SiFri-Mail: A Pertinent Communication and Categorization Tool through the E-mail Protocol using Natural Language Processing and Support Vector Mechanism

Luis Anton Imperial
De La Salle University – Dasmariñas
+63 0976 048 2659
ilp0824@dlsud.edu.ph

Christian Friolo
De La Salle University – Dasmariñas
+63 0961 816 3017
fcc2386@dlsud.edu.ph

Timothy Allen Sicad
De La Salle University - Dasmariñas
+63 0956 887 6420
stn0169@dlsud.edu.ph

ABSTRACT
Natural Language Processing (NLP) has emerged as a powerful tool for analyzing and categorizing text-based data, such as emails. The "Sifri-Mail" project is a proof-of-concept for a cross-platform, Artificial Intelligence (AI)-powered electronic mail (e-mail) client application program that will simplify the daily workflow of users by linking all their email accounts into one place before categorizing them by topic and priority through Natural Language Processing (NLP) and Support Vector Machines (SVM). It seeks to leverage these technologies to develop a platform- and provider-agnostic email categorization tool. By using NLP and SVM, Sifri-Mail aims to automatically categorize emails, enhancing efficiency and ensuring that important messages are prioritized.

Keywords
categorization, email, email client, natural language processing, social networks, support vector mechanism, language model, artificial intelligence

1. INTRODUCTION

1.1 PROJECT CONTEXT
In today's digital landscape, email has become an essential communication tool, with billions of messages exchanged daily. However, the increasing volume of emails has led to challenges in effectively managing and categorizing incoming messages. This issue is particularly problematic in professional settings, where efficient communication is crucial for maintaining productivity and ensuring that important messages are not overlooked. Studies indicate that the number of emails sent and received globally is expected to exceed 347 billion per day by 2023, underscoring the need for more advanced solutions to manage this communication flow .

To address these challenges, Natural Language Processing (NLP) has emerged as a powerful tool for analyzing and categorizing text-based data, such as emails. NLP techniques enable systems to comprehend and process human language with a high degree of accuracy, making it possible to categorize emails based on their content rather than just keywords. When combined with Support Vector Machines (SVM), which are effective for text classification, these technologies can significantly improve the accuracy and efficiency of email categorization [3].

The "Sifri-Mail" project seeks to leverage these technologies to develop a platform- and provider-agnostic email categorization tool. By using NLP and SVM, Sifri-Mail aims to automatically categorize emails, enhancing efficiency and ensuring that important messages are prioritized. This approach is particularly relevant in today's fast-paced digital environment, where efficient communication is key to organizational success. [1]

1.2 PURPOSE AND DESCRIPTION
The purpose of the SiFri-Mail project is to create an intelligent tool that streamlines email management by automatically categorizing emails based on their content. By employing advanced technologies like Natural Language Processing (NLP) and Support Vector Machines (SVM), the tool is capable of understanding and analyzing the text of incoming emails with a high degree of accuracy. This ensures that important messages are promptly identified and prioritized, while less relevant ones are effectively organized. The tool aims to improve the overall efficiency of managing large volumes of emails, reducing the time and effort required for manual sorting, and helping users maintain a more organized and functional inbox. This enhanced email management will contribute to better communication and productivity, especially in environments where the timely processing of information is critical.

1.3 STATEMENT OF THE PROBLEM
The main problem is the inefficiency and inaccuracy of current email management systems in handling the growing volume and complexity of email communications. Traditional methods, such as manual sorting and keyword-based filtering, often fail to effectively manage the diverse and voluminous nature of modern email content. This inadequacy leads to important emails being overlooked, miscategorized, or lost in the clutter, which can result in missed opportunities, decreased productivity, and communication breakdowns.

- How might the system handle ambiguous or contextually complex emails that could lead to incorrect categorization?
- What challenges could arise in ensuring that the categorization system adapts to evolving email content and user needs over time?
- How can the system balance accuracy and speed in processing large volumes of emails without compromising on either?

1.4 RESEARCH OBJECTIVE
The research objectives for SiFri-Mail aim to address the growing complexities of email management in both personal and professional environments. With the daily volume of emails reaching unprecedented levels, traditional methods of email organization, such as manual sorting and basic keyword filtering, no longer suffice. This project seeks to enhance email management by leveraging advanced technologies, specifically Natural Language Processing (NLP) and Support Vector

Mechanism (SVM), to create an intelligent, automated system capable of classifying and prioritizing emails based on their content. By doing so, SiFri-Mail strives to improve productivity, reduce the time spent on email management, and ensure that critical communications are not lost in the clutter of inboxes.

The objectives outlined below encompass both the general goal of developing a robust email categorization tool and the specific technical and functional milestones that will guide the development and evaluation of the system.

1.4.1 GENERAL OBJECTIVES

The general objective of the SiFri-Mail project is to develop a pertinent communication and categorization tool through the email protocol using natural language processing and support vector mechanism.

1.4.2 SPECIFIC OBJECTIVES

1. To design and implement a machine learning-based email categorization system using NLP and SVM that automatically sorts emails according to their content and relevance.
2. To evaluate the accuracy of the email categorization system by comparing its performance against traditional keyword-based filtering methods in real-world scenarios.
3. To ensure the scalability of the system in handling large volumes of emails without compromising processing speed or accuracy.
4. To adapt the system for evolving email content and user needs through continuous learning and updates to the classification model, ensuring long-term relevance and accuracy.
5. To conduct user testing and feedback collection to refine the tool's user interface and functionality, ensuring it meets the practical needs of its users across different platforms.
6. To minimize the manual effort involved in email management by providing users with customizable categorization options, enabling personalized organization and prioritization of emails.

1.5 SIGNIFICANCE OF THE STUDY

Due to the lack of an intuitive, affordable, AI-powered communication tool on the market, this study aims to simplify daily organizational workflows by providing our target customer base with an easy-to-use email app that incorporates natural language processing technology to minimize the busywork of categorization and sorting.

1.6 SCOPE AND DELIMITATION

The scope of this research covers the analysis of email categorization using Natural Language Processing (NLP) techniques with a specific focus on sentiment analysis. The primary objective is to assess the effectiveness of sentiment-based email classification to improve user prioritization and management of emails. This involves categorizing emails by urgency, relevance, and context using a support vector mechanism to train the sentiment analysis model.

Data collection for this research includes testing by 20 to 50 respondents, who will use the application and provide feedback. The feedback will be gathered through structured online surveys hosted on Google Forms, focusing on user experience, accuracy of email categorization, and overall satisfaction with the tool. Questions will cover the clarity of categorization, perceived accuracy, ease of use, and areas for improvement.

Delimitation of this study excludes the technical specifications of the application, such as programming languages and platforms used. The research will not explore technical development or backend processing. Additionally, email content data will be anonymized to preserve privacy, and the study will avoid any long-term monitoring of user email habits post-research. Questions unrelated to sentiment accuracy, such as in-depth technical feedback on interface design, will not be included in the survey.

2. RELATED WORK AND TERMINOLOGY

This chapter contains the researched review done by the proponents about the related ideas regarding the proposed system. It includes the differences and similarities found among other email management systems. This chapter constitutes more on the study of system literature and also covers related views and ideas presenting other email categorization systems made possible by other proponents and programmers. The review encompasses the critical aspect of the email management system study, focusing on the role of Natural Language Processing (NLP) and Support Vector Mechanism (SVM) in categorizing and managing emails.

In this chapter, the proposed study aims to address the inefficiencies of current email management systems by utilizing NLP and SVM techniques. The review includes both local and foreign literature and studies, each contributing insights into how these technologies can enhance the accuracy and speed of email categorization.

"Land Cover Classification of the Abra River Basin with Remote Sensing and Machine Learning Algorithms" by Matso (2022) uses machine learning algorithms such as Support Vector Machine and Random Forest to classify the land cover of the Abra River Basin. The support vector machine classifier produced the highest accuracy. The land cover map produced can be used as input in preparing a watershed management plan for each of the municipalities covered by the Abra river basin. [4]

"Application of Support Vector Machine in Corn Disease Detection" by Sherlyn Avendaño (2019) conducted research on detecting common corn leaf diseases using the Support Vector Machines (SVM) algorithm. Their study found that 92.73% of the total corn leaf samples were accurately classified by the system. This study highlights the potential of SVM in handling subjectivity and categorizing image data accurately. [5]

"Stressor Classification of Filipino Political Tweets Using LDA, SVM, XGBoost, Logistic Regression" by Mark Gabriel E. Edaño, Ryan Joseph S. Gonzales, Raphael Carlo B. Laguda, and Joel C. De Goma (2020) explored the use of SVM for classifying stressors in Filipino political tweets. Their study utilized various machine learning techniques, including Latent Dirichlet Allocation (LDA), SVM, XGBoost, and Logistic Regression, to categorize tweets. The results provided insights into the application of SVM for text categorization in social media contexts, relevant to email categorization tasks. [6]

"Baybayin Character Recognition Using Support Vector Machine" by Rodney Pino, Dr. Renier Mendoza, and Dr. Rachelle Sambayan (2020) focused on developing an AI-powered Baybayin translator using SVM. This project involved applying SVM for character recognition in the ancient Filipino Baybayin writing system. This study showcases SVM's versatility in text

categorization tasks across different applications, including historical and cultural contexts. [7]

An article by Dagooc (2023) describes a report done by the cybersecurity company Kaspersky back in 2022, observing 4,559,288 reported phishing incidents in the Philippines, making it the fifth ranked Southeast Asian country experiencing the most phishing attacks. These primarily come in the form of fake emails pretending to be well known delivery companies containing fraudulent links. [10]

In a November 2023 FICO survey of 1,001 adults, 98% of adults have used real-time payment services, and with it over 35% of Filipino respondents have expressed concern with being scammed into sending money to criminals. [8]

Email is ubiquitous at the workplace. By examining email traffic in a convenience sample of 55 employees who used their business email address daily during a typical workweek, results revealed that employees' workplace telepressure was positively related to their email reply quantity and, surprisingly, unrelated to their email response latency. [14]

Business email compromise (BEC) attacks are a real concern for businesses, as human error is a consistent vulnerability. Rather than target the network perimeter itself, there is a chance hackers and malicious actors might go for the users of a system instead. [9] With the rising threat of online and electronic safety and privacy risks, especially through email, ways to manage your emails become more valuable.

However, this is where concepts such as artificial intelligence come into play, which can be indispensable tools for a user or company. Through this technology, it is possible to scan through billions of transactions for suspicious activity and give real time alerts, which are concepts that can be applied for NLP for categorizing emails by topic and priority. [11]

Furthermore, it has been found that one contributor to work-related stress is the blurring of boundaries between work and home domains, known as work-home interference (WHI). In a study done by Braitwaite et al. (2024), participants who reported that email was highly important and/or felt overloaded by emails were more likely to engage with work emails during leisure time. Additionally, email engagement in leisure time was associated with poorer physical and psychological health, but not productivity. [12]

A study on chatbots by Ortiz-Garces et al. (2024) stated the importance of syntactic structure analysis in the performance of NLP models when it comes to processing ideas found inside human language, which can often be variable and ambiguous. The study found a model with advanced syntactic analysis capabilities is more able to adapt to different linguistic contexts and process it with more coherence, accuracy, and respect to real-world scenarios. [15]

A paper by Karim et al. (2019) gathered a collection of methods used to detect spam in emails and reported their findings. The section about Support Vector Machines (SVM) emphasizes the importance of optimization of the kernel type and kernel parameters. Additionally, a positive correlation has been found between improved performance of the model as the amount of features available for feature selection and extraction increases. [13]

Natural language processing (NLP) is a branch of artificial intelligence that helps computers understand, interpret and manipulate human language. It draws from many disciplines, including computer science and computational linguistics, in its pursuit to fill the gap between human communication and computer understanding. [20]

NLP goes hand in hand with text analytics, which counts, groups and categorizes words to extract structure and meaning from large volumes of content. Text analytics is used to explore textual content and derive new variables from raw text that may be visualized, filtered, or used as inputs to predictive models or other statistical methods.[20]

Text classification in NLP involves categorizing and assigning predefined labels or categories to text documents, sentences, or phrases based on their content. It aims to automatically determine the class or category to which a piece of text belongs. Text classification algorithms analyze the features and patterns within the text to make accurate predictions about its category, enabling machines to organize, filter, and understand large volumes of textual data.[19]

For instance, a form of text analytics is sentiment analysis, wherein NLP can help interpret reams of user comments, social media posts, or customer service requests. [16]

A popular algorithm for NLP is the Support Vector Machines (SVM) algorithm. [17] SVMs are a supervised learning method used to perform binary classification on data. They are motivated by the principle of optimal separation, the idea that a good classifier finds the largest gap possible between data points of different classes. [18]

7. REFERENCES

[1] Haigh, C. (2018, March 15). Because we are social beings: why our gatherings are key to our approach. The Collaborate Out Louder, via Medium. https://medium.com/the-collaborate-out-louder/because-we-are-social-beings-why-our-gatherings-are-key-to-our-approach-2e6ac6ad4d22

[2] Radicati Group. (2020). Email Statistics Report, 2020-2024. Palo Alto, CA: The Radicati Group. https://radicati.com/wp/wp-content/uploads/2020/01/Email_Statistics_Report,_2020-2024_Executive_Summary.pdf.

[3] Yan, et al. (2021). E-mail classification with machine learning and word embeddings for improved customer support. Neural Computing and Applications. https://link.springer.com/article/10.1007/s00521-020-05058-4.

[4] Matso, N. (2022). Land Cover Classification of the Abra River Basin with Remote Sensing and Machine Learning Algorithms. https://ejournals.ph/article.php?id=16975

[5] Avendaño, S. (2019). Application of Support Vector Machine in Corn Disease Detection. UE Research Bulletin, vol. 21 no. 1. https://ejournals.ph/article.php?id=17178

[6] Edaño, M.G.E., Gonzales, R.J.S., Laguda, R.C.B., & De Goma, J.C. (2020). Stressor Classification of Filipino Political Tweets Using LDA, SVM, XGBoost, Logistic Regression. https://ieomsociety.org/proceedings/2022istanbul/258.pdf?form=MG0AV3

[7] Pino, R., Mendoza, R., & Sambayan, R. (2020). Baybayin Character Recognition Using Support Vector

Machine. Retrieved from https://www.researchgate.net/publication/349323182_Optical_character_recognition_system_for_Baybayin_scripts_using_support_vector_machine/fulltext/602dbf674585158939b069b2/Optical-character-recognition-system-for-Baybayin-scripts-using-support-vector-machine.pdf.

[8] BusinessWorld Publishing. (2024, May 21). Filipinos worry about falling for financial scams - study. BusinessWorld Online. https://www.bworldonline.com/banking-finance/2024/05/22/596548/filipinos-worry-about-falling-for-financial-scams-study/

[9] CT Link Systems, Inc. (2023, December 12). Email security service philippines. CT Link. https://www.ctlink.com.ph/services/email-security-service-philippines/

[10] Dagooc, E. M. (2023, July 19). Philippines ranks fifth with most phishing attacks in 2022. Philstar.com. https://www.philstar.com/the-freeman/cebu-business/2023/07/20/2282414/philippines-ranks-fifth-most-phishing-attacks-2022

[11] Origenes, O. (2023, October 31). Fighting fraud in the Philippines: Your guide to analytics, alerts and consortiums. TransUnion. https://www.transunion.ph/blog/fighting-fraud-in-the-philippines-your-guide-to-analytics-alerts

[12] Braithwaite, E., Walker, L., Cooper, C. and Jones, M. (2024, February). Emails 24/7: Agile working or electronic leash? Associations between engaging with work emails outside of normal working hours and health and productivity. DOI:10.21203/rs.3.rs-3990832/v1. https://www.researchgate.net/publication/378842446_Emails_247_Agile_working_or_electronic_leash_Associations_between_engaging_with_work_emails_outside_of_normal_working_hours_and_health_and_productivity.

[13] Karim, A., Azam, S., Shanmugam, B., Kannoorpatti, K., & Alazab, M. (2019). A comprehensive survey for intelligent spam email detection. Ieee Access, 7, 168261-168295.

[14] Cambier, R and Vlerick, P. (2020, August). You've got mail: does workplace telepressure relate to email communication?. Cognition Technology and Work 22(3). DOI:10.1007/s10111-019-00592-1. https://www.researchgate.net/publication/335435478_You%27ve_got_mail_does_workplace_telepressure_relate_to_email_communication.

[15] Ortiz-Garces, I., Govea, J., Andrade, R. O., & Villegas-Ch, W. (2024). Optimizing chatbot effectiveness through advanced syntactic analysis: A comprehensive study in natural language processing. Applied Sciences, 14(5), 1737.

[16] Cloudflare. (n.d.). What is natural language processing (NLP)? Cloudflare. Retrieved October 1, 2024, from https://www.cloudflare.com/learning/ai/natural-language-processing-nlp/

[17] Elton. (n.d.). The Support Vector Machines (SVM) algorithm for NLP. Python Wife. Retrieved October 1, 2024, from https://pythonwife.com/the-support-vector-machines-svm-algorithm-for-nlp/

[18] McGonagle, J., & Chandak, K. (n.d.). Support Vector Machines. Brilliant. Retrieved October 1, 2024, from https://brilliant.org/wiki/support-vector-machines/

[19] Parlad. (2023, August 9). Understanding Text Classification in NLP with Movie Review Example. Analytics Vidhya. Retrieved October 1, 2024, from https://www.analyticsvidhya.com/blog/2020/12/understanding-text-classification-in-nlp-with-movie-review-example-example/

[20] SAS Institute. (2024, February 14). Natural Language Processing (NLP): What it is and why it matters. SAS Institute. Retrieved October 1, 2024, from https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html