KWAME OWUSU BOAHENE

CHAPTER ONE

APPLIED PROJECT

BACKGROUND

In recent times, the advancement of technology has brought about improvement in the lives of people. Particularly in the field of communication, technology has led to the creation of innovations in both software and hardware that fosters seamless communication between people. Amongst these innovations, is electronic mail popularly known as e-mail. One benefit e-mail provides firms and individuals is the ability to send and receive messages on their electronic devices instantly.

In 2015, 205.6 billion emails were sent/received per day and is projected reach 246.5 billion by 2019 [1]. Although this show speeding's up communication, the popularity of email as a means of communication leads to several problems. One of these problems especially for firms is an influx of email in their inboxes. With an inbox full of emails, many questions arise, who should answer these questions, how can we respond quickly to these questions and which of these emails is spam?

Thankfully considerable research has been done into developing a spam filtering system as such my project will focus on answering who should answer these questions and how can we respond quickly to these questions. These questions bring up the study of email classification that is being able to sort emails into appropriate categories per user requirement. Since this is a very subjective field, one categorization system for person A may not meet the criteria for person B; my project focuses on building an email classification system for the Support Center of Ashesi University.

PROBLEM STATEMENT

The Ashesi Support Center is the hub for solutions for all problems and questions relating to IT, facilities, logistics, and other issues on campus. Addressing all the needs of faculty, students and non-teaching staff are essential to ensuring teaching activities run smoothly.

However, the reality is the support center does not have the answer to all the questions it receives. Therefore, daily, support personnel sift through the emails the center gets and forwards it to an appropriate responder for an answer. The process results in a delayed response and is a tiresome process considering the support center receives an influx of emails.

In response to the problem, my project proposes the use of supervised and unsupervised learning approaches to automatically forward the emails the support center receives to an appropriate responder. This would ensure the mail reaches an appropriate responder as soon as possible and reduce the burden of having to sift through a pool of emails.

AIMS AND OBJECTIVES

1. Suggest methods to use when developing email classification systems
2. Build an email classification system to lessen the burden support center faces
3. Ensure that all problems are handled the appropriate responder
4. Identify and understand the various algorithms used for email classification
5. Gain more knowledge and understanding of natural language processing as a field in artificial intelligence
6. Understand how supervised and unsupervised systems work

RELATED WORKS

Since this field is a relatively old field, considerable research has been made into the development of algorithms used for email classification. Therefore, in creating an email classification system for the support center, it is imperative to review some existing literature that relates to the project.

In building an email classification system, two critical processes take place preprocessing data for the classifier and building the classifier that processes the data. These two processes are fundamental because the ability to produce clean, easy to understand data is imperative to the classifier being able to correctly group emails. Similarly, the efficiency and correctness of the algorithm used in the classifier are crucial to the results produced.

Taking this into consideration, in Grbovic et al. [2], the authors propose an approach for automatically distinguishing between human and machine-generated emails as well as classifying emails without requiring the user to predefine any folder. In determining whether an email is machine generated or human, the authors use reserved words like “do not reply,” “mailer-daemon” and the occurrence of words like “unsubscribe” in the header of the message to determine if its machine generated. To determine if an email is by a human, the authors use checks if the sender has an address of the form “<first name>. <last name>” to tell. Based on these approaches, the authors grouped the email into “60,000 human senders and 80,000 machine senders” [2]. For determining the email categories, the authors used existing user folder activities to generate a set of folders to use. For the classifier, the authors build an online classifier for scalability and an offline classifier for period classification. After creating the systems, the authors tested the system, and it achieved precision and recall rates close to 90%.

Likewise, in an unsupervised learning approach to classify and summarize emails by Ayodele et al. [4]. The authors preprocess data by removing unnecessary words like “a, the, in, at.” After preprocessing to make the emails as clean as possible. The classifier takes the data and groups it in the respective user activity. For generating summaries, the summarization algorithm examines the email selects the words with the highest frequencies and the sentence that contains the words with the highest frequency. The algorithm rearranges the selected sentence and words to generate a good summary. After testing the algorithms, the authors achieved an accuracy of 90% for their classifier and “a recall score between 50% and 100% difference in the qualities of the email summaries while the precision score is between 20% and 70%” [4].

Similarly, in the approach used by, Nenkova & Bagga [4] in developing an email classification system for a contact center. The authors clean the data by initial cleaning the data by identifying deleting signature blocks and quotes passages from emails. They further use a tool to identify nouns and verb phrases in the data. In the classifier used which is a Naïve Bayes one, word tokens like nouns, verb phrases, dictionaries from constructed dictionaries. After testing, the system achieved an accuracy of 90%.

PLAN FOR REQUIREMENT ANALYSIS

Since my applied project deals directly with the support center. The requirement analysis process involves communication directly with support center to identify the following

1. What the system requirements are
2. How the currently implemented system works
3. Which approach (supervised, unsupervised and semi-supervised learning) should be used.
4. The amount of mails received by the center daily.

REFERENCES

[1] The Radicati Group. Email Statistics Report, 2015-2019. (March 2015). Retrieved October 9, 2018 from http://www.radicati.com/wp/wp-content/uploads/2015/02/Email-Statistics-Report-2015-2019-Executive-Summary.pdf

[2] Mihajlo Grbovic, Guy Halawi, Zohar Karnin, and Yoelle Maarek. 2014. How Many Folders Do You Really Need?: Classifying Email into a Handful of Categories. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (CIKM '14), 869–878. DOI: https://doi.org/10.1145/2661829.2662018

[3] Ani Nenkova and Amit Bagga. 2003. Email classification for contact centers. In Proceedings of the 2003 ACM symposium on Applied computing (SAC '03). ACM, New York, NY, USA, 789-792. DOI: https://doi.org/10.1145/952532.952689

[4] Taiwo Ayodele, Rinat Khusainov, and David Ndzi. 2007. Email classification and summarization: A machine learning approach. In *2007 IET Conference on Wireless, Mobile and Sensor Networks (CCWMSN07)*, 805–808. DOI: https://doi.org/10.1049/cp:20070271