# Detection of Email Spam using Natural Language Processing

*Belma Šehić, Amina Srna*

*Department of Information Technologies - Faculty of Engineering, Natural and Medical Sciences - International Burch University*

*Sarajevo, Bosnia and Herzegovina*

*8/1/2024*

*belma.sehic@stu.ibu.edu.ba, amina.srna@stu.ibu.edu.ba*

## ABSTRACT

In the vast world of the internet, one persistent issue is the arrival of unwanted emails, commonly known as spam. These unsolicited digital communications can target individuals, groups, or even entire companies, posing a significant threat. The danger lies in the fact that spammers might collect email addresses used for online registrations, leaving genuine users vulnerable to various attacks. Considering the popularity of online platforms today, email spam classification is a critical task in today's digital world, where the amount of spam emails has increased dramatically. In this project, we propose to use machine learning (ML) and natural language processing (NLP). The project aims to develop an efficient spam model predictor that can accurately identify and filter spam emails from legitimate ones. The dataset used in this project will consist of a large number of email messages with their corresponding labels (1 for spam and 0 for ham). We will use NLP techniques such as tokenization, stop word removal, stemming, lemmatization, and feature extraction to preprocess the text data and extract relevant features. Regular expressions will also take place. The accuracy of the classifier will be measured using evaluation metrics such as precision, accuracy and score. Moreover, this project holds a dual focus on sentiment analysis, considering the emotional tone and context of email messages. This sentiment-oriented approach adds a layer of sophistication to the classification model, contributing not only to improved spam detection but also to a deeper understanding of user engagement and communication dynamics. In conclusion, the project's significance extends beyond the realm of email security, as it contributes to the advancement of both NLP and ML techniques for email spam classification. The proposed model's potential application in real-world scenarios promises to alleviate the burden of spam for end-users while simultaneously pushing the boundaries of technological innovation in the domain of online communication.

***Keywords: Message, Ham/Spam, Natural Language Processing, Machine Learning, Online Platform, Sentiment Analysis, Email Security, Productivity***

## 1. INTRODUCTION

Email spam has become a significant problem in today's digital age, posing challenges for individuals, businesses, and organizations alike. Spam emails are unsolicited messages that flood inboxes, wasting valuable time and resources while potentially exposing users tomalicious content or scams. To combat this issue, machine learning techniques have emerged as powerful tools for email spam detection. The consequences of spam extend beyond mere annoyance, infiltrating personal and official domains, causing economic disruptions, and tarnishing reputations. Existing strategies to distinguish emails face challenges in detecting "zero days" attacks, resulting in higher false positives rates and reduced accuracy. As the internet plays a crucial role in daily life, the paper investigates the dark side of it. The escalation of spam emails over the past decade is recognized as a significant concern, leading to unwanted storage consumption, time wastage, and bandwidth occupancy. While traditional manual blocking methods face challenges, recent advancements, particularly in machine learning, offer automated solutions for text analysis, black and white lists, and community-based methods. Email is one of the most popular communication methods, but unfortunately, it is also a common target for spam messages. Spam emails not only waste time but can also contain malicious links or attachments that can harm computer systems. As the volume of emails continues to grow, it has become challenging to identify and classify spam emails manually. Therefore, the development of machine more effectively. Practical application of research findings in the

learning (ML) and natural language processing (NLP) techniques has opened up new avenues for automated email spam classification. In this project,we aim to use ML and NLP techniques to classify emails as spam or legitimate, based on their content and other relevant features. The project involves building a model to analyze the text of emails and determine whether they are spam or legitimate. This study has the potential to provide a valuable solution to the problem of email spam and help users to manage their email

development of spam detection tools, educational outreach to raise awareness, collaboration with industry professionals, and policy recommendations for a more robust legal framework are practical outcomes aimed at addressing the spam challenge. The research's importance lies in its contribution to enhanced cybersecurity measures, protection of user data, understanding economic disruptions caused by spam, and the incorporation of cutting-edge technologies in the fight against spam.Moving beyond emails, the study delves into the realm of spam reviews in the e-commerce sector. Recognizing the pivotal role of customer reviews, the paper emphasizes the importance of detecting and preventing fake reviews to maintain trustworthiness. In conclusion, this paper aims to provide a comprehensive overview of the multifaceted challenges posed by spam across various digital platforms. By proposing a novel approach for feature selection and classification, the study seeks to contribute to the ongoing efforts in developing more effective spam detection mechanisms.

## 2. NATURAL LANGUAGE PROCESSING

Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can. NLP combines computational linguistics—rule-based modeling of human language—with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to 'understand' its full meaning, complete with the speaker or writer's intent and sentiment. NLP drives computer programs that translate text from one language to another, respond to spoken commands, and summarize large volumes of text rapidly—even in real time. There's a good chance you've interacted with NLP in the form of voice-operated GPS systems, digital assistants, speech-to-text dictation software, customer service chatbots, and other consumer conveniences. But NLP also plays a growing role in enterprise solutions that help streamline business operations, increase employee productivity, and simplify mission-critical business processes.

## 3. LITERATURE SURVEY

• *Almeida, T. A., Gómez, H. F., & Yamakami, A. (2010). Contributions to the study ofSMS spam filtering: New collection and results. Journal of Machine Learning Research, 11, 3611-3628.* -This study focuses on SMS spam filtering but provides insights into feature selection and classification algorithms applicable to email spam detection using machine learning.

• *Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., Paliouras, G., & Spyropoulos, C. D. (2000). An evaluation of naive Bayesian anti-spam filtering. In Proceedings of the Workshop on Machine Learning in the New Information Age (Vol. 1, No. 1-3, pp.9-17)***.** -This study evaluates the performance of the Naive Bayes algorithm for email spam filtering. It compares different feature representations and discusses the impact of different factors on classification accuracy.

## 4. MACHINE LEARNING IN E-MAIL CLASSIFICATION

Incorporating machine learning into email classification significantly enhances the system's ability to distinguish between spam and legitimate content. With a repertoire of ten classifiers, including diverse algorithms like Naïve Bayes, Support Vector Machines, Decision Trees, and Neural Networks, the project aims for a comprehensive analysis of email data. The adaptability of machine learning models ensures continual improvement in accuracy by learning from new data, while real-time processing capabilities minimize the risk of delayed threat identification. Ensemble learning techniques further contribute to the robustness of the classification system by combining predictions from multiple classifiers.

## 5. TECHNOLOGIES USED

This research project relies on the Python programming language, particularly emphasizing the use of NLTK in a Jupyter Notebook for natural language data processing. NLTK offers diverse tools for tokenization, stemming, tagging, and more, fostering collaboration through its global forum. NumPy is employed for mathematical functions, aiding data operations during preprocessing. Pandas proves crucial for data cleaning and preparation, addressing the initial state of the unprepared dataset. Matplotlib and Seaborn enhance graphical representation, providing functions for drawing graphs and visualizing data efficiently in the Jupyter Notebook environment.. This cohesive integration ensures a strong foundation for executing and analyzing machine learning models in the project.

## 6. GENERAL WORKFLOW

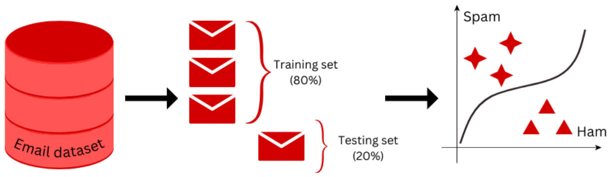


*Fig.1 - Instance Gathering, Training and Testing and Classification Process*

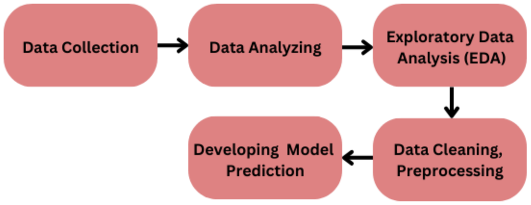

Data Collection
Data Analyzing
Exploratory Data Analysis (EDA)
Data Cleaning, Preprocessing
Developing Model Prediction

*Fig. 2 - Workflow*

## 7. DATA SET

The used dataset contains a collection of email text messages, labeled as either spam or not spam. Each email message is associated with a binary label, where "1" indicates that the email is spam, and "0" indicates that it is not spam. The dataset is intended for use in training and evaluating spam email classification models.

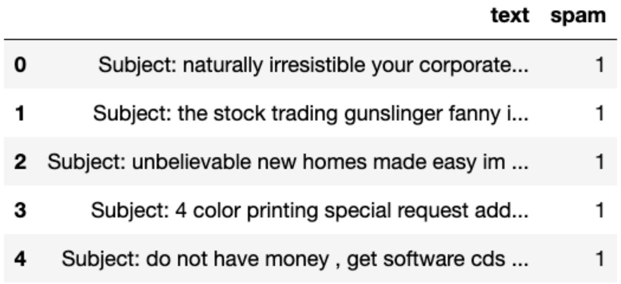
**Columns:**

**text (Text)**: This column contains the text content of the email messages. It includes the body of the emails along with any associated subject lines or headers.

**spam_or_not (Binary)**: This column contains binary labels to indicate whether an email is spam or not. "1" represents “spam”, while "0" represents “not spam/ham”.

**Usage**:

This dataset can be used for various Natural Language Processing (NLP) tasks, such as text classification and spam detection. Researchers and data scientists can train and evaluate machine learning models using this dataset to build effective spam email filters.



| | text | spam |
|---|---|---|
| **0** | Subject: naturally irresistible your corporate... | 1 |
| **1** | Subject: the stock trading gunslinger fanny i... | 1 |
| **2** | Subject: unbelievable new homes made easy im ... | 1 |
| **3** | Subject: 4 color printing special request add... | 1 |
| **4** | Subject: do not have money , get software cds ... | 1 |

*Fig.3 - Initial dataset*

## 8. EDA

Exploratory Data Analysis, simply referred to as EDA, is the step where you understand the data in detail. You understand each variable individually by calculating frequency counts, visualizing the distributions, etc. Also the relationships between the various combinations of the predictor and response variables by creating scatterplots, correlations, etc. EDA is typically part of every machine learning / predictive modeling project, especially with tabular datasets.

**The main steps of EDA are:**

1. Understand which variables could be important in predicting the Y (response).

2. Generate insights that give us more understanding of the business context and performance.

The main idea is to study the relationships between variables using various visualizations, statistical metrics (such as correlations) and significance tests, in the process, draw insights that may put the predictive modeling problem in context. Some of the insights can be eye-openers to your clients.

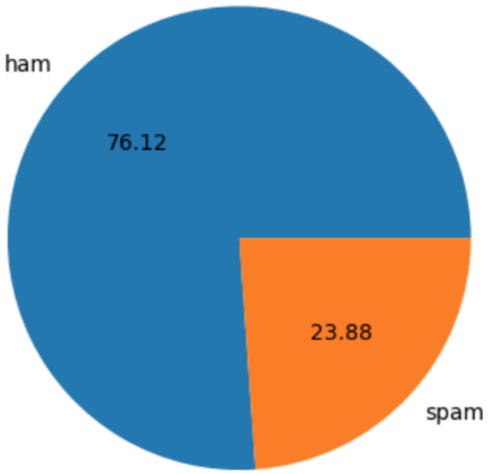


Fig. 4 - Ham/spam occurrence

**Heat Map [Fig. 5] explanation**:

1. A color gradient is used to indicate the magnitude or strength of the values.

2. Each cell in the heatmap corresponds to an intersection of two variables in the matrix.

3. The intensity or darkness of the color indicates the strength of the relationship.

4. Numerical annotations inside each cell provide the exact values of the matrix.

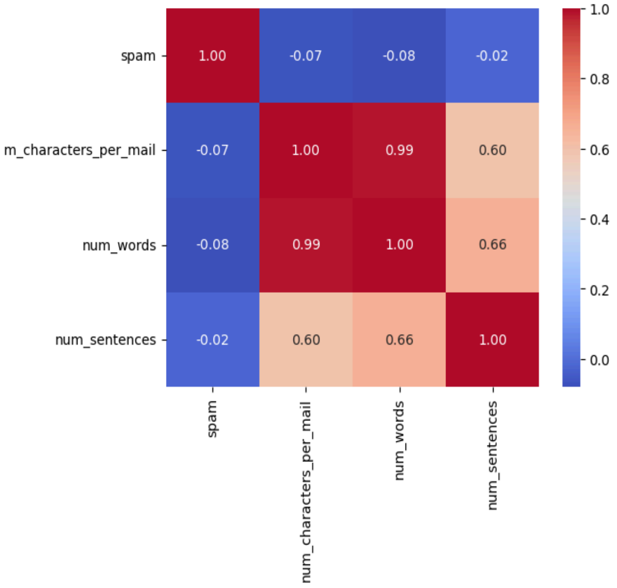


Fig. 5 - Heatmap

After counting and appearing the number of characters, words and sentences into separate columns, we can see their visual comparison representation in the figure [Fig.6] below.

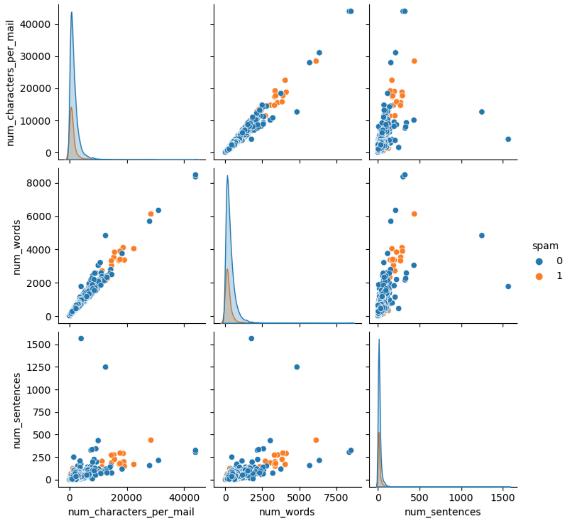
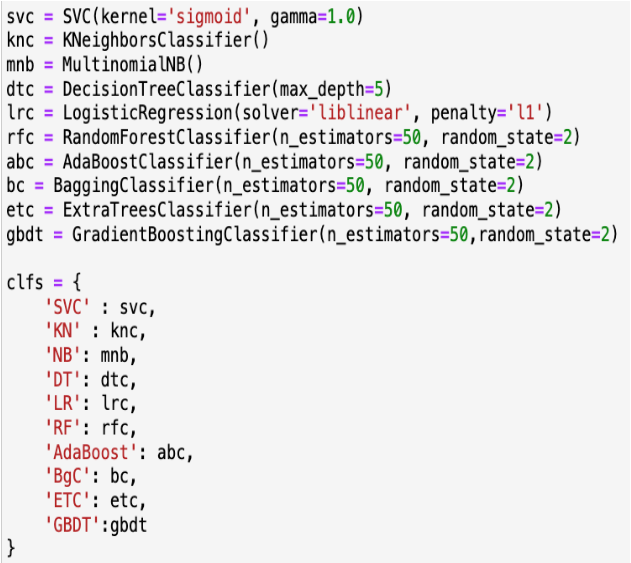


Fig. 6 - Characters, words and sentences relational graph

## 9 . DEVELOPING MODEL PREDICTION

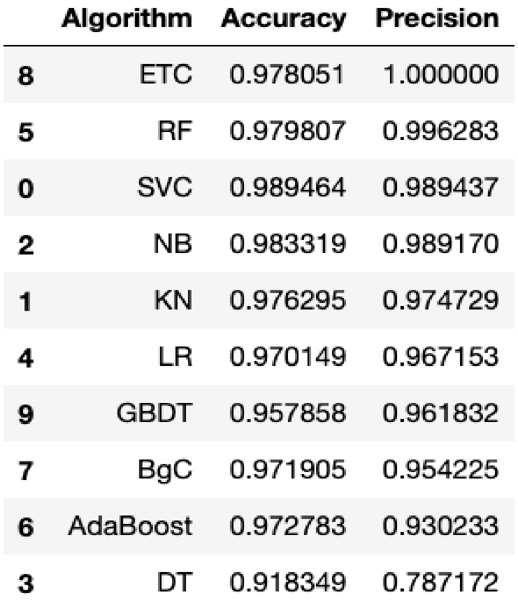
Dividing our dataset by principle 80-20 and using following classifiers on it we get accuracy and precision results as shown below [Fig. 7, Fig.8]:



```
svc = SVC(kernel='sigmoid', gamma=1.0)
knc = KNeighborsClassifier()
mnb = MultinomialNB()
dtc = DecisionTreeClassifier(max_depth=5)
lrc = LogisticRegression(solver='liblinear', penalty='l1')
rfc = RandomForestClassifier(n_estimators=50, random_state=2)
abc = AdaBoostClassifier(n_estimators=50, random_state=2)
bc = BaggingClassifier(n_estimators=50, random_state=2)
etc = ExtraTreesClassifier(n_estimators=50, random_state=2)
gbdt = GradientBoostingClassifier(n_estimators=50,random_state=2)

clfs = {
    'SVC' : svc,
    'KN' : knc,
    'NB': mnb,
    'DT': dtc,
    'LR': lrc,
    'RF': rfc,
    'AdaBoost': abc,
    'BgC': bc,
    'ETC': etc,
    'GBDT':gbdt
}
```

Fig.7 - Code snippet on classifiers

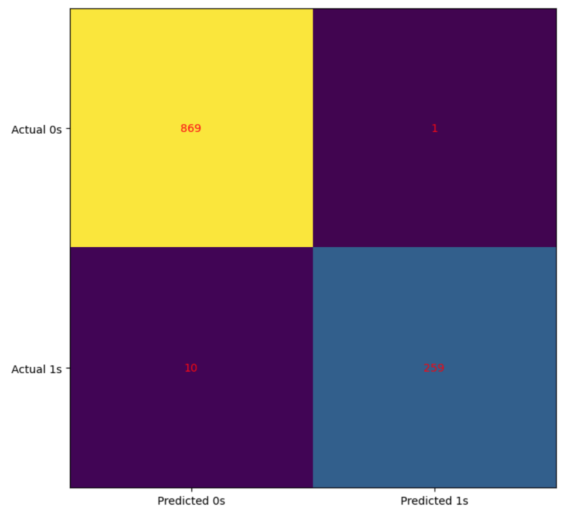


| | Algorithm | Accuracy | Precision |
|---|---|---|---|
| 8 | ETC | 0.978051 | 1.000000 |
| 5 | RF | 0.979807 | 0.996283 |
| 0 | SVC | 0.989464 | 0.989437 |
| 2 | NB | 0.983319 | 0.989170 |
| 1 | KN | 0.976295 | 0.974729 |
| 4 | LR | 0.970149 | 0.967153 |
| 9 | GBDT | 0.957858 | 0.961832 |
| 7 | BgC | 0.971905 | 0.954225 |
| 6 | AdaBoost | 0.972783 | 0.930233 |
| 3 | DT | 0.918349 | 0.787172 |

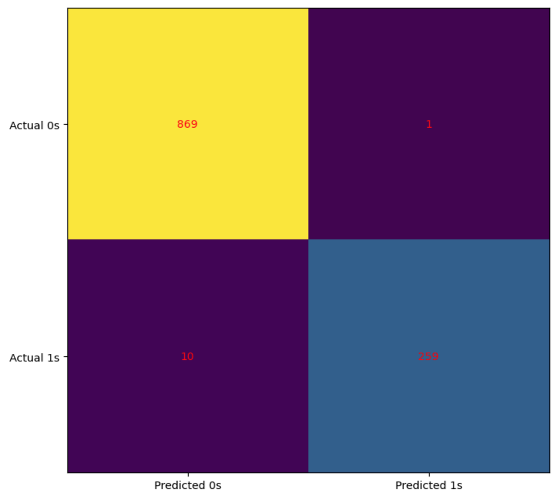
Fig. 8 - Used classifiers and their scores

## 10. CONCLUSION

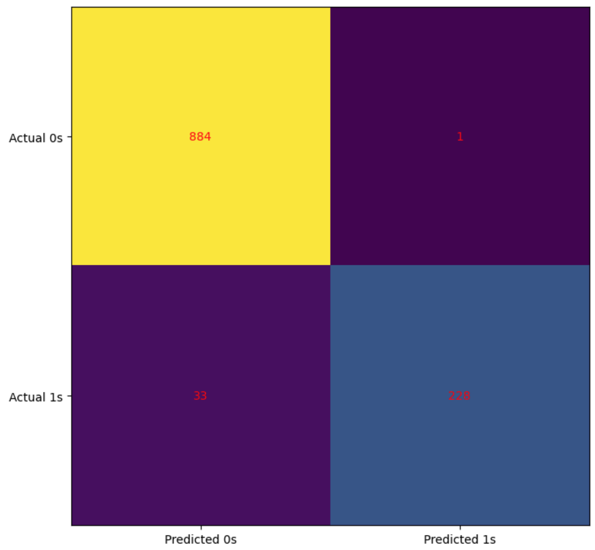
In conclusion, this research project underscores the significance of leveraging Python, particularly within the Jupyter Notebook environment, for effective natural language data processing and machine learning. The utilization of NLTK has empowered the project with tools for language-related tasks, fostering collaboration through its global forum. The inclusion of NumPy, Pandas, Matplotlib, and Seaborn has formed a robust foundation for data preprocessing, analysis, and visualization. This comprehensive integration of libraries ensures a streamlined workflow, enhancing the efficiency of machine learning models developed within the project. By addressing the complexities of email classification through the application of ten classifiers, this project aims to contribute valuable insights to the broader domain of spam detection. The diverse array of libraries employed demonstrates the project's commitment to methodological rigor and the utilization of cutting-edge tools in pursuit of accurate and efficient results. As the project unfolds, it is anticipated that these methodologies and findings will provide a meaningful contribution to the field of machine learning-based email classification, further advancing our capabilities in combating the persistent challenge of spam emails. The significance of this project lies in its potential to revolutionize email classification and spam detection through the strategic integration of machine learning techniques. In an era where digital communication is ubiquitous, the rampant increase in spam emails poses a serious threat, consuming valuable resources, time, and bandwidth. By employing a sophisticated approach with ten classifiers, this project seeks to enhance the accuracy and efficiency of spam detection, ultimately mitigating the adverse impact of unwanted emails on individuals, organizations, and online platforms. The utilization of machine learning, coupled with the extensive use of Python libraries like NLTK, NumPy, Pandas, Matplotlib, and Seaborn, signifies a cutting-edge methodology. This approach not only ensures the reliability of results but also demonstrates the adaptability of contemporary technologies to address evolving challenges. As spam emails continue to evolve in sophistication, the outcomes of this research project are poised to contribute substantially to the ongoing efforts in bolstering cybersecurity. The development of robust email classification models could lead to more secure and efficient communication channels, safeguarding users from potential threats, phishing attacks, and information breaches. Ultimately, the project's findings have the potential to reshape the landscape of email security, making a meaningful impact on the digital ecosystem and reinforcing the importance of advanced machine learning techniques in ensuring a safer online environment.



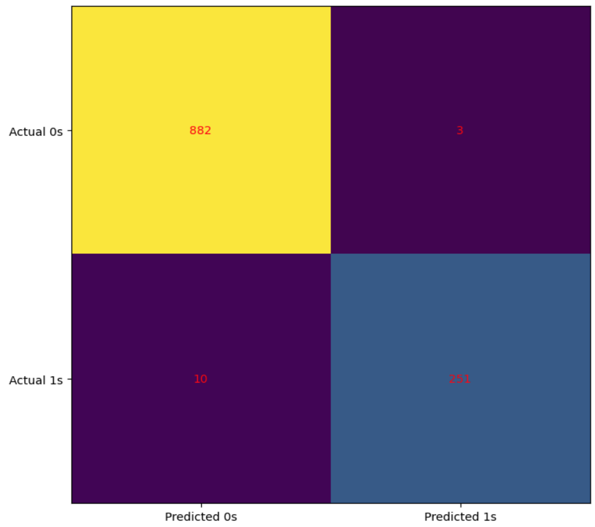
```
Testing  set score for SVC: 1.000000
```



```
Testing  set score for ETC: 1.000000
```



```
Testing  set score for AdaBoost: 0.986687
```



```
Testing  set score for Naive Bayes: 0.993889
```

[Fig. 9,10,11,12] - Classifier scores for conclusion (SVC, ETC, Adaboost, Naive Bayes)

## 11. REFERENCES

[1] Tom M. Mitchell. (n.d.). GENERATIVE AND DISCRIMINATIVE CLASSIFIERS: NAIVE BAYES AND LOGISTIC REGRESSION.

[2] Python | Stemming words with NLTK. (2023, April 15). GeeksforGeeks. https://www.geeksforgeeks.org/python-stemming-words-with-nltk/

[3] Rayan, A., & Taloba, A. I. (2021). Detection of email spam using natural language processing based random forest approach. https://doi.org/10.21203/rs.3.rs-921426/v1

[4] Isra'a AbdulNabi , Qussai Yaseen - The 2nd International Workshop on Data-Driven Security (DDSW 2021) March 23 - 26, 2021, Spam Email Detection Using Deep Learning Techniques Isra'a AbdulNabi∗ , Qussai Yaseen

[5] Jasneet Kaur - A STUDY ON SPAM DETECTION WITH SENTIMENT ANALYSIS https://www.mililink.com/upload/article/2008948707aams_vol_215_march_2022_a27_p2695-2705_jasneet_kaur.pdf

[6]What is NLP? - https://www.ibm.com/topics/natural-language-processing#:~:text=the%20next%20step-,What%20is%20natural%20language%20processing%3F,same%20way%20human%20beings%20can.

[7] Ravi Kiran SS, Atmosukarto S. Spam or not spam–that is the question. Citeseerx. 2009.

[8] Iqbal, K. and Khan, M.S. (2022), "Email classification analysis using machine learning techniques", Applied Computing and Informatics, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/ACI-01-2022-0012

[9] Selva Prabhakaran - Exploratory Data Analysis (EDA) – How to do EDA for Machine Learning Problems using Python

[10] Ezpeleta, E., Zurutuza, U., & Gómez Hidalgo, J. M. (2016). Does sentiment analysis help in Bayesian spam filtering? Lecture Notes in Computer Science, 79-90