ECE 2800

Communication Systems Filtering using NLP
Written Report

Ahmed Abdelgalil, Madison Lewis, Hafsa Traore

25 April 2022

Table of Contents

Introduction

Overview

Natural Language Processing also known as NLP is a branch of Artificial Intelligence that uses Machine Learning to process and interpret text and data. This allows computers to understand, interpret, and manipulate human language making it easier for humans to communicate with smart devices and achieve everyday tasks. Some common use cases of NLP in today's world include smart assistants such as Siri and Alexa, spell check, voice-to-text messaging, autocomplete, and many more. Furthermore, some of the common use cases for NLP include improving user experience, automate support, and monitoring as well as analyzing user/customer feedback.

Background

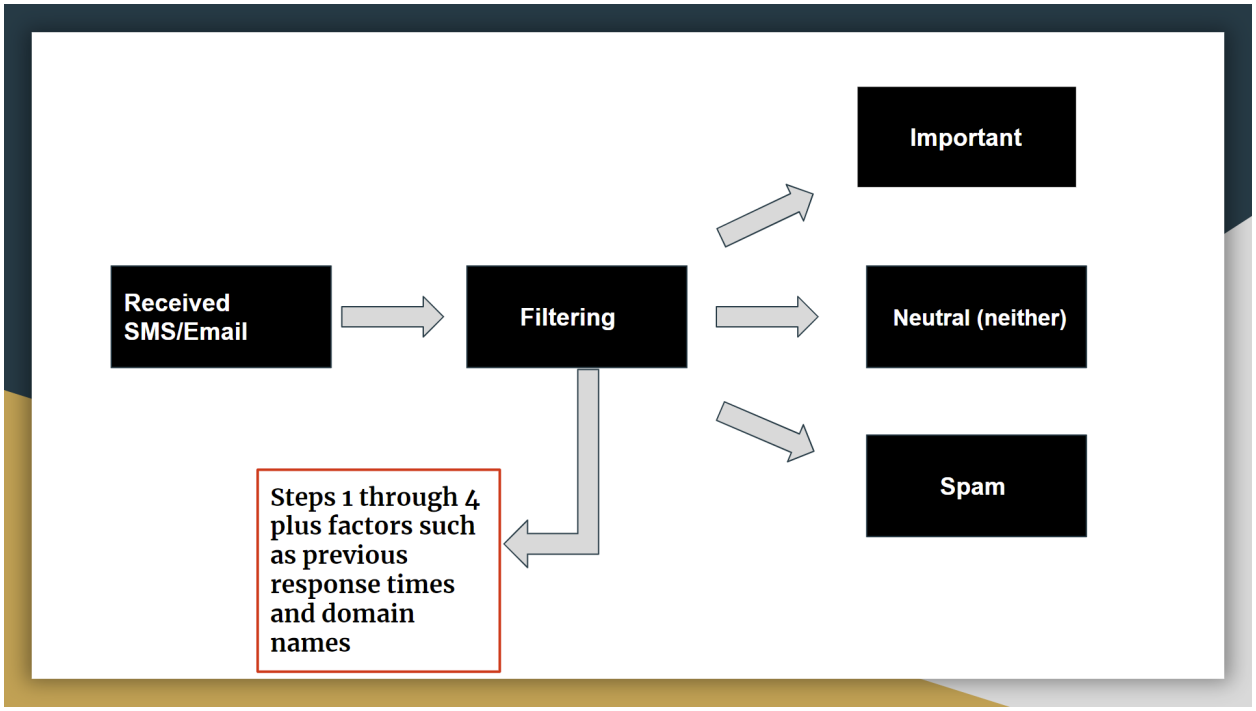
For this project, we decided to use NLP for filtering messages and emails received from communication systems. Similar to how many email applications are capable of filtering and flagging spam emails, we plan to implement a new algorithm that utilizes NLP to flag high priority/important messages. The motivation behind this is to provide quick response times which in return lead to satisfied users across many platforms. This algorithm can be applied to various forms of communications and applications such as Gmail, SMS, Slack, etc.

Objective

The objective of this project is to create a filtering algorithm using Natural Language Processing that is capable of accurately categorizing received messages/emails into one of three categories: important, spam, and neutral.

Outline & Specifications

Below is a basic outline of how the algorithm will function. The end goal of filtering is achieved in three steps. The first step is receiving the intended message/email. The second step is the filtering process where the NLP algorithm takes care of the filtering process through multiple steps explained in the methodology section. In addition to filtering, during this step, other factors such as response time and domain names are taken into consideration to further train the model and increase the accuracy of the output. The third and final step is the classification phase. In this step the message has gone through the filtering process and is ready to be classified into one of the three categories shown below.



Methodology

Feasibility Analysis

Filtering algorithms for messaging systems is a major part of improving customer experience. There are many approaches to “spam filtering”. This includes Content Filtering, which analyzes text and looks for trends that are predictable or connected to money. Such as deal offers or promoted explicit material. This approach is flawed in that it only takes into account money related scams, which could filter important messages as spam.

Another method used is Blacklist Filters. This method blocks emails that have been flagged as spam. This is not efficient, as spammers can simply change their email domain and surpass the filter. This also depends on flagging systems that are not efficient and are known to flag messages of real people as spam.

The NLP algorithm as a communication system filter accurately classifies messages and emails. The filtering step of the system includes: data preprocessing, removal of “stop words”, tokenization, lemmatization, and classification. Each of these parts ensures that the text is thoroughly prepared to be interpreted and categorized. Having three categories provides more range and improves user experience by prioritizing the messages. The algorithm is also constantly adjusting through machine learning in order to correct any wrong filtering. NLP is best suited for the complex human language and communication mechanisms.

Proposed Approach

The proposed approach for our invention sets a timeline of two years. In two years, we will accomplish specific and manageable tasks to allow for our approach to be evaluated and, if successful, form a sound basis for further development. We will provide long-term research goals from multiple short-term objectives.

Our first year will primarily be spent focusing on research to design our product and implement these findings. Once this initial implementation is to our satisfaction, we will be able to continue with our development into phase two, or the second year.

As shown within our schedule, in more detail below, the two year span from June 2021 to May 2023 will provide assurance that a subsequent amount of time for our project sequence will be accomplished in an efficient and feasible manner.

Non-Technical Aspects

In today’s world, the impact of Natural Language Processing is present amongst many of the leading industries. For instance, NLP played an important role in the healthcare industry with the recent outbreak of the COVID-19 virus. While many people were concerned about the number of patients who contracted the virus on a daily basis, the need for clear and concise documentation

of data and meta-data regarding the diagnosis, testing, and tracking of the spread of the virus grew larger. In return, many Data Scientist and Machine Learning Engineers were able to apply their expertise in NLP to large sets of unstructured data to make it better actionable and analyzable for healthcare professionals.

In addition to the healthcare industry, the use of NLP is gradually increasing in the marketing and e-commerce industry. Today, many businesses such as Amazon have adopted the use of NLP to improve customer satisfaction by automatically analyzing and sorting customer tickets and requests by topic, intent, urgency, and many other categories to achieve quick response times and therefore achieve maximum customer satisfaction. Furthermore, many of these businesses are using NLP to deploy chatbots as another form of increasing customer satisfaction.

Although NLP is proving to be effective across many industries, like any other emerging Artificial Intelligence technology it still faces ethical issues when it comes to things such as transparency, reproducibility, and plagiarism. For instance, many believe that chatbots have a tension between accuracy and robustness and believe as though they should provide responses that a moral human would.

Administration

Major Tasks

We are a small start up of twenty-five individuals (including the three founders) that was founded in 2021. Our Company consists of the following five departments: Machine Learning/ AI, Software Engineering, Research, Legal, and HR/Customer service. Individuals were assigned specific departments as needed and based on their technical backgrounds and experiences. However, it is important to note that for our technical & analytical departments, our employees are versatile and possess basic technical skills that allow them to shift work and partake in different departments when needed.

The table below shows a breakdown of our departments

Department Name	Number of Individuals
Machine Learning & Artificial Intelligence	9
Software Engineering	6
Research	2
Human Resources & Customer Service	2
Legal	3
Total Number of Employees	**22**

The first three departments will focus on our major task while the HR and Legal team will focus on other related issues and tasks that do not necessarily impact our production timeline. These majors tasks include the following:

- Research and Analysis
    - This task will include basic tasks such as researching our product and potential softwares, libraries, and algorithms that could be of good use.
- Coding and Design
    - This stage is where we begin coding and designing our machine learning models through libraries such as Tensorflow, Pytorch, etc.
- Testing
    - This is our testing stage where we perform performance tests and stress tests internally before deploying our product.
- Maintenance
    - This is an everlasting stage where we maintain our products efficiently and security while responding to customer issues.

Lastly, below is a table that shows the breakdown of the major tasks based on departments:

Major Task	**Responsible Teams**
Research & Analysis	Research & ML/AI
Coding and Design	Software, ML/AI, and Research
Testing	Software & ML/AI
Maintenance	Software, ML/AI, and Research

As one can see, all of our major tasks are split amongst at least two various teams for the sake of efficiency. Furthermore, as mentioned before, all of our technical employees possess great versatility skills that allow them to partake in different tasks led by other departments. For instance, although workers within our research team are experts when it comes to research, they are still capable of partaking in the coding stage if and when needed.

Schedule

The timeline below shows a breakdown of our anticipated schedule for the launch of our product as well as what our employees will be doing for each time period.

Budgeting and Resources

Budgeting

Based on calculations made, we are looking for a total budget of $100 million. Below is a table portraying a breakdown of how this budget will be spent:

Task	Cost (in dollars)
Payments (administration, salaries, etc.)	16 million
Office space & resources	15 million
Research	15 million
Coding & Design	10 million
Marketing	4 million
New features, updates, & launch	10 million
Testing	5 million
Regulations & Legal work	10 million
Maintenance & Security	15 million
Total	**100 million**

Resources

Furthemore, below is a list and brief explanation of the resources we will need:

- Office space
    - This includes work space, gym, and other facilities provided in similar companies.
- High performance technology
    - This includes computers, monitors, software, and other necessary equipment our employees will need to provide the best product possible.
- Licensing & patent
    - This basically covers any legal restrictions we may need to launch our product and workspace.

The three resources mentioned above are necessities I will need in order to get started. In return, they are vital resources to acquire; however, we do plan on adding other resources in the future once we hit projections and begin our expansion domestically and then internationally.

References

[1] EdPrice-MSFT. “Natural Language Processing Technology - Azure Architecture Center.” *Azure Architecture Center | Microsoft Docs*, docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/natural-language-processing.

[2] “Email Spam Filtering: Different Methods & How They Work.” *Fortinet*, www.fortinet.com/resources/cyberglossary/spam-filters.

[3] “Free From Stanford: Ethical and Social Issues in Natural Language Processing.” *KDnuggets*, www.kdnuggets.com/2020/07/ethical-social-issues-natural-language-processing.html.

[4] Future Analytica. *Ethics in Natural Language Processing*, Blogger, 11 Feb. 2022, futureanalyticaai.blogspot.com/2022/02/ethics-in-natural-language-processing.html.

[5] Gan, Sie Huai. “How To Design A Spam Filtering System with Machine Learning Algorithm.” *Medium*, Towards Data Science, 16 May 2021, towardsdatascience.com/email-spam-detection-1-2-b0e06a5c0472.

[6] “What Is Natural Language Processing?” *SAS*, www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html.