

A Mini Project Synopsis on

E-Mail Spam Detection

T.E. - I.T Engineering

Submitted By

Sameer Sawant 19104054

Raj Shisode 19104070

Siddhart Chhoriya 20204005

Under The Guidance Of

Prof. Apeksha Mohite



DEPARTMENT OF INFORMATION TECHNOLOGY

A.P.SHAH INSTITUTE OF TECHNOLOGY

G.B. Road, Kasarvadavali, Thane (W), Mumbai-400615

UNIVERSITY OF MUMBAI

Academic year : 2021-22

CERTIFICATE

This to certify that the Mini Project report on E-Mail Spam Detection has been submitted by Sameer Sawant(19104054),Raj Shisode(19104070) and Siddhart chhoriya (20204005) who are a Bonafede students of A. P. Shah Institute of Technology, Thane, Mumbai, as a partial fulfilment of the requirement for the degree in **Information Technology**, during the academic year **2021-2022** in the satisfactory manner as per the curriculum laid down by University of Mumbai.

Ms Apeksha Mohite
Guide

Prof. Kiran Deshpande
Head Department of Information Technology

Dr. Uttam D.Kolekar
Principal

External Examiner(s)

- 1.
- 2.

Place:A.P.Shah Institute of Technology, Thane

Date:

ACKNOWLEDGEMENT

This project would not have come to fruition without the invaluable help of our guide **Prof. Apeksha Mohite** and **Prof. Roshini Singh** Expressing gratitude towards our HOD, **Prof. Kiran Deshpande**, and the Department of Information Technology for providing us with the opportunity as well as the support required to pursue this project. We would also like to thank our teacher **Prof. Nahid Shaikh** who gave us her valuable suggestions and ideas when we were in need of them. We would also like to thank our peers for their helpful suggestions.

TABLE OF CONTENTS

1. Introduction.....	1
1.1.Purpose.....	1
1.2.Objectives.....	1
1.3.Scope.....	2
2. Problem Definition.....	3
3. Proposed System.....	4
3.1. Features and Functionality.....	4
4. Project Outcomes.....	7
5. Software Requirements	8
6. Project Design.....	9
7. Project Scheduling.....	15
8. Screenshot of Application.....	16
9. Conclusion.....	17

References

Acknowledgement

Chapter 1

Introduction

Apart from the numerous benefits and conveniences people around the world can enjoy due to the Internet, there are also multiple drawbacks. Not all of them are obvious to an average user, and perhaps only professional IT workers face them from time to time. However, there is a problem almost any Internet user has encountered at least once in a lifetime. Unlike many people might think, spam is not just an annoying but harmless email message; in fact, spam can be a dangerous tool capable of harming its recipients, and should be outlawed.

Even though spam messages usually have an “unsubscribe” link, getting off a spammer’s list requires a number of actions, such as visiting the website, acknowledging unsubscription, sending confirmation letters, typing the captcha, and so on. This might be not a problem in the case of being a target of several spammers; however, usually Internet users receive dozens of spam messages daily; unsubscribing from each of them is almost impossible.

1.1. Purpose:

Implementing spam filtering is extremely important for any organization. Spam emails, also known as junk email involves nearly identical messages sent to numerous recipients by email

Not only does spam filtering help keep garbage out of email inboxes, it helps with the quality of life of business emails because they run smoothly and are only used for their desired purpose. Spam filtering is essentially an anti-malware tool, as many attacks through email are trying to trick users to click on a malicious attachment, asking them to supply their credentials, and much more.

1.2 Objectives:

- Email Spam Detector should be able Detect content of the given message instantaneously without any delay and provide at most accurate or answer which the users were looking for.
- And to alert the user about suspicious emails that are sent by unauthorized users. To identify whether an email is a spam or a ham.
- Also one of the main objectives of this project was to learn about the use of machine learning algorithm in E mail spam detection process.

1.3 Scope:

The project provides sensitivity to the client and adapt to the future spam techniques. It also considers a complete message instead of single words with respect to its content. It reduces the IT administration workload as now they will have to deal with less important emails and more of important emails that require urgent attention. Email Spam detection method can also be used to reduce cyber criminal activities because the email spam detector may also filter out those messages that contain malicious links that may automatically download dangerous software in your system or device.

Chapter 2

Problem Definition

Junk emails waste a bunch of time and effort that could've been used for something more productive—but that's not even the worst part. Spam is also a popular means of transferring harmful malware and electronic viruses. And in an age where hacking tools and techniques grow more and more sophisticated by the minute, spam-instigated security attacks become a perpetual threat.

Spam emails are also an avenue for marketers to exploit your data's privacy. Responding to just one unsolicited email could put you in the mailing lists of many other companies. Before you know it, your spam emails would've already multiplied tenfold.

Phishing scams have also gained traction through spamming. A spam email that was made to look like it came from a legitimate entity that you trust (like your bank or someone you know)

could end up stealing sensitive information if you're not careful. You could be a victim of identity theft, or you could lose all your money if you mistakenly hand in your bank details.

Chapter 3

Proposed System

E-mail Spam detector for healthcare will use Machine Learning technique like Natural Language processing to understand the user input and will predict output as per the trained model and response as per the best matching reply it find.

In proposed System we will be having an interface wherein we would just have to copy and paste the email spam message after few seconds of processing data with the help of naïve Bayes algorithm, The system will come up with a message shoeing whether an email is a spam or a ham.

3.2 Features and Functionality

1. Give response instantly – The Email Spam detector is able to provide response on whether an email is a spam or a ham on the user input very quickly as possible.
2. Easy to understand User Interface – The Graphic User Interface on the email spam predictor should be clean and easy to understand able and have flexible display window
3. Accurate response –An Email spam detection system should be able to predict the response according to the content used in the email as accurate as possible to give answer which are needed in engaging and interactive manner.

3.2 Literature Survey:

Title	Author/s	Findings
Study of Spam E mails	Prince Raj, 2Justice Kumar, 3Mr. Ashish Tiwary	The author has worked with different machine learning algorithms for email classification such as Neural Network (NN), Support Vector Machine (SVM), J48 Decision Tree based classifier, Naïve Bayes. The dataset used by the author was Spam Base dataset. In this paper work, the author didn't mention advantages and disadvantages of any algorithm.
A systematic literature review on spam content detection and classification .	Abbas.M and Ibrahim.M (2020)	Based on our research objective, the initial search keywords were carefully chosen. Following an initial search, new words discovered in several related articles were used to generate several keywords. These keywords were later trimmed to fit the research's objectives. We chose certain search keywords based on the goal of our survey work, and after performing an initial search on those words, several keywords were derived from selected articles.

Comprehensive Survey for Intelligent Spam Email Detection	Kalaivani.K.M, Dharinishree.K.A, Gowsalya.B, Punitha.P, Dr. Kalaikumaran.T (2019)	(2013) This research initiative is to address a gap that has risen over time in the field of spam email detection. The current solutions are mostly lagging behind the innovativeness the spammers are constantly bringing in, which heavily justifies the emergence of machine learning based anti-spam propositions
A Survey of Existing E-Mail Spam Filtering Methods Considering Machine Learning Techniques	Shridevi.A.Mali, Sravanthi.G, Siva Subbha Rao.P, Rajja.S, Sushma.S.J, Reddy.A.R (2017)	This survey paper elaborates different Existing Spam Filtering system through Machine learning techniques by exploring several methods, concluding the overview of several Spam Filtering techniques and summarizing the accuracy of different proposed approach regarding several parameters. Moreover, all the existing methods are effective for email spam filtering. Some have effective outcome and some are trying to implement another process for increasing their accuracy rate.

Chapter 4

Project Outcomes

- The outcome of our project is to provide a system where in it should be able to respond with correct and accurate response after it process the data/Information provided from the given input and notify the user with a response whether an email is a spam or not.
- . Offer as a place to any user a place can quickly enter all the necessary data and with some minor concern over their health and get a solution correspondingly without visiting doctor, thus also helping hospitals also on avoiding large appointment on minor problem but also save the user time and money in the long run.
- This system, to solve the problem of spam, the spam classification system is created to identify spam and non-spam.
- Since spammers may send spam messages many times, it is difficult to identify it every time manually. So This project will be using some of the strategies in our proposed system to detect the spam.

Chapter 5

Software Requirements

Software Used –

- Python (greater than 3.6),
- Jupyter Notebook,
- Pycharm,
- NLTK (Natural Language Tool Kit),
- Numpy,
- Streamlit,
- Sklearn,
- Any IDE which is able run python application.

Chapter 6

Project Design

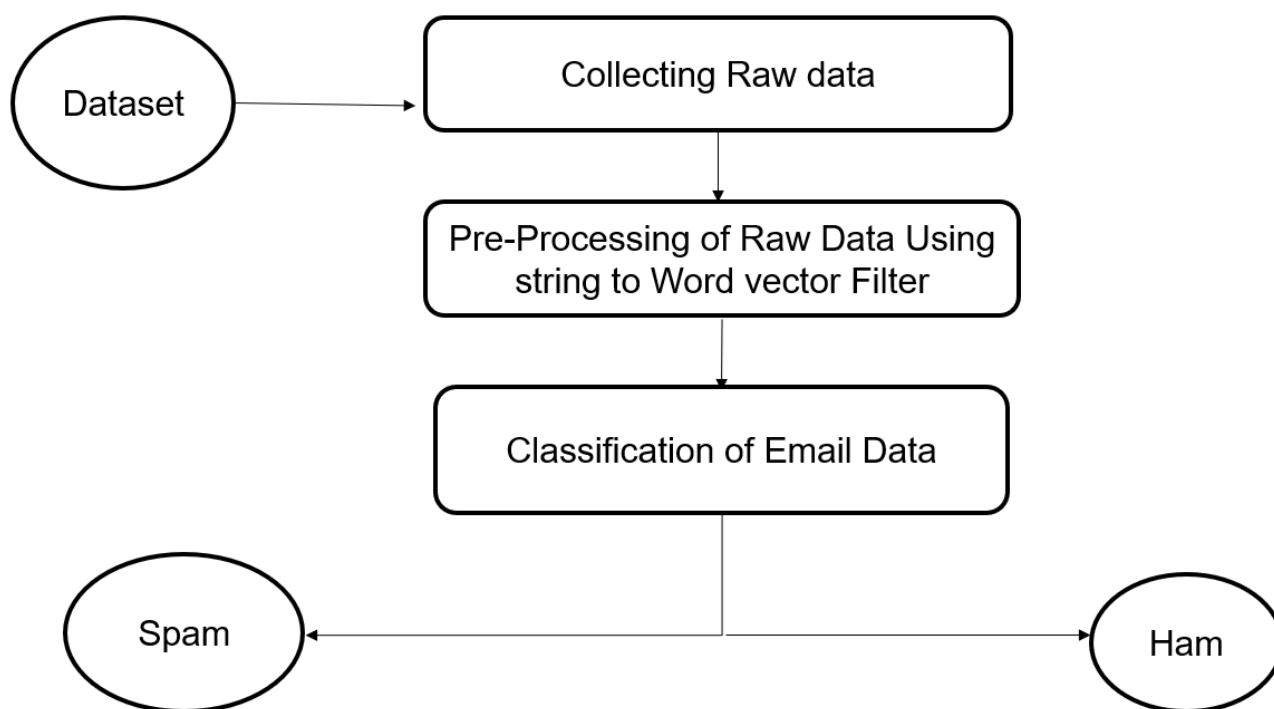


Figure: Flow of Modules

For our project the main machine learning code is done using python, Jupyter Notebook and using pycharm, numpy and nltk libraries. The flow of modules in our project is that at first we have to train the data which is available SMS Spam Collection Dataset at kaggle which

is consist of spam and ham email examples which is used to predict emails. As for training the data first nltk tokenizes the tags,pattern and response and forms array and then stemming of porter type is done to the words, converting to lower case, removing special characters, stop words and punctuations is done so that the words should be form of basic words meaning rather than any complex. So after all this words are made into bag of words so that words are converted into number as the model is not able to understand words.

Model is made on Jupyter Notebook so the processing of data gets as accurate result as possible with least loss possible in the prediction. Next process is to connect to the user interface which is made using Streamlit and Heroku. Vectorization and sklearn models are used to train and test the data. Voting and stacking is performed to test accuracy and precision of each models and combination of models. The website asks for input mail message, after pressing the predict button the website predicts email is spam email or ham email and displays it as the output.

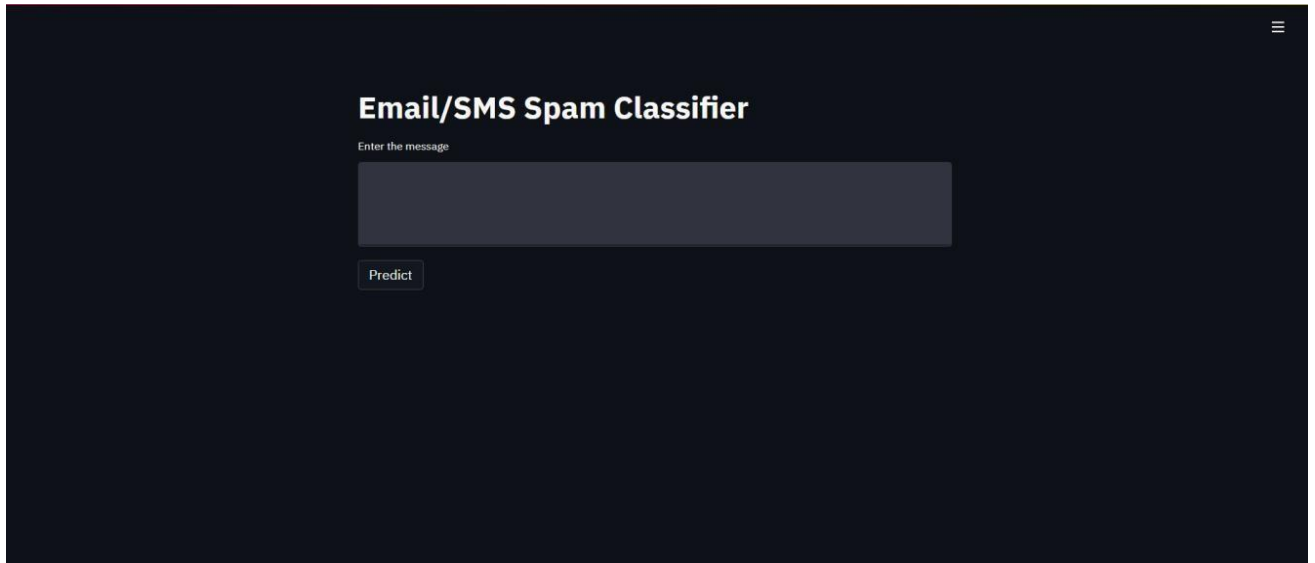
Chapter7

Project Scheduling Template

Sr. No	Group Member	Time duration	Work to be done
<u>1</u>	Sameer Sawant	1 st week of February	Implementing 1 st module/ functionality (<i>nltk.py/ By making on this module we were able make the input and output of words in array, by tokenizing and stemming and converting it into numbers</i>)
		2 nd week of February	Testing 1 st module (<i>train.py/ By using intents.json we were able process the data and also train the model and get loss and accuracy</i>)
<u>2</u>	Raj Shisode	3 rd week of February	Implementing 2nd module/ functionality (<i>model.py/ We build a neural network to predict and train and test model and get data preprocessed accordingly</i>)
<u>3</u>	Siddhart Chhoriya	By the end of march month	Implementing 3rd module/ functionality (<i>chat.py/Making the gui and connecting the different module to each other, and displaying the result as working chatbot</i>)

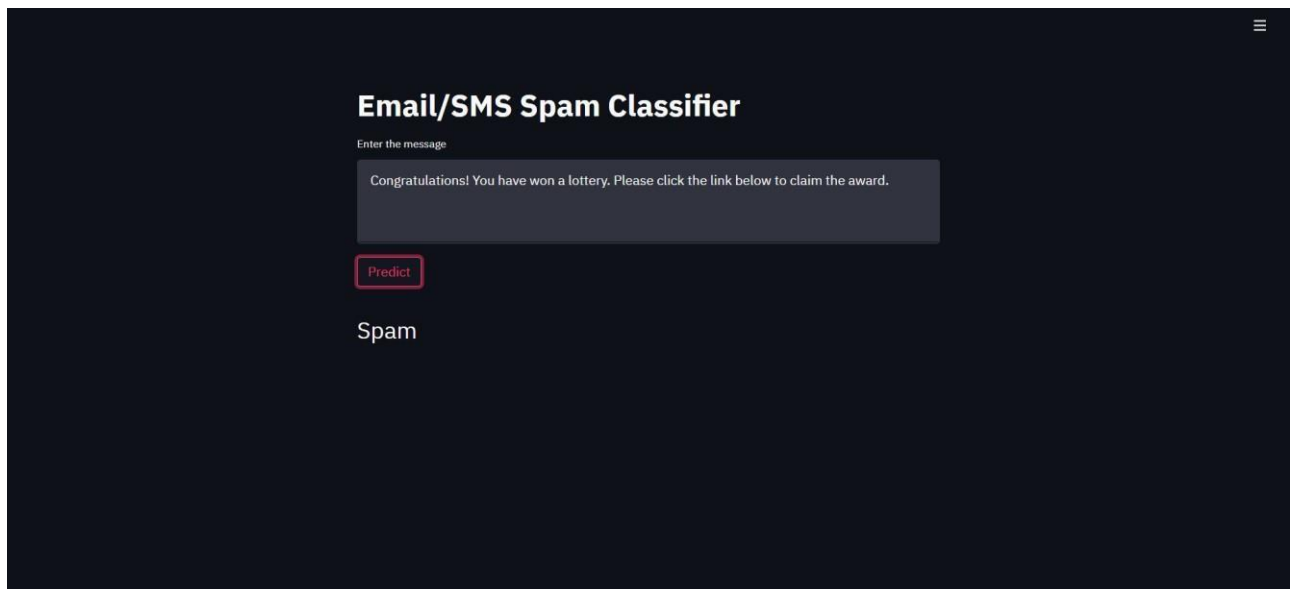
Chapter 8

Screenshot of Website



The screenshot shows a web application titled "Email/SMS Spam Classifier" on a dark background. Below the title is a label "Enter the message" followed by a large, empty text input field. Below the input field is a button labeled "Predict".

Figure1.



The screenshot shows the same web application as Figure 1, but with the input field containing the text "Congratulations! You have won a lottery. Please click the link below to claim the award." The "Predict" button is now highlighted with a red border. Below the button, the word "Spam" is displayed, indicating the classification result.

Figure 2. Email Detected Spam

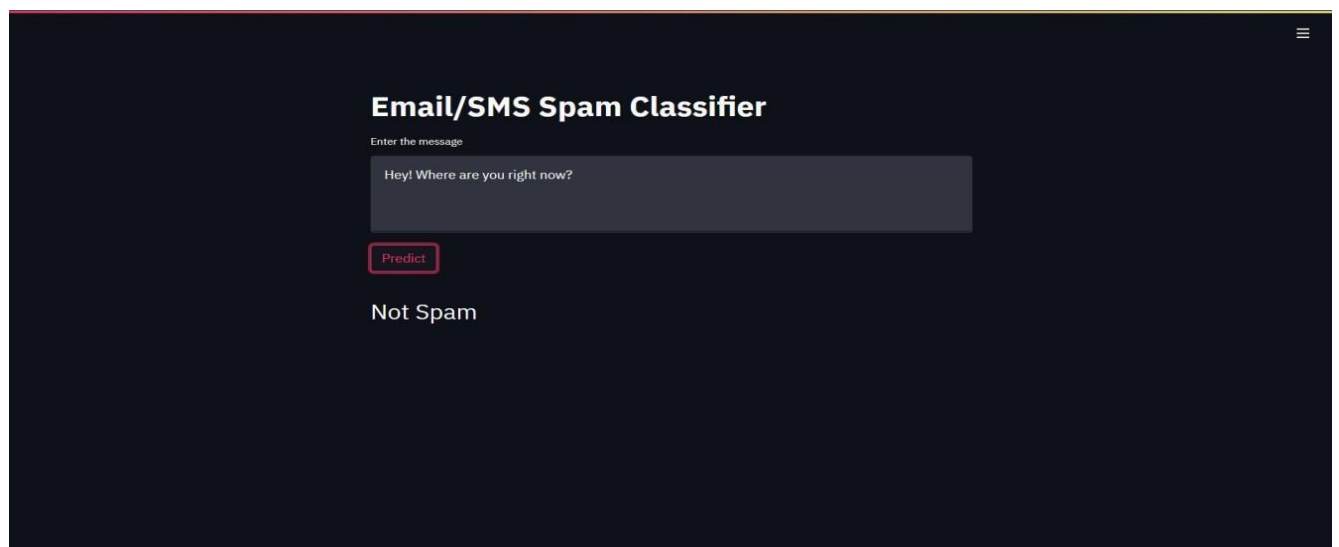
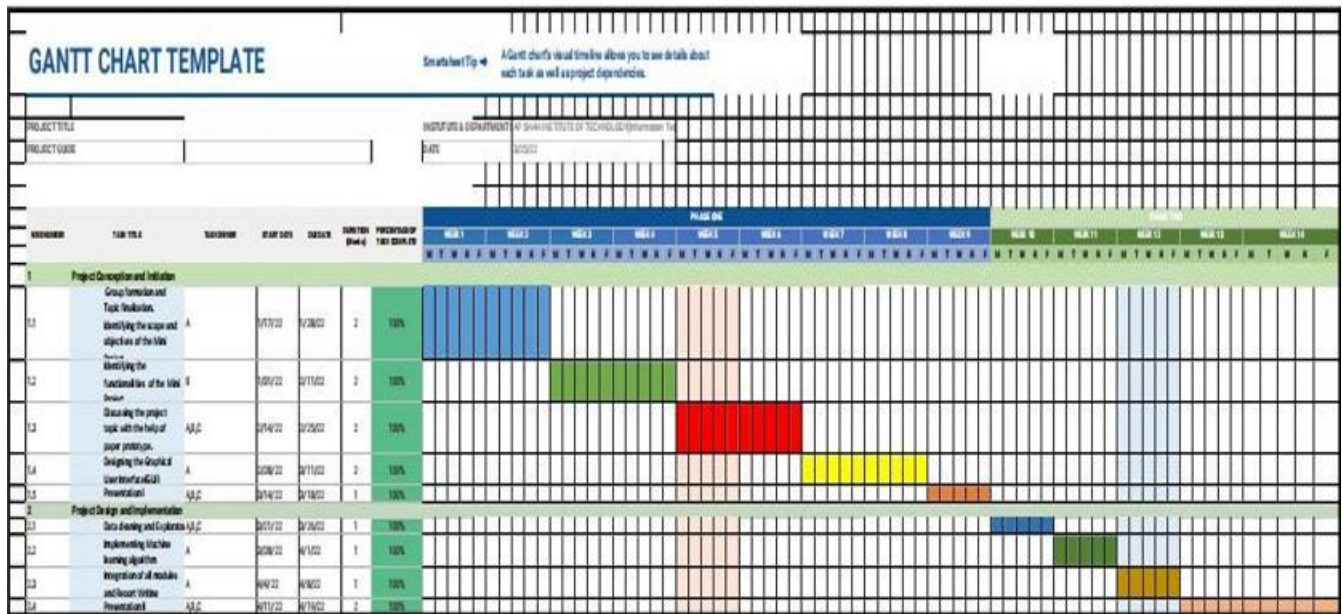


Figure: Email detected not spam

GANTT CHART



Gantt chart help teams to plan work around the deadlines and properly allocate resources. Projects planners also use Gantt charts to maintain a bird's eye view of projects. They depict, among other things, the relationship between the start and end dates of tasks, milestones, and dependent tasks. Modern Gantt chart programs such as Jira Software with Roadmaps and Advanced Roadmaps synthesize information and illustrate how choices impact deadlines.

Chapter 9

Conclusion

An E-mail spam detector is a great tool for saving time and identifying the original and authentic email it's developed to provide an accurate response of input in easy and engaging manner. This project help us to understand training and testing of model along with response and pattern data processing for email spam detector. While enabling email filtering can prevent phishing and scam emails from victimizing you, sometimes, a few sophisticated ones can make their way through. Email filters are not enough to protect you from being victimized by a scam. You must learn how to recognize the signs of an email scam if you wish to prevent yourself from falling for one. There are a few common signs that an email may be a scam or phishing attempt. The first step is knowing the difference between a phishing email and a spam email.

Working on this project also helps us to provide brief exposure in the domain of Machine Learning and how it will shape our lives in future. Machine learning is now becoming a main part of every technology based enterprise from very large to small on getting the user interact with website in more than few ways to increase engagement on their website and also to give user more information of their product or services in interact manner while having user attention in larger span.

References

- [1] <https://www.soscanhelp.com/blog/what-is-email-spam-filtering-and-how-does-it-work>.
- [2] <https://www.mailchannels.com/what-is-spam-filtering/>.
- [3] Documentation of Pytorch and Nltk
- [4] https://www.researchgate.net/publication/342113653_Email_based_Spam_Detection