTION

The growth of social media users in Indonesia, particularly Twitter, has brought about a significant transformation in digital communication patterns. However, this ease of sharing

information has also triggered an increase in the spread of negative content, such as hate speech *and* abusive language *[* 1], [5]. This phenomenon not only triggers social unrest but can also escalate into physical or psychological conflict if not handled properly [2], [4]. Automatic identification of this dangerous communication is crucial to maintaining the integrity of the digital space and protecting users, especially vulnerable groups [3].

Several previous studies have explored hate speech detection in English with significant results [11], [13]. However, applying these methods to the Indonesian context faces unique challenges due to the complex language structure, use of slang , and cultural diversity [6], [7]. Research by Alfina et al. [7] and Ibrohim & Budi [9] has made an important step by providing an initial dataset and preliminary study on multi-label detection on Indonesian Twitter.

Although binary classification is often used, the reality in the field shows that a single post often contains several label categories at once, such as hate speech directed at an individual while also containing elements of racism [8], [12]. Therefore, a multi-label classification approach is considered more representative [21], [22]. Various machine learning algorithms such as *Support Vector Machine* (SVM), *Naive Bayes* (NB), and *Random Forest Decision Tree* (RFDT) have been tested for their reliability in handling large-scale textual data [10], [17].

In addition to algorithms, classification success is highly dependent on data preprocessing, including text normalization and feature extraction such as TF-IDF or *Deep Learning*- based techniques [15], [20]. The use of data transformation methods such as *Binary Relevance* (BR), *Label Power-set* (LP), and *Classifier Chains* (CC) has become standard in handling dependencies between labels in multi-label classification [21].

This study aims to analyze the performance of various machine learning algorithms in detecting hate speech and offensive language on Indonesian Twitter. The main focus of this study is to evaluate the effectiveness of *Label Power-set* transformation combined with RFDT, as well as to underline the importance of text normalization in improving detection accuracy [14], [18]. The results of this study are expected to contribute to the development of content moderation systems that are more intelligent and adaptive to local language characteristics.

2. RESEARCH METHODOLOGY

This research methodology is designed to handle multi-label classification of Indonesian Twitter textual data. The workflow consists of several main stages: data collection, preprocessing, annotation, and machine learning algorithm implementation.

2.1 Data Collection

The dataset used in this study is sourced from previous research conducted by Ibrohim et al. [9] and enriched with additional data retrieved using the Twitter Search API through the Tweepy library. Data collection was conducted over a seven-month period (March 20 to September 10, 2018) to capture various trending expressions of hate speech. The final dataset consists of thousands of unique tweet entries representing various categories of hate speech and offensive language in Indonesia.

2.2 Data Preprocessing

Preprocessing is a crucial step in improving data quality before entering the classification stage. The steps involved include:

1. Data Cleaning: Removes non-alphanumeric characters, URLs, *hashtags* , and anonymized *usernames* (@user).
2. Case Folding: Changes all text to *lowercase* .
3. Text Normalization: Convert non-standard words, abbreviations, and slang *into* standard form using the normalization dictionary (*new_kamusalay.csv*) [5].

4. Handling Abusive Words: Identify abusive words using a reference list (*abusive.csv*) to strengthen the classification features.

2.3 Annotation Process

The annotation process is carried out in stages using a *crowdsourcing* method . To maintain quality and objectivity, annotators are selected based on certain criteria:

1. Diversity: Includes a variety of age ranges and educational backgrounds.
2. Language Skills: Have a good understanding of local language dialects and contexts.
3. Experience: Active Twitter users who understand the platform's culture of engagement. Each tweet is assessed by at least two annotators to determine a label (HS, Abusive, or other specific category) to minimize subjective bias.

2.4 Multi-label Classification Approach

Since a single tweet can have more than one label (for example, containing both hate speech and offensive language), this study applies multi-label data transformation techniques [21], [22]:

1. Binary Relevance (BR): Treats each label as an independent binary classification problem.
2. Label Power-set (LP): Converts each unique combination of labels into a single, unique class label.
3. Classifier Chains (CC): Forms a chain of classifiers where the predictions of the previous label become features for the next label.

2.5 Algorithm and Feature Extraction

This study compares the performance of three popular algorithms:

1. Support Vector Machine (SVM): Known to be effective for high-dimensional feature spaces [11].
2. Naive Bayes (NB): Used as an efficient *baseline for text classification [17].*
3. Random Forest Decision Tree (RFDT): Used because of its ability to handle non-linear data and prevent *overfitting* .

Features are extracted using the Term Frequency-Inverse Document Frequency (TF-IDF) method to represent the importance weight of a word in a document against the entire data corpus [20].

2.6 Evaluation Metrics

Model performance was evaluated using standard classification metrics, namely Accuracy , Precision , Recall , and F1-Score . Testing was performed using a *cross-validation* scheme to ensure the stability of the results across different data subsets.

3. RESULTS AND DISCUSSION

This section presents the results of multi-label classification experiments for hate speech and offensive language detection, as well as an analysis of the performance of the algorithms used.

3.1 Data Distribution Analysis

Based on the labeling results on the Indonesian Twitter dataset, it was found that the label distribution was uneven. The labels "HS" (Hate Speech) and "Abusive" appeared predominantly. Specifically, the most frequently appearing hate speech category related to sentiments against certain individuals and groups (SARA). This indicates that the polarity of conversations on Indonesian Twitter is heavily influenced by issues of identity and personality.

3.2 Model Performance Comparison

Experimental results using three machine learning algorithms with three multi-label transformation techniques show significant performance variations. Table 1 summarizes the accuracy results from these tests.

Table 1. Comparison of Algorithm Accuracy and Transformation Techniques

Algorithm	Binary Relevance (BR)	Label Power-set (LP)	Classifier Chains (CC)
Naive Bayes (NB)	68.4%	70.1%	69.2%
Support Vector Machine (SVM)	74.2%	76.5%	75.8%
Random Forest (RFDT)	77.8%	81.2%	79.5%

Based on Table 1, the Random Forest Decision Tree (RFDT) algorithm combined with the Label Power-set (LP) technique provides the highest accuracy of 81.2% . This is due to RFDT's ability to handle non-linear relationships between word features, while the LP technique successfully captures correlations between labels by treating the combination of labels as a single, unique class [22].

3.3 Impact of Text Normalization

Analysis shows that the preprocessing stage, specifically text normalization using *new_kamusalay.csv* , improves model accuracy by approximately 5-8%. The massive use of slang and abbreviations on Indonesian Twitter (such as "yg" becoming "yang", "tdk" becoming "tidak") previously often caused *misclassification* because word features were considered different even though they actually had the same meaning [5], [20].

3.4 Error Analysis

Although the RFDT-LP model performed best, it still misclassified tweets containing sarcasm or subtle innuendo. The model struggled to detect hate speech that did not explicitly (implicitly) use offensive language. This suggests the need for further research to develop deeper context-based features, such as the use of *Word Embedding* (Word2Vec or BERT), to capture broader semantic meaning [14], [15].

3.5 Discussion on Specific Labels

In specific tests against target labels, the model performed very well in detecting the “Abusive” label due to its highly distinctive abusive features. However, for specific labels such as “HS_Religion” or “HS_Race”, the model requires more balanced training data to avoid bias towards the majority label “HS_Individual”. Overall, the multi-label approach proved to be much more effective in mapping the complexity of language in social media than simple binary classification [9].

4. CONCLUSION

This research has successfully implemented a multi-label classification method to detect hate speech *and* abusive language *on* Twitter in Indonesia. Based on the experimental results and analysis, the following key conclusions can be drawn:

First, the use of a multi-label classification approach has proven to be more effective and representative in mapping the complexity of language in social media compared to binary classification, because it is able to identify several label categories in one tweet simultaneously [9], [22].

Second, from a comparison of various machine learning algorithms, Random Forest Decision Tree (RFDT) with Label Power-set (LP) transformation technique showed the best performance with an accuracy level reaching 81.2% . The LP technique excels because of its ability to consider the correlation between labels as a unique class unit, which is very relevant to the characteristics of hate speech data in Indonesia [21].

Third, the data preprocessing process, especially text normalization using a slang dictionary , plays a vital role in increasing model accuracy by up to 8%. This proves that handling variations in informal writing on Indonesian Twitter is crucial before the data enters the TF-IDF feature extraction stage [20].

Finally, although the model has shown good results, error analysis indicates challenges in detecting implicit hate speech or using sarcasm. This provides an opportunity for future research to explore the use of deeper semantic features such as *Word Embedding or Transformer* architecture to improve understanding of language context [14], [15].

5. RECOMMENDATION

Based on the research findings and limitations identified during the process of classifying hate speech on Indonesian Twitter, the author offers several recommendations for future research development.

1. Exploration of Advanced NLP Methods: Given that traditional machine learning models still struggle with implicit context and sarcasm, it is advisable to use advanced *Deep Learning architectures like Bidirectional Encoder Representations from Transformers* (BERT) or large language models *that* are specifically trained on the Indonesian language corpus (IndoBERT). This approach is expected to better capture the semantic meaning and nuances of slang accurately [14].
2. Data Source Expansion: Further research should not only be limited to the Twitter platform, but also expand to other social media platforms such as Instagram, TikTok, or Facebook which have different linguistic characteristics and audiences, in order to test the scalability of the developed model.
3. Dataset Development and Annotation: A more balanced dataset is needed, especially for less frequently occurring label categories (such as hatred of disability or sexual orientation), to prevent the model from being biased toward the majority label. Furthermore, the use of *active learning* techniques in the annotation process can help improve the efficiency of selecting high-quality data.
4. Multimodal Feature Integration: Considering that hate speech is often conveyed in the form of images (memes) or videos, future detection system development needs to consider a multimodal approach that combines text analysis with visual analysis to obtain more comprehensive detection results.
5. Implementation of Real-time Systems: Recommends the implementation of the developed model in the form of an API (*Application Programming Interface*) or real - time monitoring tool that regulators or platform providers can use to carry out content moderation automatically and quickly.

ACKNOWLEDGEMENT

The author expresses his deepest gratitude to all parties who have contributed to the completion of this research. Special thanks go to:

1. Muhammad Okky Ibrohim , for providing the *"id-multi-label-hate-speech-and-abusive-language-detection"* dataset which is the main basis for the analysis and findings of this research.
2. Wahyu Andhika Rizaldi and Athallah Narda Wiyoga , for their assistance and expertise in providing data visualization code references on the Kaggle platform, which was very helpful in presenting the data for this study.
3. The authors and researchers whose work is cited in this paper, for their insights and scientific contributions which serve as valuable references in the development of hate speech detection studies.

REFERENCES

[1] National Commission on Human Rights, 2015, 2015 Human Rights Violations Report,https://komnasham.go.id.
[2] Davidson, T., Warmsley, D., Macy, M., and Weber, I., 2017, Automated Hate Speech Detection and the Problem of Offensive Language, Proc. 11th Int. AAAI Conf. on Web and Social Media (ICWSM).
[3] Chen, Y., Zhou, Y., Zhu, S., and Xu, H., 2012, Detecting Offensive Language in Social Media to Protect Adolescent Online Safety, Int. Conf. on Privacy, Security, Risk, and Trust.
[4] Stanton, G. H., 2009, The Eight Stages of Genocide,http://genocidewatch.net.
[5] Hernanto, I., and Jeihan, F., 2018, Language and Vocabulary Development in Social Media, Journal of Language Research.
[6] Nurasijah, Y., 2018, Language Challenges in the Digital Era: Analysis of Social Media, Language and Communication Journal.
[7] Alfina, I., Mulia, R., Fanany, M. I., and Ekanata, Y., 2017, Hate Speech Detection in Indonesian Language: A Dataset and Preliminary Study, Proc. 9th ICACSIS.
[8] Putri, R., 2018, Sentiment Analysis and Hate Speech Detection on Twitter Social Media, Undergraduate Thesis, University of Indonesia.
[9] Ibrohim, M. O., and Budi, I., 2018, Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter, Proc. ALW2.
[10] Setyadji, G., 2019, Identification of Hate Speech on Indonesian-Language Social Media using SVM, Journal of Information Technology, Vol. 5, No. 2, pp. 45-52.
[11] Waseem, Z., and Hovy, D., 2016, Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter, Proc. NAACL Student Research Workshop.
[12] Schmidt, A., and Wiegand, M., 2017, A Survey on Hate Speech Detection using Natural Language Processing, Proc. 5th Int. Workshop on Natural Language Processing for Social Media.
[13] Nobata, K., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y., 2016, Abusive Language Detection in Online User Content, Proc. 25th Int. Conf. on World Wide Web.
[14] Mozafari, M., Farahbakhsh, R., and Crespi, N., 2019, A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media, Int. Conf. on Complex Networks and Their Applications.
[15] Badjatiya, P., Gupta, S., Gupta, M., and Varma, V., 2017, Deep Learning for Hate Speech Detection in Tweets, Proc. 26th Int. Conf. on World Wide Web.
[16] Fauzi, M. A., 2019, Random Forest Regression for Sentiment Analysis on Indonesian Presidential Election, Journal of Computer Science and Information Technology.

[17] Santoso, L., and Adji, TB, 2020, Textual Analysis of Indonesian Language Tweets on Political Issues Using Naive Bayes, National Journal of Electrical Engineering and Information Technology.
[18] Zhang, Z., and Luo, L., 2019, Hate Speech Detection: A Solved Problem? The Challenging Case of Long-tail Data, Proc. 23rd Conf. on Computational Natural Language Learning (CoNLL).
[19] Mubarak, H., Darwish, K., and Magdy, W., 2017, Abusive Language Detection on Arabic Social Media, Proc. 1st Workshop on Abusive Language Online.
[20] Gunawan, W., and Sarno, R., 2018, Feature Selection using Information Gain for Hate Speech Detection in Indonesian Language, International Seminar on Application for Technology of Information and Communication.
[21] Ramadhani, A. M., and Goo, H. S., 2017, Multi-label Classification for Indonesian News Articles, Proc. International Conference on Data and Software Engineering (ICoDSE).
[22] Tsoumakas, G., and Katakis, I., 2007, Multi-label Classification: An Overview, International Journal of Data Warehousing and Mining, Vol. 3, No. 3, hal 1-13.