Predicting SMS Categories with Machine Learning and Text Analysis

Jiaqin Wu

1 Summary

In this research, text-processing techniques and TF-IDF were employed to convert SMS text into matrices, facilitating the use of supervised machine-learning models for categorization. These models consistently achieved an accuracy rate of over 97%, with the random forest model standing out as the top performer at 99.17%.

2 Research Objective

In our rapidly evolving technological landscape, the specter of data privacy concerns looms large, casting a shadow over our digital lives (Mai, 2016). Each day, the average person's inbox is inundated with a deluge of unsolicited Short Message Services (SMS). While some of these messages may be benign, others harbor insidious risks, potentially leading to fraudulent activities that can wreak havoc in our lives (Prusty, Sainath, Jayasingh, & Mantri, 2022). What compounds this issue is the unsettling fact that a majority of individuals remain blissfully ignorant of the intricate pathways through which their personal information is surreptitiously disseminated across the vast expanse of the internet (Malandrino et al., 2013). Despite the growing awareness of these concerns, they persist as a daunting challenge.

In the contemporary landscape, where we navigate through a multitude of applications and websites, our mobile phones have become indispensable tools for identity verification. Yet, it is disheartening to acknowledge that nefarious actors often exploit this very process, operating in the shadows as they covertly trade our sensitive information for financial gain. The magnitude of this problem extends beyond mere inconvenience; it holds profound societal implications (Athey, Catalini, & Tucker, 2017).

Within this context, this research project seeks to harness the informational content embedded within these messages, with the primary objective of accurately classifying them as either spam or ham. By delving deep into the intricacies of these messages, our aim is to unveil the underlying patterns and distinct characteristics that distinguish spam messages from legitimate ones. Through a comprehensive examination of the textual and contextual aspects, we endeavor to shed light on these clandestine operations. By doing so, we aspire to empower individuals with the knowledge and awareness needed to safeguard their personal information and navigate the digital world with confidence and resilience. In an era where data privacy concerns continue to evolve, this endeavor is not only timely but essential for the well-being of individuals and the security of our interconnected society (Androulidakis, 2016).

3 Data Sources

This research delves into the intricate interplay between text-based elements and their influence on the classification of SMS messages. The dataset under scrutiny originates from the UC Irvine Machine Learning Repository (Almeida & Hidalgo, 2012), encompassing a corpus of 5574 SMS texts from diverse sources. Within this dataset, we initially encountered a pair of variables: one for text content and the other for labeling messages as either spam or ham. Before embarking on the more granular aspects of our research, a preliminary data check was imperative. Fortunately, no missing values were uncovered, but our dataset harbored 403 duplicated records. To ensure data integrity, these duplications were duly purged, leaving us with a clean dataset of 5171 SMS texts.

The ensuing analysis calls for a broader perspective, where the distribution of SMS text types comes into play (Figure 1). It becomes evident that ham SMS messages substantially outnumber their spam counterparts within our dataset. This distribution sets the stage for our investigation into text length and its implications.

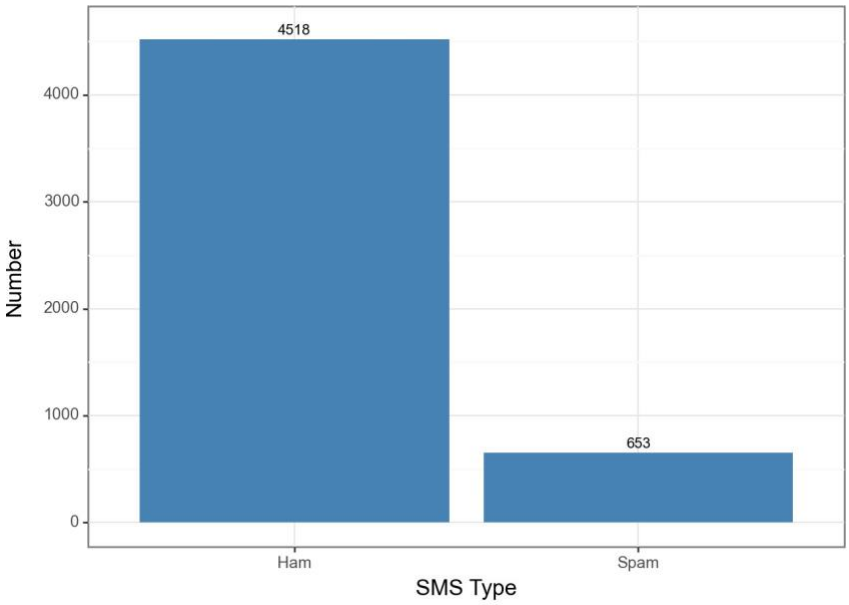

Figure 1: Distribution of SMS Text Types

To facilitate further analysis, we generated a new variable: ”text length,” derived from the contents of the messages. Descriptive statistics (Table 1) were then employed to characterize the distribution of text lengths. Intriguingly, the descriptive table hints at the presence of potential outliers, inviting a deeper examination. This suspicion led to the creation of a boxplot (Figure 2), vividly illustrating the distribution of text length values. The plot reveals that most text lengths fall within the range of 0 to 50 words, yet outliers with lengths exceeding 50 words are conspicuous.

Table 1: Descriptive Information for Text Length (N=5171)

Mean	Std	Min	25%	Med.	75%	Max
15.43	11.10	1	7	12	22	171

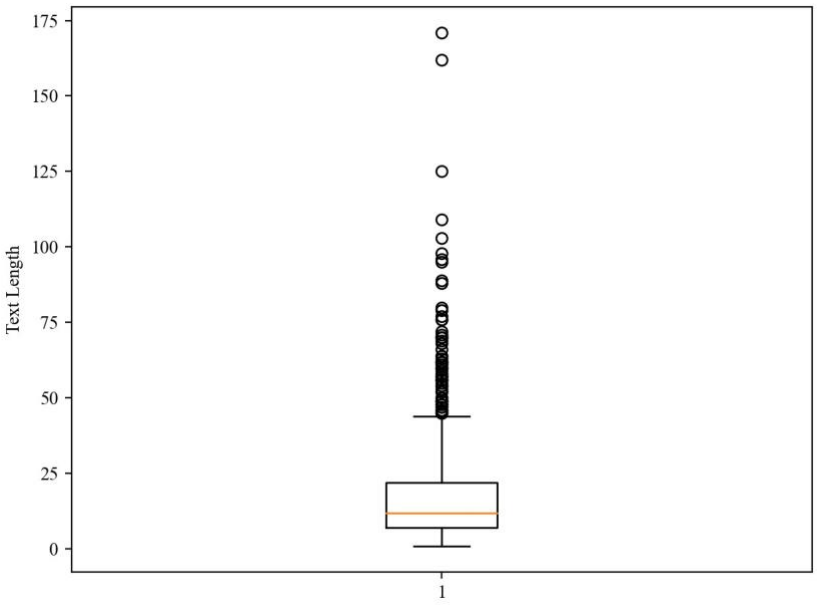

Figure 2: Distribution of SMS Text Length

Our inquiry extends to a comparison between the text lengths of spam and ham SMS messages. This investigation is furthered by density plots, with one representing spam lengths and the other ham lengths (Figure 3). Notably, the density plot discloses that spam SMS messages predominantly exhibit lengths of around 25 words, while ham SMS messages tend to cluster around 8 words. However, it is worth noting that some ham SMS messages extend beyond 50 words, thus contributing to the dataset's outlier values.

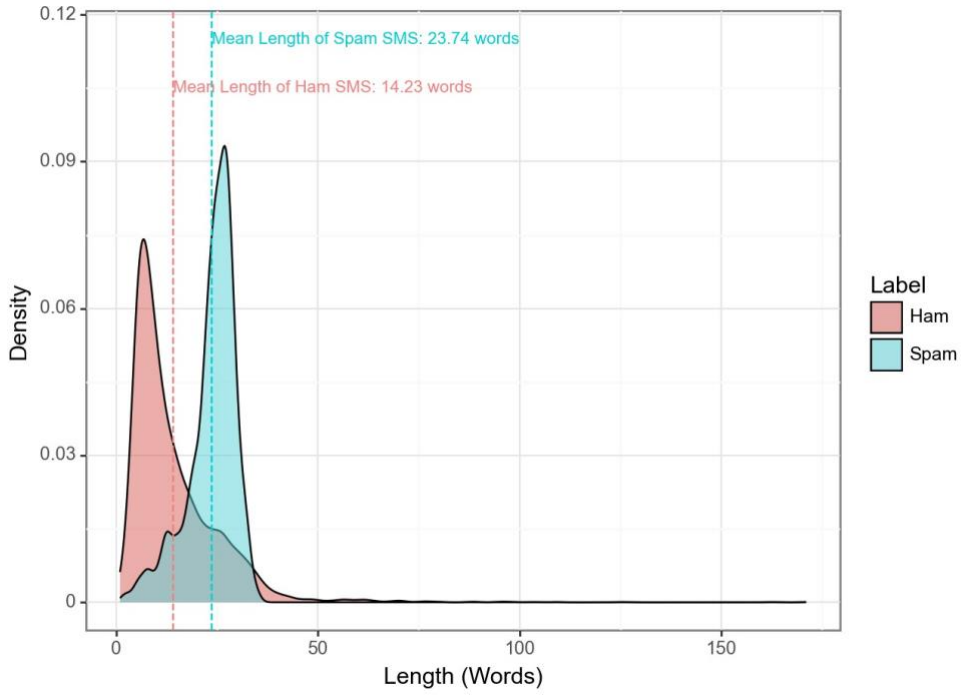

Figure 3: Distribution of SMS Text Length By SMS Types

The quest for differentiation between key terms in spam and ham SMS messages leads us to employ word cloud visualizations (Figure 4). These word clouds vividly illustrate the most frequently occurring terms within the original text contents. In the ham message word cloud, colloquial terms such as "u," "ur," and "Ok" prevail. Conversely, the spam message word cloud showcases more directive language, featuring terms like "please call," "call," and "reply," as well as enticing phrases like "attractive prizes," including keywords like "prize," "free," "win," and "won." These distinctive terms equip us with valuable insights for discerning between spam and ham SMS messages.

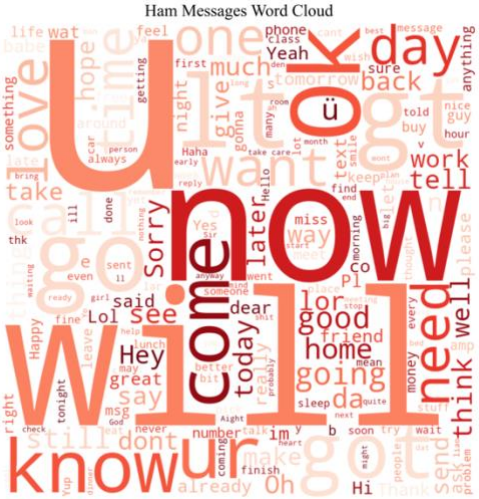

Figure 4: Word Cloud Figures

In order to prepare our data for classification prediction, a series of preprocessing steps were applied to refine the original texts, which often contained various extraneous symbols and characters. These steps aimed to streamline the data into a more manageable format. Initially, all text was converted to lowercase to ensure uniformity during subsequent comparisons and analyses. Following this, the text was tokenized, breaking it down into individual words by considering spaces and punctuation marks as word separators. To enhance the precision of our analysis and account for words sharing common roots, we performed lemmatization to reduce each word to its base form. Additionally, to eliminate irrelevant terms that added little to the text's meaning, common stopwords were removed from the original texts.

To facilitate the application of supervised machine learning models for prediction, we adopted the TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer for text transformation. TF-IDF was chosen for its ability to balance the importance and frequency of words within the text. To account for the significance of both single words (unigrams) and word pairs (bigrams) in the text context, our vectorizer was configured to generate both unigrams and bigrams during the transformation process. However, the resulting matrices would contain 39,001 columns without any restrictions in the vectorizer, including infrequent terms that contribute little to the analysis. To address this issue, we set the "max features" parameter to 3000 during transformation, ensuring that only the most frequently occurring terms were retained. As a result of these transformations, our final prediction matrix possessed a dimension of 5171 rows by 3000 columns.

4 Techniques Applied

In this research, a series of crucial techniques were applied to effectively process and analyze the text data for the classification task, ensuring that the models could make accurate predictions.

- Text Preprocessing: To ensure the quality and reliability of our textual data, we initiated a meticulous text preprocessing stage. This critical step was implemented to rectify and enhance the overall cleanliness, structure, and interpretability of the textual content.

- TF-IDF Vectorization: To represent the text data in a format suitable for machine learning models, the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization technique was utilized. TF-IDF offers a method to transform the original text into matrices, where each row corresponds to a document, and each column corresponds to a unique term or phrase. Notably, the "sublinear" scaling and normalization techniques were intentionally omitted during TF-IDF vectorization. This decision aligns with the research focus on keyword-based analysis, prioritizing the identification of keywords and key phrases that strongly indicate whether a message is spam or ham. The omission of "sublinear" scaling and normalization was intentional as it preserves the emphasis on keyword importance within the text.

- Data Splitting and Balancing: Prior to model training, the dataset was divided into training and testing sets. Given the substantial class imbalance, where ham messages significantly outnumber spam messages, a strategic approach was taken to balance the proportions in both the training and testing sets. To address this, upsampling was applied to the original dataset, effectively mitigating class imbalance. This balancing technique was crucial to ensure that the machine learning models could make accurate predictions and maintain fairness in their assessments.

- Model Selection and Evaluation: In the process of selecting machine learning models, we explored a diverse array of algorithms, encompassing both parametric and non-parametric approaches. This comprehensive approach led us to consider models such as logistic regression, a parametric model, in conjunction with non-parametric alternatives like Naive Bayes and decision trees. To ensure optimal model performance, we utilized grid search in hyperparameter fine-tuning across all the models and selected the model with the best performance in the final evaluation. For instance, within the logistic regression model, we experimented with different regularization techniques and penalty scores to gauge their impact on performance. Similarly, in the case of decision trees, we fine-tuned parameters such as maximum depth, minimum leaf size, and criterion.

  Our commitment to thorough evaluation extended to the assessment of model performance. This evaluation was conducted through a rigorous cross-validation procedure, employing a k-fold approach

with k set to 5. By doing so, we were able to comprehensively gauge each model's ability to handle previously unseen data, ensuring that our selected models were robust and capable of generalizing effectively.

5 Findings

5.1 Key Insights

This research has yielded compelling insights, primarily regarding the outstanding performance of our chosen supervised models in predictive tasks. These models consistently achieved remarkably high accuracy rates, with the majority surpassing the notable 97% threshold and some even exceeding 99%, shown in Figure 4. This revelation underscores the substantial distinctions between the prevalent word patterns in spam SMS messages and those within ham SMS communications.

Among the array of models utilized for prediction, the K-Nearest Neighbors (KNN) model, XGBoost, and Random Forest model emerged as top-performing choices. The success of these models is attributed to their distinct working mechanisms.

The KNN model exhibited exceptional performance following fine-tuning, where it proved its efficacy by accurately classifying data points based on their proximity to others. Specifically, by setting the number of neighbors to 11 and opting for the Manhattan distance metric, the KNN model excelled in delineating patterns within our dataset. This approach leveraged the spatial relationships between data points, ensuring precise label classification based on their distances.

On the other hand, the XGBoost and Random Forest models belong to the category of decision tree classifiers. These models operate by scrutinizing the textual patterns within the dataset and subsequently

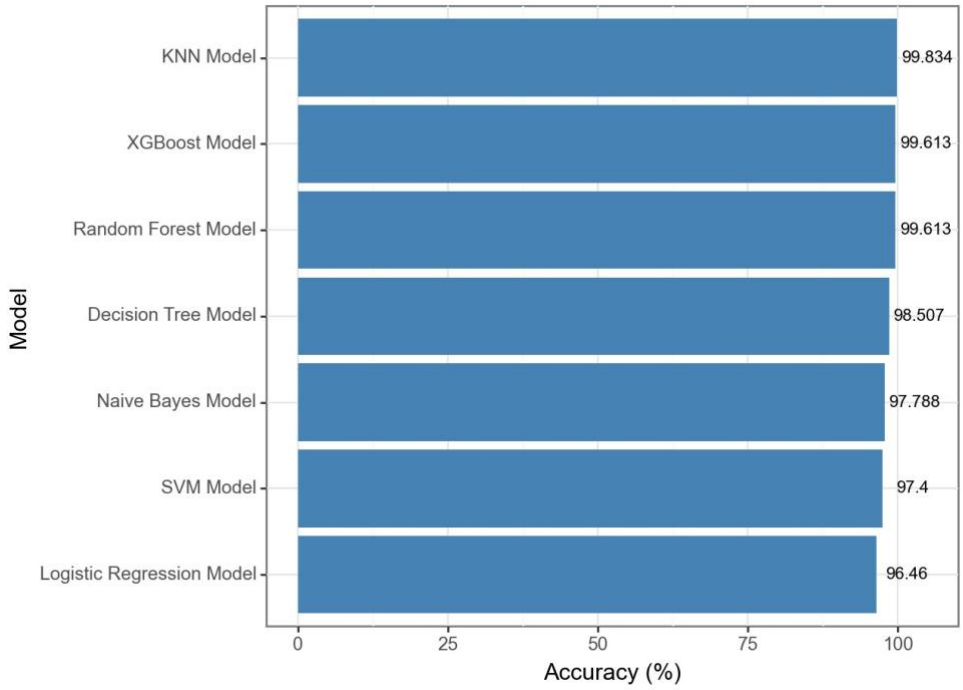

Figure 5: Comparison of Accuracy of Different Models

assigning each SMS message to its appropriate group. This text pattern analysis allows for the effective categorization of SMS messages, a testament to the strength of decision tree-based classifiers in handling complex data patterns and improving prediction accuracy.

Subsequently, we proceeded to identify the model with the highest prediction accuracy and meticulously assessed its performance using a confusion matrix. In Figure 5, we gain deeper insights into the model's proficiency. Notably, the selected model adeptly discriminates between ham and spam SMS, signifying distinct stylistic disparities between the two categories. While the model's performance is commendable, a

closer examination reveals a nuanced aspect. In the case of the KNN model, a total of 3 spam SMS instances are misclassified as ham SMS. This phenomenon sheds light on the existence of a very small subset of exceptionally crafty spam messages that mimic the characteristics of legitimate ham SMS, making their detection a challenging task. Confronting this challenge underscores the importance of bolstering the model's discriminatory abilities in identifying these subtle nuances. This can be achieved by augmenting the model's training dataset with an expansive and diverse array of samples.

we then went a step further by extracting pivotal features, namely word terms, from advanced treebased models such as decision trees, random forests, and XGBoost. These models excel in classification tasks, leveraging the significance of individual word terms to achieve remarkable predictive accuracy. Our objective was to unveil the key contributors influencing the models' judgments, specifically in discerning whether a message falls into the spam or ham category. The insightful findings, presented in Figure 6, underscore the prominence of certain words. Notably, terms like "call," "free," "win," and essential indicators such as "www," "http," and "sms" emerged as highly influential. This exploration sheds light on words that significantly impact the classification process, serving as a valuable resource to prompt individuals to exercise vigilance when receiving messages from unfamiliar sources. Such awareness is instrumental in discerning and avoiding potentially harmful or unwanted messages.

5.2 Implications

The insights derived from this research carry significant implications, particularly in the realm of policy formulation and decision-making. With an in-depth understanding of the noteworthy distinctions between spam and legitimate SMS communications, policymakers can design more targeted and effective interventions to combat unsolicited and potentially harmful messages. The high accuracy rates achieved by the supervised models, particularly the Random Forest and SVM, indicate their potential utility in real-world applications such as spam message filtering and prevention. These models could serve as invaluable tools

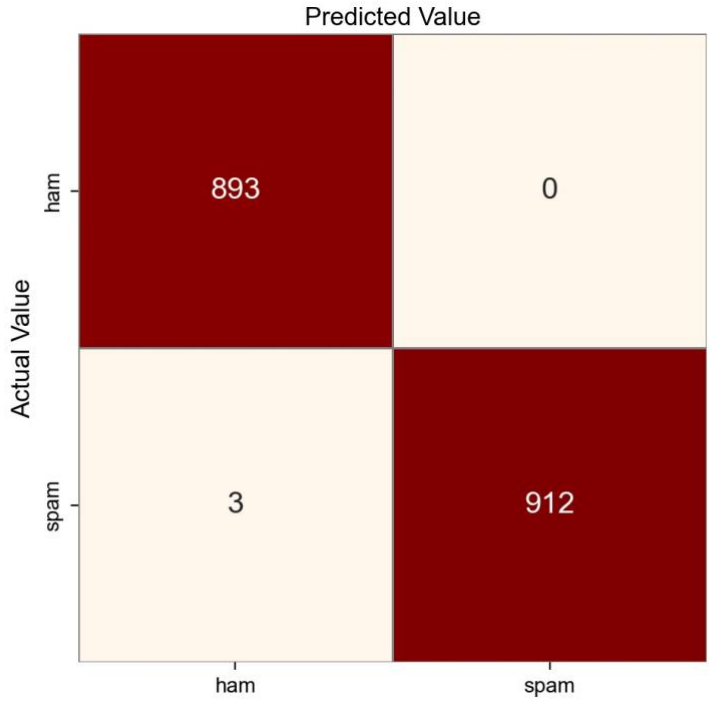

Figure 6: Confusion Matrix of the KNN Model

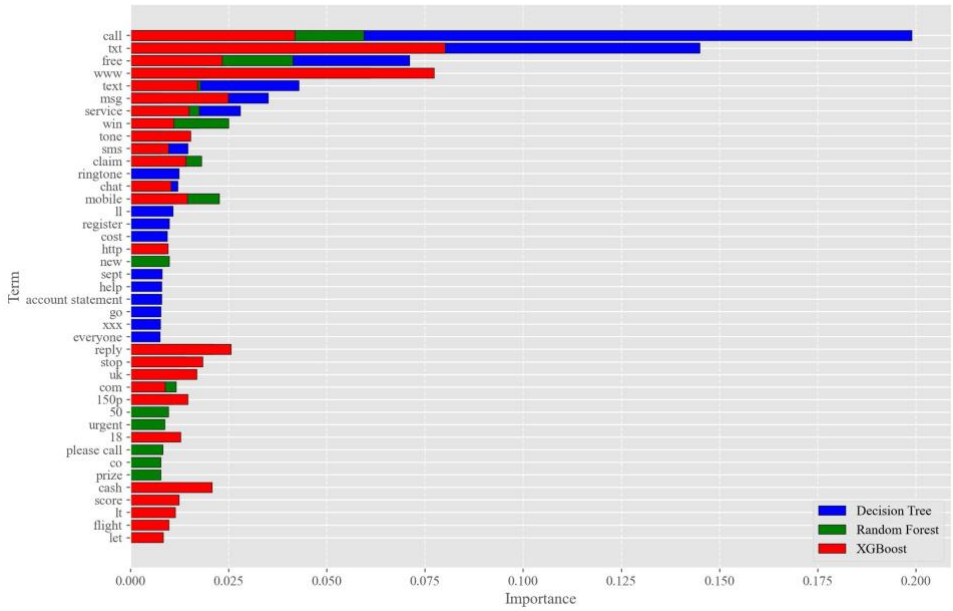

Figure 7: Comparison of Top 25 Feature Importance of Tree-based Models

for telecommunications regulatory authorities, enabling the development of robust anti-spam measures. Furthermore, the research's emphasis on algorithm refinement, including hyperparameter tuning, highlights the importance of optimizing machine learning techniques to enhance their predictive capabilities. The insights garnered here underscore the promise of machine learning in the realm of telecommunications regulation, paving the way for more efficient and effective policies to protect users from unwanted and malicious messages.

5.3 Limitations and Further Considerations

Nonetheless, it is imperative to acknowledge the temporal constraints of our dataset, which contains text contents dating back to 2012. In the swiftly changing landscape of technology and messaging, nearly a decade has transpired since the inception of this data. Over this period, it is reasonable to assume that fraudulent and advertising activities have evolved, ushering in new keywords, tactics, and trends in the realm of SMS-based spam. While our research provides invaluable insights into historical spam message patterns, it may not comprehensively encapsulate the dynamic nature of contemporary spam messages.

To address this limitation, future investigations should aim to employ more recent datasets, thereby offering a more accurate reflection of the current state of SMS-based spam. However, it is crucial to highlight that the enduring principles and strategies elucidated in our research transcend temporal boundaries. These timeless insights empower users to adeptly recognize and combat spam messages, providing a valuable resource for users navigating the evolving landscape of spam communication. As such, our findings remain relevant and applicable, regardless of the ever-changing tactics employed by spammers.

References

Almeida, T., & Hidalgo, J. (2012). *SMS Spam Collection.* UCI Machine Learning Repository. (DOI: https://doi.org/10.24432/C5CC84)

Androulidakis, I. I. (2016). SMS security issues. In *Mobile phone security and forensics* (pp. 71–86). Cham: Springer International Publishing.

Athey, S., Catalini, C., & Tucker, C. E. (2017, April). The digital privacy paradox: Small money, small costs, small talk. *SSRN Electron. J.*.

Mai, J.-E. (2016, May). Big data privacy: The datafication of personal information. *Inf. Soc.*, *32*(3), 192–199.

Malandrino, D., Petta, A., Scarano, V., Serra, L., Spinelli, R., & Krishnamurthy, B. (2013, November). Privacy awareness about information leakage. In *Proceedings of the 12th ACM workshop on workshop on privacy in the electronic society.* New York, NY, USA: ACM.

Prusty, S. R., Sainath, B., Jayasingh, S. K., & Mantri, J. K. (2022). SMS fraud detection using machine learning. In *Intelligent systems* (pp. 595–606). Singapore: Springer Nature Singapore.