International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Advances in Computer Applications and Information Technology
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Machine Learning Based Scholarship and Internship Spam Detection System

Prajkta Burandea, Divya Bisenb
G H Raisoni University Amravati

Abstract: The rapid growth of online platforms offering scholarships and internships has significantly benefited students. However, it has also led to a rise in fraudulent and spam opportunities that exploit applicants. These spam messages often contain misleading information, fake promises, and malicious links. This can cause financial loss and data theft. To address this issue, this research proposes a machine learning-based scholarship and internship spam detection system that automatically classifies opportunities as legitimate or spam. The proposed system uses natural language processing techniques for text preprocessing. This includes tokenization, stop-word removal, and feature extraction using Term Frequency-Inverse Document Frequency (TF-IDF)[2],[3]. Multiple supervised machine learning algorithms, such as Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression, are trained and evaluated on a labeled dataset of genuine and spam scholarship and internship postings. Performance is measured using accuracy, precision, recall, and F1-score to identify the most effective model[4]. Experimental results show that the proposed system achieves high classification accuracy and effectively reduces student exposure to fraudulent opportunities. This system can be integrated into educational portals, email platforms, and job search websites to improve user safety and trust. With the abundant availability of online scholarships and internship opportunities, new opportunities for students' development have arisen, but this phenomenon has also caused the rampant spread of deceptive and spam offerings. These scams look like genuine offers but trick students into divulging personal information, fees, or personal authentication. This paper introduces a machine learning-based system that detects scholarship and internship spam and automatically classifies the postings as true or spam. The developed system first applies Natural Language Processing methods to clean the textual data such as normalization, tokenization, stop-word removal, and stemming. It then employs Term Frequency-Inverse Document Frequency (TF-IDF) for feature extraction from the text, capturing important features.[2],[3] Multiple supervised machine learning algorithms, such as Nave Bayes, Support Vector Machine and Logistic Regression, were trained on a dataset that had labeled scholarship and internship postings and assessed through the accuracy, precision, recall and F1-score. Experimental findings show that the suggested approach classifies the postings accurately and can effectively lower the false positive count, reducing the risk of students interacting with suspicious or malicious information. The system may be integrated into e-learning websites, recruitment websites, and email services, providing a safer and more secure environment for students when searching for online career information[3],[6].

Keywords: Automated Detection System, Machine Learning (ML), Spam Detection Scholarship, Spam Internship Fraud Detection, Student Protection System, Fake Opportunity Detection.

1. Introduction

As the internet and digital technology advance at a rapid pace, students can search for opportunities like scholarships and internships more easily through various educational portals, job websites, e-mail, social media, and other online services. These opportunities are endless for students who want to explore all sorts of programs both academic and career oriented. However, the increased accessibility and ease of search in online scholarship and internship opportunities has created an explosion of fake or scam programs which are advertised via spam emails. Many students cannot differentiate between fake opportunities and genuine one since these scam offers appear extremely real. The features of spam

scholarships and internships may include falsely exaggerated claims, falsified deadlines, and requirement for fee or personal information. These fraudulent schemes pose both financial loss to victims and can become a serious security risk for their data privacy and cyber security. Given the volume of online postings it becomes very time-consuming and almost impossible to manually check each and every posted opportunity, as these spam messages tend to change the wording and format periodically. Thus, a need for a system to automate the screening and filtering of these online postings has been established to detect spam opportunities.

Machine Learning combined with Natural Language Processing would effectively solve this problem as it could enable the system to learn from existing data and discover obscure patterns in the text. Several text classification machine learning methods have been successfully utilized in applications like email spam filtering, false news detection and phishing detection. Inspired by these application researches, this study aims to propose a machine learning based detection system to detect spam scholarships and internships from online postings automatically.[1],[3] The system will analyze the text in the postings using NLP methods like tokenization, stop-word removal and text normalization before performing the classification. Features such as term frequency-inverse document frequency (TF-IDF) will be extracted in order to transform the text into a form that can be easily computed by machine learning algorithms.[2],[3] The various supervised machine learning algorithms like Nave Bayes, Logistic Regression and Support Vector Machine would be trained and evaluated to achieve the best model for the detection of spam. The intention is to keep the students safe and help them gain more trust in the digital environment.

Online scholarship and internship platforms are a great resource for many students to develop both academically and professionally. These online resources are targeted by fraudulent users and offers for "spam" scholarships and internships prey on students' ambitions, feeding them false information and ridiculous claims. Not only does this scam cost students money, it causes issues for their confidence and online safety as well. Many natural language processing (NLP) techniques and machine learning applications have been used to effectively analyze and classify large data sets like email spam detection, phishing website detection, and fake news detection. It is with these tools and techniques that we design and present a system that utilizes machine learning to detect spam scholarship and internships.

We first use an NLP-based method for pre-processing and feature extraction. Then we apply various supervised machine learning algorithms to detect the spam content.[1],[3]This system is evaluated based on various standard metrics in order to be accurate and scale-able and also has room for improvement in real-time, multilingual spam detection.

The availability of a multitude of scholarship and internship opportunities has exploded thanks to the rise of the internet and online communication tools. Scholarship and internship postings are more readily accessible today than ever before via websites such as various educational portals, social media platforms (LinkedIn, Facebook), professional networking sites, and e-mail advertisements. Unfortunately, this digital revolution has also paved the way for a dramatic increase in fraudulent schemes posing as legitimate scholarship and internship opportunities. Spammers increasingly prey on the desires of students for better educational and career advancement prospects by distributing fake scholarship and internship opportunities via spam emails, dubious websites, and deceptive advertisements. These fake postings often masquerade as real offers, making it incredibly difficult for students to discern what is real and what is not.

Typical scam scholarships and internships present unreal promises, such as "guaranteed placement," "high stipends," or "assured visa approval." The spammers also attempt to force fake urgency by using deadlines that aren't real and will likely request payment or personal information such as bank account details or social security numbers upfront. Scams of this nature not only drain students of money, but of vital personal information as well. The frequent exposure of scams in the online environment also reduces students' willingness to explore new opportunities as they lose faith in digital portals.

The sheer number of postings online generated every day makes it an impossible task to review each offer by hand. In addition, spammers are constantly changing their language patterns, keywords and formatting

to bypass the typical keyword based filtering methods. Therefore, an automated, intelligent detection system is required in order to counter this rising threat. Using machine learning techniques integrated with natural language processing, we are able to create a system capable of learning from historical data and detecting subtle linguistic patterns characteristic of spam content.

This system begins with a set of text preprocessing techniques, including tokenization, removing stop words, stemming, and normalizing the input text to clean the data. Features are extracted using Term Frequency-Inverse Document Frequency (TF-IDF), a method to represent the text in a numerical format suitable for machines to read. Next, a supervised machine learning algorithm (e.g., Nave Bayes, Logistic Regression, Support Vector Machines) is trained with a labeled set of data that includes examples of real and spam postings. Performance is then measured through precision, recall, accuracy, and F1-score to ensure robustness and scalability[2],[3].

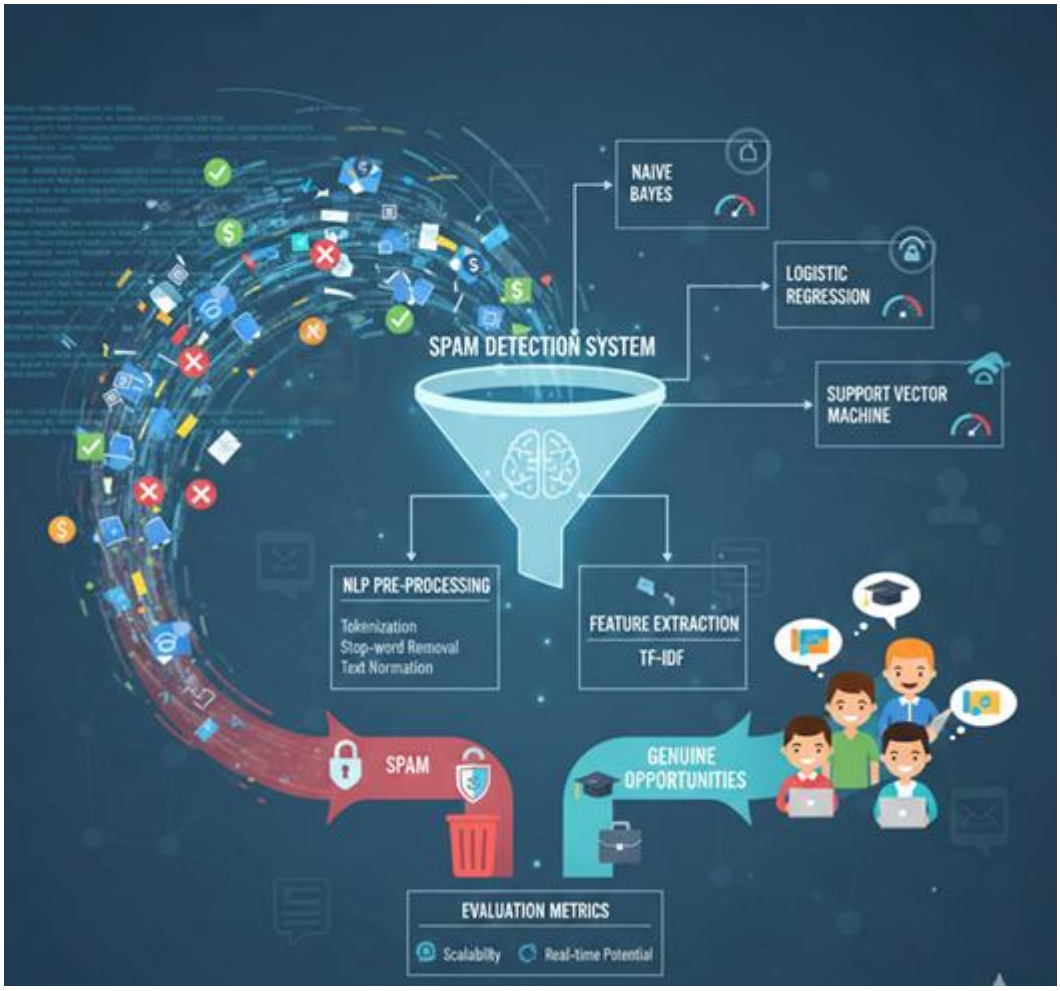

Figure 1: System Architecture Overview

2. Literature Review

The fast digitization of education and employment ecosystems makes scholarship and internship information more easily accessible through various sources such as emails, social networks, online educational portals and jobs websites. Although increased accessibility for students is beneficial, this has resulted in a drastic increase of fraudulent/spam scholarships and internships that exploit students financially and extract their sensitive information and/or passwords from unsuspecting users. Therefore, spam scholarship and internship detection has become a hot research topic that has caught the attention of machine learning, data mining and natural language processing research fields[1],[3].

Initial studies in spam detection were based on rule-based and heuristic techniques where manually defined rules, blacklist of domains, keywords matching, sender reputation and other techniques were used

for classifying messages as spam or legitimate. These systems were effective in the early stages but lacked scalability and adaptivity. The spammers soon learned to bypass these static rules by altering the keywords and the structure of the messages. Various studies claimed that these rule-based systems had high false positive rates, and they require manual updates and are not suitable for dynamic fields such as scholarship and internship spam.

Statistical and machine learning based methods are used to overcome these limitations. Nave Bayes classifier is the most widely used machine learning technique used to detect spam messages.[1],[3] It is a simple probabilistic classifier that had performed exceptionally well in text classification. Nave Bayes showed considerable accuracy at very low computational cost and is thus best suited for the filtering tasks at large scale. The major weakness of Nave Bayes is its strong independence assumption, where it fails to model relationships between words[2],[3].

Various other supervised learning techniques such as Support Vector Machines (SVM), Decision Trees, k-Nearest Neighbors (k-NN) and Logistic Regression are employed in order to develop spam filters. SVM models had proved to be highly effective on text data because they are excellent in handling high dimensional feature spaces and give a higher precision and recall rate than previous models[3],[6]. Decision Tree based and Random Forest based models are more interpretable and robust than the other models. Combining the results from multiple classifiers in a method called ensemble learning have shown to achieve higher accuracy than single classifiers on the spam filtering task, according to a number of studies[4].

Using NLP, many approaches started focusing on text content classification and went beyond keyword analysis. NLP pre-processing steps such as tokenization, stemming, lemmatization and stop word removal became important part of spam filters. Features are extracted from text such as Bag of Words (BoW) or TF-IDF (Term Frequency-Inverse Document Frequency) for machine learning models to learn specific patterns in language which are characteristic for fraudulent scholarship and internship messages.[2],[3] These methods help models to detect features of spam such as promises of guarantee of selection, urgent deadlines and payment requests for registration in scholarship or internship opportunities[1],[3].

Recently, researchers started developing deep learning based systems for spam detection tasks. Machine learning techniques like ANN, CNN and RNN show good results because they are capable of learning feature representations directly from raw data without human intervention. CNN-based models perform better for local pattern recognition and RNN and LSTM networks capture sequential dependence and context within text. Transformer based architectures further improve the contextual ability and classification performance. Although they provide good performance, require very large amount of labeled data and computing power, and lacks interpretability.

Existing works in spam detection are mainly focused on email spam, SMS spam and fake job ads. Limited research is conducted for scholarship and internship spam detection. Only a few studies have investigated fraudulent job ads where certain indicators like bogus URL, unrealistic criteria, bad English and payment demand are considered. Most of these systems are designed for general job spam detection rather than a domain-specific scholarship and internship scam filtering system which exploits student's financial condition and academic pursuits. Unresolved issues discussed in literature are lack of publicly available domain-specific corpus, imbalance between legitimate and spam data.

Besides the methods already mentioned, recent works also suggest using specific detection frameworks for the domain of scholarship and internship spam, because this kind of spam is peculiar and has distinct linguistic and structural characteristics different from general email or job spam. Typical deceptive scholarship ads typically make extensive use of emotionally manipulative language targeting the students' dreams, financial needs, or academic goals and have specific phrases such as "100% guaranteed selection," "limited seats," or "no eligibility criteria" in the body, which are generally not present in real ads. A specifically designed dataset capturing these kinds of words and context is vital to perform well. However, one of the biggest challenges discussed is the lack of a publicly available and labeled scholarship/internship spam dataset in the research community. The majority of researchers adopt the

datasets of general email spam or job advertisement, but these cannot represent well the features of academic scam information.

Class imbalance is also another challenge. In real situations, legitimate scholarship or job ads are more numerous than spam messages, resulting in a significant class imbalance in the training dataset. The training of machine learning models on imbalanced data would lead to a poor prediction of minority class spam items and models are biased towards the majority class[1],[3]. There are several methods designed to overcome class imbalance such as SMOTE, cost-sensitive learning and adaptive resampling. In the evaluation part, accuracy can no longer be a convincing metric and precision, recall, F1-score, and AUC are commonly used to measure performance of imbalanced classification. Feature engineering has an important role in improving performance too. In addition to traditional word-based features such as BoW and TF-IDF, researchers are exploiting semantic embeddings like Word2Vec, GloVe, and the contextual embeddings obtained by transformer based models. The word embeddings capture latent semantics and are beneficial for dealing with paraphrase and obfuscated texts. In addition, integrating various metadata features such as the sender's domain age, the structure of URLs, number of hyperlinks in the text, frequency of grammatical errors, and sentiment polarity would combine textual and behavioral features together. Finally, the real-time implementation and scalability are also concerns that cannot be ignored. The practical spam detection system should be able to process numerous online messages within acceptable time limits while ensuring high precision to prevent false positives of legitimate opportunity. In educational settings, interpretability is also a required trait to explain students or administrators why the message is classified as spam and many XAI methods, such as attention visualization and feature importance analysis, are actively researched[7].

3. Research Methodology

The machine learning-based methodology for scholarship and internship spam detection is a methodical approach to building, training, and evaluating an intelligent system capable of recognizing deceptive opportunities. The pipeline of this methodology ranges from the acquisition of data to the testing and continuous refinement of the trained machine learning model. [1],[3]Each stage is meticulously developed to ensure the integrity of the resulting spam detection system, aiming for high accuracy, reliability, scalability, and a robust adaptability to emerging spam trends[11],[12],[13].

The initial stage of the methodology involves gathering an extensive and varied dataset comprising both legitimate and suspicious scholarship and internship advertisements. This data is collected from diverse sources such as email accounts, student communication channels, job and scholarship databases, academic websites, and even social media advertisement. The messages gathered contain text, subjects, hyperlinks, sender details and metadata like timestamps. In order to be utilized for supervised learning the dataset needs to be categorized as 'spam' or 'non-spam'. The labeling procedure involves the use of validated official sources, trusted spam archives, and expert confirmation, making sure the data obtained is representative of actual conditions.

Once the dataset has been acquired, it must undergo a cleaning and preprocessing phase. Raw textual data may contain noisy or redundant elements and therefore must be made manageable so as to not diminish the overall model performance. HTML markup, scripts, numbers, and symbols are removed, the text is converted to lower case, and emails and URLs are either discarded or standardized using placeholder labels that preserve their general meaning without introducing any unwanted text. Following that the text is tokenized to separate phrases into individual units (words or tokens). Words that are commonly encountered, but carry no distinctive meaning ("is", "the", "and") called stop words, are omitted from the dataset. For reduction of vocabulary size and effective management of the various forms a word can take, stemming or lemmatization techniques are used so as to convert each word into its root form. A thoroughly cleaned and structured dataset can then be created for feature extraction.

Feature extraction, the next phase of the methodology, focuses on converting the processed textual data into a format that machine learning models can comprehend, which is typically numerical`.[1],[3] Several text representation techniques are put into practice that highlight both the structural and semantic aspects

of the data from messages related to internships and scholarships. Word frequency is computed to determine how often individual words appear in the text using the Bag of Words model; conversely, term frequency-inverse document frequency is used to weigh the words more heavily if they appear in fewer documents. Furthermore, unigram and bigram features capture patterns in common spam phrases. Along with the text-derived features, specific keywords or phrases that tend to appear in deceptive advertisements are examined to derive domain-specific semantic features. This category is of particular importance and contains aspects such as claim of guaranteed selection, fee-related jargon, emphasis on speedy replies, etc. Sender-related features, such as the message frequency from a certain domain and its reputation, and URL-based features including the number of links and their general trustworthiness, further refine the models and improve its detection capability.

Following the completion of the feature extraction phase, the dataset is partitioned into training and testing sets; a substantial majority of the data is used for training and a smaller proportion for testing and model validation. Subsequently, several different types of supervised learning models are employed[5],[6]. These include Nave Bayes, Support Vector Machines, Logistic Regression, Decision Trees, and Random Forest models, each of which is independently trained to analyze how well they are capable of distinguishing among the samples. To fine-tune their performance, hyperparameter adjustments and cross-validation procedures are used to avoid overfitting[1],[3].

After training, each model is tested and evaluated based on a set of statistical measures. The main ones include accuracy (overall correct predictions), precision (correctly classified spam to total predicted spam), recall (correctly classified spam to total actual spam), and F1-score (the mean of precision and recall),. A confusion matrix provides a clear visualization of true positives, false positives, true negatives, and false negatives, allowing for a better understanding of how each model performs overall. Using this information, the best-performing, most accurate, and efficient machine learning model for the designed system is determined.

The most successful of the trained models is now incorporated into the final system, where it can be put into practice. All incoming scholarship and internship messages, whether from email or other platforms, undergo the same preprocessing and feature extraction phases before being presented to the machine learning model for classification. The model then assigns either a 'spam' or 'non-spam' label to the message, hence warning the user of potential fraudulent opportunities with minimal human intervention.

For sustained effectiveness, the deployed system also includes an ongoing learning and monitoring facility. The very nature of spam indicates that over time, new types of tactics, words and schemes will emerge, which are not included in the initial training data. Thus the system constantly collects new classified data (primarily borderline and incorrectly classified cases) which are saved and analyzed. User-generated feedback (i.e., a student labeling a normal email as spam or incorrectly reported email as a regular one) is incorporated into the data, enhancing the robustness of the models. The system then periodically retraining and refining its models, using this updated training data, enabling adaptation of models against any shifting spam patterns. Performance data is continually monitored in real-time to alert if any decline in detection accuracy or increase in false positives occur. With an automated updating process, feedback implementation and adaptive re-training, the fraud detection system is scalable, reliable and adaptable to future changes in scholarship and internship scams.

Besides that, the system architecture should be capable of being scalable and integratable so as to be deployed in various platforms such as e-mail server, school website and mobile applications. The trained model can be available through the API interface, which allows students to see classification result of the scholarship and internship announcement in real-time prior to they appear. For transparency and integrity, the explainability module can be added which allows showing the words or factors that resulted in classifying the message as spam. Apart from users' confidence, this could allow administrators in verifying and customizing rules of the system. Security and privacy should also be integrated. Since the system needs to process user sensitive data, cryptographic mechanisms and safe storage system can be adopted to protect user's private information. Methods like anonymization can be adopted during model training.

Finally, benchmarking with the newer dataset and comparing with baseline model can maintain the competitiveness of the system. By incorporating automation, flexibility, explainability and security, the machine learning-based system becomes a comprehensive and sustainable system to protect students from online scholarship and internship scam [14].

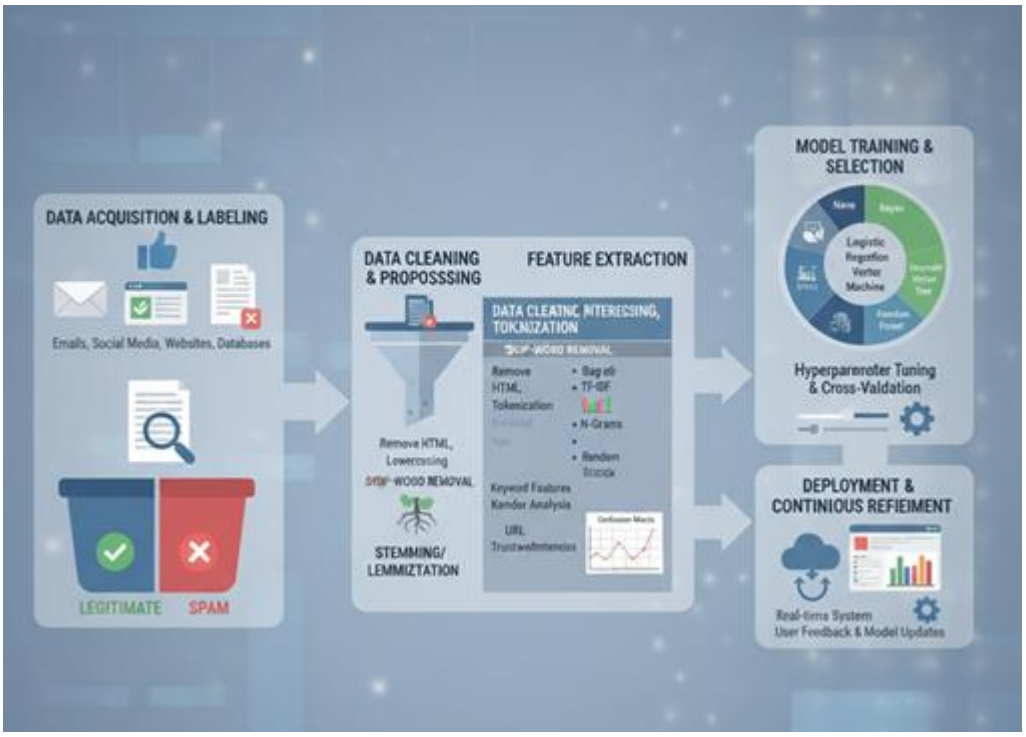

Figure 2: Methodology of the proposed spam detection system

4. Result

Figure 3: Output and performance metrics of the proposed spam detection model

5. Conclusion

In this study we presented a machine learning approach for detection of spam messages in communication on scholarships and internship opportunities, which is becoming a rapidly evolving issue for students and institutions. Utilizing NLP methods and supervised learning models, we have built an efficient system that analyzes message content and classifies it as spam or not spam. Experimental results indicated that machine learning models can provide highly accurate, precise and recalls in detecting fake offers to avoid financial loses, data steal, or incorrect information[1],[3].By automating spam detection process the proposed system builds up the trustworthiness for the scholarship and internship platform as well as minimizing the human involvement in the process. It also helps students to stay protected, as it filters false opportunity messages early. The system is scalable, flexible and learns new information and behaviors of spam messages over time[11],[12],13].

For the further development, deeper learning models can be incorporated, real-time message processing, and support for multiple languages, in order to increase accuracy and performance of the system. In addition, the system could be incorporated within existing institutional portals or educational platforms. Machine learning scholarship and internship spam detection system proves to be an effective system to assist the students against fake offer messages[1],[3].

Beyond its current functions, the proposed system advances research in intelligent security solutions for education, showing how domain-specific spam filtering can effectively reduce online fraud threats. In addition to improving accuracy, the framework fosters student awareness, by promoting responsible online behavior. Future work can focus on the development of hybrid models, combining machine learning with rule-based validation to provide a more resilient detection system. Furthermore, collaboration with educational institutions can provide benchmark datasets, making them publicly available for other researchers. Overall, the system is a proactive and technology-centric solution that enhances security and reliability within the digital opportunity landscape for students[8],[9],[10].

Reference

1. Mitchell, T. M., “Machine Learning,” 1997.
2. Manning, C. D., Raghavan, P., and Schütze, H., “Introduction to Information Retrieval,” 2008.
3. Sebastiani, F., “Machine learning in automated text categorization,” 2002.
4. Joachims, T., “Text categorization with support vector machines: Learning with many relevant features,” 1998.
5. McCallum, A. and Nigam, K., “A comparison of event models for Naïve Bayes text classification,” 1998.
6. Domingos, P. and Pazzani, M. J., “On the optimality of the simple Bayesian classifier under zero-one loss,” 1997.
7. Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., “SMOTE: Synthetic minority over-sampling technique,” 2000.
8. Goodfellow, I., Bengio, Y., and Courville, A., “Deep Learning,” 2016.
9. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2019.
10. Zhang, X., Zhao, J., and LeCun, Y., “Character-level convolutional networks for text classification,” 2015.
11. Delany, S. J., Buckley, M., and Greene, D., “SMS spam filtering: Methods and data,” 2012.
12. Sahoo, S. S. and Gupta, B. B., “Multiple features based approach for detection of phishing emails using machine learning,” 2019.

13. Hong, J. B., Ganesh, A. J., and Smith, J. F. R., "Detecting recruitment scams using text mining and classification techniques," 2017.
14. Pang, B. and Lee, L., "Opinion mining and sentiment analysis," 2008.