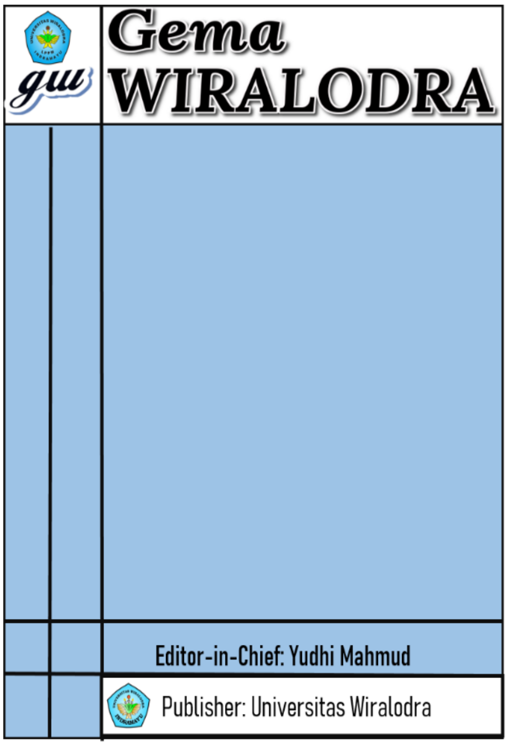
Gema Wiralodra

Publication details, including instructions for authors and subscription information:
https://gemawiralodra.unwir.ac.id


Gema
WIRALODRA
Editor-in-Chief: Yudhi Mahmud
Publisher: Universitas Wiralodra

Waste Pollution Classification in Indonesian Language using DistilBERT

Bambang Nursandi[a]*, Abba Suganda Girsang[b]

[a]Universitas Bina Nusantara, Indonesia, bambang.nursandi@binus.ac.id
[b]Universitas Bina Nusantara, Indonesia, agirsang@binus.edu

To cite this article:
Nursandi, B & Girsang, A.S. (2024). Waste Pollution Classification in Indonesian Language using DistilBERT. *Gema Wiralodra*, *15*(1), 404-413
To link to this article:
https://gemawiralodra.unwir.ac.id/index.php/gemawiralodra/issue/view/24
Published by:
Universitas Wiralodra
Jln. Ir. H. Juanda Km 3 Indramayu, West Java, Indonesia

p – ISSN: **1693 -7945**
e – ISSN: **2622 - 1969**

Original Article

Waste Pollution Classification in Indonesian Language using DistilBERT

Bambang Nursandi[a*], Abba Suganda Girsang[b]
[a]Universitas Bina Nusantara, Indonesia, bambang.nursandi@binus.ac.id
[b]Universitas Bina Nusantara, Indonesia, agirsang@binus.edu

Corresponding Author: bambang.nursandi@binus.ac.id

Abstract

In Indonesia, waste pollution poses pressing environmental and health challenges, making accurate classification vital for targeted mitigation efforts. Our research aimed to extract relevant data from Twitter to address these issues and assess how effectively the DistilBERT model can classify text in the Indonesian language regarding waste pollution. DistilBERT, a streamlined counterpart to the acclaimed BERT architecture, is designed to mirror BERT's advanced linguistic comprehension but with reduced computational demands. Leveraging the essence of transfer learning, proposed method using DistilBERT benefits from extensive textual datasets, making it ideal for scenarios with limited data accessibility. We adopted DistilBERT for the niche challenge of classifying waste types using a constrained dataset derived from Twitter conversations in Indonesian—a medium known for its concise and often ambiguous content. Despite the dataset's limited scope and the inherent noise in Twitter data, the research result using the DistilBERT demonstrated astounding efficacy, achieving Precision: 98%, Recall: 98% and F1-Score: 98%. This outcome underscores DistilBERT's ability to navigate and discern complex textual nuances in data-restricted environments. Our research also included a comparative analysis with other methods, further highlighting the significance of transfer learning in addressing natural language processing challenges, particularly in critical contexts like Indonesia's waste management efforts.

Keyword: Distilbert, Text Classfication, Natural Language Processing, Data Mining

1. Introduction

Waste pollution is a major ecological concern in Indonesia, with the nation producing roughly 7.8 million tons of plastic refuse each year, and out of that, 4.9 million tons are not properly managed (World Bank, 2021). The buildup of non-reusable garbage in dumps, combined with the extensive degradation period of many materials, poses considerable risks to both nature and public health (Solla, 2022). A considerable portion of the debris retrieved from the waterways and containment centers in Jakarta, Indonesia's main city, is made up of plastics, as highlighted in recent on-ground studies (Gokkon, 2022). Industrial waste is a pressing issue in the modern era, necessitating proactive waste management from the onset of production (Nasir et al., 2015). In rural area, the textile industry's wastewater reported has affected nearby villages (Komarawidjaja, 2016). Discussions and reports about waste contamination in Indonesia are prevalent on Twitter. The Twitter user base in Indonesia has seen a surge, reaching 24 million by the beginning of 2023 (Kemp, 2023; Statista, 2023).

Text categorization is an effective NLP method that allows for the grouping of text segments into multiple established tags (Tran-Thien, 2023). When addressing waste pollution in Indonesia, text categorization can serve to sort waste types into various classifications like plastic, paper, and biodegradable waste. The method leverages algorithms to understand, interpret, and categorize text based on its content. This automated process not only enhances efficiency but also supports the implementation of more effective recycling and waste management strategies. By accurately categorizing waste types, Indonesia can better target recycling efforts, reduce landfill use, and address environmental pollution more effectively. This approach aligns with broader environmental protection goals, promoting sustainability and reducing the negative impact of waste on natural ecosystems.

In the realm of NLP, numerous classification techniques have been employed to tackle diverse challenges, including algorithms like naive Bayes and support vector machines (SVM)(Lie, 2018). The Bag of Words (BOW) method offers a straightforward mechanism for text classification, building from a foundational vocabulary list. Furthermore, Transformer-based architectures, especially in recent times, have demonstrated exemplary performance in text categorization undertakings within natural language processing. These architectures stand at the forefront, driving advancements in contemporary NLP solutions (Shaheen et al., 2020).

Our study seeks alternative strategies to achieve precise text categorization, particularly when constrained by data acquisition capabilities, with a focus on leveraging DistilBERT for Indonesian language, Bahasa. DistilBERT boasts remarkable efficiency, encompassing 40% fewer parameters compared to bert-base-uncased and delivering a 60% speed advantage(Masoumi & Bahrani, 2022; William, 2021). Consequently, the model we introduce in this paper is primarily grounded in Twitter content, specifically addressing waste classifications in Indonesian language, namely 'Limbah Padat', 'Limbah Cair', 'Limbah Gas', and 'Limbah Suara'. Additionally, this research furnishes a thorough juxtaposition of the efficacy of SVM, Naïve Bayes, and RoBERTa in related tasks.

This research paper aims and objectives to explore the dynamic and increasingly relevant field of environmental discourse on social media, with a specific focus on waste pollution topics as discussed on Twitter in the Indonesian language. The primary aim is to meticulously extract and analyze data from Twitter, aiming to understand the public perception and narrative surrounding waste issues. This involves employing sophisticated data extraction methods to gather relevant tweets and analyze their content, providing a comprehensive view of the public discourse on waste.

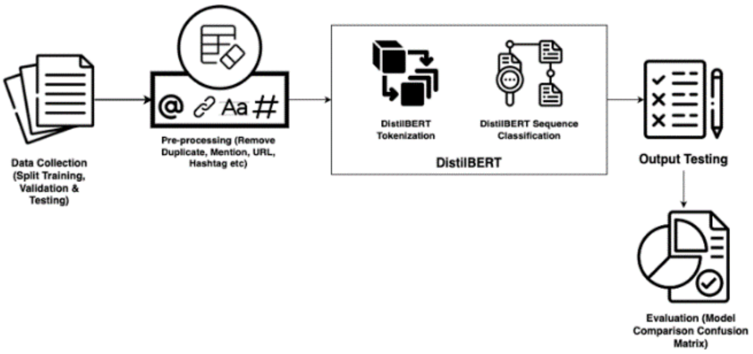
A pivotal part of the study is the implementation of the DistilBERT model, a cutting-edge natural language processing tool, for classifying discussions related to waste pollution. By training the DistilBERT model on Indonesian-language tweets, the research seeks to offer an objective and in-depth insight into how waste pollution is discussed in the digital sphere. This approach is not only about classifying data accurately but also understanding the nuances of public opinion and awareness regarding waste pollution.

Furthermore, the research aims to position DistilBERT in the wider context of data analysis methodologies by comparing it with other classification methods. This comparative analysis will assess DistilBERT's effectiveness, particularly in terms of accuracy and efficiency, against traditional natural language processing models and other data analysis approaches. Such a comparison is vital for identifying the strengths and limitations of DistilBERT and providing a robust recommendation for the most effective method for analyzing environmental discourse on social media platforms like Twitter, especially in the context of waste pollution. Through this integrative approach, the research endeavors to contribute significantly to the understanding of social media discourse on environmental issues and the development of efficient tools for data analysis in this domain.

2. Method

In this section, Authors provide a comprehensive explanation of the proposed method on DistilBERT in text classification, in the context of Twitter data.

Figure 1
Proposed Method using DistilBERT

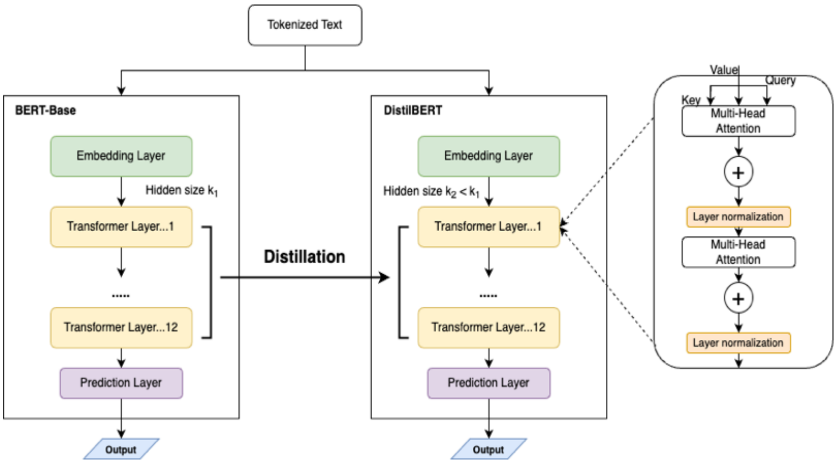

Data Collection (Split Training, Validation & Testing)
@ Aa #
Pre-processing (Remove Duplicate, Mention, URL, Hashtag etc)
DistilBERT Tokenization
DistilBERT Sequence Classification
DistilBERT
Output Testing
Evaluation (Model Comparison Confusion Matrix)

In Figure 1, the dataset creation process involving two main steps: data cleansing and waste pollution data reduction. Twitter data is collected using keywords that encompass four waste categories. Following this, the data cleansing phase is undertaken to filter and select data that is relevant and of high quality. Subsequently, in the data reduction stage, the dataset is further refined to focus on data that is specific and relevant to the research. The Twitter dataset, once cleansed and reduced, is then divided into two parts: data for training and data for testing the waste pollution classification model. The research proceeds with tokenization and training the model using DistilBERT, so that, by the end of the study, the model's output will serve as a reference for implementing waste pollution classification.

DistilBERT Architecture

As illustrated in Figure 2, BERT-base features 12 transformer layers, whereas DistilBERT has only 6 transformer layers. This reduction in the number of layers renders DistilBERT lighter and faster in computation. The size of the hidden layer in the embedding for BERT is indicated by k1, while for DistilBERT, it is denoted as k2 ≤ k1, suggesting that the hidden layer size in DistilBERT could be smaller than or equal to that of BERT. Consequently, DistilBERT has fewer parameters in each layer.

Figure 2
DistilBERT Architecture



DistilBERT is trained using a technique known as distillation, where knowledge from a larger model (BERT) is transferred to a smaller model (DistilBERT) (Adel et al., 2022). This is typically achieved by training the smaller model to mimic the output distribution of the larger model. DistilBERT is considered more efficient in certain contexts due to its smaller size, requiring less computational resources to operate. Despite its reduced size, DistilBERT retains most of the performance capabilities of BERT, as the distillation process ensures that the smaller model learns a distilled version of the knowledge from the larger model.

Data Collection

This research utilizes information derived from Twitter posts acquired via the platform's API. The information gathered comes from streaming through one of Twitter's APIs, focusing on specific keywords and user handles. The dataset, curated from Twitter posts, is subsequently segmented into training, testing, and validation cohorts for waste pollution categories. This corpus is then subjected to an exhaustive analytical procedure to ascertain classifications aligned with specific waste type. Authors use a selection of keywords to extract posts that might encompass information details to the waste pollution highlighted issue. In total, 39 keywords were utilized, encompassing terms of categories in ‘limbah padat’ or solid waste, ‘limbah cair’ or liquid waste, ‘limbah gas’ or gas waste, and ‘limbah suara’ or noise waste.

Pre-processing

Pre-processing of Twitter data, especially when dealing with a language as rich and diverse as Indonesian, is a critical step to ensure the data is suitable for analysis(Ramdhani et al., 2020). The post often contains URLs, emojis, numbers, and special characters that might not add significant value. Many languages, often use slang, abbreviations, or informal terms. Mapping these to their standard equivalents or understanding their context is important for analysis. For example, 'jgn' could be expanded to 'jangan' (don't). Incorporating these pre-processing steps, tailored to the Indonesian language, can significantly improve the quality of data analysis, making the derived insights more accurate and actionable.

Reduction

The data reduction phase for waste pollution-related tweets aims to ensure that each tweet is unique and pertains to a mentioned waste pollution. During this stage, a manual inspection is conducted using a Python codes. If a tweet lacks related information, it will be removed from the dataset. Additionally, tweets consisting of fewer than four words will also be excluded.

Cleansing

The objective of this refinement is to ensure the text is predominantly noise-free, leading to a standardized presentation, thereby facilitating better analysis for computational engines. Integral steps in this process include removing duplicate, converting text to lowercase, excising symbolic characters, such as hashtag, emoji, mention/username symbol, and eliminating punctuations as well as embedded hyperlinks from the tweet's post. Girsang et al. (2021) Utilizing Python's foundational *regex* codes facilitates the meticulous cleaning procedure. A typical tweet content, as illustrated below, will undergo this cleaning to achieve a more cleaned format. *“Ciptakan Lingkungan Yang Bersih dan Nyaman, Polsek Depok Polresta Cirebon Bergotong Royong Bersama Masyarakat Membersihkan Sampah Plastik. #gotongroyong #PolriPresisi https://t.co/5vzmdCaPE8”*

Following the cleaning procedure, tweet above will change as follows: *“ciptakan lingkungan yang bersih dan nyaman, polsek depok polresta cirebon bergotong royong bersama masyarakat membersihkan sampah plastik gotongroyong polripresisi”*

Labeling

The labeling phase is carried out manually based on four main categories of waste pollution in Indonesian language: ‘limbah padat’ or solid waste, ‘limbah cair’ or liquid waste, ‘limbah gas’ or gas waste, and ‘limbah suara’ or noise waste. The term ‘limbah padat’ encompasses waste such as household waste, organic waste, plastic waste, carcasses, burning residue, medical waste, used cans, used bottles, used clothes, ash, and scrap metal. The ‘limbah cair’ refer to waste likedirty water, wastewater, polluted rivers, used oil, waste oil, residue, pesticides, toxins, sludge, factory discharge, sedimentation, chemical waste, chemical liquids, and ipal. Then ‘limbah gas’ refer to waste like emissions, carbon dioxide, carbon monoxide, greenhouse gases, air pollution, smoke, fires, and trash smells. On the other hand, ‘limbah suara’ refer to waste like noise, noise pollution, noise disturbance, exhaust sound, and sirens. Initially, Authors identifies tweets pertaining to this waste pollution. During this labeling process, these 39 distinct wastes serve as the foundational reference, subsequently bifurcating them into the four overarching categories as illustrated in Table 1. Authors further refines this classification manually, examining each tweet within *prodi.gy annotator* for accuracy(prodigy, 2017). Examples of this labeling methodology are showcased in Table 2.

Table 1
Classification Keywords

Classification Label	Keywords
Solid Waste	household waste, organic waste, plastic waste, carcasses, burning residue, medical waste, used cans, used bottles, used clothes, ash, scrap metal
Liquid Waste	dirty water, waste water, polluted rivers, used oil, waste oil, residue, pesticides, toxins, sludge, factory discharge, sedimentation, chemical waste, chemical liquids, ipal
Gas Waste	emissions, carbon dioxide, carbon monoxide, greenhouse gases, air pollution, smoke, fires, trash smells
Voice Waste	noise, noise pollution, noise disturbance, exhaust sound, noise, sirens

Table 2
Waste Pollution Labeling Example

Tweet	Label
create a clean and comfortable environment, Depok Police Cirebon Police work together with the community to clean up plastic waste mutual cooperation Polripresisi	Solid Waste

Tokenization

In the intricate process of tokenization, each tweet undergoes a transformation where it's deconstructed into its constituent words. This step is paramount in understanding and analyzing the granular elements of the content. DistilBERT comes equipped with its dedicated tokenization method called DistilBertTokenizer(Sanh et al., 2019). This tokenizer leverages the pre-trained 'distilbert-base-uncased' model to ensure that the tokenization aligns with contemporary linguistic patterns. Table 3 showcase a representative sample of the tokenization outcome.

Table 3
Waste Pollution Classification Tokenization

Tweet	Label	Tokenization
create a clean and comfortable environment, Depok Police Cirebon Police work together with the community to clean up plastic waste mutual cooperation Polripresisi	Solid Waste	['create', 'environment', 'yang', 'clean', 'and', 'comfortable', 'polsek', 'depok' 'polresta', 'cirebon', 'bergotong', 'royong', 'together ', 'society', 'cleaning', 'trash', 'plastic', 'mutual cooperation', 'polripresisi']

Training

Author compiled a training dataset for waste pollution by ensuring an equal number of tweets(Kim & Hwang, 2022; Werner de Vargas et al., 2023), 10,000 each and a testing dataset is 1000 each, for ‘limbah padat’,’limbah cair’, ‘limbah gas’ and ‘limbah suara’ categories.

The training phase is conducted using the Python programming language by invoking the DistilBertForSequenceClassification function from the DistilBERT library (Kici et al., 2021). The data is partitioned into an 80% segment for training purposes and a 20% allotment for validation.

Evaluation

The evaluation will employ the confusion matrix model to determine the values of true positive, true negative, false positive, and false negative by comparing the actual values and predicted values(Gaye & Wulamu, 2019). This will help in assessing the performance of DistilBERT in the classification accuracy of waste pollution.

3. Results and Discussion

Descriptive Analysis

The research leveraged a substantial dataset collected over various periods to analyze conversations about waste pollution on Twitter, specifically focusing on the Indonesian language. The dataset comprises two main segments: the training dataset and the testing dataset. The training dataset, used to teach the model how to classify tweets accurately, consisted of 561,666 tweets collected from October 2022 to June 2023. The testing dataset, used to evaluate the model's performance, included 213,406 tweets gathered between August 2022 and September 2022. This broad collection strategy aimed to capture a wide array of public discussions on waste pollution without limiting the sources to specific journalistic accounts, thereby ensuring an authentic and comprehensive representation of public discourse.

The initial dataset of 775,072 tweets underwent rigorous preprocessing, including reduction and cleansing, to refine the dataset's quality and relevance. This process reduced the dataset to approximately 27% of its original volume, resulting in 215,376 tweets, split into 159,139 for training and 56,237 for testing. Further refinement was undertaken through manual filtering to isolate tweets exclusively related to waste pollution, significantly narrowing down the volumes to 40,000 tweets for training and 4,000 tweets for testing. This step was crucial to ensure the model's focus on accurately identifying and classifying discussions related to waste pollution.

To aid in the model's learning process and validation, the training dataset was segmented further. As shown in Table 4, a validation set, constituting 20% of the training data, was established to fine-tune the model and adjust its parameters for optimal performance. Consequently, the final dataset allocations were 32,000 tweets for training, 8,000 tweets for validation, and 4,000 tweets for testing. This structured dataset distribution supports the model's development and evaluation, ensuring it effectively classifies waste pollution-related discussions within the vast sea of social media conversations.

Table 4

Dataset Distribution

Dataset	Count	Percentage
Training Dataset	40,000 tweets x 80% = 32,000 tweets	72.7%
Validity Dataset	40,000 tweets x 20% = 8,000 tweets	18.1%
Testing Dataset	4,000 tweets	9.2%
Total	44,000 tweets	100%

The study demonstrated that the DistilBERT model achieved significant and notably good results in classifying tweets into various categories related to waste pollution. Specifically, the model excelled in identifying tweets pertaining to categories such as recycling initiatives, plastic pollution, and community clean-up efforts. The accuracy and precision in these categories were markedly high, showcasing DistilBERT's capability to discern nuanced discussions within the broader theme of waste pollution.

Training Model Analysis

The training process is carried out using configuration set to optimize its training and evaluation processes(Silva Barbon & Akabane, 2022). Following Table 5, the batch size, which dictates the number of samples processed before the model is updated, is set to 8 for both training batch_size and evaluation batch_size. The training is set to iterate over the entire dataset ten times, as indicated by ‘num_train_epochs=10’. Evaluation and logging are configured to occur at the end of each epoch Both training and evaluation are enabled, as denoted by ‘do_train=True’ and ‘do_eval=True’. The ‘no_cuda=False’ means that GPU acceleration, if available, will be used. Additionally, the model is set to load the best version at the end of the training, ensuring that the most optimized iteration is retained. Upon the completion of this training regimen, the waste pollution classification model is evaluated, with its accuracy being gauged based on the testing data.

Table 5

DistrilBERT Training Config

Configuration	Value
per_device_traing_batch_size	8
per_device_eval_batch_size	8
num_train_epoch	10
evaluation_strategy, save_strategy	true
do_train, do_eval	true
load_best_model_at_end	true

As Table 6, The DistilBERT model completed its training with a total of 40,000 data. Over the course of the training, it achieved a training loss of 0.0453, indicating the difference between the model's predictions and the actual results. This low loss suggests that the model was able to fit the training data effectively.

The training runtime was recorded at approximately 1,762.16 seconds. The efficiency of the model can be gauged by the samples processed per second, which stood at 181.595. And total floating-point operations (FLOPs) during training were about 1.639 x 10^16. Furthermore, the model averaged 22.699 steps per second.

Table 6

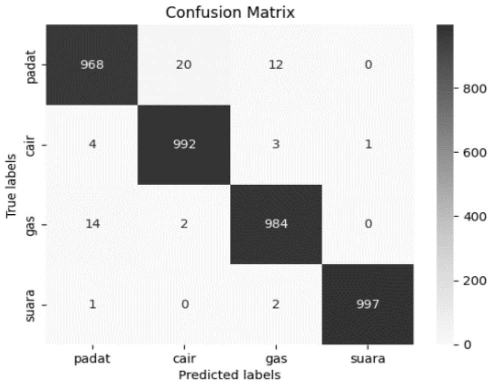
DistilBERT Training Epoch

Epoch	Training Loss	Validation Loss
1	0.073200	0.049474
2	0.152400	0.053678
3	0.136600	0.053864
4	0.004200	0.049860
5	0.096200	0.053140
6	0.269200	0.049278
7	0.000300	0.054914
8	0.150500	0.061521
9	0.062500	0.061789
10	0.030700	0.065411

For overall performance in Figure 3, The DistilBERT model showcased an impressive performance with an overall accuracy of 99% across 4,000 testing data. This high accuracy rate demonstrates the model's efficiency in classifying the given data into its respective categories.

Figure 3

DistilBERT Confusion Matrix



Both the macro average and the weighted average for precision, recall, and F1-score were consistently at 0.99. These averages further emphasize the model's robustness and its ability to maintain a high standard of accuracy across diverse categories as seen in Table 7.

Table 7

Classification using DistilBERT Evaluation

	Precision	Recall	F1-score	Support
Congested	0.98	0.97	0.97	1000
Liquid	0.98	0.99	0.99	1000
Gas	0.98	0.98	0.98	1000
Voice	1.00	1.00	1.00	1000
Accuracy			0.98	4000
macro avg	0.98	0.98	0.98	4000
weigthed avg	0.98	0.98	0.98	4000

On notes, in overall of waste pollution categories, the "suara" (voice) category stands out due to its impeccable identification. One possible reason for this unmatched accuracy could be the limited vocabulary in the Indonesian language that commonly denotes something as waste. In languages, certain terms might have distinct and unambiguous meanings, making them easier to classify with high precision. Furthermore, the volume of data plays a significant role in determining the quality of the outcomes. A substantial dataset can provide a more comprehensive representation, allowing for more accurate predictions and classifications. However, it's also essential to recognize that not just the quantity, but the quality and diversity of data can influence the results. In the case of "suara," it appears that both the linguistic characteristics of the Indonesian language and the dataset's properties worked in tandem to produce such impeccable results.

Model Comparison

A comparison between models that are commonly employed for basic text classification, such as TF-IDF SVM and Naive Bayes (Colas & Brazdil, 2006). Additionally, Authors evaluated one of the transformer-based models, RoBERTa, ensuring that its configuration was standardized to provide a fair comparison. The results of this comparative analysis are presented in Table 8.

Table 8
Model Comparison

Model	Precision	Recall	F1-Score
DistilBERT	98%	98%	98%
TF-IDF SVM	97%	97%	97%
TF-IDF Naïve Bayes	49%	47%	47%
RoBERTa	93%	93%	93%

The results show that the TF-IDF SVM has a relatively high performance compared to the TF-IDF Naïve Bayes. In contrast, among the transformer models tested in this study, BERT variations like RoBERTa scored slightly lower than DistilBERT in terms of accuracy. Further analysis was conducted using Cross Validation to assess model fitting, specifically to determine if the models were well-fitted, over-fitting, or under-fitting. The researchers employed cross-validation with 7 epochs to compare the accuracy of Training and Validation. The outcomes, as detailed in Table 9, indicate that the models are approaching a state of over-fitting.

Table 9
Cross Validation

	Accuracy
Training	99.37%
Validity	99.01%

4. Conclusion

Through Twitter, communities have the capability to swiftly share and access information regarding waste pollution in their local environments. This demonstrates that data gathered from Twitter can serve as a viable source for text classification processes related to waste pollution. The efficacy of the text classification, particularly using DistilBERT in this study, is influenced by multiple parameters optimized for the system. One of the key parameters is the writing behavior associated with waste terminology. The way waste is described significantly impacts the accuracy of waste pollution category identification. Even though the model has been trained to recognize characteristics specific to the Indonesian language, the incorporation of colloquial expressions or slang can pose challenges, potentially leading to misinterpretations. Furthermore, when benchmarked against other text classification models, DistilBERT showcased superior performance, achieving a remarkable F1-score of 99%. This top-tier accuracy, compared to other models, underscores its potential in enhancing the reporting and analysis of waste pollution. Such advancements can be pivotal for both communities and governmental bodies, enabling them to take proactive measures based on real-time reports and thorough waste pollution analyses.

5. References

Adel, H., Dahou, A., Mabrouk, A., Elsayed Abd Elaziz, M., Kayed, M., El-henawy, I., Alshathri, S., & Ali, A. (2022). Improving Crisis Events Detection Using DistilBERT with Hunger Games Search Algorithm. *Mathematics*, *10*, 447. https://doi.org/10.3390/math10030447

Colas, F., & Brazdil, P. (2006). Comparison of SVM and some older classification algorithms in text classification tasks. *IFIP International Conference on Artificial Intelligence in Theory and Practice*, 169–178.

Gaye, B., & Wulamu, A. (2019). Sentiment Analysis of Text Classification Algorithms Using Confusion Matrix. *Cyberspace Data and Intelligence, and Cyber-Living,* *Syndrome, and Health: International 2019 Cyberspace Congress, CyberDI and CyberLife, Beijing, China, December 16–18, 2019, Proceedings, Part I 3*, 231–241.

Girsang, A. S., Isa, S. M., & Fajar, R. (2021). Implementation of a Geocoding in Journalist Social Media Monitoring System. *International Journal of Engineering Trends and Technology*, *69*(12), 103–113. https://doi.org/10.14445/22315381/IJETT-V69I12P212

Gokkon, B. (2022). *Plastic accounts for three-fourths of waste choking Jakarta's rivers*. Mongabay.Com.

Kemp, S. (2023). *DIGITAL 2023: INDONESIA*. Dataportal.

Kici, D., Malik, G., Cevik, M., Parikh, D., & Basar, A. (2021). A BERT-based transfer learning approach to text classification on software requirements specifications. *Canadian Conference on AI*, *1*, 042077.

Kim, M., & Hwang, K.-B. (2022). An empirical evaluation of sampling methods for the classification of imbalanced data. *PLOS ONE*, *17*(7), e0271260-. https://doi.org/10.1371/journal.pone.0271260

Lie, S. (2018). *Multi-Class Text Classification Model Comparison and Selection*. Toward Data Science.

Masoumi, F. S., & Bahrani, M. (2022). *Utilizing distilBert transformer model for sentiment classification of COVID-19's Persian open-text responses*.

Nasir, M., Saputro, D. E. P., & Handayani, S. (2015). Manajemen Pengelolaan Limbah Industri. *Jurnal Manajemen Dan Bisnis*, *19*(2), 143–149.

Prodigy. (2017). *Text Classification*. Https://Prodi.Gy/Docs/Text-Classification.

Ramdhani, M. A., Ramdhani, M. A., Maylawati, D. S. adillah, & Mantoro, T. (2020). Indonesian news classification using convolutional neural network. *Indonesian Journal of Electrical Engineering and Computer Science*, *19*(2), 1000–1009. https://doi.org/10.11591/ijeecs.v19.i2.pp1000-1009

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, *abs/1910.01108*. http://arxiv.org/abs/1910.01108

Shaheen, Z., Wohlgenannt, G., & Filtz, E. (2020). *Large Scale Legal Text Classification Using Transformer Models*. http://arxiv.org/abs/2010.12871

Silva Barbon, R., & Akabane, A. T. (2022). Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study. *Sensors*, *22*(21), 8184.

Solla, D. G. (2022). *Advanced waste classification with Machine Learning: Improving waste classification techniques*. Towards Data Science.

Statista. (2023). *Number of Twitter users in Indonesia from 2014 to 2019*. Statista.Com.

Tran-Thien, V. (2023). *7 Text Classification Techniques for Any Scenario*. Dataiku.Com.

Wage Komarawidjaja. (2016). Sebaran Limbah Cair Industri Tekstil Dan Dampaknya Di Beberapa Desa Kecamatan Rancaekek Kabupaten Bandung. *Jurnal Teknologi Lingkungan*, *17*(2).

Werner de Vargas, V., Schneider Aranda, J. A., dos Santos Costa, R., da Silva Pereira, P. R., & Victória Barbosa, J. L. (2023). Imbalanced data preprocessing techniques for machine learning: a systematic mapping study. *Knowledge and Information Systems*, *65*(1), 31–57. https://doi.org/10.1007/s10115-022-01772-8

William, R. (2021). *Hugging Face Transformers: Fine-tuning DistilBERT for Binary Classification Tasks*. Towards Data Science.

World Bank. (2021). *Plastic Waste Discharges from Rivers and Coastlines in Indonesia*. www.worldbank.org