International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 01 | Jan 2025 www.irjet.net p-ISSN: 2395-0072

Enhancing Email Forensic Analysis with Deep Learning and Semantic Techniques: A Literature Review

K V Neha[1], Jinsu Anna John[2], Vanimol Sajan[3]

Dept. of Computer Science and Engineering
College of Engineering, Kallooppara, Thiruvalla

Abstract - *Email data analysis stands as a critical pillar within the realm of digital communication, holding significant importance in both cybersecurity and the comprehension of customer sentiment. In this paper, we embark on a comprehensive literature survey that delves into the intricate landscape of email data analysis, with a specific focus on exploring cutting-edge techniques and methodologies. Traditional approaches to email analysis often struggle to capture the subtle semantic nuances inherent in email exchanges. However, recent strides in deep learning, machine learning, and natural language processing have opened up promising avenues for addressing these challenges. Leveraging these advancements, we aim to enhance the efficacy and accuracy of email data analysis methodologies. This literature review involves an exhaustive examination of the latest methodologies in the field. Through this process, we meticulously scrutinize notable studies, each offering unique insights and advancements in email analysis. By distilling the essence of these studies, we gain a comprehensive understanding of the current techniques in email data analysis. A pivotal aspect of our survey is the comparative analysis of the reviewed methodologies. This comparative examination allows us to elucidate the distinguishing features and descriptions of each approach, providing valuable insights for practitioners and researchers alike. Furthermore, our synthesis of existing studies contributes to the ongoing discourse surrounding robust analysis techniques in digital communication. By highlighting the potential of advanced technologies, particularly in the realm of machine learning, we aim to underscore the significance of adopting these methodologies for efficient and accurate email data analysis. Looking ahead, we identify several promising future research directions. For instance, we advocate for the exploration of deep learning techniques for multiclass email classification, which holds the potential to further advance email forensic analysis capabilities and bolster cybersecurity defense mechanisms. As we navigate the ever-evolving landscape of digital communication, our paper seeks to pave the way for continued innovation and progress in the field of email data analysis.*

Key Words: LSTM, CNN, Semantics

1.INTRODUCTION

In today's rapidly evolving landscape of digital communication, the significance of analyzing and organizing email data cannot be overstated. This paper addresses the critical role email data plays across various sectors, including cybersecurity and understanding customer sentiment. With the continuous growth in the volume and complexity of digital data, there is an escalating need for advanced methodologies capable of effectively managing and interpreting textual information.

This study embarks on a comprehensive exploration of state of-the-art techniques and methodologies for scrutinizing email data, with a primary focus on improving efficiency, accuracy, and resilience through the integration of advanced technologies. The proliferation of digital communication platforms has introduced new challenges in forensic analysis, cybersecurity, and content filtering. Traditional approaches to email scrutiny often fall short, relying on basic keyword-based techniques that may overlook the nuanced semantic intricacies inherent in email exchanges. However, recent advancements in deep learning, machine learning, and natural language processing offer promising solutions to overcome these challenges.

Before finalizing the technological frameworks and algorithms for this study, an extensive literature review was conducted to identify the latest methodologies and their contributions to the field of email analysis. Numerous notable studies were examined, each providing unique insights and advancements. This literature review culminates in a comparative analysis of the reviewed methodologies, outlining their features and descriptions. It provides valuable insights for informed decision-making in email data analysis.

Through the synthesis of existing studies, this paper aims to contribute to the ongoing discourse on robust analysis techniques, highlighting the potential of advanced technologies in machine learning for efficient and accurate analysis of email data. By addressing the complexities and challenges in email analysis, this study seeks to advance the understanding and application of methodologies in this domain of digital communication.

2. LITERATURE SURVEY

An extensive review of numerous research papers and articles was carried out, focusing on identifying innovative solutions driven by advanced technologies and highlighting

the most efficient approaches. Presented below is a compilation of some of these noteworthy contributions.

2.1 SeFACED: SEMANTIC-BASED FORENSIC ANALYSIS AND CLASSIFICATION OF E-MAIL DATA USING DEEP LEARNING [3]

SeFACED (Semantic-Based Forensic Analysis and Classification of E-Mail Data Using Deep Learning) represents a cutting-edge approach in the realm of digital forensics, particularly in the context of email data analysis. This innovative framework leverages the power of deep learning to enhance the semantic understanding of email content, providing a more sophisticated and nuanced approach to forensic analysis. The key strength of SeFACED lies in its ability to extract and comprehend the semantic nuances within email communications, enabling a more refined classification of content. By harnessing deep learning techniques, SeFACED transcends traditional keyword-based methods, offering a more accurate and context-aware mechanism for forensic analysis.

The integration of deep learning algorithms facilitates the automatic identification of patterns, relationships, and anomalies within email data, empowering investigators with a more comprehensive understanding of the communication context. This not only enhances the efficiency of forensic analysis but also contributes to the overall accuracy of classification outcomes.

The potential applications of SeFACED are vast, ranging from investigating cybercrimes such as phishing, fraud, and identity theft, to monitoring and analyzing corporate communications for signs of data breaches or insider threats. With its ability to process large volumes of email data efficiently and accurately, SeFACED stands to significantly improve the speed and effectiveness of email forensics. Its deep learning-driven approach ensures that it can continuously adapt and improve as new patterns and tactics emerge in the evolving landscape of cyber threats. Furthermore, its ability to perform context-aware analysis makes it particularly valuable in scenarios where understanding the intent and underlying meaning of communication is critical for accurate forensic investigations.

Moreover, SeFACED's emphasis on semantic-based anal ysis aligns with the evolving nature of cyber threats and the increasing sophistication of malicious activities in the digital landscape. By addressing the limitations of conventional forensic approaches, SeFACED stands as a promising advancement in the field, offering a more robust solution for the identification, analysis, and classification of email data.

2.2 SPAM REVIEW DETECTION USING DEEP LEARNING[7]

The IEEE paper on 'Spam Review Detection Using Deep Learning' constitutes a significant contribution to the field of cybersecurity and online content analysis. online reviews have become the most important resource of customers' opinions. These reviews are used increasingly by individuals and organizations to make purchase and business decisions.

Focused on the pervasive issue of spam reviews, the paper leverages deep learning techniques to enhance the accuracy and efficiency of detection mechanisms. The key strength of this study lies in its adoption of deep learning, a powerful paradigm that excels in learning intricate patterns and representations from vast datasets. By applying deep learning models to the specific task of identifying spam reviews, the paper demonstrates a notable improvement in the precision and recall of detection algorithms compared to traditional methods.

The research showcases the adaptability of deep learning in tackling evolving challenges in online environments, where spam reviews often employ sophisticated tactics to evade detection. The integration of deep learning allows the system to autonomously learn and adapt to new patterns, making it a robust solution for real-time detection in dynamic online platforms.

Additionally, the paper discusses the potential implications of spam reviews on various domains, emphasizing the importance of accurate detection in maintaining the credibility of online platforms. The findings presented contribute valuable insights to the broader literature on spam detection, providing a nuanced understanding of the role deep learning can play in addressing contemporary challenges.

2.3 EMAIL SPAM DETECTION USING INTEGRATED APPROACH OF NAIVE BAYES AND PARTICLE SWARM OPTIMIZATION[1]

The paper presents a compelling and insightful exploration into the ever-evolving realm of email security, with a particular focus on the pressing issue of email spam. It introduces a sophisticated and innovative approach by integrating two distinct methodologies, creating a powerful synergy in the fight against spam. The fusion of Naive Bayes, a widely recognized probabilistic classification algorithm, with Particle Swarm Optimization (PSO), a potent optimization technique inspired by natural systems, marks a significant advancement in spam detection strategies. This novel combination not only aims to tackle the complexity of email spam but also seeks to improve the efficiency and accuracy of filtering unwanted messages in a highly dynamic environment.

The integration of Naive Bayes and PSO is a noteworthy and strategic move. By combining the probabilistic power of Naive Bayes with the optimization strengths of PSO, the paper introduces a robust method aimed at improving the accuracy and efficiency of spam detection systems. Naive Bayes, known for its simplicity and efficiency in classification tasks, particularly excels in dealing with high-dimensional data, which is characteristic of email content. On the other hand, PSO, with its ability to find optimal solutions through swarm intelligence, adds a layer of sophistication by enhancing the feature selection process. This synergy not only seeks to improve spam detection rates but also aims to make the system more adaptive and responsive to evolving spam tactics.

The experimental results presented in the paper provide strong evidence of the effectiveness of this integrated approach. The results demonstrate significant improvements in spam detection rates compared to traditional standalone methods, proving that the combination of Naive Bayes and PSO can offer superior performance in terms of both accuracy and efficiency. These findings are particularly promising in the context of spam detection, where new and increasingly deceptive methods are constantly being developed by spammers to circumvent conventional filters.

Furthermore, the paper delves into the interpretability of the combined Naive Bayes and PSO approach, offering valuable insights into how this integration contributes to a more transparent and understandable spam detection system. Interpretability is a crucial aspect of machine learning models, particularly in real-world applications where users and system administrators need to understand how decisions are made by the system. This paper highlights how the combination of Naive Bayes and PSO enhances the transparency of the model, making it easier for end-users to comprehend how spam is detected and why certain emails are flagged as spam. Such interpretability is vital for ensuring trust and confidence in the system, particularly in sensitive domains such as email security.

In conclusion, the paper offers a novel and practical solution to the ongoing challenge of email spam by integrating Naive Bayes and PSO in a way that enhances both detection accuracy and system transparency. The promising experimental results, coupled with the optimization of feature selection and the focus on interpretability, underscore the potential of this integrated approach.

2.4 SPAM FILTERING EMAIL CLASSIFICATION (SFECM) USING GAIN AND GRAPH MINING ALGORITHM[2]

The paper introduces a compelling strategy to tackle the ongoing challenge of effectively categorizing spam emails. It proposes an innovative approach that integrates a context based email classification model with a spam filter grounded in information gain calculation. This amalgamation aims to bolster the accuracy of email classification processes.

Commencing with an overview of the prevalent issue of misclassified emails, especially in the realm of spam filtering, the study emphasizes the urgency for a more robust classification framework. Leveraging established methodologies like the LingerIG spam filter, the authors advocate for a novel solution that encompasses several key stages: email pre-processing, feature extraction, and classification.

Through rigorous experimentation, the authors validate the effectiveness of their proposed solution. They achieve a re markable milestone of 100% accuracy in distinguishing spam emails, marking a notable advancement compared to prior methods. This success underscores the potential efficacy of their hybrid model in real-world applications.

Furthermore, the paper conducts insightful comparisons with alternative classification approaches, underscoring the superiority of the proposed solution in accurately identifying spam versus legitimate emails. By furnishing detailed accounts of implementation strategies and experimental findings, the authors provide valuable guidance for both researchers and practitioners seeking to refine email classification systems.

The ”SFECM” approach emerges as a promising avenue for enhancing the accuracy of spam filtering, addressing a critical requirement within the domain of email classification. The comprehensive nature of the proposed methodology, underpinned by empirical evidence, highlights its significance in propelling advancements in email classification practices.

2.5 SOCIAL ISSUES SENTIMENT ANALYSIS USING PYTHON [8]

The paper provides a concise and effective overview of its methodology, emphasizing the application of Python for sentiment analysis. The strength of this research lies in its practical approach, utilizing Python, a widely adopted programming language, to perform sentiment analysis on social issues. The paper showcases the accessibility and applicability of Python in processing and analyzing large datasets, making it a valuable resource for both researchers and practitioners in the field of social sentiment analysis.

The paper appropriately highlights the significance of sentiment analysis in understanding public perspectives on social issues, recognizing the impact of sentiment on decision making processes and public discourse. By addressing social issues, the research contributes to the broader discourse on the intersection of technology and societal concerns.

Moreover, the paper's focus on sentiment analysis aligns with the growing importance of leveraging computational methods to extract meaningful insights from social media and other online platforms. The practical implementation of sentiment analysis using Python demonstrated in the paper serves as a useful guide for those seeking to apply similar methodologies in their research or applications.

2.6 COMPREHENSIVE REVIEW ON EMAIL SPAM CLASSIFICATION USING MACHINE LEARNING ALGORITHMS[4]

The paper offers an in-depth exploration of the field of email spam classification through machine learning. They eloquently outline the urgent necessity for effective email management solutions due to the escalating volume of spam messages inundating users globally. The paper meticulously scrutinizes a variety of machine learning algorithms utilized in spam detection, providing valuable insights into their effectiveness and suitability.

A notable strength of this review is its comprehensive analysis of supervised, unsupervised, and semi-supervised machine learning methodologies. Through a detailed examination of each approach's intricacies and performance metrics, the authors offer invaluable guidance for practitioners and researchers alike. Particularly noteworthy is the focus on supervised learning, which emerges as the predominant choice due to its consistently high accuracy rates, substantiated by thorough statistical analysis.

Furthermore, the paper delves into specific algorithms employed, such as Artificial Neural Networks (ANN), Naive Bayes, Support Vector Machine (SVM), and Decision Trees, providing a nuanced understanding of their capabilities and limitations. This thorough exploration enables readers to discern the most appropriate algorithmic framework for their spam classification. A significant contribution of the paper is its identification of future research avenues and challenges. The authors aptly emphasize the need for real-time spam classification systems, dynamic feature updates, and reduction of system processing time. Additionally, they advocate for diversifying email feature analysis beyond traditional methods like Bag of Words (BoW) and email body text, underscoring the importance of addressing evolving spamming techniques.

2.7 EMAIL SPAM DETECTION USING BIDIRECTIONAL LONG SHORT TERM MEMORY WITH CONVOLUTIONAL NEURAL NETWORK[5]

This approach integrates with a method for email spam detection, utilizing sentiment analysis of email content Word Embeddings and Bidirectional Long Short Term Memory (Bi LSTM) networks, complemented by Convolutional Neural Network (CNN) techniques to expedite training and enhance feature extraction. Performance evaluation of their model incorporates precision, recall, and F-score metrics, utilizing datasets such as lingspam and spam text message classification datasets. Notably, their model achieves an exceptional accuracy rate of approximately 98-99% surpassing benchmarks set by prevalent machine learning classifiers and cutting-edge methodologies.

This paper introduces a fresh perspective on spam detection by amalgamating sentiment analysis with sophisticated neural network architectures. Through comprehensive performance evaluations, the authors demonstrate the superior accuracy of their model compared to conventional classifiers and existing methodologies. Leveraging datasets like lingspam and spam text message classification ensures the robustness and adaptability of their proposed approach. Their integration of sentiment analysis, CNN, and Bi-LSTM networks presents a promising avenue for bolstering email security and enhancing user experience. Through meticulous comparisons with established methods, the authors substantiate the effectiveness and superiority of their model in accurately identifying spam emails

3. COMPARISON

An extensive review of numerous research papers and articles was carried out , focusing on identifying innovative solutions driven by advanced technologies and highlighting the most efficient approaches. Presented below is a compilation of some of these noteworthy contributions.

Sr.no	Studies	Mining Technique used	Feature Selection	Accuracy	Precision	Recall	F1
1	Maryam Hina ,Mohsin Ali, Abdul Rehman Javed, Fahad Ghabban , Liaqat Ali Khan, Zunera Jalil (2021)	Long Short-Term Memory (LSTM) based Gated Recurrent Neural Network (GRU)	TF-IDF(Term Frequency Inverse Document Frequency), Word2Vec,and word embedding	95.0%	95.0%	95.1%	95.1%
2	G.M. Shahariar, Swapnil Biswas, Faiza Omar, Faisal Muhammad Shah,Samiha Binte Hassan(2019)	Text mining, Specifically, Natural language processing (NLP)	TFIDF, n-grams and Word Embeddings (Word2Vec) techniques	—	—	—	—
3	Kriti Agarwal, Tarun Kumar(2019)	Naive Bayes algorithm	correlation-based feature selection(CFS)	95.50%	96.42%	94.50%	95.45 %
4	M.K.Chae, Abeer Alsadoon, P.W.C.Prasad, A.Elchouemi (2017)	Graph Mining Algorithm	Information gain calculation	100%	—	—	—
5	Chhinder Kaur, Anand Sharma(2020)	Data mining techniques such as data preprocessing, cleaning, and extraction were employed, but specific mining algorithms or methodologies beyond data collection are not discussed.	Not mentioned any specific feature selection technique. It primarily focuses on data collection, preprocessing, sentiment analysis, and model building using machine learning algorithms.	—	—	—	—
6	Mansoor RAZA, Nathali Dilshani Jayasinghe, Muhana Magboul Ali Muslam (2021)	Artificial Neural Networks (ANN), Naive Based Machine Learning Algorithm, Support Vector Machine(SVM),Decision Tree (DT), K-nearest Neighbour machine learning algorithm(KNN)	Information Gain(IG)and Chi-square(^{2})	96.42%	98.13%	98.38%	98.25%
7	Sefat E Rahman, Shofi Ul-lah(2020)	Bidirectional Long Short Term Memory(Bi-LSTM) and Convolutional Neural Network(CNN)	Word Embeddings, Bidirectional Long Short-Term Memory (Bi-LSTM) network, and Convolutional Neural Network(CNN)	98%	98%	98%	98%

Table -1: Comparison Table

4. CONCLUSIONS

In conclusion, this extensive literature survey has delved into the realm of robust analysis through semantic modeling, with a specific focus on the powerful combination of RoBERTa-LSTM and integrated sentiment analysis. The exploration of existing studies has revealed the evolution of semantic mode ling techniques and the integration of RoBERTa-LSTM, show casing their potential in enhancing the robustness of data analysis. The synthesis of semantic modeling techniques provides a nuanced understanding of textual data, enabling a more comprehensive and context-aware analysis. The integration of RoBERTa-LSTM, leveraging pre-trained models and long short-term memory networks, contributes to the model's ability to capture intricate patterns and dependencies in the data, further improving the accuracy and robustness of analysis. Additionally, the incorporation of integrated sentiment analysis adds a layer of depth to the study, allowing for the extraction of emotional nuances from the analyzed content. This nuanced sentiment analysis, when combined with semantic modeling, enhances the capacity to derive meaningful insights from diverse sources of textual data.

As we navigate the ever-evolving landscape of natural language processing and data analysis, the amalgamation of RoBERTa-LSTM and integrated sentiment analysis emerges as a promising approach. The survey highlights not only the strengths and potential of these methodologies but also acknowledges the challenges, such as the need for large and diverse data sets and continuous refinement of model hyper parameters. In conclusion, this literature survey underscores the significance of adopting a holistic approach to robust analysis, acknowledging the synergies between semantic modeling, RoBERTa-LSTM, and integrated sentiment analysis. This research direction holds promise for applications in various domains, offering a sophisticated means to extract meaningful insights from textual data with heightened accuracy and robustness.

5. FUTURE SCOPE

The exploration of deep learning techniques for multiclass email classification presents an opportunity to enhance email content analysis in addressing cyber threats. Moving forward, there's room for further refinement and optimization of model architectures to boost classification accuracy. Additionally, efforts to expand the availability of diverse email datasets, particularly those focusing on criminal activities, will be essential for strengthening cybersecurity research. Advanced sentiment analysis methods could also be integrated to provide a more nuanced understanding of email content. Overall, ongoing research and development in this field promise to advance email forensic analysis capabilities and improve defense mechanisms against cyber threats.[6] [9]

REFERENCES

[1] Kriti Agarwal and Tarun Kumar. Email spam detection using integrated approach of na¨ ıve bayes and particle swarm optimization. In 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), pages 685–690. IEEE, 2018.

[2] MK Chae, Abeer Alsadoon, PWC Prasad, and Sasikumaran Sreedharan. Spam filtering email classification (sfecm) using gain and graph mining algorithm. In 2017 2nd International Conference on Anti-Cyber Crimes (ICACC), pages 217–222. IEEE, 2017.

[3] Maryam Hina, Mohsin Ali, Abdul Rehman Javed, Fahad Ghabban, Li aqat Ali Khan, and Zunera Jalil. Sefaced: Semantic-based forensic analysis and classification of e-mail data using deep learning. IEEE Access, 9:98398–98411, 2021.

[4] RAZA Mansoor, Nathali Dilshani Jayasinghe, and Muhana Magboul Ali Muslam. A comprehensive review on email spam classification using ma chine learning algorithms. In 2021 International Conference on Information Networking (ICOIN), pages 327–332. IEEE, 2021.

[5] Sefat E Rahman and Shofi Ullah. Email spam detection using bidirectional long short term memory with convolutional neural network. In 2020 IEEE Region 10 Symposium (TENSYMP), pages 1307–1311. IEEE, 2020.

[6] Mallikka Rajalingam. Text Segmentation and Recognition for Enhanced Image Spam Detection: An Integrated Approach. Springer Nature, 2020.

[7] GMShahariar, Swapnil Biswas, Faiza Omar, Faisal Muhammad Shah, and Samiha Binte Hassan. Spam review detection using deep learning. In 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pages 0027–0033. IEEE, 2019.

[8] ASharma and C Kaur. Social issues sentiment analysis using python. In Proceedings of the 5th International Conference on Computing, Commu nication and Security (ICCCS), 2020.

[9] Kian Long Tan, Chin Poo Lee, Kalaiarasi Sonai Muthu Anbananthen, and Kian Ming Lim. Roberta-lstm: a hybrid model for sentiment analysis with transformer and recurrent neural network. IEEE Access, 10:21517–21525, 2022