



OPEN Key insights into recommended SMS spam detection datasets

Mohammad Firdaus Johari¹, Kang Leng Chiew^{1✉}, Abdul Razak Hosen¹, Kelvin S.C. Yong², Adnan Shahid Khan¹, Irshad Ahmed Abbasi^{3✉} & Daniel Grzonka⁴

Short Message Service (SMS) spam poses significant risks, including financial scams and phishing attempts. Although numerous datasets from online repositories have been utilized to address this issue, little attention has been given to evaluating their effectiveness and impact on SMS spam detection models. This study fills this gap by assessing the performance of ten SMS spam detection datasets using Decision Tree and Multinomial Naïve Bayes models. Datasets were evaluated based on accuracy and qualitative factors such as authenticity, class imbalance, feature diversity, metadata availability, and preprocessing needs. Due to the multilingual nature of the datasets, experiments were conducted with two stopwords removal groups: one in English and another in the respective non-English languages. The key findings of this research have led to the recommendation of Dataset 5 for future SMS spam detection research, as evidence from the dataset's high qualitative assessment score of 3.8 out of 5.0 due to its high feature diversity, real-world complexity, and balanced class distribution, and low detection rate of 86.10% from Multinomial Naïve Bayes. Recommending a dataset that poses challenges for high model performance fosters the development of more robust and adaptable spam detection models capable of handling diverse forms of noise and ambiguity. Furthermore, selecting the dataset with the highest qualitative score enhances research quality, improves model generalizability, and mitigates risks related to bias and inconsistencies.

Keywords SMS spam detection, Dataset evaluation, Stopwords removal, Machine learning, Dataset recommendation

Short Message Service (SMS) is the most prevalent mode of communication in today's interconnected society. Developed by German engineer Friedrich Hillebrand and his colleague Bernard Ghillebaert in 1984, SMS was conceived as a method for transmitting messages via the telephone network using GSM standards, a vision that came to fruition in the 1990s¹. SMS is distinct from Multimedia Messaging Service (MMS) due to its character limitations and its incapacity to transmit videos, audios, and images over traditional cellular networks^{1–3}.

While SMS offers numerous societal benefits, such as facilitating communication, it also creates opportunities for cybercriminals to innovate their deceptive tactics. One such tactic is the dissemination of spam messages, which are unsolicited electronic communications^{4,5}. The appeal of SMS for spam lies in the availability of unlimited pre-paid SMS packages in countries like India, Pakistan, China, and increasingly, the United States. Additionally, SMS spam often achieves higher response rates compared to email spam, as SMS is a trusted service that users frequently rely on for confidential exchanges. Consequently, SMS spam has emerged as a significant issue, imposing substantial costs related to lost productivity, network bandwidth consumption, management overhead, and compromised personal privacy⁶.

The incidence of SMS spam remains persistently high, particularly within the United States⁷, observed a notable uptick in spam text activity, with a 58.0% increase reported in 2022. Additionally⁷, noted that 1 in every 3 Americans encountered fraudulent schemes through spam texts, with 65% of these individuals only realizing they had been deceived after the fact. These findings are corroborated by⁸, who documented a staggering 157.0% surge in spam texts among Americans, amounting to 225 billion such messages in 2022 alone. Furthermore, Orred's research revealed that individuals between the ages of 18 and 44 are particularly vulnerable to financial losses from phone scams, with 55.6% being male, 42.2% female, and 2.3% identifying as non-binary. The severity of the issue is graphically depicted in Fig. 1, as presented by⁹.

¹Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Kota Samarahan, Sarawak 94300, Malaysia. ²Faculty of Engineering, Computing and Science, Swinburne University of Technology, Sarawak Campus, Jalan Simpang Tiga, Kuching, Sarawak 93350, Malaysia. ³Department of Computer Science and Artificial Intelligence, College of Computing and Information Technology, University of Bisha, Bisha 67714, Saudi Arabia. ⁴Department of Computer Science, Cracow University of Technology, Krakow, Poland. ✉email: klchiew@unimas.my; aabasy@ub.edu.sa

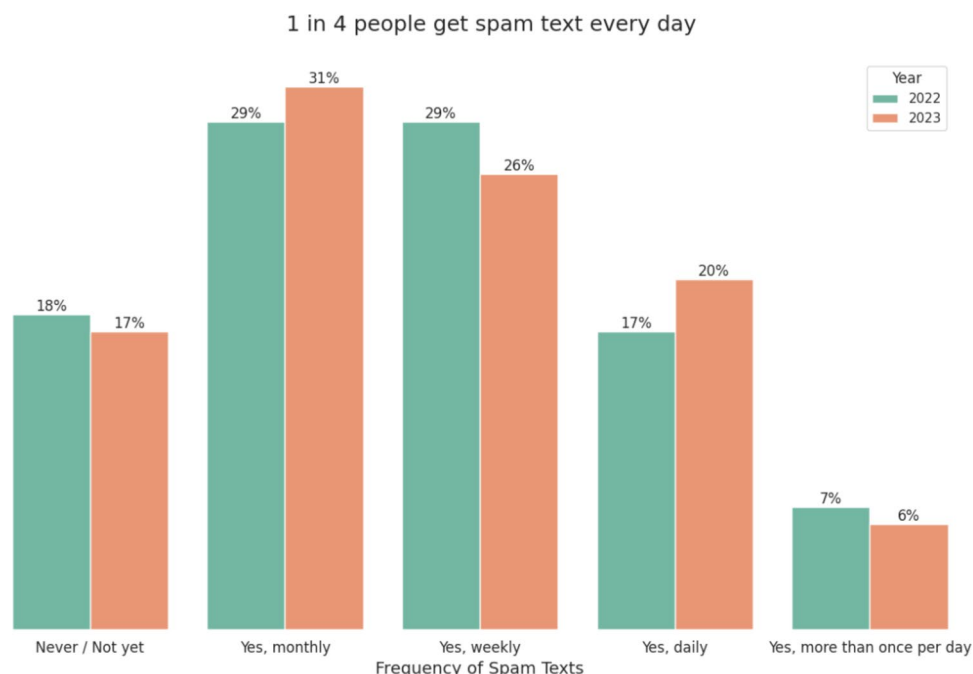


Fig. 1. The Growth of Spam Texts Among the Americans by⁹.

The detection of SMS spam messages offers a myriad of advantages, chief among them being the mitigation of financial losses and the restoration of trust in communication channels within society. SMS spam typically encompasses fraudulent schemes devised to deceive recipients into making payments or providing services, thereby compromising the integrity of the communication medium. Over time, such deceptive practices erode the credibility and reliability of SMS as a communication channel, resulting in decreased engagement with vital messages¹⁰. Furthermore, the identification of SMS spam messages enhances user privacy and safeguarding of sensitive information, thereby enhancing overall user experience and mitigating the intrusion of unsolicited content^{11,12}.

Previous studies on SMS spam detection have employed a variety of algorithms, each demonstrating notable algorithmic performance. For instance¹³, had tested the performance of Support Vector Machine models, achieving accuracies of 97.8%⁵. fine-tuned the hyperparameters of a Convolutional Neural Network, achieving an accuracy of 99.44%¹⁴., on the other hand, utilized multiple models with various word embedding techniques, finding that LSTM achieved the highest accuracy at 98.5%. Furthermore¹⁵, experimented with a modified transformer model for SMS spam detection, reaching an accuracy of 98.9% through hyperparameter tuning.

While these studies highlight the potential of advanced machine learning algorithms in detecting SMS spam, the effectiveness of these models heavily depends on the quality of the datasets used for training and evaluation. SMS spam messages pose significant challenges globally, with billions of fraudulent messages sent annually, leading to financial losses and privacy concerns. To combat this issue, researchers rely on publicly available datasets to train machine learning models for spam detection. However, while these datasets are widely used, their quality and suitability for robust spam detection remain underexplored. The performance of spam detection models is highly dependent on the characteristics of the datasets used, such as class balance, noise, and feature diversity. Yet, to the best of the authors' knowledge, no comprehensive study has evaluated the quality of SMS spam detection datasets to understand their impact on model performance.

This study aims to fill this gap by providing insights and recommendations on ten publicly available SMS spam detection datasets. Rather than focusing solely on classifier accuracy, this research emphasizes understanding dataset characteristics and their influence on spam detection performance. Specifically, this study evaluates datasets based on quantitative and qualitative metrics, analyze the performance of Decision Tree and Multinomial Naïve Bayes models, and recommend datasets for future research based on their challenge level and quality. It is hypothesized that the characteristics of the dataset has a significant impact on the model performance, with Multinomial Naïve Bayes expected to outperform Decision Tree due to its robustness in handling high-dimensional text data. Additionally, it is anticipated that the most challenging dataset, which will be most recommended dataset, will provide valuable testbeds for improving model adaptability and robustness, offering insights into dataset selection for future SMS spam detection research. The criteria of the recommended dataset must present a significant challenge for models to achieve high accuracy compared to other datasets and must attain the highest average Likert score in qualitative assessments to ensure the quality and credibility of the recommendation.

By offering a structured framework for dataset evaluation, this study contributes to the development of more robust machine learning models and informs future research on spam detection systems. The primary contributions of this work include:

1. A comparative analysis of Decision Tree and Multinomial Naïve Bayes model performance across all datasets.
2. An evaluation of the quality of ten publicly available SMS spam detection datasets using quantitative and qualitative metrics.
3. Dataset recommendation for advancing spam detection research using datasets with varying complexity and characteristics.

Literature review

Previous work on SMS spam detection

Throughout previous studies on SMS spam detection, a dataset named SMS Spam Collection v.1 has been previously utilized by^{15–18}. This dataset comprises 5,574 text messages, with 4,827 classified as ham (legitimate) and 747 classified as spam.

Utilizing the SMS Spam Collection v.1 and UtkML's Twitter Spam Detection Competition dataset^{15,19} introduced a modified Transformer model for detecting SMS spam messages. This model was compared against traditional machine learning algorithms and deep learning algorithms, such as LSTM (Long Short-Term Memory) and CNN-LSTM (Convolutional Neural Network - Long Short-Term Memory). The authors found that their proposed model outperformed all other compared models across all tested metrics. Specifically, it achieved an accuracy of 98.92% on the SMS Spam Collection v.1 dataset and 87.06% on UtkML's Twitter dataset.

¹⁶ concentrate on enhancing SMS spam detection through the introduction of new content-based features. Their aim is to bolster the performance of spam detection methods by integrating semantic categories of words as features and reducing the feature space. Utilizing a dataset similar to those used in previous studies, namely the SMS Spam Collection v.1, the authors employ a variety of models, including Naive Bayes (NB), k-Nearest Neighbors (kNN), kNN45, Support Vector Machine (SVM), Information Theoretic Co-Training, and Boosted-Random Forest. The findings indicate that the Boosted-Random Forest algorithm attained the highest accuracy of 98.47% and the highest Matthews Correlation Coefficient of 0.934 among all algorithms evaluated. Furthermore, the Boosted-Random Forest exhibited the highest Sensitivity at 89.1% and the lowest Balanced Hit Rate at 0.1%. Overall, the Boosted-Random Forest algorithm surpassed other models in terms of accuracy, Matthews Correlation Coefficient, and sensitivity, establishing itself as the most effective model for SMS spam detection in this study.

¹⁷ employed this dataset along with SMS Spam Corpus v.0.1 Big to enhance SMS spam detection on mobile phones by integrating FP-Growth for frequent pattern mining and Naive Bayes Classifier for classification. They found that the SMS Spam Collection v.1 dataset alone yielded the highest accuracy of 98.51% with a 9% minimum support, while the SMS Spam Corpus v.0.1 Big dataset's accuracy improved by 1.15% when utilizing FP-Growth. Moreover, combining both datasets resulted in an accuracy of 98.47% with a 6% minimum support. The authors concluded that the quantity and quality of the dataset significantly influence the accuracy of the models, noting an inverse relationship between dataset quantity and the minimum support parameter. Specifically, as the dataset size increases, the required minimum support parameter for optimal performance decreases. Conversely, smaller datasets require higher minimum support values to avoid generating excessive or irrelevant features, which could negatively impact the classification model's accuracy.

In an effort to detect SMS spam messages using the H2O framework¹⁸, employed the SMS Spam Collection v.1 dataset to evaluate the performance of various machine learning algorithms, including Random Forest, Deep Learning, and Naïve Bayes, for SMS classification. Emphasizing the significance of features such as the number of digits and the presence of URLs in accurately identifying SMS spam messages, the authors found that Random Forest recorded the highest precision, recall, F-measure, and accuracy scores compared to the other models, albeit with the slowest runtime. Conversely, while Naïve Bayes achieved lower scores in precision, recall, F-measure, and accuracy, it excelled in terms of runtime efficiency.

Previous work on data curation and data quality assessment

²⁰ defines data curation as the process of organizing and managing large volumes of data to streamline the annotation process. Data curation determines the starting and labelling points to ensure efficient resource use when handling extensive datasets. The author emphasizes that data curation is crucial for businesses aiming to optimize their data processes, save significant time for machine learning engineers, focus more on model development, and facilitate the integration of models into business workflows. Essentially, data curation involves managing, annotating, and organizing data to ensure it is of the highest quality, accessible, and usable, which is essential for achieving optimal model performance²⁰.

Before delving into the previous practice of data curation, it is essential to understand the issues related to data quality²¹. have identified a range of issues that can compromise dataset quality, thereby affecting the integrity and effectiveness of machine learning models. In addition to common data quality issues such as spelling errors, duplicate records, conflicting fields, and inconsistencies, the authors also highlighted other problems including insufficient metadata, labelling errors, pre-processing challenges, dataset biases, and low annotation quality from crowdsourced data, a claim that is supported by²².

In light of these data quality issues²¹, have reported the risks associated with the aforementioned dataset quality problems. Poor data quality can lead to a decline in model performance, evidenced by decreased metrics such as accuracy, precision, and recall. Additionally, model reliability and stability are compromised, as poor-quality datasets can render models unreliable or unstable, diminishing their practical utility²¹. Furthermore, these issues can result in incorrect or misleading conclusions, posing risks and potentially causing losses in business decisions. They also present security threats, such as privacy breaches and susceptibility to malicious attacks²¹.

To mitigate these risks, various data curation criteria and quality assessments have been proposed²², for example, outlined a four-step process: assessing datasets, identifying data quality issues, evaluating metadata, and preparing a metadata report. Although the data evaluation framework proposed by²³ is aimed at Intrusion Detection System (IDS) datasets, it also validates the data quality assessment framework proposed by²². For instance²³, evaluated the IDS dataset based on its completeness in network configuration and traffic to accurately represent real-world scenarios and simulate genuine attack behavior. The authors also assessed dataset reliability by examining the accuracy of tagging and labelling. Furthermore²³, emphasized the importance of comprehensive data documentation, or metadata, noting that insufficient metadata reduces a dataset's usability for other researchers.

Additionally²³, introduced several unique data evaluation criteria, such as anonymity and heterogeneity. The authors highlighted the importance of privacy protection in datasets as well as balancing privacy concerns to maintain the dataset's usefulness. Moreover, the authors had also pointed out that incorporating heterogeneous data sources can enhance detection capabilities, making the dataset more valuable to the research community.

Looking at higher-dimensional data evaluation metrics²¹, utilized the concept of the dataset lifecycle, which comprises several processes from data collection to data destruction, as illustrated in Fig. 2. As visualized in the figure, each of these processes has its own quality evaluative metrics which contributes to the overall quality of the data.

The overall work of²¹ is particularly beneficial for individuals interested in generating their own datasets. For those who, like in the current research, are focused on collecting and utilizing publicly available datasets, only data testing portion of²¹ is pertinent, in which this research will employ. Additionally, the metrics involved in data collection and data annotation can optionally be used to enhance the interpretability of the results and to ensure that the outcomes are more reliable and accurate.

Methodology

Methodology description

In this research, a comparative analysis – based approach is utilised to evaluate the quality of ten available SMS spam detection dataset. This methodology is presented in Fig. 3, and is a similar replication of the methodology used by²⁴. It consisted of several phases: Problem Understanding, Data Collection, Data Understanding, Data Preparation, Modelling, and Evaluation.

The first phase involved identifying the problem and the objectives of the research. This included choosing a topic of interest and pinpointing relevant issues that require further study in the area of SMS spam detection. Additionally, research objectives were formulated to define the scope of the research. A thorough literature review of SMS spam detection was conducted, with particular reference to the work of²⁵. The work of²⁵ was instrumental in studying the standard structure of SMS spam datasets and the spam detection code used. Furthermore, previous data curation practices from the past few years has also been reviewed to provide more insight on the process to be implemented in this research.

The second phase was data collection, which involved sourcing previously published research articles in the area of SMS spam detection that included the authors' data and spam detection code. This step was crucial as it provided an opportunity to study the multiple view of the structure of SMS spam datasets and the associated

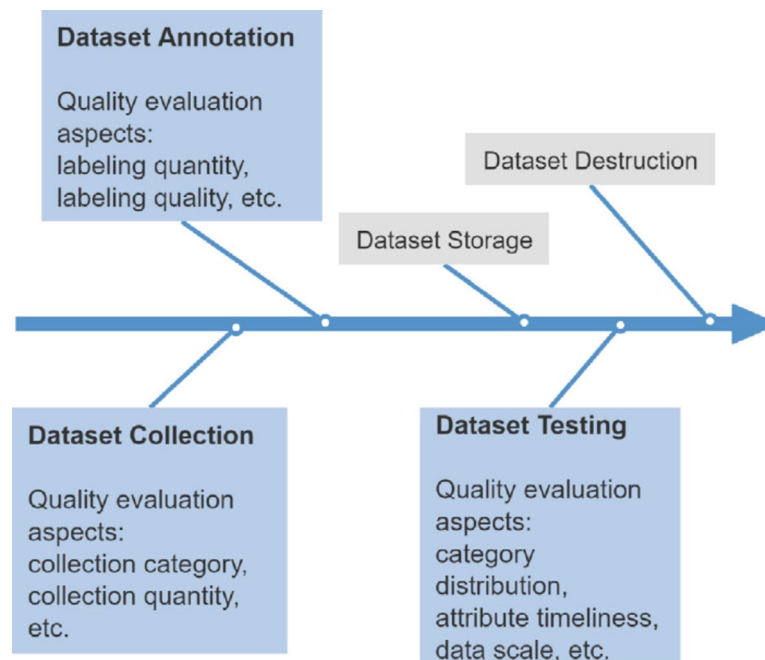


Fig. 2. Data Lifecycle, as proposed by²¹.

The Methodology Adopted For This Research

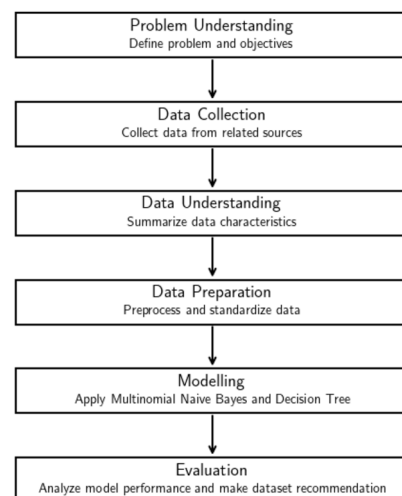


Fig. 3. The methodology used in this research as replicated from²⁴.

detection code. Following this, additional SMS spam datasets were collected from various platforms, including GitHub, Kaggle, and Google Dataset Search. Collecting datasets from these diverse and reputable sources was essential to ensure a comprehensive and robust dataset for this research. The dataset selection process is not restricted to English-language datasets. Any dataset will be utilized in this research unless it lacks essential components, such as raw text messages or labels (spam/ham), rendering it unsuitable for analysis.

The third phase involves understanding the collected data by summarizing the characteristics and attributes of each dataset. Akin to the dataset testing phase proposed by²¹, this phase includes identifying the number of rows and columns, the language used as well as the similarities and differences between each dataset. This phase concurrently involves identifying the quality criteria of the dataset. As discussed in more detail in later section, the criteria consist of the authenticity of the source, class imbalance, the diversity of features, the availability of metadata, and the data preprocessing requirement. This leads to the understanding the quality level of the data which helps to identify the subsequent steps needed for analysis. This step is crucial for the qualitative section of the discussion, where the correlation between the structure of the dataset and the model performance and data reusability is discussed. Additionally, this phase involves understanding the spam detection code used by²⁵ for the subsequent phase. Acknowledging the choice of classification models by²⁵ had led to the formulation of a hypothesis that serves as an expectation of the outcome of the research.

The next phase is data preparation, a crucial step to prepare the data for modelling, which involves data preprocessing. Each dataset is then standardized to only two columns; v1 (category of message) and v2 (text message). This aligns with the authors' spam detection code, which only accepts these columns of data. Additionally, data integration was applied, since some datasets contains a collection of data in separated Excel files.

The fifth phase is to execute the spam detection code by²⁵ with the previously prepared data. The current research uses Google Colaboratory as the Integrated Development Environment for executing the code. Google Drive was also utilized and mounted to the Google Colaboratory notebook to mount the dataset to the spam detection code. The code uses two classification models; Decision Tree and Multinomial Naïve Bayes. Additionally, the code involves splitting the data into training and testing set in a portion of 80:20, respectively.

The final phase is analysing the experimental results. The generated evaluation metrics and confusion matrix from both classification models were documented and studied to understand its performance. The research is then iterated by using different datasets to study the performance of the models. The most accurate and reliable models was documented during each execution of different dataset. Additionally, the experiment was iterated by modifying the code to remove the language stopwords of the datasets with language other than English, and the result was recorded, compared and justified, allowing for an unbiased recommendation of datasets.

Dataset description

The current study utilizes 10 publicly available datasets on SMS spam detection, obtained from relevant journal articles and dataset repositories. The datasets are in different languages or is a transliteration of language other than English. Each dataset exhibits varying levels of class imbalance. For instance, datasets 1, 3, and 9 display highly imbalanced class distributions, while datasets 6 and 8 have a perfectly balanced class distribution. The remaining datasets have moderately imbalanced distributions. In terms of source, most datasets (2, 5, 6, and 9) were collected from GitHub, while datasets 3, 8, and 10 were retrieved from Kaggle. Dataset 1 and dataset 7 were sourced from Archive and Zenodo, respectively. It is important to note that since Dataset 7 is non-proprietary, the collection for the dataset entails contacting the authors of the dataset through Zenodo before the authors grant access to the download link of the dataset. Unfortunately, the link to the source of dataset 4 and its associated metadata is unavailable; it was retrieved for this research on February 20, 2024. Table 1

Dataset	Citations	Class distribution	Gini Coefficient	Language	Source
1	26	Spam – 747 Ham – 4825	0.2179	English	https://archive.ics.uci.edu/dataset/228/sms+spam+collection
2	27	Spam – 2523 Ham – 2128	0.4998	Turkish	https://github.com/onrksy/TurkishSMS-Collection
3	28	Spam – 2241 Ham – 14,460	0.2911	English, German and French	https://www.kaggle.com/datasets/debapampal2002/sms-dataset1
4	Not accessible	Spam – 217 Ham – 286	0.4998	Bengali	Not accessible
5	29	Spam – 1571 Ham – 2456	0.4999	English	https://github.com/AbayomiAlli/SMS-Spam-Dataset
6	30	Spam – 1000 Ham – 1000	0.5000	English, and Hindi (Transliterated)	https://github.com/princebari/-SMS-Spam-Classification-on-Indian-Dataset-A-Crowdsourced-Collection-of-Hindi-and-English-Messages/blob/main/README.md
7	31	Spam – 2130 Ham – 2193	0.4999	Persian	https://zenodo.org/records/7832188
8	32	Spam – 574 Ham – 569	0.5000	Indonesian	https://www.kaggle.com/code/gevabriel/indonesian-sms-spam-detection-using-indobert/input
9	33	Spam – 74 Ham – 30	0.4999	Hindi (Transliterated)	https://github.com/paulpriyam/spamTransliteration/tree/master
10	34	Spam – 107 Ham – 77	0.4999	English and Hindi (Transliterated)	https://www.kaggle.com/datasets/uds5501/sms-dataset/data

Table 1. Overview of the dataset, including the class distribution, Gini coefficient, Language used, and source.

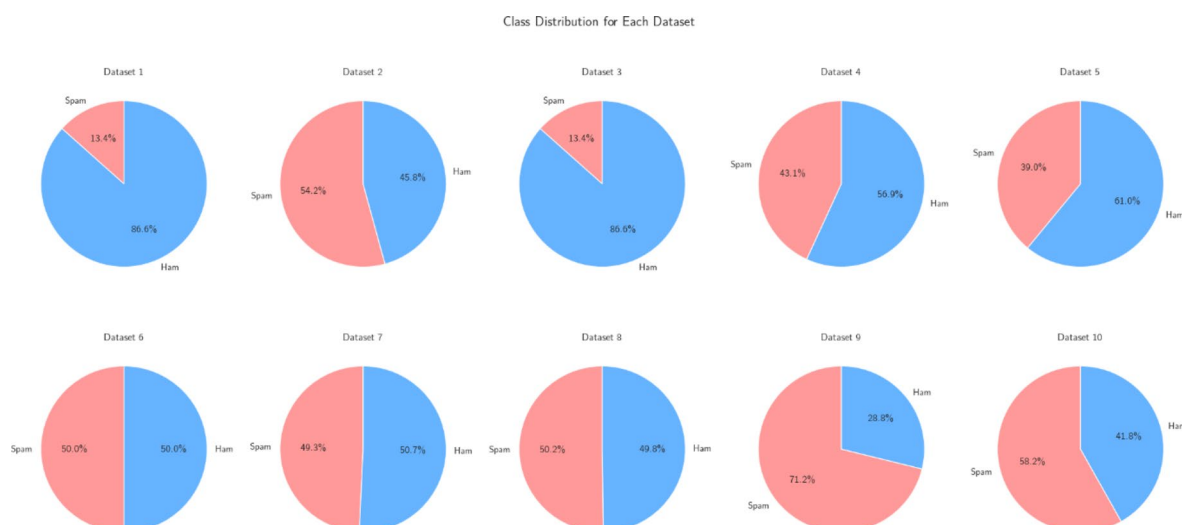


Fig. 4. Class Distribution for each dataset.

provides additional information on each dataset. Additionally, Fig. 4 illustrate the visualization of the class distribution for each dataset. As seen in Table 1, the datasets are either in English language (Dataset 1 and Dataset 5), monolingually non-English language (Dataset 2, Dataset 4, Dataset 7, and Dataset 8), transliterated non-English language (Dataset 9), monolingually non-English languages combined with English language (Dataset 3), or transliterated non-English language combined with English language (Dataset 6 and Dataset 10). The transliterated non-English language datasets are different from direct translation of non-English language datasets in a sense that the text structure is generally preserved in the transliterated dataset as opposed to being transformed entirely.

Dataset issues

These datasets contain many issues. Aside from duplicated and missing values, certain datasets had some spelling errors in the labelling column; ham or spam. Another problem in this column is the existence of white spaces after the label, which is difficult to detect when filtering the labels in Microsoft Excel. In addition, some of the data are made available in separate CSV files, which requires the integration of these data into a single file. Some dataset either has no header or has a header that uses different terms. Furthermore, certain datasets are in language other than English, which brings the problem of inconsistent data, especially since the dataset uses different categorical terms, such as 'Normal' and 'Spam' instead of ham or spam. Moreover, certain datasets lack

sufficient background information or has no metadata, posing a risk of providing limited contextual information to researchers who intend to utilize these datasets.

Risk associated with dataset issues

The aforementioned dataset issues pose significant risks to this research, particularly in terms of model performance and result reliability. Poor data quality can severely impact the effectiveness of SMS spam detection models, leading to inaccuracies, inconsistencies, and biased predictions. According to the Garbage-In-Garbage-Out (GIGO) principle, flawed input data inevitably produces flawed outputs, making it critical to address these data quality concerns.

Several common data quality problems influence model performance, including spelling errors, duplicate records, field conflicts, and inconsistencies. Spelling errors in spam messages can lead to misclassification, particularly when models rely on text-based features. Duplicate records can skew dataset distributions, resulting in biased model training and overfitting. Inconsistencies in labeling, such as the difference in the spelling of “Spam” and “SPAM”, or also known as field conflicts can cause confusion in classification models, reducing their ability to learn correct patterns, and producing unwanted results (non-binary results).

Other critical data quality issues that disturb the model performance include insufficient metadata, labeling errors, preprocessing challenges, dataset bias, and low-quality crowdsourced data. Insufficient metadata reduces dataset interpretability, making it difficult for researchers to understand the context and characteristics of the data, which in turn affects the interpretation of the model's output. Additionally, the lack of metadata reduces the datasets' reusability, as it impedes researchers' ability to fully comprehend and effectively utilize the data²². Labeling errors can lead to incorrect training signals, weakening model reliability and reducing classification accuracy. Preprocessing challenges, such as missing values and inconsistent column structures, will result in an increase effort required to prepare the datasets and may introduce additional errors. Dataset bias, particularly in unbalanced class distribution will result in models that favor the dominant class, which leads to poor generalization. Additionally, low-quality crowdsourced data often suffer from unreliable annotations and inconsistencies, which further deteriorate the model performance and also reduces the reusability of the dataset.

Moreover, flawed datasets can cause execution failures in machine learning pipelines, or even when successfully executed, produce misleading conclusions. When datasets contain errors or biases, model predictions may not reflect real-world SMS spam detection challenges, leading to false confidence in model performance. This discrepancy can result in ineffective spam filters, which may allow harmful messages to bypass detection or incorrectly classify legitimate messages as spam. Given these risks, the assurance of high-quality datasets is crucial for developing reliable and robust SMS spam detection models.

Preprocessing steps applied

To preprocess the datasets, all columns were removed except for two: labels and text/messages. These columns were then renamed to v1 and v2, respectively. In some datasets, all features were consolidated into a single cell. An example can be observed in Dataset 2, whereby each of the text message, the numerical category of the message (1 represent spam and 2 represent ham), and textual category of the message (spam or ham) is fitted into its individual cell, with semicolon being the separated value. To address this, the Text-to-Column feature in Microsoft Excel was used to split the single column into multiple columns.

Additionally, data integration was performed to handle the issue of separate data files, resulting in the consolidation of all data files into a single CSV file, a process particularly relevant for datasets 3 and 5. Outliers from the dataset were removed by identifying rows with missing values before deleting them to ensure data integrity. Other than that, trailing white spaces in the labels were trimmed to ensure correct binary output for the models. Furthermore, any rows in the label column containing unintended data, such as labels other than ‘spam’ or ‘ham’, were detected and removed. However, given the objective of this research, no steps were taken to address the issue of missing or insufficient metadata.

No synthetic data sampling techniques were applied to improve the dataset with class imbalance. However, the dataset underwent multiple preprocessing steps, including label encoding to convert categorical labels into numeric values, feature engineering by adding a character length column, and extensive text cleaning. Additionally, to remove the noises in the dataset, text preprocessing techniques were applied, which involved regular expression-based replacements for email addresses, URLs, currency symbols, phone numbers, and numerical values, followed by removing non-word characters. To further remove the noises, the textual data was normalized through lowercasing, tokenization, stopword removal, and stemming.

For stopword removal, different tools were used depending on the dataset's language. The Natural Language Toolkit (NLTK)³⁵ was used for English (Datasets 1, 3, and 5) and Indonesian (Dataset 8). SnowballStemmer (TurkishStemmer)³⁶ handled stopwords in Turkish (Dataset 2), while the Bengali Natural Language Processing Toolkit (bnlp_toolkit)³⁷ was used for Bengali (Dataset 4). Hazm³⁸ was applied for Persian (Dataset 7), and stopwords-iso³⁹ was utilized for Hindi (Datasets 6, 9, and 10). Finally, the cleaned text data was vectorized using CountVectorizer, the target labels were encoded with LabelEncoder, and the dataset was split into training and testing sets with an 80–20 ratio. Table 2 maps the previously discussed issues in 3.3 and preprocessing steps applied to each dataset, except Dataset 6 since it does not contain any of the aforementioned issues.

While preprocessing is a crucial step in addressing data quality, it is not sufficient on its own. A more systematic approach to data evaluation and quality assessment is essential to enhance dataset usability and improve model performance. Drawing on established data quality assessment frameworks²², outlines a four-step process that includes assessing the dataset to understand its structure, content, and potential gaps; identifying data quality issues such as duplicates, inconsistencies, missing values, and inaccuracies; evaluating metadata completeness and accessibility using a five-star scale; and compiling a comprehensive report documenting the dataset's quality, identified issues, and improvement recommendations for future researchers.

Dataset	Issues	Preprocessing steps applied
1	Imbalanced distribution of spam and ham	No preprocessing steps were applied
2	Inconsistent data due to usage of different categorical terms and existence of white spaces after the label, and header uses different terms	Standardize the categorical terms to 'spam' and 'ham', remove white spaces after each label, and standardize the terms for the included header
3	Data of different languages are stored in a separated CSV file, each header uses different terms, and imbalanced distribution of spam and ham	Integrate the data from each file into a single CSV file, and standardize the terms for the included header
4	Absence of metadata, has no header, and lost source of dataset	A header was added with standardized terms for the included column
5	Data are separated in different CSV files, and has no header	Integrate the data from each file into a single CSV file, and a header was added with standardized terms for the included column
7	Contains no header	A header was added with standardized terms for the included column
8	Header uses different terms and has no metadata	Standardize the terms for the included header
9	Header uses different terms and has no metadata	
10	Header uses different terms and lacks significant metadata information	

Table 2. Issues and preprocessing steps applied to each dataset.

Model	Hyperparameters
Decision Tree	max_depth = 5, 6, or 7, min_samples_split = 2, 3, or 4, min_samples_leaf = 2 or 3, criterion = gini, max_features = sqrt
Multinomial Naïve Bayes	alpha = 1, 2, or 5

Table 3. Hyperparameters used for each model.

For intrusion detection system datasets²³, proposes additional evaluation criteria to ensure high-quality datasets. These include having a “complete traffic” or balanced representation of legitimate (ham) and spam messages, as well as a clearly labeled dataset where messages are accurately categorized as “spam” or “not spam.” Stratified sampling can be employed to ensure the dataset includes a proportional mix of spam and non-spam messages across different demographics, regions, and languages to ensure the dataset is not skewed toward specific types of messages or user groups.

The authors also emphasize the importance of “attack diversity,” which refers to the inclusion of various spam types, such as phishing, promotional spam, scams, and malicious links. To enhance this diversity, diverse data collection sources should be leveraged, including multiple mobile carriers, regions, and user demographics. This is possible through a collaborative data sharing through partnerships with mobile carriers, app developers, or organizations which can provide access to a broader range of SMS data and further enrich the dataset’s heterogeneity. Techniques like federated learning can also be employed to train models on decentralized data sources without sharing raw data in order to ensure privacy while maintaining diversity.

The authors had emphasized that importance of “heterogeneity” to ensure the dataset reflects global SMS traffic. This involves incorporating messages from different carriers, regions, and user demographics. However, to protect user privacy, privacy-preserving techniques such as anonymization, differential privacy, and tokenization should be applied. For example, personally identifiable information (PII) like names, phone numbers, and addresses can be obfuscated or replaced with placeholders, while still retaining the structure and meaning of the text.

To capture the full context of SMS exchanges, the dataset should exhibit “complete interaction,” including details such as sender, receiver, timestamps, and associated metadata. This metadata can be anonymized through the aforementioned privacy-preserving techniques to protect the user’s privacy while still providing valuable context for analysis. Finally, a “complete network configuration” is necessary, which involves incorporating a diverse set of SMS sources, such as personal, promotional, and transactional messages, to ensure the dataset accurately represents real-world scenarios. The integration of these criteria will ensure the dataset can better capture the complexity and variability of real-world SMS traffic, ultimately improving the performance and reliability of spam detection models while protecting the privacy of the user and increasing the research value of the dataset.

Model hyperparameter tuning

The current study employs two traditional machine learning models, Decision Tree (DT) and Multinomial Naïve Bayes (MNB). To optimize the performance of the models, hyperparameter tuning was conducted using GridSearchCV with a five-fold cross-validation strategy. Both models were trained with a range of parameter values, and the best-performing hyperparameters were selected based on their respective datasets. Table 3 provides an overview of the training hyperparameters used for each model.

Experimental result

The current research serves to evaluate ten publicly available SMS Spam datasets by using Multinomial Naïve Bayes and Decision Tree. The experimental results of this study can be categorized into two groups: Removal of

Decision Tree							
Dataset	Accuracy	Precision (Ham)	Precision (Spam)	Recall (Ham)	Recall (Spam)	F1 - Score (Ham)	F1 - Score (Spam)
1	96.86%	0.97	0.95	0.99	0.83	0.98	0.89
2	98.28%	0.98	0.98	0.98	0.99	0.98	0.98
3	98.35%	0.99	0.94	0.99	0.94	0.99	0.94
4	78.22%	0.80	0.76	0.82	0.73	0.81	0.74
5	76.55%	0.87	0.65	0.73	0.82	0.80	0.72
6	92.50%	0.92	0.93	0.93	0.92	0.93	0.92
7	89.94%	0.89	0.91	0.91	0.89	0.90	0.90
8	89.96%	0.88	0.92	0.90	0.90	0.89	0.91
9	76.19%	0.64	0.90	0.88	0.69	0.74	0.78
10	86.49%	0.76	0.95	0.93	0.83	0.84	0.88

Table 4. Comparison results of various evaluation metrics from decision tree with english Language stopwords removal.

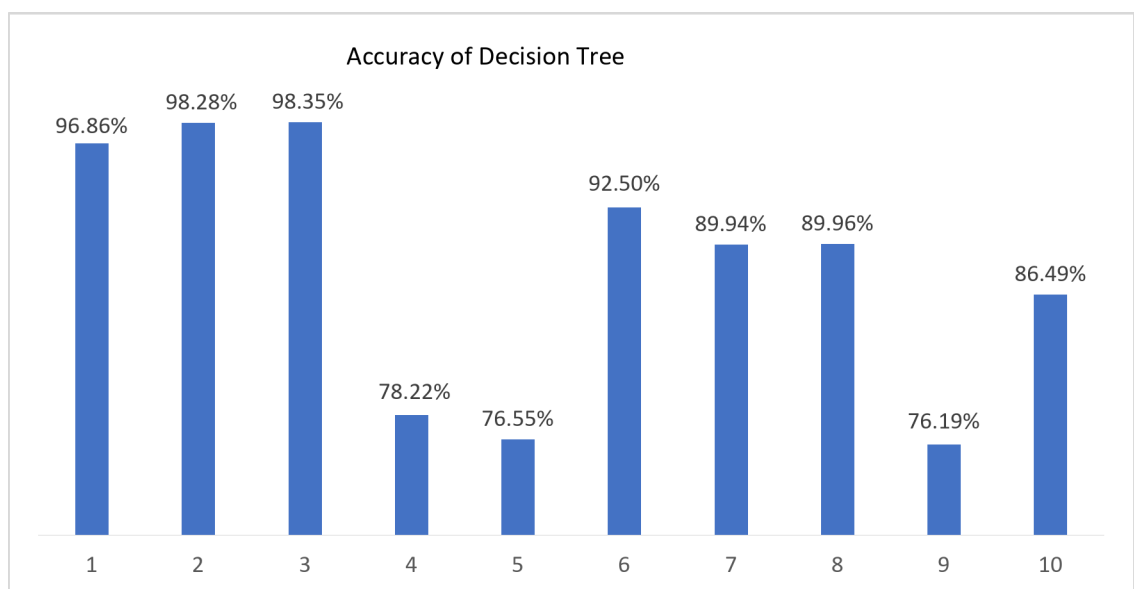


Fig. 5. The accuracy of Decision Tree with English language stopwords removal.

English stopwords only and removal of respective non-English language stopwords. Both sets of experimental results are presented in this section and will be further discussed in Section V.

Baseline analysis

SMS Spam Detection with English Language Stopwords Removal.

Experimental result from decision tree

The goal of this analysis was to evaluate how effectively the Decision Tree model could classify SMS messages as spam or non-spam when English stopwords were removed from the datasets. Stopwords are common words that can introduce noise in text-based models, and their removal is a standard preprocessing step to improve classification accuracy. Table 4 summarizes the experimental results, while Fig. 5 visually compares the accuracy achieved across the datasets.

The Decision Tree demonstrates high accuracy with Dataset 1 (96.86%), Dataset 2 (98.28%), and Dataset 3 (98.35%). The precision scores for 'ham' range from 0.97 to 0.99, while for 'spam', they range from 0.94 to 0.98. The recall scores for 'ham' are 0.99, 0.98, and 0.99 for Dataset 1, Dataset 2, and Dataset 3, respectively. In contrast, the recall scores for 'spam' are 0.83, 0.99, and 0.94 for the respective datasets. The F1-score for 'ham' consistently stands at 0.98 across all dataset 1 and dataset 2 but the value increase to 0.99 in dataset 3, whereas the F1-score for 'spam' varies between 0.89 and 0.98 across the datasets.

The Decision Tree exhibits moderate accuracy for Dataset 6, Dataset 7, Dataset 8, and Dataset 10. Among these, Dataset 6 achieves the highest accuracy at 92.50%, while Dataset 10 has the lowest at 86.49%. The recall scores for 'ham' range from 0.76 to 0.92, and for 'spam', from 0.91 to 0.95. Precision scores for 'ham' vary from

Multinomial Naïve Bayes							
Dataset	Accuracy	Precision (Ham)	Precision (Spam)	Recall (Ham)	Recall (Spam)	F1 - Score (Ham)	F1 - Score (Spam)
1	98.48%	0.99	0.95	0.99	0.95	0.99	0.95
2	99.03%	0.99	0.99	0.99	0.99	0.99	0.99
3	98.29%	0.99	0.92	0.99	0.96	0.99	0.94
4	89.11%	0.88	0.90	0.93	0.84	0.91	0.87
5	86.10%	0.93	0.77	0.85	0.89	0.88	0.83
6	96.00%	0.96	0.96	0.96	0.96	0.96	0.96
7	93.76%	0.94	0.93	0.93	0.94	0.94	0.94
8	95.20%	0.94	0.96	0.95	0.95	0.95	0.96
9	90.48%	0.80	1.00	1.00	0.85	0.89	0.92
10	83.78%	0.72	0.95	0.93	0.78	0.81	0.86

Table 5. Comparison results of various evaluation metrics from multinomial Naïve Bayes with english Language stopwords removal.

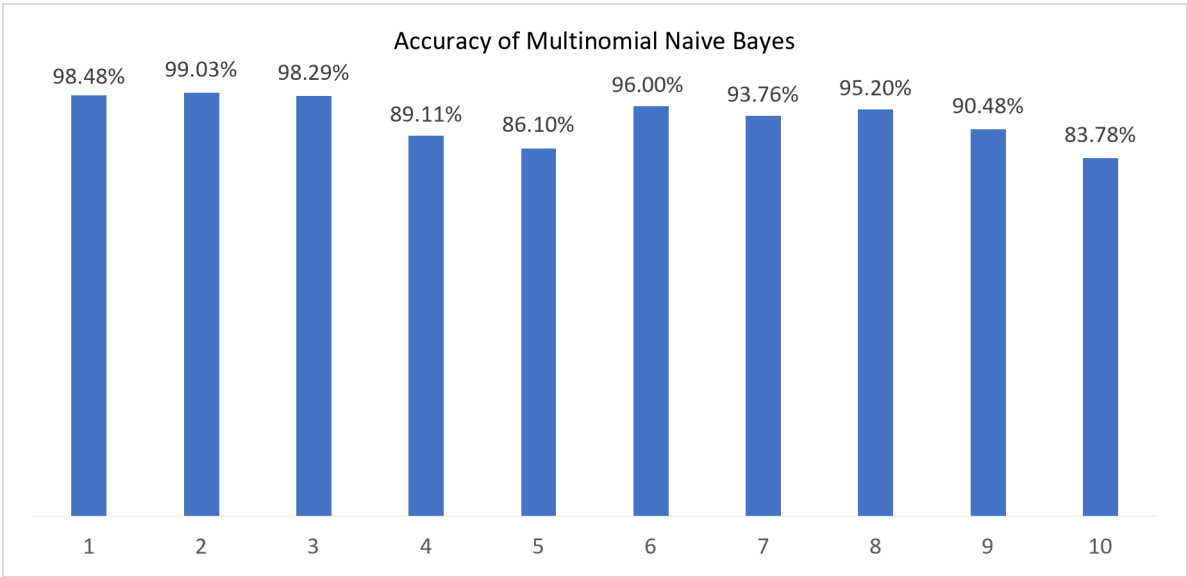


Fig. 6. The accuracy of Multinomial Naïve Bayes with English language stopwords removal.

0.90 to 0.93, and for ‘spam’, from 0.83 to 0.92. The F1-scores for ‘ham’ range between 0.84 and 0.93, and for ‘spam’, between 0.88 and 0.92.

The datasets with the lowest accuracy in this group are Dataset 4, Dataset 5, and Dataset 9. Dataset 4 has the highest accuracy among these at 78.22%, while Dataset 9 has the lowest accuracy at 76.19%. The recall scores for ‘ham’ range from 0.64 to 0.87, and for ‘spam’, from 0.65 to 0.90. Precision scores for ‘ham’ span from 0.73 to 0.88, and for ‘spam’, from 0.69 to 0.82. The F1-scores for ‘ham’ vary between 0.74 and 0.81, and for ‘spam’, between 0.72 and 0.78.

Results from multinomial naïve bayes

This analysis evaluates the performance of Multinomial Naïve Bayes (MNB) in classifying the SMS spam messages with only the English stopwords removed across all ten datasets, which are presented in Table 5. Figure 6 shows the accuracy of the model throughout each dataset, which shows a variation of performance across the datasets.

Datasets 1, 2, 3, and 6 had achieved an accuracy rate of 98.48%, 99.03%, 98.29%, and 96.00%, respectively. These datasets show strong performance across all metrics, with Dataset 2 notably achieving near-perfect scores in precision, recall, and F1-scores for both Ham and Spam.

Datasets 7, 8, and 9 shows accuracies of 93.76%, 95.20%, and 90.48%, respectively. These results indicate a balanced performance, maintaining high precision and recall scores for both ham and spam, with Dataset 8 showing particularly robust precision and F1-scores for Spam.

Datasets 4, 5, and 10 exhibit lower accuracy rates of 89.11%, 86.10%, and 83.78%, respectively. Despite the lower overall accuracy, some metrics such as precision for Ham in Dataset 5 and recall for Ham in Dataset 10 remain relatively high, suggesting that specific areas of performance are still strong.

Decision Tree							
Dataset	Accuracy	Precision (Ham)	Precision (Spam)	Recall (Ham)	Recall (Spam)	F1 - Score (Ham)	F1 - Score (Spam)
2	97.85%	0.99	0.97	0.97	0.99	0.98	0.98
4	81.19%	0.81	0.82	0.88	0.73	0.84	0.77
6	91.75%	0.92	0.92	0.92	0.92	0.92	0.92
7	91.33%	0.92	0.91	0.91	0.92	0.91	0.91
8	88.21%	0.84	0.92	0.91	0.86	0.87	0.89
9	76.19%	0.80	0.75	0.50	0.92	0.62	0.83
10	78.38%	0.69	0.86	0.79	0.78	0.73	0.82

Table 6. Comparison results of various evaluation metrics from decision tree with non-English Language stopwords removal.

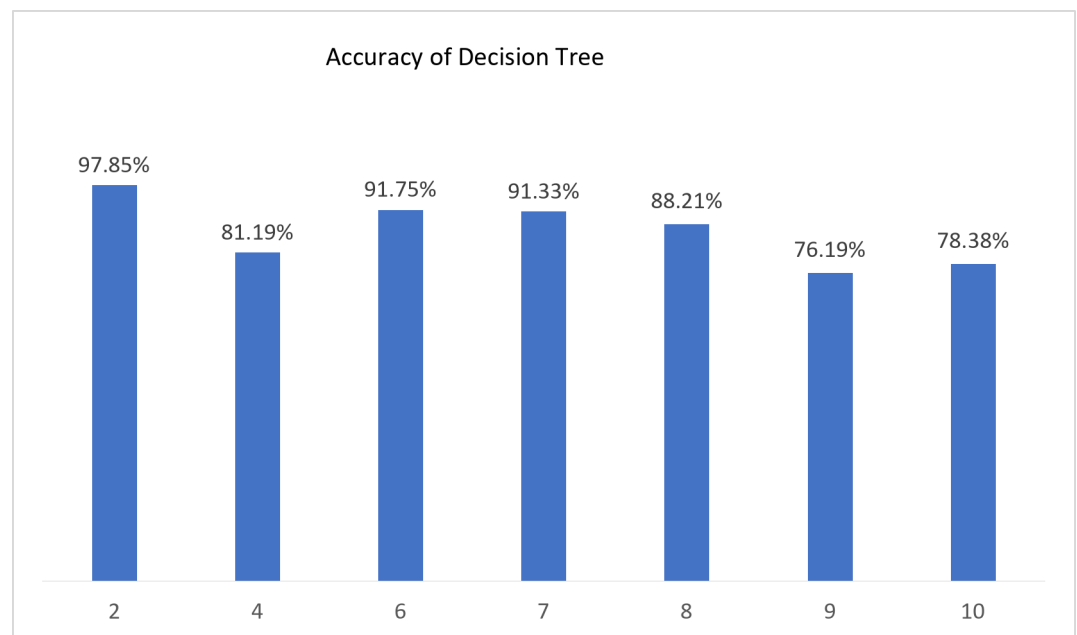


Fig. 7. The accuracy of Decision Tree with non-English language stopwords removal.

These results underscore the effectiveness of the Multinomial Naïve Bayes classifier in handling SMS spam detection without the removal of non-English language stopwords, particularly in datasets where high accuracy is achieved. The variations in performance across different datasets highlight the importance of dataset characteristics on classifier efficacy, which will be explored further in subsequent sections of the research.

Enhanced analysis

SMS Spam Detection with Non-English language Stopwords Removal.

Experimental result from decision tree

In this study, ten publicly available SMS spam datasets are evaluated on DT and MNB, eight of which are either monolingually non-English language (Dataset 2, Dataset 4, Dataset 7, and Dataset 8), transliterated non-English language (Dataset 9), monolingually non-English languages combined with English language (Dataset 3), or transliterated non-English language combined with English language (Dataset 6 and Dataset 10). The first group of the experiment only tests the performance of DT and MNB when English stopwords are removed from all datasets.

In the second group of the experiment, the SMS spam detection code is modified to remove respective non-English language stopwords for only one language. The analysis of this second group of experiment discard Dataset 1, and Dataset 5, since both datasets uses English language. Additionally, Dataset 3 is also discarded since it is a multilingual dataset, which requires multiple stopwords removal toolkits to remove German and French stopwords. The findings from Decision Tree in this group of experiment revealed differing levels of performance across the datasets, as shown in Table 6 whereas Fig. 7 shows the accuracy of this model in each dataset.

Multinomial Naïve Bayes							
Dataset	Accuracy	Precision (Ham)	Precision (Spam)	Recall (Ham)	Recall (Spam)	F1 - Score (Ham)	F1 - Score (Spam)
2	98.28%	0.99	0.98	0.97	0.99	0.98	0.98
4	90.10%	0.89	0.93	0.95	0.84	0.92	0.88
6	96.75%	0.96	0.97	0.97	0.96	0.97	0.97
7	92.95%	0.94	0.92	0.92	0.94	0.93	0.93
8	96.94%	0.99	0.95	0.94	0.99	0.97	0.97
9	90.48%	0.80	1.00	1.00	0.85	0.89	0.92
10	86.49%	0.80	0.91	0.86	0.87	0.83	0.89

Table 7. Comparison results of various evaluation metrics from multinomial Naïve Bayes with non-English Language stopwords removal.

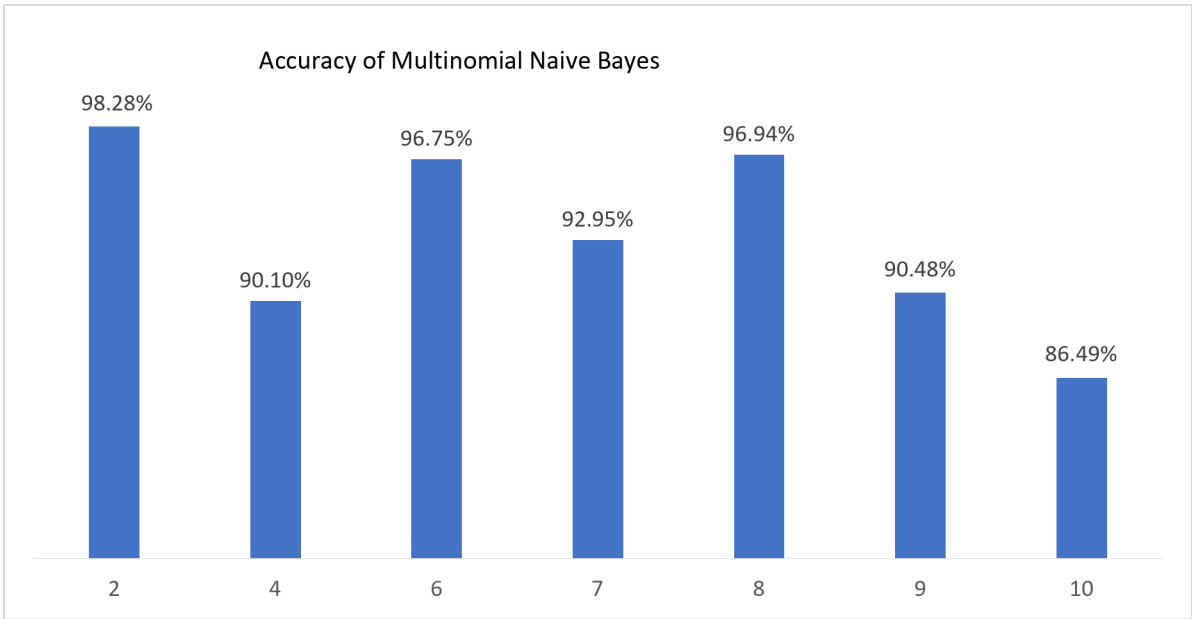


Fig. 8. The accuracy of Multinomial Naïve Bayes with non-English language stopwords removal.

Datasets 2, 6, and 7 shows promising result, achieving accuracy rates of 97.85%, 91.75%, and 91.33% respectively. Dataset 2 shows particularly strong performance, with high precision, recall, and F1 scores for both ham and spam, achieving a precision of 0.99 for ham and 0.97 for spam, and an F1 score of 0.98 for both ham and spam. Similarly, Datasets 6 and 7 exhibit balanced performance with precision, recall, and F1 scores of 0.92 for both ham and spam, indicating consistent detection capabilities.

Datasets 4 and 8 achieve moderate accuracy rates of 81.19% and 88.21% respectively. Dataset 4 demonstrates a precision of 0.81 for ham and 0.82 for spam, with recall values of 0.88 for ham and 0.73 for spam, resulting in F1 scores of 0.84 for ham and 0.77 for spam. Dataset 8 shows a precision of 0.84 for ham and 0.92 for spam, with recall values of 0.91 for ham and 0.86 for spam, leading to F1 scores of 0.87 for ham and 0.89 for spam.

Datasets 9 and 10 shows suboptimal result with accuracy rates of 76.19% and 78.38% respectively. Despite the lower overall accuracy, certain metrics remain noteworthy. For instance, Dataset 9 exhibits a high recall of 0.92 for spam, although its precision for ham is relatively lower at 0.80, resulting in an F1 score of 0.62 for ham and 0.83 for spam. Dataset 10 shows a precision of 0.69 for ham and 0.86 for spam, with recall values of 0.79 for ham and 0.78 for spam, leading to F1 scores of 0.73 for ham and 0.82 for spam.

Experimental result from multinomial naïve bayes

This analysis evaluates the performance of MNB in classifying SMS spam messages when the stopwords from each respective Non-English language is removed. The experimental result of MNB, which is shown in Table 7, showed that MNB indeed outperformed the model performance of DT across all Non-English datasets. Figure 8 shows a visual illustration of MNB performance for this group of experiment.

Datasets 2, 6, and 8 displays strong result, achieving accuracy rates of 98.28%, 96.75%, and 96.94% respectively. Dataset 2 shows exceptional performance, with high precision, recall, and F1 scores for both ham and spam, achieving a precision of 0.99 for ham and 0.98 for spam, and an F1 score of 0.98 for both ham and spam. Dataset 6 demonstrates balanced performance with precision, recall, and F1 scores of 0.96 for ham and 0.97 for spam.

Similarly, Dataset 8 achieves a precision of 0.99 for ham and 0.95 for spam, with recall values of 0.94 for ham and 0.99 for spam, resulting in F1 scores of 0.97 for both ham and spam.

In addition, Datasets 4, 7, and 9 shows satisfactory result with accuracy rates of 90.10%, 92.95%, and 90.48% respectively. Dataset 4 shows a precision of 0.89 for ham and 0.93 for spam, with recall values of 0.95 for ham and 0.84 for spam, leading to F1 scores of 0.92 for ham and 0.88 for spam. Dataset 7 demonstrates balanced performance with a precision of 0.94 for ham and 0.92 for spam, and recall values of 0.92 for ham and 0.94 for spam, resulting in F1 scores of 0.93 for both ham and spam. Dataset 9, while having a moderate accuracy rate, exhibits a precision of 0.80 for ham and 1.00 for spam, with recall values of 1.00 for ham and 0.85 for spam, leading to F1 scores of 0.89 for ham and 0.92 for spam.

It was found that Dataset 10 exhibits a weak performance with an accuracy rate of 86.49%. Despite the lower overall accuracy, certain metrics remain noteworthy. Dataset 10 shows a precision of 0.80 for ham and 0.91 for spam, with recall values of 0.86 for ham and 0.87 for spam, resulting in F1 scores of 0.83 for ham and 0.89 for spam.

Discussion

This study serves to evaluate ten publicly available SMS spam datasets by using Decision Tree (DT) and Multinomial Naïve Bayes (MNB). This section of the paper serves to further discuss the possible explanation for the obtained result in Section IV. The discussion is structured based on a series of questions that were derived during the analysis of the results.

Comparative result of multinomial naive bayes and decision tree Across different datasets

An overall analysis of the result from Decision Tree and Multinomial Naïve Bayes from both groups can be shown the superior performance of the Multinomial Naïve Bayes (MNB) model over Decision Trees (DT) in detecting SMS spam texts, which can be attributed to the intrinsic characteristics of these algorithms.

MNB is exceptionally well-suited for text data due to its foundational assumptions of word independence and frequency^{40,41}. This characteristic aligns seamlessly with the bag-of-words model commonly employed in text classification. The probabilistic framework of MNB, based on Bayes' Theorem, enables it to manage word distribution in text data with high efficacy⁴². MNB calculates the probability of a message being spam based on word frequencies, assuming that the presence or absence of a particular word in a message is independent of any other word. While this assumption simplifies the modeling process and allows for efficient classification, it also means that MNB may struggle with capturing deeper contextual, non-linear relationships or complex dependencies between words, which could limit its effectiveness in datasets where spam messages exhibit more sophisticated linguistic structures.

Conversely, Decision Trees (DT) are more apt for structured data characterized by explicit feature-value relationships. Unlike MNB, DT can model complex decision boundaries and capture nonlinear relationships in data, making them highly interpretable and adaptable to structured classification problems⁴³. However, in high-dimensional textual data, the model's reliance on discrete feature splits can lead to inefficiencies, particularly when words are sparsely distributed across messages. Effective application of DT to textual data necessitates intricate feature engineering, such as feature scaling, weighting, or dimensionality reduction, to mitigate the challenges posed by the high cardinality of unique words⁴³.

MNB can effectively manage high-dimensional sparse data⁴⁴, a common characteristic of text classification tasks. This is because MNB can deal with the sparsity of word occurrences across documents. The model's probabilistic approach is robust in learning from the frequencies of words⁴⁴. On the other hand, DT can struggle with sparse data because they have to split nodes based on the presence or absence of words. This often leads to overfitting and reduced generalization^{45,46}. The tree structure in DT becomes too complex and specific to the training data when dealing with such high-dimensional sparse datasets⁴⁷.

Regarding the bias-variance trade-off, MNB has high bias but low variance⁴⁸, meaning that while it makes strong simplifying assumptions, it tends to generalize well across different datasets without drastic performance fluctuations. This robustness is particularly beneficial when working with datasets that contain noise or inconsistencies, such as Dataset 5, where extensive preprocessing is required. However, MNB's inability to capture complex interactions between words may lead to performance limitations, particularly in cases where spam messages contain subtle linguistic variations or semantic patterns that require contextual understanding.

In contrast, DT tends to have lower bias but higher variance, making it more flexible in capturing complex decision boundaries but also more prone to overfitting, especially when applied to high-dimensional datasets⁴⁹. Overfitting occurs when DT learns from noise or idiosyncratic patterns in the training data rather than capturing generalizable trends, leading to weaker performance on unseen data. This issue is particularly evident in datasets with high complexity, such as Dataset 5, where variations in spam message structures may cause DT to create overly specific decision rules that do not generalize well across different spam categories.

The superior performance of Multinomial Naïve Bayes (MNB) over Decision Trees (DT) observed in this study aligns with findings from prior research. A comparative analysis of the model's highest accuracy, as presented in Table 8, further corroborates these results in relation to previous studies.

Factors influencing performance variability in experimental runs

Upon a closer inspection of the model performance of DT and MNB, it can be concluded that both models exhibit different performance across all datasets used in this study. Performance variability refers to the differences in the experimental results obtained from each dataset, which can be influenced by factors such as dataset characteristics, class distribution, and the presence of noisy data.

The inherent complexity and structure of the data can significantly influence model performance. DT may encounter difficulties with datasets that exhibit complex and nuanced patterns, which probabilistic models like

Research	Accuracy of MNB	Accuracy of DT
Current research	99.03%	98.35%
⁵⁰	97.87%	93.52%
⁵¹	95.93%	93.52%

Table 8. The superior accuracy of MNB over DT, as supported by prior studies.

MNB are better equipped to handle. This is because DT tend to overfit on the intricate details of the training data, leading to varied performance when tested on different datasets⁵². In addition, language variations across datasets can influence model effectiveness. Each dataset representing different languages or transliterations can affect MNB’s performance based on linguistic features unique to each language⁵². Word frequencies, sentence structures, and common spam words vary across languages, contributing to performance variability⁵².

Class imbalance is another critical factor affecting performance variability. Variations in the distribution of classes within the training and testing sets can cause significant fluctuations⁵³. The size, quality, and spam-to-spam ratio of each dataset profoundly impact MNB’s performance. Larger datasets generally provide a more robust learning base, while imbalanced datasets can skew results⁵⁴. If a particular class is underrepresented in a split, the model may struggle to recognize it, thereby affecting overall performance.

The presence of random noise and outliers in the data can further contribute to performance variability, leading to inconsistent outcomes^{52,55}. When noise affects different parts of the data during each run, the model’s training process can be disrupted in varying ways⁵⁵. For example, datasets with high noise levels or irrelevant features may cause DT to grow overly complex, leading to poor generalization⁵⁶.

Dataset characteristics significantly influenced the performance of both models. Dataset 2, which achieved the highest accuracy for MNB (99.03%), exhibited a nearly balanced class distribution and contained minimal noise, allowing MNB’s probabilistic approach to effectively distinguish between spam and non-spam messages. Similarly, Dataset 3, which recorded the highest accuracy for DT (98.35%), contained multilingual text suggesting that DT was better able to leverage structured word patterns across different languages. However, MNB’s word independence assumption likely contributed to its slightly lower performance in this dataset.

In contrast, Dataset 5, which recorded one of the lowest accuracies for both models (MNB: 86.10%, DT: 76.55%), demonstrated high feature diversity and real-world complexity. The dataset’s high qualitative assessment score indicates a rich set of spam characteristics, yet the presence of noise, feature redundancy, and potential label inconsistencies may have contributed to lower classification performance. DT’s tendency to overfit was particularly evident in Dataset 5 and Dataset 9, both of which have high Gini coefficients (0.4999), indicating a more balanced distribution of spam and ham messages. However, the high feature complexity and presence of transliterated text in Dataset 9 negatively impacted both models, particularly DT, which struggled to create meaningful feature splits.

Assessment of datasets

A high-quality dataset can be evaluated quantitatively or qualitatively. Quantitative evaluation of the datasets used in this research has been thoroughly discussed in previous sections. These datasets were then subjected to a qualitative assessment to determine its reliability, effectiveness, and reusability, which encompasses several critical assessments, such as the authenticity of the source, class imbalance, the diversity of features, the availability of metadata, and the data preprocessing requirement. Each of the datasets are evaluated based on the five-level evaluation points for each assessment, which coincide with the five Likert scale points defined in each of the respective table.

A dataset retrieved or downloaded from an authentic source offers various benefits, including high accuracy and reliability, which reduces the likelihood of errors and inconsistencies during data analysis²⁰. Additionally, authentic datasets are often accompanied by clear documentation to facilitate reusability, allowing other researchers to replicate the study and verify the results⁵⁷. Furthermore, data from reliable sources typically exhibit consistency in format, structure, and quality, simplifying data preprocessing and analysis, which reduces the need for extensive cleaning and transformation²⁰. To assess the authenticity of the data source, this research followed the work of⁵⁸ and constructed a five-level evaluation system as follows, which directly influence the qualitative assessment of the dataset’s source authenticity.

- 1. Is the dataset published in a peer-reviewed journal, a conference paper, or been used in a competition?
- 2. Is there clear documentation about how the data was collected?
- 3. Is there a clear history of the data, including any transformations or processing steps it has undergone?
- 4. Is the publication or release date of the dataset clearly stated?
- 5. Is the dataset internally consistent with no unexplained variations?

Table 9 shows the score assigned to each dataset for their authenticity of source. The scores are assigned based on the given evaluative points above. For instance, Dataset 1 receives the highest Likert value of 5 because the dataset was published in a peer-reviewed journal, has clear documentation about how the data was collected, clear history of the data, has clearly stated the release date of the dataset, and is internally consistent with no unexplained variations.

The class distribution of a dataset is important as it can directly and indirectly influence the obtained results. A balance class distribution improved the model’s performance and help the model to generalize better to new, unseen data, as they are less likely to be biased towards the more frequent classes⁵⁹. On the other hand,

		Very poor	Poor	Fair	Good	Excellent
Source Authenticity		1	2	3	4	5
	Dataset 1					√
	Dataset 2			√		
	Dataset 3	√				
	Dataset 4	√				
	Dataset 5			√		
	Dataset 6				√	
	Dataset 7				√	
	Dataset 8		√			
	Dataset 9			√		
	Dataset 10	√				

Table 9. Scores assigned to each dataset for source authenticity.

		Very poor	Poor	Fair	Good	Excellent
Class Imbalance		1	2	3	4	5
	Dataset 1	√				
	Dataset 2				√	
	Dataset 3	√				
	Dataset 4				√	
	Dataset 5				√	
	Dataset 6					√
	Dataset 7					√
	Dataset 8					√
	Dataset 9		√			
	Dataset 10			√		

Table 10. Scores assigned to each dataset for class imbalance.

imbalance class distribution will result in poor model performance, especially because it causes low recall for minority classes, and skewed accuracy⁶⁰. To assess the balance of class distribution within the dataset, this research proposed a five-level evaluation system to quantify the severity of class imbalance, inspired by the imbalance ratio (IR) discussed in prior work^{61,62}. This framework allows for a more nuanced understanding of imbalance severity and informs the selection of appropriate techniques, such as threshold adjustment⁶³ to improve model performance. The proportions of spam and non-spam instances are computed and converted into percentages. These percentages are then evaluated using a rating scale ranging from 1 to 5, as delineated by the subsequent criteria:

1. (Very Poor): The dataset is highly imbalanced, with one class comprising 90% of the data.
2. (Poor): The dataset is significantly imbalanced, with one class comprising 80% of the data.
3. (Fair): The dataset is moderately imbalanced, with one class comprising 70% of the data.
4. (Good): The dataset is slightly imbalanced, with one class comprising 60% of the data.
5. (Excellent): The dataset is perfectly or nearly perfectly balanced, with classes having equal or nearly equal representation.

Table 10 shows the score assigned to each dataset for their class imbalance. The scores are assigned based on the given based on the five-level classification system above. For instance, Dataset 6 receives the highest Likert value of 5 because the dataset has a perfectly balanced distribution of spam and non-spam.

The incorporation of diverse features in SMS spam detection research serves to enhance detection accuracy and the model's robustness. Varied feature types capture distinct facets of spam messages, enabling the detection algorithm to make more nuanced decisions. This aligns with the findings of⁶⁴, who noted that diversity in datasets helps models generalize their learnings to new and unseen cases. Similarly, incorporating diverse feature types ensures that the model can handle a wider range of spam characteristics, thus improving its robustness and adaptability. To measure feature diversity within datasets, this research proposes a five-level evaluation system that quantifies diversity purely based on the number of features present rather than the specific type of features. The rationale for this classification is as follows:

- Datasets with fewer features (Scores 1–2) contain only basic text and labels, limiting their ability to provide meaningful distinctions between spam and ham messages.

- Datasets with moderate features (Score 3) begin to introduce additional attributes, offering minor improvements to model learning.
- Datasets with high feature diversity (Scores 4–5) provide richer insights by incorporating multiple attributes, which significantly enhance model performance and generalizability.

The five-level evaluation system is as follows:

1. (Very poor): The dataset only contains two features; the raw text messages and its label (e.g., spam/ham).
2. (Poor): The dataset contains three features; raw text messages, labels, and an additional attribute.
3. (Fair): The dataset contains four features; raw text messages, labels, and two additional attributes.
4. (Good): The dataset contains five features; raw text messages, labels, and three additional attributes.
5. (Excellent): The dataset contains six or more features, providing a diverse range of attributes that significantly improve spam detection.

Table 11 shows the score assigned to each dataset for their diversity of features. The scores are assigned based on the given based on the five-level classification system above. For instance, Dataset 5 receives the highest Likert value of 5 because the dataset has five additional attributes in addition to raw text messages and labels.

Metadata encompasses supplementary details associated with a text message beyond its actual content. These additional pieces of information furnish crucial contextual insights about the dataset, encompassing its origin, purpose, and structural characteristics, which plays a pivotal role in accurately interpreting the outcomes derived from the dataset²². Furthermore, robust metadata practices foster data sharing and collaborative endeavours among researchers by simplifying the comprehension and utilization of shared datasets⁵⁸. The clarity inherent in meticulously documented metadata enhances communication and collaboration across disciplinary and institutional boundaries. To evaluate metadata availability, this research proposes a five-level evaluation system that assesses metadata richness based on the number of metadata fields present and their degree of exposure. The rationale for this classification is as follows:

- Datasets with minimal metadata (Scores 1–2) provide little to no contextual information, reducing their applicability for advanced analysis.
- Datasets with moderate metadata (Score 3) include some metadata fields but may have missing values or limited exposure.
- Datasets with high metadata availability (Scores 4–5) provide structured, comprehensive metadata, improving interpretability and dataset usability.

The five-level evaluation system is as follows:

1. (Very poor): The dataset contains only text messages and labels (spam/ham) with no additional metadata.
2. (Poor): The dataset includes 1–2 metadata fields, but exposure is limited or inconsistent.
3. (Fair): The dataset contains 3–4 metadata fields, offering some context but lacking full exposure.
4. (Good): The dataset contains 5–6 metadata fields, with structured exposure of metadata across most records.
5. (Excellent): The dataset contains 7 or more metadata fields, providing fully detailed and consistently structured metadata across all records.

Table 12 shows the score assigned to each dataset for their metadata availability. The scores are assigned based on the given based on the five-level classification system above. For instance, Dataset 5 receives the highest Likert value of 5 because the dataset has a fully detailed and consistently structured metadata across all records.

The assessment of data preprocessing pertains to the extent of preparatory measures needed to render the data compatible for model ingestion. This preparatory phase encompasses both data cleansing and integration procedures. The degree of preprocessing varies across datasets, with certain datasets necessitating more extensive preprocessing efforts than others. Consequently, datasets requiring extensive preprocessing impose a higher computational burden, impacting the feasibility of research workflows.

		Very poor	Poor	Fair	Good	Excellent
		1	2	3	4	5
Diversity of features	Dataset 1	√				
	Dataset 2	√				
	Dataset 3			√		
	Dataset 4	√				
	Dataset 5					√
	Dataset 6	√				
	Dataset 7	√				
	Dataset 8	√				
	Dataset 9	√				
	Dataset 10	√				

Table 11. Scores assigned to each dataset for features diversity.

		Very poor	Poor	Fair	Good	Excellent
		1	2	3	4	5
Metadata availability	Dataset 1		√			
	Dataset 2		√			
	Dataset 3		√			
	Dataset 4	√				
	Dataset 5					√
	Dataset 6		√			
	Dataset 7		√			
	Dataset 8		√			
	Dataset 9	√				
	Dataset 10		√			

Table 12. Scores assigned to each dataset for metadata availability.

		Minimal	Low	Moderate	High	Extensive
		1	2	3	4	5
Data Preprocessing Requirement	Dataset 1					√
	Dataset 2		√			
	Dataset 3			√		
	Dataset 4				√	
	Dataset 5		√			
	Dataset 6					√
	Dataset 7		√			
	Dataset 8					√
	Dataset 9					√
	Dataset 10					√

Table 13. Scores assigned to each dataset for data preprocessing requirements.

To evaluate the effort required for data preprocessing, this research proposes a five-level evaluation system based on the complexity of preprocessing tasks required. The rationale for this classification is as follows:

- Datasets requiring minimal preprocessing (Scores 1–2) are well-structured and nearly ready for use, with only minor cleaning needed.
- Datasets requiring moderate preprocessing (Score 3) contain minor inconsistencies that necessitate text normalization and label standardization.
- Datasets requiring extensive preprocessing (Scores 4–5) are highly unstructured, with significant noise, missing values, and imbalanced data, requiring multiple preprocessing steps.

The five-level evaluation system is as follows:

1. (Minimal): The dataset requires little or no preprocessing.
2. (Low): The dataset requires minor formatting adjustments.
3. (Moderate): The dataset requires the application of standard preprocessing steps.
4. (High): The dataset requires significant preprocessing steps.
5. (Extensive): The dataset requires multiple significant preprocessing steps.

Table 13 shows the score assigned to each dataset for their requirement of data preprocessing. The scores are assigned based on the given based on the five-level classification system above. For instance, Dataset 2 receives the low Likert value of 2 because the dataset requires minor formatting adjustments.

Factors contributing to accuracy variations in dataset 4 and dataset 7 for both group of experiment

This research employs ten publicly available SMS spam detection dataset. Among the ten datasets used in this study, Dataset 2, Dataset 4, Dataset 7 and Dataset 8 are presented in their natural linguistic form, rather than being transliterated as compared to other non-English language datasets. Yet, an interesting observation arise when we compare the trend of model performance of DT and MNB in the first and second group of experiment for Dataset 7. While the model performance for DT and MNB increase in the second group of experiment as compared to the first group of experiment for Dataset 2, Dataset 4 and Dataset 8, only DT showed an increase model performance, while MNB showed a decline in model performance. To ensure a smooth and logical

discussion, Dataset 4 is randomly chosen among the other monolingually non-English language dataset to be compared against Dataset 7.

Table 14 shows accuracy achieved by DT and MNB for Dataset 4 and Dataset 7 in both group of the experiment. From Table 14, it can be seen that while the accuracy for DT and MNB increases for Dataset 4 when comparing their performance of the first group of experiment with the second group of experiment, only the accuracy for DT increases for Dataset 7 whereas the accuracy for MNB decreases when comparing the accuracy of the model of the first group of experiment with the second group of experiment. The differences in performance between Dataset 4 and Dataset 7 in both group of experiment can be attributed to several factors: The relevance of stopwords and sensitivity of the models on both datasets.

The improvement observed in Dataset 4 when Bengali stopwords are removed can be attributed to the specific linguistic features of Bengali. The removal of Bengali stopwords likely reduced noise and irrelevant features, improving the quality of the features available for both models⁶⁵. For other non-English datasets, transliteration issues and language-specific nuances could impede the effectiveness of stopwords removal. Inconsistent transliteration and variations in spelling can leave noise in the data, limiting the improvement in model performance⁶⁶. Additionally, the quality of the stopwords list plays a crucial role in model performance. According to⁶⁴, an incomplete or inaccurate stopwords list can limit the expected improvement in performance. In this study, the experimental results for both MNB and DT showed improvements after removing Bengali stopwords for the Dataset 4, suggesting that the stopwords list used was comprehensive. The distinct separation between spam and non-spam messages in Dataset 4, after stopwords removal, highlights the efficacy of this preprocessing step.

Conversely, the original language in Dataset 7 might rely heavily on stopwords to convey essential context. Removing these stopwords can disrupt the contextual integrity required for MNB to perform effectively, as this model relies on the word frequency distribution to make accurate predictions. In this case, stopwords carry significant meaning within the language structure, and their removal can negatively impact MNB's performance. However, DT benefits from the removal of stopwords in Dataset 7, as this model can better handle a reduced feature set by focusing on the remaining words, suggesting that the stopwords in Dataset 7 were adding unnecessary complexity and noise in DT which hinders its decision-making process.

The sensitivity of MNB and DT to stopwords removal also plays a crucial role in the observed accuracy variations. For MNB, the performance in Dataset 4 improves with the removal of stopwords, as it reduces noise and enhances the signal, allowing the model to focus on more informative words. In contrast, for Dataset 7, the removal of stopwords disrupts the probability calculations that MNB relies on, thereby reducing its accuracy. In Dataset 4, DT similarly benefits from reduced complexity and less noise, leading to clearer decision boundaries and improved accuracy. On the other hand, DT shows an increase in accuracy when stopwords are removed from Dataset 7, indicating that stopwords in this dataset were acting as noise, and their removal helped the DT model create more accurate splits.

It is important to note that while MNB showed an accuracy trend that does not align with most of the other observation with other datasets when comparing the model's accuracy of the first group of experiment with the second group of experiment in Dataset 7 (refer to Table 14), the class imbalance between Dataset 4 and Dataset 7 could explain the differences of the accuracy observed. As shown in Table 9, Dataset 4, rated 4 on the Likert scale, is more imbalanced than Dataset 7, rated 5, resulting in misleadingly high accuracy for Dataset 4 due to data imbalance. This underscores the issue of skewed accuracy and highlights the critical role of stopwords in addressing such challenges.

Factors Contributing to Enhanced Accuracy in Dataset 3 Relative to Other Datasets

Dataset 3, which consists of SMS messages in English, German, and French, exhibited a unique performance trend in which the Decision Tree (DT) model outperformed Multinomial Naïve Bayes (MNB). This performance difference can be attributed to several factors, including the impact of the language feature independence assumption, handling of class imbalance, the incomplete stopwords removal process in the first group of experiments, and overfitting and variance.

One key factor influencing the model's performance is the language feature independence assumption inherent to MNB. This model assumes that word occurrences are independent, meaning that each word's probability is calculated separately from others. While this assumption often works well for monolingual datasets, it becomes problematic in multilingual datasets like Dataset 3, where word meanings and distributions vary across different languages. For example, common spam-related words in English are expressed differently in French and German. Since MNB aggregates word frequencies across all three languages without distinguishing between them, it fails to recognize spam indicators effectively across multiple linguistic structures. Conversely, DT does not rely on the independence assumption and instead recursively splits the dataset based on the most informative features^{64,65}. This flexibility allows DT to adapt to language-specific spam indicators, making it more effective in handling multilingual datasets like Dataset 3.

Dataset	Decision Tree		Multinomial Naïve Bayes	
	First Group of experiment	Second Group of experiment	First Group of experiment	Second Group of experiment
4	78.22%	81.19% ↑	89.11%	90.10% ↑
7	89.94%	91.33% ↑	93.76%	92.95% ↓

Table 14. The accuracy of DT and MNB for dataset 4 and dataset 7 for both group of the experiment.

Another crucial factor affecting the models' performance is class imbalance. Dataset 3 contains 2,241 spam messages compared to 14,460 non-spam messages, creating a significant imbalance that influences model learning, as indicated by its rating on the Likert scale in Table 9. As a probability-based classifier, MNB struggles with imbalanced data because its probability estimates are naturally skewed in favor of the majority class (non-spam messages). In other words, MNB may not handle class imbalance as effectively unless specific techniques like class weighting or resampling are employed⁴⁴. As a result, spam messages are often misclassified due to lower word frequencies. In contrast, DT is more resilient to class imbalance as it learns decision rules based on how well each feature (word) separates spam from non-spam. Instead of relying solely on word occurrence probabilities, DT dynamically adjusts its decision boundaries, allowing it to classify spam messages more effectively even when they are the minority^{64,65}.

The stopword removal strategy in the first group of the experiment also played a significant role in model performance differences. In the first group of the experiment, only English stopwords were removed while German and French stopwords remained in the dataset. This had a disproportionate effect on MNB, as it relies heavily on word frequency distributions. The presence of frequent yet uninformative German and French stopwords had introduced noise into MNB's probability calculations, reducing its ability to differentiate between spam and non-spam. Since MNB assigns equal importance to all words, these stopwords diluted the significance of actual spam-related terms, leading to lower classification accuracy. In contrast, DT naturally selects the most important words for classification through its recursive splitting process, meaning it was less affected by the presence of unremoved stopwords. This allowed DT to remain more robust despite the incomplete stopword removal process.

Additionally, the consideration of overfitting and variance may provide further explanation. DT, while prone to overfitting, can perform exceptionally well, capturing patterns effectively without overfitting if the dataset is not too noisy⁶⁷. In contrast, MNB, generally less prone to overfitting due to its simplicity, might ignore some intricate patterns that a DT could capture⁴⁴.

Dataset recommendation

In the present research, ten SMS spam detection datasets were analyzed. Each dataset is characterized by its distinct attributes, which exert influence on the performance of the employed models: Decision Tree and Multinomial Naïve Bayes. The primary objective of this investigation is formulating dataset recommendations predicated upon model performance. Specifically, these recommendations are anchored in the accuracy metrics generated by the models. Since MNB consistently outperformed DT in the experiments, dataset evaluations are based on MNB's results to provide more reliable and accurate insights for future research.

The dataset recommendation will be made based on the quantitative result (model accuracy) and qualitative assessment. To facilitate the quantitative recommendation process, a set of grading criteria is introduced, contingent upon the accuracy levels attained by MNB. These criteria are stratified into three distinct categories: high accuracy ($\geq 95\%$), moderate accuracy (90 – 94.99%), and low accuracy ($< 90\%$). Given the absence of established industry benchmarks and previous studies providing thresholds for SMS spam detection, the categorization in this study serves as a means to interpret model performance across datasets. This exploratory approach may offer guidance for future research in developing more definitive benchmarks for SMS spam detection. Furthermore, given the delineation of the research into two experimental groups, the recommendations will encompass MNB performance from both groups. The guideline of categorization criteria for each dataset, classified as most challenging, moderately challenging, or least challenging, are presented in Table 15. It is important to note that these criteria do not apply to Dataset 1, Dataset 3, and Dataset 5, as they were not included in the second experiment group. Their categorization is therefore determined based solely on MNB's accuracy in the first group.

In this research, a challenging dataset is one in which the models exhibit lower accuracy, not because of flaws, but due to its diverse spam patterns, real-world complexity, and feature richness. While noise and ambiguity may contribute to difficulty, such datasets encourage the development of more adaptable and generalizable models. Additionally, the recommended dataset is the one with the highest overall qualitative assessment score, ensuring it is well-documented, diverse, and beneficial for advancing spam detection research.

Recommending the most challenging dataset is beneficial as it highlights dataset complexity, thereby driving the development and refinement of more robust and sophisticated models. Furthermore, challenging datasets promote the advancement of algorithms capable of greater adaptability and resilience to diverse forms of noise and ambiguity. Additionally, recommending the dataset with the highest average score across qualitative factors improves research quality, model performance, and usability, while mitigating risks related to bias, data inconsistencies, and unnecessary complexity.

First group	Second group	Categorization
High accuracy	High accuracy	Least challenging
High accuracy	Moderate accuracy	Least challenging
High accuracy	Low accuracy	Moderately challenging
Moderate accuracy	Moderate accuracy	Moderately challenging
Moderate accuracy	Low accuracy	Most challenging
Low accuracy	Low accuracy	Most challenging

Table 15. The guideline of categorization criteria for each dataset.

Dataset	MNB Accuracy	Category of accuracy
Dataset 1 (English)	98.48%	High accuracy
Dataset 2 (Turkish)	99.03%	High accuracy
Dataset 3 (English, French, and German)	98.29%	High accuracy
Dataset 4 (Bengali)	89.11%	Low accuracy
Dataset 5 (English)	86.10%	Low accuracy
Dataset 6 (Hindi)	96.00%	High accuracy
Dataset 7 (Persian)	93.76%	Moderate accuracy
Dataset 8 (Indonesian)	95.20%	High accuracy
Dataset 9 (Hindi)	90.48%	Moderate accuracy
Dataset 10 (Hindi)	83.78%	Low accuracy

Table 16. The category of challenges of each dataset with the removal of english Language stopwords.

Dataset	MNB Accuracy	Category of accuracy
Dataset 2 (Turkish)	98.28%	High accuracy
Dataset 4 (Bengali)	90.10%	Moderate accuracy
Dataset 6 (Hindi)	96.75%	High accuracy
Dataset 7 (Persian)	92.95%	Moderate accuracy
Dataset 8 (Indonesian)	96.94%	High accuracy
Dataset 9 (Hindi)	90.48%	Moderate accuracy
Dataset 10 (Hindi)	86.49%	Low accuracy

Table 17. The category of challenges of each dataset with the removal of the respective non-English Language stopwords.

Dataset	Category of challenges
Dataset 1 (English)	Least challenging
Dataset 2 (Turkish)	Least challenging
Dataset 3 (English, French, and German)	Least challenging
Dataset 4 (Bengali)	Most challenging
Dataset 5 (English)	Most challenging
Dataset 6 (Hindi)	Least challenging
Dataset 7 (Persian)	Moderately challenging
Dataset 8 (Indonesian)	Least challenging
Dataset 9 (Hindi)	Moderately challenging
Dataset 10 (Hindi)	Most challenging

Table 18. Overall category of challenges for each dataset.

Based on the aforementioned distinct categories, the category of accuracy for each dataset with the removal of English language stopwords and with the removal of respective non-English language stopwords is summarized in Tables 16 and 17, respectively. Table 18 shows the overall category of challenges for each dataset based on the delineated dataset criteria.

Based on Table 18, Datasets 1, 2, 3, 6, and 8 are identified as the least challenging for MNB, making them high-quality datasets suitable for baseline comparison studies or testing new models due to their consistent performance. In contrast, Datasets 7 and 9 present moderate challenges, making them useful for assessing model robustness and refining algorithms or feature engineering techniques. Datasets 4, 5, and 10 are the most challenging, offering valuable testbeds for developing and evaluating novel methodologies to enhance model performance under more complex conditions. These recommendations are grounded in the observed model performance.

By averaging the Likert score assigned to each dataset from Tables 9, 10, 11, 12 and 13, an additional recommendation emerges based on qualitative assessments; dataset authenticity, class imbalance, feature diversity, metadata availability, and preprocessing quality. As indicated in Table 19, Dataset 5 has the highest average score, making it the most recommended dataset. Given that Dataset 5 is also one of the most challenging for MNB, Dataset 5 is strongly recommended for future SMS spam detection research. This is followed by Dataset 6 and Dataset 8, particularly for evaluating model performance in complex scenarios.

Dataset	Average value
Dataset 1	2.8
Dataset 2	2.4
Dataset 3	2.4
Dataset 4	2.2
Dataset 5	3.8
Dataset 6	3.4
Dataset 7	2.8
Dataset 8	3.0
Dataset 9	2.4
Dataset 10	2.4

Table 19. Average value of each dataset based on the likert values from Tables 9 to Table 13.

Since the recommended dataset, Dataset 5, has one of the highest average Likert values from Table 18, it can be an ideal testbed for driving algorithm development and enhancing the adaptability and robustness of SMS spam detection models. For example, since Dataset 5 did not score the highest value for the qualitative assessment of class imbalanced, it necessitates the integration of advanced resampling techniques, such as SMOTE or undersampling during the development of algorithm to ensure fair model training and evaluation. Future development of algorithm tuned to Dataset 5 will have account for the introduction of inconsistencies due to data integration from multiple sources which require models to be resilient to noisy, incomplete, and heterogeneous data.

Leveraging Dataset 5 for model training encourages the development of more adaptable algorithms capable of handling diverse spam message structures. Feature engineering techniques, such as extracting semantic patterns, contextual embeddings, and n-grams, can be integrated during algorithm development to further enhance model effectiveness in identifying spam characteristics that may not be explicitly labeled. Additionally, the dataset's challenging characteristics encourages the experimentation with hybrid and ensemble learning approaches that improve model generalization in order to ensure higher performance across different SMS datasets. Moreover, transfer learning can be explored by fine-tuning models trained on Dataset 5 and applying them to different datasets, which reinforces the model's ability to generalize across various SMS spam detection tasks.

Conclusion

This research focuses into the efficacy of two machine learning models, namely Decision Tree and Multinomial Naïve Bayes, in detecting SMS spam messages across ten openly accessible datasets. Among these datasets, two are in English, while the remainder are in various non-English languages. Consequently, the investigation is divided into two experimental group: The first group involves solely the removal of English stopwords, while the second group entails the removal of respective non-English language stopwords from the non-English language datasets. Across both experimental groups, Multinomial Naïve Bayes consistently surpasses Decision Tree across most of the datasets. Additionally, in the second experimental group, Multinomial Naïve Bayes exhibits a decline in accuracy for Dataset 7, partly attributable to the fact that Datasets 4 and 7 are presented in their native languages, whereas the other non-English language datasets have undergone transliteration. Each dataset underwent a qualitative evaluation utilizing a Likert scale ranging from 1 to 5, based on diverse criteria including source authenticity, class imbalance, feature diversity, metadata accessibility, and the presence of data quality issues. Leveraging the accuracy outcomes of Multinomial Naïve Bayes, dataset criteria categorization was devised to recommend datasets for future SMS spam detection research. Datasets 1, 2, 3, 6, and 8 are classified as least challenging, while Datasets 7 and 9 exhibit moderate challenges, and Datasets 4, 5, and 10 are categorized as the most challenging datasets. These latter datasets are recommended for future research in SMS spam detection, as they possess the potential to foster innovation and boost the development of model's ability at handling intricate datasets characterized by noise and ambiguity. By averaging the Likert value of each dataset, it was found that Dataset 5 had garnered the highest average qualitative score across all assessed criteria, rendering it the most recommended dataset for deployment in future SMS spam detection research. The findings of this study can guide researchers in selecting appropriate datasets by emphasizing the importance of datasets that introduce sufficient complexity to enhance model robustness. Rather than choosing datasets that allow models to achieve artificially high accuracy, researchers should prioritize those that include variations in message structure, noise, and linguistic diversity. Additionally, datasets with rich metadata improve interpretability and reusability, making them more valuable for real-world applications. However, this study has limitations. First, it focuses on a limited set of machine learning models (Decision Tree and Multinomial Naïve Bayes), which may not fully capture the potential of deep learning approaches for spam detection. Second, while the evaluation considered qualitative aspects of datasets, it did not explore advanced data augmentation techniques that could mitigate class imbalance. Future research should investigate multilingual spam datasets and apply synthetic data generation or augmentation techniques to address dataset biases. Additionally, evaluating deep learning-based methods, such as transformer-based architectures, on these datasets could provide deeper insights into their effectiveness.

Data availability

The SMS datasets are publicly accessible at the following URL:

<https://archive.ics.uci.edu/dataset/228/sms+spam+collection>
<https://github.com/onrkrsy/TurkishSMS-Collection>
<https://www.kaggle.com/datasets/debapampal2002/sms-dataset1>
<https://github.com/AbayomiAlli/SMS-Spam-Dataset>
<https://github.com/princebari/-SMS-Spam-Classification-on-Indian-Dataset-A-Crowdsourced-Collection-of-Hindi-and-English-Messages/blob/main/README.md>
<https://zenodo.org/records/7832188>
<https://www.kaggle.com/code/gevabriel/indonesian-sms-spam-detection-using-indobert/input>
<https://github.com/paulpriyam/spamTransliteration/tree/master>
<https://www.kaggle.com/datasets/uds5501/sms-dataset/data>

Received: 8 December 2024; Accepted: 26 February 2025

Published online: 10 March 2025

References

1. Lemzy, A. SMS, MMS, and text messages: What are the differences, Textmagic. Accessed: Dec. 05, 2024. [Online]. Available: <https://www.textmagic.com/blog/what-is-sms-and-how-is-it-different-from-text-messages/>
2. Wilkinson, D. SMS vs MMS: Key Differences, SimpleTexting. Accessed: Dec. 05, 2024. [Online]. Available: <https://simpletexting.com/blog/sms-vs-mms-messages/>
3. Liu, E. & SMS vs MMS: Key Differences You Need to Know - OpenPhone. Accessed: Dec. 05, 2024. [Online]. Available: <https://www.openphone.com/blog/sms-vs-mms/>
4. Delany, S. J., Buckley, M. & Greene, D. SMS spam filtering: methods and data. *Expert Syst. Appl.* **39** (10), 9899–9908. <https://doi.org/10.1016/j.eswa.2012.02.053> (Aug. 2012).
5. Roy, P. K., Singh, J. P. & Banerjee, S. Deep learning to filter SMS spam. *Future Generation Comput. Syst.* **102**, 524–533. <https://doi.org/10.1016/j.future.2019.09.001> (Jan. 2020).
6. Abdulhamid, S. M. et al. A review on mobile SMS spam filtering techniques. *IEEE Access* **5**, 15650–15666. <https://doi.org/10.1109/ACCESS.2017.2666785> (2017).
7. Beckman, J. Spam Text Statistics (Growth and Severity of Fraud in 2024), Techreport. Accessed: Dec. 05, 2024. [Online]. Available: <https://techreport.com/statistics/cybersecurity/spam-text-statistics/>
8. Orred, K. 2023 Spam Text Statistics: Are Spam Texts on the Rise? Accessed: Dec. 05, 2024. [Online]. Available: <https://www.text-e-m-all.com/blog/spam-text-statistics>
9. Crestodina, A. Is spam getting worse? Here's the data for 2023, Is spam getting worse? Here's the data for 2023. Accessed: Dec. 05, 2024. [Online]. Available: <https://www.linkedin.com/pulse/spam-getting-worse-heres-data-2023-andy-crestodina/>
10. Chaturved, S. A. & Purohit, L. Spam Message Detection: A Review, *IJCDs*, vol. 12, no. 3, pp. 439–451, Aug. (2022). <https://doi.org/10.12785/ijcds/120135>
11. Poomka, P., Pongsena, W., Kerdprasop, N. & Kerdprasop, K. SMS Spam Detection Based on Long Short-Term Memory and Gated Recurrent Unit, *IJFCC*, vol. 8, no. 1, pp. 11–15, Mar. (2019). <https://doi.org/10.18178/ijfcc.2019.8.1.532>
12. Rajput, J. The Art of SMS Spam Filtering. Accessed: Dec. 05, 2024. [Online]. Available: <https://www.linkedin.com/pulse/art-sms-spam-filtering-janhavi-raiput-dv5tf/>
13. Navaney, P., Dubey, G. & Rana, A. SMS Spam Filtering Using Supervised Machine Learning Algorithms, in *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida: IEEE, Jan. pp. 43–48. (2018). <https://doi.org/10.1109/CONFLUENCE.2018.8442564>
14. Gadde, S., Lakshmanarao, A. & Satyanarayana, S. SMS Spam Detection using Machine Learning and Deep Learning Techniques, in *7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India: IEEE, Mar. 2021, pp. 358–362. (2021). <https://doi.org/10.1109/ICACCS51430.2021.9441783>
15. Liu, X., Lu, H. & Nayak, A. A spam transformer model for SMS spam detection. *IEEE Access* **9**, 80253–80263. <https://doi.org/10.1109/ACCESS.2021.3081479> (2021).
16. Karami, A. & Zhou, L. (PDF) Improving Static SMS Spam Detection by Using New Content-based Features. Accessed: Dec. 05, 2024. [Online]. Available: https://www.researchgate.net/publication/275521521_Improving_Static_SMS_Spam_Detection_by_Using_New_Content-based_Features
17. Delvia Arifin, D., Shaufiah & Bijaksana, M. A. Enhancing spam detection on mobile phone Short Message Service (SMS) performance using FP-growth and Naive Bayes Classifier, in *IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, Bandung, Indonesia: IEEE, Sep. 2016, pp. 80–84. (2016). <https://doi.org/10.1109/APWiMob.2016.7811442>
18. Suleiman, D. & Al-Naymat, G. SMS spam detection using H2O framework. *Procedia Comput. Sci.* **113**, 154–161. <https://doi.org/10.1016/j.procs.2017.08.335> (2017).
19. Bhidya, M. UtkML's Twitter Spam Detection Competition. Accessed: Dec. 05, 2024. [Online]. Available: <https://kaggle.com/twitter-r-spam>
20. Kniazieva, Y. What Is Data Curation in Today's Information Landscape? | Label Your Data. Accessed: Dec. 05, 2024. [Online]. Available: <https://labeyourdata.com/articles/what-is-data-curation>
21. Gong, Y., Liu, G., Xue, Y., Li, R. & Meng, L. A survey on dataset quality in machine learning. *Information and Software Technology* vol **162**, 107268. <https://doi.org/10.1016/j.infsof.2023.107268> (Oct. 2023).
22. Rosli, M. M., Tempero, E. & Luxton-Reilly, A. Evaluating the quality of datasets in software engineering. *adv. Sci. Lett.* **24** (10), 7232–7239. <https://doi.org/10.1166/asl.2018.12920> (Oct. 2018).
23. Gharib, A., Sharafaldin, I., Lashkari, A. H. & Ghorbani, A. A. An Evaluation Framework for Intrusion Detection Dataset, in *International Conference on Information Science and Security (ICISS)*, Pattaya, Thailand: IEEE, Dec. 2016, pp. 1–6. (2016). <https://doi.org/10.1109/ICISSEC.2016.7885840>
24. Kou, G. et al. Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. *Appl. Soft Comput.* **86**, 105836. <https://doi.org/10.1016/j.asoc.2019.105836> (Jan. 2020).
25. Prasad, J. K., Christy, S. & SMS Spam Detection Using Multinational Naive Bayes Algorithm Compared with Decision Tree Algorithm. *BALTIC J. LAW POLITICS*, **15**, 4, doi: <https://doi.org/10.2478/bjlp-2022-004037>. (2022).
26. Tiago Almeida, J. H. SMS spam collection. *UCI Mach. Learn. Repository*. <https://doi.org/10.24432/C5CC84> (2011).
27. Karasoy, O. & Balli, S. Spam SMS Detection for Turkish Language with Deep Text Analysis and Deep Learning Methods, *Arab J Sci Eng.* vol. 47, no. 8, pp. 9361–9377, Aug. (2022). <https://doi.org/10.1007/s13369-021-06187-1>
28. Pal, D. SMS Spam Detection. Accessed: Dec. 05, 2024. [Online]. Available: <https://www.kaggle.com/datasets/debapampal2002/sms-dataset1>

29. Onashoga, A. S., Abayomi-Alli, O. O., Sodiya, A. S. & Ojo, D. A. An Adaptive and Collaborative Server-Side SMS Spam Filtering Scheme Using Artificial Immune System, *Information Security Journal: A Global Perspective*, vol. 24, no. 4–6, pp. 133–145, Dec. (2015). <https://doi.org/10.1080/19393555.2015.1078017>
30. Yadav, K., Kumaraguru, P., Goyal, A., Gupta, A. & Naik SMS Spam Classification on Indian Dataset A Crowdsourced Collection of Hindi and English Messages, GitHub. Accessed: Dec. 05, 2024. [Online]. Available: <https://github.com/princebari/-SMS-Spam-Classification-on-Indian-Dataset-A-Crowdsourced-Collection-of-Hindi-and-English-Messages/blob/main/README.md>
31. Salari, M. & Shayegan, M. A. PSSC dataset: Improving spam detection in Persian SMS by providing a comprehensive dataset, Zenodo. Accessed: Dec. 05, 2024. [Online]. Available: <https://zenodo.org/records/7832188>
32. Gevabriel Indonesian SMS Spam Detection using IndoBERT. Accessed: Dec. 05, 2024. [Online]. Available: <https://kaggle.com/code/gevabriel/indonesian-sms-spam-detection-using-indobert>
33. Priyam, P. spamTransliteration/Personal SMS Data-set, GitHub. Accessed: Dec. 05, 2024. [Online]. Available: <https://github.com/paulpriyam/spamTransliteration/blob/master/Personal%20SMS%20Data-set%20.csv>
34. Singh, U. SMS dataset, Kaggle. Accessed: Dec. 05, 2024. [Online]. Available: <https://www.kaggle.com/datasets/uds5501/sms-dataset>
35. Bird, S., Klein, E. & Loper, E. *Natural Language Processing With Python* 1st edn (O'Reilly, 2009).
36. Ç. Evren (Kapusuz), Stemming Turkish Words Using Snowball. C/S Enformatik Teknolojileri, Dec. 21, 2006. [Online]. Available: <http://cs.com.tr>
37. Sarker, S. & BNLP. Natural language processing toolkit for Bengali language,; arXiv. (2021). <https://doi.org/10.48550/ARXIV.2102.00405>
38. Khalash, M. & Imani, M. Hazm - Persian NLP Toolkit, GitHub. [Online]. Available: <https://github.com/roshan-research/hazm?tab=readme-ov-file#hazm--persian-nlp-toolkit>
39. Diaz, G., Suriyawongkul, A., Pukhalskyi, M. & Solomon, B. stopwords-iso, GitHub. [Online]. Available: <https://github.com/stopwords-iso/stopwords-iso>
40. McCaffrey, J. Multinomial Naive Bayes Classification Using the scikit Library -- Visual Studio Magazine. Accessed: Dec. 05, 2024. [Online]. Available: <https://visualstudiomagazine.com/articles/2023/04/17/multinomial-naive-bayes.aspx>
41. Bhakta, S. Multinomial Naive Bayes, GeeksforGeeks. Accessed: Dec. 05, 2024. [Online]. Available: <https://www.geeksforgeeks.org/multinomial-naive-bayes/>
42. Shriram, M. N. B. & Explained Function, Advantages & Disadvantages, Applications, upGrad blog. Accessed: Dec. 05, 2024. [Online]. Available: <https://www.upgrad.com/blog/multinomial-naive-bayes-explained/>
43. Saini, A. What is Decision Tree? [A Step-by-Step Guide], Analytics Vidhya. Accessed: Dec. 05, 2024. [Online]. Available: <https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/>
44. Abbas, M. et al. Multinomial Naive Bayes Classification Model for Sentiment Analysis, *Unpublished*. (2019). <https://doi.org/10.13140/RG.2.2.30021.40169>
45. Ravindran, R. Overfitting and Pruning in Decision Trees — Improving Model's Accuracy | by Rishika Ravindran | Nerd For Tech | Medium. Accessed: Dec. 05, 2024. [Online]. Available: <https://medium.com/nerd-for-tech/overfitting-and-pruning-in-decision-trees-improving-models-accuracy-fdb9ecd1160>
46. pramod om Decision Trees, Medium. Accessed: Dec. 05, 2024. [Online]. Available: <https://medium.com/@ompramod9921/decision-trees-8e2391f93fa7>
47. Kumar, S. 3 Techniques to Avoid Overfitting of Decision Trees | by Satyam Kumar | Towards Data Science. Accessed: Dec. 05, 2024. [Online]. Available: <https://towardsdatascience.com/3-techniques-to-avoid-overfitting-of-decision-trees-1e7d3d985a09>
48. Ray, S. & Naive Bayes Classifier Explained With Practical Problems., Analytics Vidhya. Accessed: Dec. 05, 2024. [Online]. Available: <https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>
49. Singh, S. Understanding the Bias-Variance Tradeoff, Towards Data Science. Accessed: Dec. 05, 2024. [Online]. Available: <https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229>
50. S. B.R, D. R.M, and A. L., A Collective Approach to SMS Spam Detection: Harnessing the Power of Multimodal Features, Machine Learning, and Advanced Classifiers, in 2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIIE), Ballari, India: IEEE, pp. 1–8. Nov. (2023). <https://doi.org/10.1109/AIKIIE60097.2023.10390514>
51. Agrawal, N., Bajpai, A., Dubey, K. & Patro, B. An Effective Approach to Classify Fraud SMS Using Hybrid Machine Learning Models, in 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), Lonavla, India: IEEE, Apr. pp. 1–6. (2023). <https://doi.org/10.1109/I2CT57861.2023.10126300>
52. Amro, A., Al-Akhras, M., Hindi, K. E., Habib, M. & Shawar, B. A. Instance Reduction for Avoiding Overfitting in Decision Trees, *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 438–459, Jan. (2021). <https://doi.org/10.1515/jisys-2020-0061>
53. Ramakrishnan, K., Balakrishnan, V. & Govaichelvan, K. Linguistic Feature-based Classification for Anger and Anticipation using Machine Learning, in *Proceedings of the 3rd International Conference on Deep Learning Theory and Applications*, Lisbon, Portugal: SCITEPRESS - Science and Technology Publications, pp. 140–147. (2022). <https://doi.org/10.5220/0011289300003277>
54. Tabrez, S. Distribution of Test Data vs. Distribution of Training Data, Tutorialspoint. Accessed: Dec. 05, 2024. [Online]. Available: <https://www.tutorialspoint.com/distribution-of-test-data-vs-distribution-of-training-data>
55. Jeni, L. A., Cohn, J. F. & De La Torre, F. Facing Imbalanced Data—Recommendations for the Use of Performance Metrics, in *Humaine Association Conference on Affective Computing and Intelligent Interaction*, Geneva, Switzerland: IEEE, Sep. 2013, pp. 245–251. (2013). <https://doi.org/10.1109/ACII.2013.47>
56. Vankov, I. I. The hazards of dealing with response time outliers. *Front. Psychol.* **14**, 1220281. <https://doi.org/10.3389/fpsyg.2023.1220281> (Aug. 2023).
57. Hasebe, K. et al. Aug., The Effect of Noise on Deep Learning for Classification of Pathological Voice, *The Laryngoscope*, vol. 134, no. 8, pp. 3537–3541, (2024). <https://doi.org/10.1002/lary.31303>
58. Habermann, T. Metadata and reuse: antidotes to information entropy. *Patterns* **1** (1), 100004. <https://doi.org/10.1016/j.patter.2020.100004> (Apr. 2020).
59. Karcher, S., Kirilova, D., Pagé, C. & Weber, N. How Data Curation Enables Epistemically Responsible Reuse of Qualitative Data, *TQR Jun* doi: <https://doi.org/10.46743/2160-3715/2021.5012>. (2021).
60. Olamendy, J. C. & Practical, M. L. Addressing Class Imbalance, Medium. Accessed: Dec. 05, 2024. [Online]. Available: <https://medium.com/@juanc.olamendy/practical-ml-addressing-class-imbalance-25c4f1b97ee3>
61. Brownlee, J. Failure of Classification Accuracy for Imbalanced Class Distributions, MachineLearningMastery.com. Accessed: Dec. 05, 2024. [Online]. Available: <https://www.machinelearningmastery.com/failure-of-accuracy-for-imbalanced-class-distributions/>
62. Haibo He & Garcia, E. A. Learning from Imbalanced Data, *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. (2009). <https://doi.org/10.1109/TKDE.2008.239>
63. Weiss, G. M. & Provost, F. Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction, *jair*, vol. 19, pp. 315–354, Oct. (2003). <https://doi.org/10.1613/jair.1199>
64. Provost, F. & Fawcett, T. Robust classification for imprecise environments. *Mach. Learn.* **42** (3), 203–231. <https://doi.org/10.1023/A:1007601015854> (2001).
65. Klosterman, S. Why decision trees are more flexible than linear models, Steve Klosterman. Accessed: Dec. 05, 2024. [Online]. Available: <http://www.stevекlosterman.com/flexible-decision-trees/>
66. Splitting in Decision Tree, Shiksha Online. Accessed: Dec. 05, 2024. [Online]. Available: <https://www.shiksha.com/online-courses/articles/splitting-in-decision-tree/>

67. Alharbi, A. A. Classification Performance Analysis of Decision Tree-Based Algorithms with Noisy Class Variable, *Discrete Dynamics in Nature and Society*, vol. pp. 1–10, Feb. 2024, (2024). <https://doi.org/10.1155/2024/6671395>

Acknowledgements

The authors are thankful to the Deanship of Graduate Studies and Scientific Research at the University of Bisha for supporting this work through the Fast-Track Research Support Program.

Author contributions

F.J.M. and H.A.R. performed data collection, experimentation and analysis. F.J.M., K.L.C., and K.S.C.Y. wrote the manuscript. A.S.K., I.A.A. and D.G conducted review and editing. All authors read, revised and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.L.C. or I.A.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025