Next-Generation Spam Filtering: Comparative Analysis of Text Classification Approaches

*Dr. Akhil Pandey , Dr. Vishal Srivastava , Er. Sangeeta Sharma*, , *Banti Lal* **,** Professor, Assistant Professor, B. Tech Scholar, Department of Computer-Science Of Engineering, Artificial Intelligence And Data Science Arya College of Engineering and IT, Kukas, Jaipur, Affiliated with Rajasthan Technical University(RTU), Kota
Email:Vishal Shrivastava sangeetayuwansh1@gmail.com, **neeleshbhardwaj381@gmail.com**

Abstract

The explosive growth of text data poses daunting management and analysis challenges, especially in terms of high storage and processing requirements1. Spam dissemination across multiple digital media requires advanced and adaptive filtering mechanisms. This paper presents a thorough survey and experimental investigation of different text classification methods, conventional machine learning methods, and deep learning methods, for spam filtering. A new research domain taxonomy is introduced to classify these methods, providing a high-grained overview of their strengths and weaknesses. The paper also discusses the application of advanced language models, such as GPT-4, BERT, and Roberta, with Convolutional Neural Networks (CNNs) in email spam filtering, establishing the effectiveness of this fine-tuning of these models to develop effective defense against adaptive spam and phishing attacks2. AI-generated manipulative content detection opportunities and challenges are discussed, with emphasis on the significance of ethical considerations and the application of successful detection methods, such as reasonable AI (XAI) and ensemble learning methods.

This book is a contribution to the development of spam filtering techniques and is the basis for secure email defense mechanisms against constantly changing attacks2.

Keywords:

Text categorization, spam message filtering, artificial intelligence, deep neural networks, computational linguistics, ethical artificial intelligence, explainable artificial intelligence, collective learning, and swarm-based intelligence.

I. Introduction

The increasing volume of unwanted and malicious content, also known as spam, poses a severe threat to individuals, organizations, and society2. Spam email overloads email users, consumes network bandwidth, occupies disk spaces, and is frequently packed with malware4. In 2022, the average total cost of a data breach globally was USD 4.35 million, and stolen credentials triggered 19% of breaches and phishing triggered 16%4. Traditional spam filtering techniques are typically not capable of keeping up with evolving techniques followed by spammers and need the creation and implementation of more intelligent and adaptive techniques. Text classification, a fundamental problem in natural language processing (NLP), is a potential solution to automatically identify and block spam on various platforms, such as email, social media, and online reviews1.

Existing survey articles on text classification have a tendency to have big groups of algorithms, which is not precise. As a solution to this limitation, this paper suggests a high-resolution and fine-grained taxonomy particularly tailored to research areas of text classification. The taxonomy is organized in hierarchical stages: research field-based category, research field-based sub-category, methodology technique, methodology sub-technique, and research field applications. With a two-way direction evaluation framework—empirical and experimental—a high-resolution and fine-grained characterization of text classification algorithms and applications is provided. The framework helps to categorize techniques better and precisely, which helps researchers to make more wise decisions based on accurate, field-specific information1.

The primary contributions of the paper are listed as follows:

•An innovative taxonomy for text classification model research areas1.

•An elaborate explanation of the traditional machine learning and deep learning approaches to spam filtering1.

•Empirical comparison of state-of-the-art language models, such as GPT-4, BERT, and RoBERTa, for spam email classification2.

•Discussion of challenges and opportunities in detection of AI-generated false content and integration of XAI as a means for increasing transparency and trustworthiness2.

II. Contextual Framework and Relevant Literature

A. Spam Detection Methods

Spam filtering has become more sophisticated, using different techniques to detect and filter spam emails2. A broad categorization of anti-spam methods classifies these methods into two types: static and dynamic4. Rule-based and blacklists in conventional methods can be easily bypassed by spammers4. Pre-configured white listing and blacklisting filtering methods are not effective as spammers use newly infected machines as relays, changing IPs and email addresses4. Machine learning methods like Naïve Bayes, Support Vector Machines (SVM), and K-Nearest Neighbours (KNN) have been promising to learn automatically from labelled data for spam detection2. Dynamic methods analyze the message content using text modelling methods learned with statistical or machine-learning approaches to detect an email as spam4. Deep learning methods like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been observed to perform well in learning relationships in text data for better spam detection2.

B. Text Classification Algorithms

Text classification models must sort and categorize humongous quantities of text data1. They are usually classified into traditional machine learning methods and deep learning methods.

1.Traditional Machine Learning Methods: Naïve Bayes2., using Bayes' theorem with strong independence assumptions among features; SVM2., which identifies the maximum-margin separating hyper plane between classes; and KNN2., classifying a case based on the most common class among the k-nearest neighbors. Kim and Lee14 proposed a novel attribute weighting method to improve Naive Bayesian Classification performance using non-linear optimization to acquire attribute weights. Yu et al.19 applied text categorization of KNN using K-Means for class imbalanced issues.

2.Deep Learning Methods: These include CNNs2, which learn local text features through Convolutional layers; RNNs2, which handle sequential data by retaining a hidden state with some memory of what has been fed so far; and Transformers2., which use attention mechanisms to weigh relative importance of various parts of the input sequence. To illustrate, the Back Propagation Lion (BP Lion) algorithm is used to enhance neuron weight updates21. Dong et al.22 used pre-trained BERT models used for text feature extraction, word, and label embeddings with a Self-Interaction Attention Mechanism23.

C. Hybrid Methodologies

Machine learning and met heuristics are currently combined in research to improve the detection of spam using diverse strategies2. Adaptive algorithms and hybrid models are being created for maximum efficiency. Ensemble learning, in which multiple models are merged for maximum overall performance, is also applied in spam detection2. Bakunin et al., for example, introduce a novel system for spam email detection using hybridization of machine learning models and a sine cosine swarm intelligence algorithm optimized to escape the limitations of the current approaches24. This entails assuming a novel sine cosine for training logistic regression and tuning XGBoost models as parts of the hybrid machine learning-met heuristics model24.

III. Taxonomy of Text Classification Based on Research Field

To facilitate a more systematic and sophisticated description of text classification algorithms, the present work suggests a new taxonomy based on different research domains. The proposed taxonomy is organized through hierarchical layers, as shown in Figure 1:

•Field-Based Category Research

•Research Field-Based Sub-Category

•Methodological Framework

•Methodological Sub-Technique

•Research Field Applications

This top-down structure makes possible increasingly finer classification of methodologies, thereby enabling researchers to make informed decision-making grounded in true, accurate, precise observations related to their domain1.

IV. Methodology

A. Data Preparation and Collection

Two data sets were used to compare a number of text classification algorithms for spam filtering: the "NLP—SPAM/HAM Email collection"2. The "Spam Emails" data set is posted on Kaggle with a usability of 10 and 5572 unique rows and Message and Category labels25. The clean email has 86.59% in the original data set and spam email 13.41%25. The longest message row has 888 characters with an average of 7525. Partitioning, cleaning, and pre-processing the data set is the first operation.

B. Algorithm Training and Deployment

The algorithms selected for application in this research were executed under Python 3.6 and Tensor Flow 2.10.0. Where the existing literature did not provide publicly available code, the authors wrote their own code with the help of Tensor Flow. The models were optimized via the Adam optimizer. For example, Roumeliotis et al.2 tackled the problem of spam detection and classification by employing four varied models: GPT-4, BERT, RoBERTa, and CNN. Each model was fine-tuned on separate datasets, followed by validation procedures that made use of unique validation sets. Models were trained to predict on the test sets to label email messages as spam or not spam, with the outcomes stored in unique CSV files for analysis26.

C. Assessment Criteria

The performance of the algorithms was tested on a range of different measures, including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC)2. These are all global measures of the performance of the algorithms at labeling correctly as spam and not making false positives or false negatives. In addition, Bacanin et al.24 used accuracy,

Precision, recall, and F1 score, and some other classification metrics are relevant to assess the performance of their hybrid method.

D. Swarm Intelligence Optimization

Bacanin et al.24 used a diversity-oriented Sine Cosine Algorithm (DOSCA) in two distinct machine learning configurations for spam classification. DOSCA was used particularly to tune the weights and bias of the hidden neurons of the XGBoost model and to aid in training the logistic regression (LR) model. The algorithm is designed to overcome the shortcomings experienced in conventional SCA implementations to improve exploration and exploitation27. The method combines machine learning algorithms with a refined sine cosine swarm intelligence to alleviate the shortcomings of current methods24.

V. Experimental Test and Evaluation Results

A. Comparison of Performance of Text Classifying Methods

Experimentation outcomes confirm the validity of different models of text classifying in spam2 identification. Use of deep learning processes like CNNs and RNNs is superior to traditional machine learning algorithms like Naïve Bayes and SVM2. Performance will, however, be experiment setup and dataset dependent2. Chandan et al.10, for example, compared BERT with the traditional machine learning models on a Kaggle dataset, and the outcome confirmed that BERT achieved a highest test accuracy of 98% that was superior to Logistic Regression, multinomial Naïve Bayes, SVM, and Random Forest algorithms.

B. The Refining of Advanced Language Models

GPT-4, BERT, and Roberta fine-tuning for spam filtering is promising2. The models can identify implicit relationships in text data and are thus best utilized to counter sophisticated spam and phishing attacks2. Roumeliotis et al.3 described how fine-tuned models could be run in the background as intermediaries between email servers and users and could label emails in milliseconds.

C. Opportunities and Challenges of Detecting AI-Generated Deceptive Content

One of the biggest challenges to spam filtering2 is growing AI-based content complexity. There must be strong AI detector models created that can identify electronically created content and distinguish it from human-created content2. Federated learning, differential privacy, and secure multi-party computation are a few of the techniques that can be employed to solve the problem of privacy2. Techniques such as XAI can be employed to provide transparency to the decision-making process of such models, thereby imparting them with trust and transparency2.

D. Results Achieved from Swarm Intelligence-Optimized Models

Bacanin et al.'s work24 illustrated the potential of Logistic Regression (LR) and XGBoost models trained on Diversity-Oriented Sine Cosine Algorithm (DOSCA) with English and Turkish databases.

Logistic Regression (LR): LR-DOSCA model trained with DOSCA was extremely accurate and useful for Turkish and English datasets28. For English dataset of 500 features, LR-DOSCA achieved accuracy of 96.9977%, precision of 0.969976, recall of 0.969977, and F1 score of 0.96982628. Accuracy was 97.4596%, precision was 0.974643, recall was 0.974596, and F1 score was 0.97446829 for 1000 features.

•XGBoost: The XGB-DOSCA also performed similarly30. For English dataset with 500 features, the optimal error rate was 0.01385730. For 1000 features, the optimal error rate was 0.01385731. For Turkish dataset with 1000 features, the optimal error rate was 0.06666732.

The research establishes the effectiveness of DOSCA in optimizing machine learning models for spam filtering with superior performance measures compared to other met heuristic algorithms24.

VI. Conclusion

This article proposed a massive taxonomy exclusively for text classification in research areas1. Experimental findings indicate the performance of various text classification models in spam filtering, and deep learning models and fine-tuned language models worked effectively2. Conquering the challenges and potential of AI-generated manipulative content detection is crucial to maintaining trust and authenticity in online communication2.

Future research involves developing privacy-protecting algorithms, developing ethical standards for the application of text classification technology, and researching the application of adversarial machine learning for the improvement of the resilience of spam filters2. Swarm intelligence and machine learning are future research areas, particularly to improve the flexibility and effectiveness of the spam filtering system more24. Future tests should continue to test on large real-world datasets with the hope of building confidence in the models before developing to apply them in real-world systems involved in spam detection and Internet security, and in other networks using email services33.