IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, June 2025
Impact Factor: 7.67

ML-Based Classification of Malicious and Legitimate Messages

Dr. Rachana P, Akash Pujari, Surabhi, Asha H. D, Guruprasad
Department of Information Science and Engineering
Alvas Institute of Engineering and Technology, Mijar, Mangalore, India

Abstract: *The fact that SMS spam is still a major problem highlights the need for study into creating systems that may thwart the evasive tactics utilized by spammers. Protecting the public from the negative impacts of SMS spam requires this kind of study. The main obstacles in the existing SMS spam detection and filtering environment are highlighted in this study. We present a new SMS dataset with around 68,000 messages, 39% of which are classified as spam and 61% as valid (ham) in order to aid study in this field. Notably, this dataset which has been made freely available for research purposes represents the biggest dataset of SMS spam that has been made available thus far. To investigate how spam strategies have changed over time, we do a longitudinal analysis. Furthermore, in order to assess and contrast the effectiveness of various SMS spam detection models—from conventional shallow machine learning techniques to sophisticated deep neural networks we extract both semantic and syntactic data. Our study evaluates how effectively these models and well-known commercial antispam services withstand typical spammer evasion techniques. The findings show that most shallow learning methods and existing anti-spam programs have trouble correctly identifying spam communications, particularly when confronted with complex obfuscation techniques.*

Keywords: detection of SMS spam, anti-spam services, evasive tactics, robustness study of machine learning, spamdataset, and development of SMS spam

I. INTRODUCTION

Even after over 20 years of research, SMS spam detection is still a major problem in contemporary digital cultures. According to estimates, SMS spam has increased to dangerous proportions, with losses in the United States in 2022 expected to exceed USD $330 million, more than twice as much as in 2021. ScamWatch, an Australian organization, also noted that losses from scams increased from AUD 175 million in 2020 to AUD 323 million in 2021. In February 2022 alone, there were over 8,835 instances of SMS fraud, increasing from 32,337 to 67,180 that year, making SMS the most popular way for scams to be distributed.

In the fight against SMS spam, this study highlights four major obstacles. First, there aren't many real-world, annotated datasets available. Due to these restrictions, models are unable to generalize and identify hidden spam patterns. The absence of benchmark datasets is the second issue, which makes it challenging to objectively assess how well different suggested detection techniques work. This discrepancy has caused disjointed research projects with ambiguous outcomes. Third, current machine learning models frequently lack resilience to spammers' evasive tactics. Even sophisticated spam filters are susceptible to evasion techniques like encoded URLs and obfuscation, as well as minor text changes. When evaluating models, these changing dangers are frequently overlooked. Concept drift, which occurs when models trained on historical data are unable to adjust to more recent fraud trends, is the fourth problem.

This paper offers various significant contributions to the solution of these problems. Using data from ScamWatch and Action Fraud, it presents a fresh, extensive dataset of 67,018 tagged SMS messages, of which 60.9% are real and 39.1% are spam. The scope and significance of this dataset, which spans the years 2012–2023, exceeds all other publicly available datasets on SMS spam. Preprocessing was done extensively, which included spam classification, imagebased text translation, and deduplication. To promote more study, the dataset and underlying code are made freely available.

DOI: 10.48175/IJARSCT-27305

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, June 2025
Impact Factor: 7.67

Additionally, supervised learning, deep learning, one-class learning, and positive-unlabeled learning are among the machine learning models that are evaluated in this study for SMS spam detection. Word2Vec allowed positive-unlabeled algorithms to attain up to 79% F1 score, which is similar to more conventional supervised methods like SVM. The assessment encompassed a variety of feature extraction methods, from semantic embeddings like Word2Vec, fastText, GloVe, BERT, RoBERTa, DistilBERT, and ELMo to non-semantic models like Count Vectorization and TFIDF. With the exception of one-class learning environments, semantic embeddings continuously improved performance.

II. LITERATURE REVIEW

Machine learning (ML), natural language processing (NLP), and statistical methods have been the mainstays of research on SMS (Short Message Service) spam detection for almost 20 years. Using deep learning techniques, traditional classification algorithms, and, more recently, transformerbased architectures, several research have put forth spam detection models. Notwithstanding these initiatives, SMS spam is still a major issue, especially given the speed at which spam strategies are developing, the paucity of data, and the absence of practical assessment.

The 2012 publication of the SMS Spam Collection dataset by Almeida et al. [4] is among the most well-known datasets in SMS spam research. Out of the 5,574 messages, 747 have been classified as spam. This dataset's tiny size and out-ofdate content severely limit its relevance to contemporary spam patterns, despite the fact that it has been used extensively for benchmarking. Other databases, such as the SpamHunter dataset [14], only include 947 spam messages despite efforts to increase the number and quality of data. These datasets are frequently quite unbalanced, which makes training and assessing models even more difficult.

To detect spam, a variety of machine learning techniques have been used. Simple text characteristics like bag-ofwords (BoW) and TF-IDF were used in early research [1], [2], and [5] along with algorithms like Naïve Bayes, Support Vector Machines (SVM), and Decision Trees. Recent research has examined deep learning models, including hybrid models that integrate syntactic and semantic data, Convolutional Neural Networks (CNNs), and Long ShortTerm Memory networks (LSTMs) [16]–[20]. Though their performance has increased, these models are frequently only tested on static datasets that do not account for changing spam strategies.

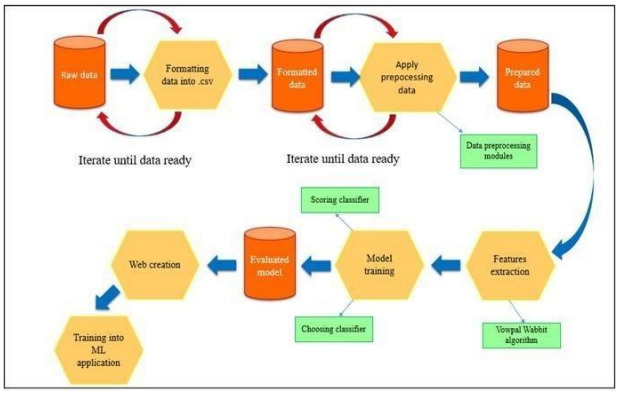

Fig: spam message and mail classification

III. SMS SPAM DATASET

To facilitate study in the field of SMS spam identification, a variety of SMS datasets have been made available. Nevertheless, a low percentage of spam messages, obsolete information, and tiny size are some of the major drawbacks of the majority of these datasets. Some of the most well-known datasets on SMS spam are included in Table 1 (not displayed here), which spans from early contributions in 2012 to more current initiatives in 2022.

The 2012 edition of the SMS Spam Collection [4] is among the first and most popular datasets. Since a large portion of the spam messages in this dataset predate 2010, it is currently regarded as outdated. It also has a class imbalance, with just 747 out of 5,574 messages being spam.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, June 2025
Impact Factor: 7.67

The National University of Singapore created the NUS SMS Corpus [31], another dataset that has a higher number of texts (67,093) but was last updated in March 2015. Its effectiveness in training algorithms to detect contemporary, changing spam techniques is limited by the absence of recent spam data, despite its advantageous size.

In order to overcome the lack of current datasets, a more recent project called SpamHunter [14] gathered publicly published SMS screenshots from Twitter. With this method, about 25,889 texts in various languages were collected between 2018 and 2022. Notwithstanding its novel approach, noise significantly impacts the SpamHunter dataset. This includes duplicates, non-spam awareness messages, and a high rate of OCR-related mistakes brought on by text extraction from photos.

IV. SMS SPAM DETECTION

Numerous deep learning (DL) and classical machine learning (ML) techniques have been put forth to address the issue of SMS spam. For example, Almeida et al. [4], [41] compared a number of machine learning classifiers and found that Support Vector Machines (SVM) were the most successful. They did not investigate any deep learning models, though, and their investigation was restricted to utilizing solely word frequency characteristics. Similarly, employing TF-IDF characteristics, Gupta et al. [6] examined eight distinct machine learning methods, including SVM and a single deep learning model (CNN). Their assessment lacks variety in terms of model kinds and attributes.

By comparing CNN and LSTM models with conventional binary classifiers in a stacked ensemble style, Roy et al. [42] expanded on this investigation. The DL models, especially CNN and LSTM, showed better performance. Nevertheless, their research was limited to traditional two-class classification and omitted contemporary transformer-based models, which have proven to be more effective in a range of NLP applications.

Although they used LSTM architectures with various word embedding strategies, Jain et al. [43] only looked at Word2Vec embeddings, ignoring the possibilities of contextual word embeddings and transformer models like BERT or RoBERTa. This restricts how well their findings may be applied to current spam detection problems.

V. DATA COLLECTION AND AUGMENTATION

In order to facilitate our investigation into SMS spam identification, we conducted a comprehensive survey to locate, collect, and aggregate publically accessible SMS datasets. Our technique comprised using specific terms such "SMS dataset," "text messages," "spam SMS," and "short message service" to search via a variety of platforms, including GitHub, Google Scholar, and general online sources. Only resources that provided publicly available datasets for scholarly or research purposes were screened and chosen. Consequently, we compiled an extensive dataset of 179,440 SMS texts from various public sources in various languages.

In order to improve our corpus's freshness and diversity, we also focused on social media sites. We specifically looked for spamrelated information uploaded as screenshots of SMS on Twitter. These were gathered from tweets that were published between January 2012 and December 2017 and again between August 2022 and July 2023, expanding the time frame that the current SpamHunter dataset covers. Official scam reporting websites like Scamwatch and Action Fraud were among the other sources from which we obtained screenshots of publicly reported SMS scams.

Our goal was to develop a strong spam detection algorithm tailored to SMS texts sent in English. We used a two-stage filtering approach during preprocessing to guarantee linguistic consistency. The first step was identifying and removing non-English texts using the langdetect Python package. We ensured a clean, language-consistent dataset in the second step by validating the remaining messages and removing any leftover non-English content using the GoogleTrans module.

To convert photos into text format, we utilized the Pytesseract OCR library for SMS screenshots that were gathered from social media and websites that report scams. After language filtering, duplicate removal, and initial categorization, we combined 62,114 distinct English-language messages from pre-existing datasets with 4,904 from our fresh data set.

Of these, 60,032 messages were not classified at first. We came up with a set of nine guidelines (shown in Table 2) to help with manual categorization into "Ham" or "Spam" categories. These guidelines were developed through a review of prevalent scam patterns, conversations among the research team, and in-depth examination of previous datasets. To

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, June 2025
Impact Factor: 7.67

assure labeling accuracy, three researchers independently evaluated each communication based on the established criteria, and conflicts were settled cooperatively.

A high-quality, labeled dataset that we call the "Super Dataset" is the end product of our data collecting and augmentation procedure. 40,837 (60.9%) of the 67,018 SMS messages in it are classified as authentic (ham), while 26,181 (31.1%) are classified as spam.

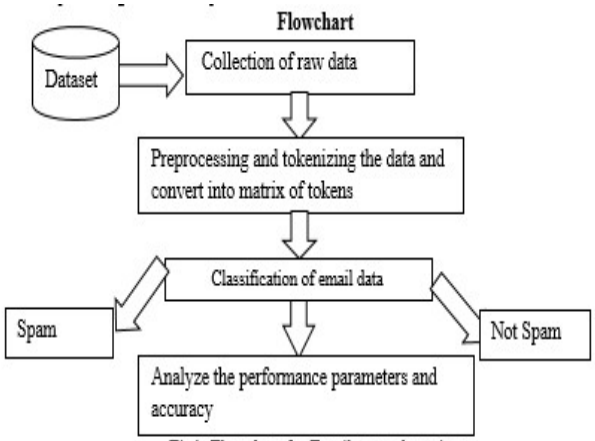

Fig1: Flowchart for Email spam detection

Fig . FLOW DIAGRAM

VI. EVOLUTION OF SMS SPAM: AN ANALYSIS OF CHANGING CHARACTERISTICS

We provide a longitudinal analysis of SMS spam in this part to look at how its characteristics and tactics have changed over time. We looked at temporal patterns, message attributes, and the evolving strategies used by spammers using a large dataset covering the years 2012–2023.

In order to make this analysis easier, we first timestamped every SMS message according to the date it was initially published or collected. After that, the entire dataset was arranged chronologically and divided into two separate subsets: DS_Legacy, which included 37,615 SMS messages gathered between 2012 and 2017 (including spam from Twitter) and DS_Latest, which included 29,403 SMS messages gathered from datasets and usercontributed collections between 2018 and 2023.

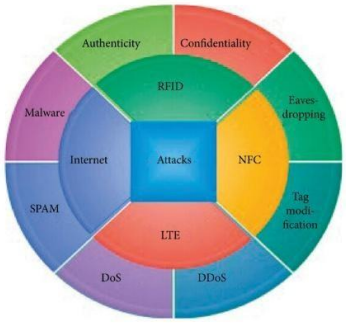

Fig. Technology used for classification

This historical separation made it possible for us to spot distinct trends and changes in spam tactics and content. Our study reveals significant changes in sender behavior, themes, and language structure across time. We found that previous spam communications tended to be more generic and used highfrequency promotional language, but more contemporary mailings use more advanced strategies such URL obfuscation, tailored content, and taking advantage of current affairs or emergencies.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, June 2025
Impact Factor: 7.67

The necessity for flexible, current spam detection algorithms is further supported by these findings, which offer insightful information on how spammers constantly modify their strategies to evade detection. The variations that have been seen highlight how dynamic SMS spam is and how crucial it is to keep a current, varied dataset for efficient detection.

VII. CONCLUSION

With potential consequences ranging from user discomfort to financial fraud, SMS spam remains a persistent problem in the field of digital communication. In order to address the problem, this research combined a variety of publicly accessible SMS datasets to produce an extensive and varied "Super Dataset" that contains both recent and historical spam messages.

We examined the long-term patterns in SMS spam and found that spam strategies have changed significantly, with contemporary spam messages becoming more individualized, dishonest, and challenging to identify with conventional techniques. By going beyond out-of-date datasets and constrained feature sets, our work also addressed the shortcomings of previous studies by assessing a wide variety of machine learning and deep learning models using both syntactic and semantic information.

REFERENCES

[1] J. Buchanan and A. J. Grant, "Investigating and prosecuting Nigerian fraud," U.S. Att'ys Bull., vol. 49, pp. 39–47, Nov. 2001.

[2] L. Zhang, J. Zhu, and T. Yao, "An evaluation of statistical spam filtering techniques," ACM Trans. Asian Lang. Inf. Process., vol. 3, no. 4, pp. 243–269, Dec. 2004.

[3] G. L. Wittel and S. F. Wu, "On attacking statistical spam filters," in Proc. Conf. Email and Anti-Spam (CEAS), 2004. [4] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, "Contributions to the study of SMS spam filtering: New collection and results," in Proc. 11th ACM Symp. Document Eng., Sep. 2011, pp. 259–262.

[5] T. Almeida, J. M. Hidalgo, and T. Silva, "Towards SMS spam filtering: Results under a new dataset," J. Inf. Secur. Syst., vol. 2, no. 1, pp. 1–18, 2013.

[6] M. Gupta, A. Bakliwal, S. Agarwal, and P. Mehndiratta, "A comparative study of spam SMS detection using machine learning classifiers," in Proc. 11th Int. Conf. Contemp. Comput. (IC3), Aug. 2018, pp. 1–7.

[7] S. Rojas-Galeano, "Using BERT encoding to tackle the mad-lib attack in SMS spam detection," arXiv preprint arXiv:2107.06400, 2021.

[8] FCC, "The top text scams of 2022," Jun. 2023. [Online]. Available: https://www.ftc.gov/news-events/datavisualizations/data-spotlight/2023/06/iykyk-top-text-scams-2022

[9] ACCS, "Scam statistics," 2022. [Online]. Available: https://www.scamwatch.gov.au/scam-statistics

[10] M. A. Abid, S. Ullah, M. A. Siddique, M. F. Mushtaq, W. Aljedaani, and F. Rustam, "Spam SMS filtering based on text features and supervised machine learning techniques," Multimedia Tools Appl., vol. 81, no. 28, pp. 39853–39871, Nov. 2022. [11] I. Ahmed, R. Ali, D. Guan, Y.-K. Lee, S. Lee, and T. Chung, "Semi-supervised learning using frequent itemset and ensemble learning for SMS classification," Expert Syst. Appl., vol. 42, no. 3, pp. 1065–1073, Feb. 2015.

[12] C. Oswald, S. E. Simon, and A. Bhattacharya, "SpotSpam: Intention analysis-driven SMS spam detection using BERT embeddings," ACM Trans. Web, vol. 16, no. 3, pp. 1–27, Aug. 2022. [13] S. Y. Yerima and A. Bashar, "Semi-supervised novelty detection with one class SVM for SMS spam detection," in Proc. 29th Int. Conf. Syst., Signals Image Process. (IWSSIP), Jun. 2022, pp. 1–4.

[14] S. Tang, X. Mi, Y. Li, X. Wang, and K. Chen, "Clues in tweets: Twitter-guided discovery and analysis of SMS spam," in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., Nov. 2022, pp. 2751–2764.

[15] A. van der Schaaf, C.-J. Xu, P. van Luijk, A. A. van't Veld, J. A. Langendijk, and C. Schilstra, "Multivariate modeling of complications with data driven variable selection: Guarding against overfitting and effects of data set size," Radiother. Oncol., vol. 105, no. 1, pp. 115–121, Oct. 2012.