EMAIL SPAM DETECTION USING LLM- LIBRARY

Dhavakumar. P, R.Sai Lakshmi ,M. Divya Dharshini

Vellore Institute of Technology, Chennai.

Dhavakumar.p@vit.ac.in

Sailakshmi.r2022@vitstudent.ac.in

Divyadharshini.m2022@vitstudent.ac.in

ABSTRACT:

One of the biggest problems on the internet is spam email. Spam is used for phishing, fraud, and illegal and unethical behavior. Spammers send malicious links through emails by faking email accounts and profiles. This causes financial harm to businesses as well as annoyance and frustration for individual email users. The purpose of this work is to present a machine learning-based method for distinguishing between valid (ham) emails and spam emails using the Sckit-LLM library. The Sckit-LLM library, which combines Scikit-Learn with potent language models like ChatGPT, is a game changer in text analysis. We can find context, sentiment, and hidden patterns in a variety of textual data sources by using Sckit-LLM.

Keywords:

Machine Learning, Sckit – LLM , Chatgpt , Sklearn, Zero ShotGPTClassifier.

INTRODUCTION:

Through the use of the Scikit-Learn library, we are able to import the ZeroShotClassifier algorithm into our dataset. This experiment is carried out with the ZeroShotClassifier algorithm, which is capable of classifying text without requiring special training. Using this algorithm, we are able to achieve an accuracy of 75% based on the Scikit-Learn library. The practice of "using email to send unsolicited emails or advertising emails to a group of recipients" is known as email spam, or electronic mail. When an email is sent that is not requested, the recipient has not given permission to receive it.

Since last year, using spam emails has become more and more common. 10 years. On the internet, spam has grown to be really unfortunate. Spam is a time, storage, and message speed waster. Although automatic email filtering is perhaps the best way to identify spam, spammers may now easily get around all of these spam filtering programs. A machine learning approach will be used to detect spam. The majority of approaches adopted closer to junk mail filtering include "text analysis, white and blacklists of domain names, and community-primarily based techniques."

Text assessment of mail contents is a widely used method to the spam; many answers deployable on server and purchaser aspects are available. Zeroshotclassifier is an algorithm which was import from the Sckit - LLM Library. This Sckit - LLM is a combination of Chatgpt + scikit-learn. This Sckit- LLM ensures that the response it receives actually contains a valid label. If not, Scikit-LLM will select a label at random while taking into account the likelihoods associated with the labels' occurrences in the training set. To put it another way, Sckit - LLM takes care of the API issues and ensures that you receive useful labels. It even chooses a label for you in the event that a response is missing one, taking into account how frequently the response occurred in the training set.

On the other hand, labeled data is not even necessary for model training. All we have to do is submit a list of labels, such as "ham" or "spam." Spam and Ham: People generally don't realize they just signed in those mailers when they click their link or download any free services, software or updating. Through this zeroshot classifier we detect this spam or ham emails.

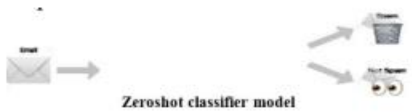


Fig: 1. Zeroshot Classifier Model

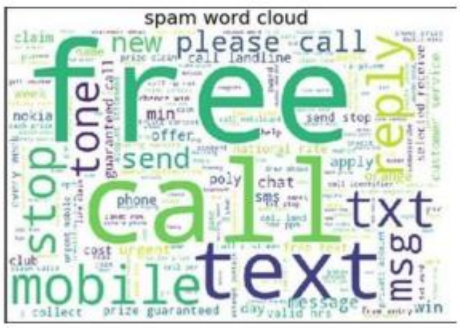


Fig:2. spam word cloud

Fig: 3. ham word cloud

LITERATURE SURVEY:

There is some related works that are apply machine learning methods in email spam detection. But by using Sckit – LLM library there is no related works that are done in email spam detection. This Sckit – LLM is a library which was developed by Iryna Kondrashchenko Data Scientist at Tractive & Msc student at JKU Linz and was launched this Sckit – LLM library on may (2023).

METHODOLOGY:

Data processing:

Perform data cleaning by removing irrelevant information such as null values and duplicate entries. Normalize the text data for consistency and improved processing. An extremely large data set with a huge number of rows and columns will always be observed when the data is taken into consideration. However, this isn't necessarily the case; the data could exist in a variety of formats. Audio, video, and image files tables with structure, etc.

The machine cannot read text data as it is, photos, or videos. A machine can only interpret 1 and 0. the team involved in data processing. data integration: The process of merging and integrating data from severalsources to offer a thorough perspective for analysis, reporting, and decision-making is known as data integration. This is especially crucial in situations when businesses store data in many forms or systems. Aunified perspective of the data is provided by effective data integration, empowering organizations to make decisions based on a thorough understanding of theirinformation landscape. It improves data quality, cuts down on duplication, and facilitates more dependable and effective analysis.

Data transformation:

The process of normalization and grouping were accomplished to scale to a particular value. The aspects of data transformation: Normalization and Standardization, Cleaning and Validation, Encoding of Categorical Data, Feature Engineering, Grouping and Aggregation, Managing Time and Date, Re scaling Processing Text Data; Processing Skewed Data; Processing Unbalanced Data. Similar to what was mentioned in the previous response, data transformation can be a part of data reduction as well. This includes normalizing, standardizing, or transforming variables to simplify analysis.

Compression of information: This part extracts a brief overview of the dataset, which is very tiny in size but yields the same analytical conclusion thus far. The particular objectives of the investigationand the properties of the dataset should be taken into consideration while selecting data reduction strategies. Achieving a balance between simplifying and preserving the crucial data required for significant analysis or modeling is crucial. The methods used in Compression of information are Feature selection, Dimensionality reduction, Aggregation, Binning or Histogramming, Filtering and Smoothing, and Clustering stop word: Any word that doesn't significantly deepen a sentence's meaning is a stop word. The sentence's meaning can be preserved even if they are disregarded. These include some of the most often used short function terms for certain search engines, like the, is, at, which, and on.

Here, stop words might be problematic when looking for phrases that contain them, especially in names like "The Who" or "Take That." Tokenization: In natural language processing (NLP), the process of tokenization entails dividing a text into smaller pieces known as tokens. Usually, tokens are words, phrases, symbols, or other significant components. Tokenization's primary objective is to make text processing and analysis easier.

Types of tokenization are Sentence Tokenization, Word Tokenization, Whitespace Tokenization, Punctuation Tokenization, Morphological Tokenization For examples: input: email spam detection using machine learning Output: “email”,” spam “,” detection”, ” using”,” machine”,” learning”. Removing Redundant Information: Redundant information, which does not contribute significantly to the analysis, can be removed. This might involve eliminating duplicate records or features that are highly correlated.

Data Understanding:

In machine learning, data understanding refers to the process of gaining insightsand knowledge about the dataset that will be used for training a model. One of the most important phases in the larger process of creating and implementing a machine learning model is "data understanding" in machine learning. Gaining knowledge about the composition, traits, and connections in the dataset that will be used to train and assess the machine learning model is known as data understanding. Understanding data paves the way for next phases of the machine learning pipeline, such as model selection, evaluation, and data preprocessing. Throughout the model constructionprocess, it facilitates data scientists and machine learning professionals in making well□informed judgments. In Our dataset there are 5171 records and 2 labels.

Checking with Null Values:

Checking for null values is an essential step in the data understanding phase ofmachine learning. In our dataset there is no null values are there. Null values may have an adverse effect on how well machine learning models function and produce biased or inaccurate outcomes. Identifying and managing null values correctly is crucial, taking into account the objectives of the study as well as the characteristics of the data. Several techniques, including imputation, removing rows or columns, or more complex approaches, may be used, depending on the amount of missing data.

Checking with Duplicates:

Checking for duplicates in our dataset is an important step in data understanding. Duplicate values may affect the accuracy and dependability of the results and cause bias in the model's performance. It's crucial to carefully examine duplicate handling in light of your machine learning task's particular requirements. You can decide to eliminate duplicates, retain only the initial instance, or handle them in a way that supports the objectives of your study, depending on the circumstances. Duplicates can skew the analysis and modeling process, leadingto inaccurate results. In our dataset there is no duplicate values are there.

Exploratory data analysis for target variable using pie chart

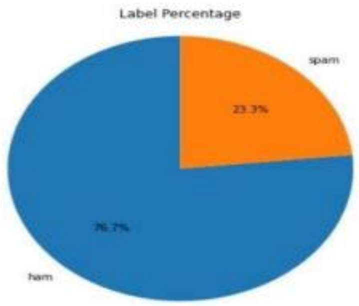
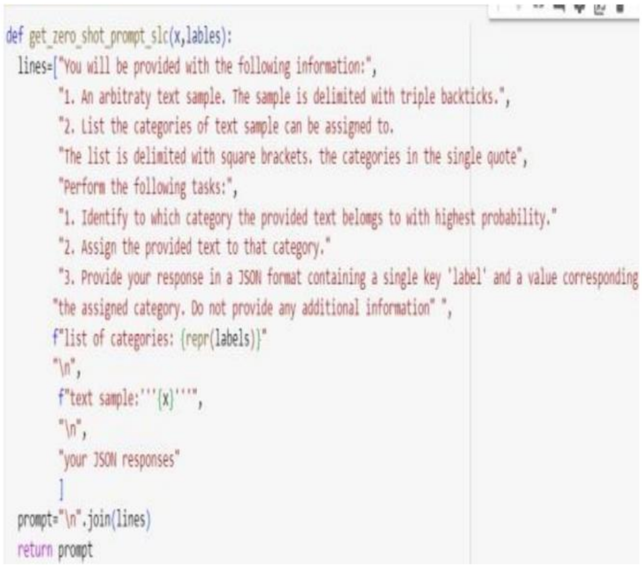


Fig.4. Percentage of Spam and ham in Pie chart

With the reference ofthe above pie chart we came to know that there are 30% (6) of spam and 70% (14) of ham data are there which showsthat the dataset are imbalance.

ZEROSHOTGPT CLASSIFIER MODEL:

The Sckit - LLM Library is the source of the ZeroShotGPTClassifier model import.Text categorization with zero shots is the purpose of the ZeroShotClassifier module. It lets you categorize text into several pre-established classes without requiring unique training examples for every class. Rather than requiring samples of a certain class to be seen during training, you give the model a list of candidate labels or classes, and it uses them to forecast the likelihood that the input text belongs to each class. The ZeroShotClassifier module creates embeddings for the input text and the candidate labels by using huge pre-trained language models such as GPT. In order to generate predictions, it then computes how similar each label is to the text. Usually, you instantiate a pre-trained model (such GPT-3 or GPT-4) and load it using the Transformers library, then use the ZeroShotClassificationPipeline class to use the ZeroShotClassifier module. Pre-trained, this ZeroShotGPTClassifier model is a language model that operates depending on the promp.



```
def get_zero_shot_prompt_slc(x,lables):
  lines=["You will be provided with the following information:",
        "1. An arbitraty text sample. The sample is delimited with triple backticks.",
        "2. List the categories of text sample can be assigned to.
        "The list is delimited with square brackets. the categories in the single quote",
        "Perform the following tasks:",
        "1. Identify to which category the provided text beloмgs to with highest probability."
        "2. Assign the provided text to that category."
        "3. Provide your response in a JSON format containing a single key 'label' and a value corresponding
       "the assigned category. Do not provide any additional information" ",
       f"list of categories: {repr(labels)}"
       "\n",
        f"text sample:'''{x}'''",
        "\n",
        "your JSON responses"
        ]
  prompt="\n".join(lines)
  return prompt
```

Fig.5. ZeroShot GPT classifier prompt

RESULT:

For increased accuracy, our model has been pre-trained to verify and contrast the outcomes. The user will receive the evaluated results from each classifier. The user can compare the results with other results to determine whether the data is "spam" or "ham" once all of the classifiers have returned their findings. For easier comprehension, graphs and tables will be displayed for each classification result. For training, the dataset is downloaded from the "Kaggle" website. "spam_ham_dataset.csv" is the name of the used dataset.

Accuracy Score

	precision	recall	f1-score	support
Spam	0.00	0.00	0.00	0
ham	0.97	0.94	0.95	32
spam	0.00	0.00	0.00	8
accuracy			0.75	40
macro avg	0.32	0.31	0.32	40
weighted avg	0.77	0.75	0.76	40

Fig.6. ZeroShot GPT classifier prompt

CHECKING OUR MODEL:

This is an intriguing theory that suggests the model won't function when examined separately. It functions with two reviews minimum. Predictions on individual samples are not supported by the ZeroShotGPTClassifier from skllm. Therefore, in order to forecast the corresponding labels, we must provide at least two emails.

```
input_text1 = input("Enter review: ")
input_text2 = input("Enter review: ")

df4 = pd.DataFrame({'Review':[input_text1, input_text2]})

# Make predictions on the input text
predicted_label = clf.predict([df4])


df5 = pd.DataFrame({'Review':[input_text1, input_text2],'predicted label': predi

print(df5)

Enter review:
```

Fig: 7. Prediction

Result from Model

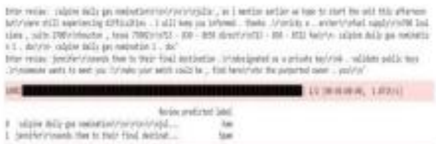


Fig: 8.Result

Overall Result

	Email Notification	predicted label	Original label
0	Subject: enron methanol ; meter # : 988291\r\n...	ham	ham
1	Subject: hpl nom for january 9 , 2001\r\n(see...	ham	ham
2	Subject: neon retreat\r\nho ho ho , we ' re ar...	ham	ham
3	Subject: photoshop , windows , office . cheap ...	Spam	spam
4	Subject: re : indian springs\r\nthis deal is t...	ham	ham
5	Subject: ehronline web address change\r\nthis ...	ham	ham
6	Subject: spring savings certificate - take 30 ...	Spam	ham
7	Subject: looking for medication ? we ' re the ...	Spam	spam
8	Subject: noms / actual flow for 2 / 26\r\nwe a...	ham	ham
9	Subject: nominations for oct . 21 - 23 , 2000\...	ham	ham

Fig.9. Comparison of Predicted label and original label based on model performance.

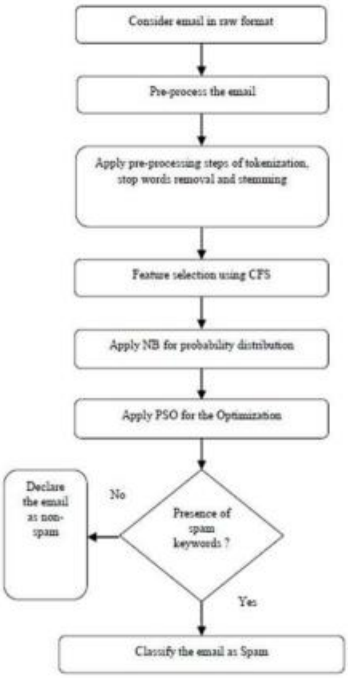


Fig: 10. model flow chart

CONCLUSION:

Our model's output allows us to quickly identify spam emails using chatGPT, an AI tool. Our project has a great deal of room for improvement. The following enhancements are possible: "Filtering of spams can be done on the basis of the trusted and verified domainnames." "The spam email classification is very significant in categorizing e-mails and to distinct e-mails that are or non-spam." "This method can be used by the big body to differentiate decent mails that are only the emails they wish to obtain.

REFERENCE:

[1] Mangena Venu , Sagar Pande, Pooja Umekar , Tushar Mahore. Comparative Analysis of Detection of Email Spam With the Aid of Machine Learning Approaches (2020).

[2] Thashina Sultana , K A Sapnaz , Fathima Sana , Jamedar Najath. Email basedSpam Detection (2020).

[3] Manoj Sethi, Sumesha Chandra, Vinayak Chaudhary, Yash.

[4] Email Spam Detection using Machine Learning and Neural Networks(2021).

[5] Sanjay Malik, Pooja Malhotra. Spam Email Detection Using Machine Learning and Deep Learning Techniques (2022).

[6] Hari k.c. comparative analysis and prediction of spam email classification using supervised machine learning techniques (2021).

[7] Rajesh Kumar J, Sudarshan P, Mahalakshni G. Email Spam Detection using Machine Learning Techniques. (2023).

[8] Nikhil Kumar, Sanket Sonowal, Nishant. Email Spam Detection Using Machine Learning Algorithms (2022).

[9] Shehan Sanjula. Spam Detection in Email using Machine Learning (2022).

[10] Hrithik Vohra , Manoj Kumar . Email Spam Detection Using Naive Bayes(2023).

[11] Khalid Iqbal, Muhammad Shehrayar Khan. . Email Classification analysisusing Machine Learning techniques. (2022)