Intelligent Systems

ITCS 6150 Spring'22

Spam Filtering API for Text/Email Messages

Submitted By:

Dayakar Ravuri – 801170140.
Yashwanth Kuricheti – 801202893.
Tejaswini Reddy Kolli – 801218910.
Naveen Kumar Kannegundla – 801210433.
Bhargava Ram Bonala – 801208206.
Santhosh Rayasam – 801262383.

Abstract:

The main motive of our project is to build a spam filtering system which when provided with message the system will predict the spam in the given message as this is a generalized creation of spam filtering, we can use anywhere like emails or message service providers etc. E-mail and messages are becoming one of the most widely used modes of communication. Spam e-mails, on the other hand, cause traffic congestion, lower productivity, and phishing, which has become a serious problem in our society. Every year, the amount of spam e-mails increases. As a result, spam e-mail filtering is an important, meaningful, and difficult problem to address.

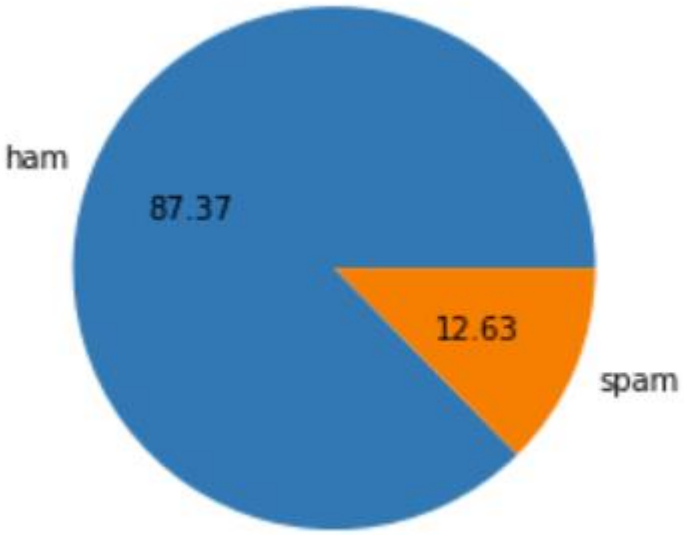
Introduction:

In a broad way, spam-filtering system is nothing but algorithms that are used for finding the spam messages in the messages this API can be used on for finding spam messages in email or messages etc. The Algorithm is creation of model based on the data provided and this model is used to predict when new text messages is provided. To create a model the data, need to be cleaned which is removing of all unnecessary or null values for the data. Once the data is cleaned Exploratory Data Analysis (EDA) is a used to analyze the data using visual techniques. In the next step of the algorithm Data preprocessing is done, Data preprocessing is the process of transforming raw data into an understandable format. Preprocessed data is being used to create the model with training and testing data, A machine learning model is built by learning and generalizing from training data, then applying that acquired knowledge to new data it has never seen before to make predictions and fulfill its purpose. Once the model is built which can be used in the real world for spam-filtering. We took the following dataset from Kaggle: **https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset**

Data Cleaning:

The process of ensuring that data is correct, consistent, and useable is known as data cleansing. You can clean data by looking for faults or corruptions, repairing or eliminating them, or manually processing data as needed to avoid repeating the same mistakes. The data set we used had 5 columns out of which there are 3 columns are NaN – not a null value as a data cleaning process we had removed these columns.

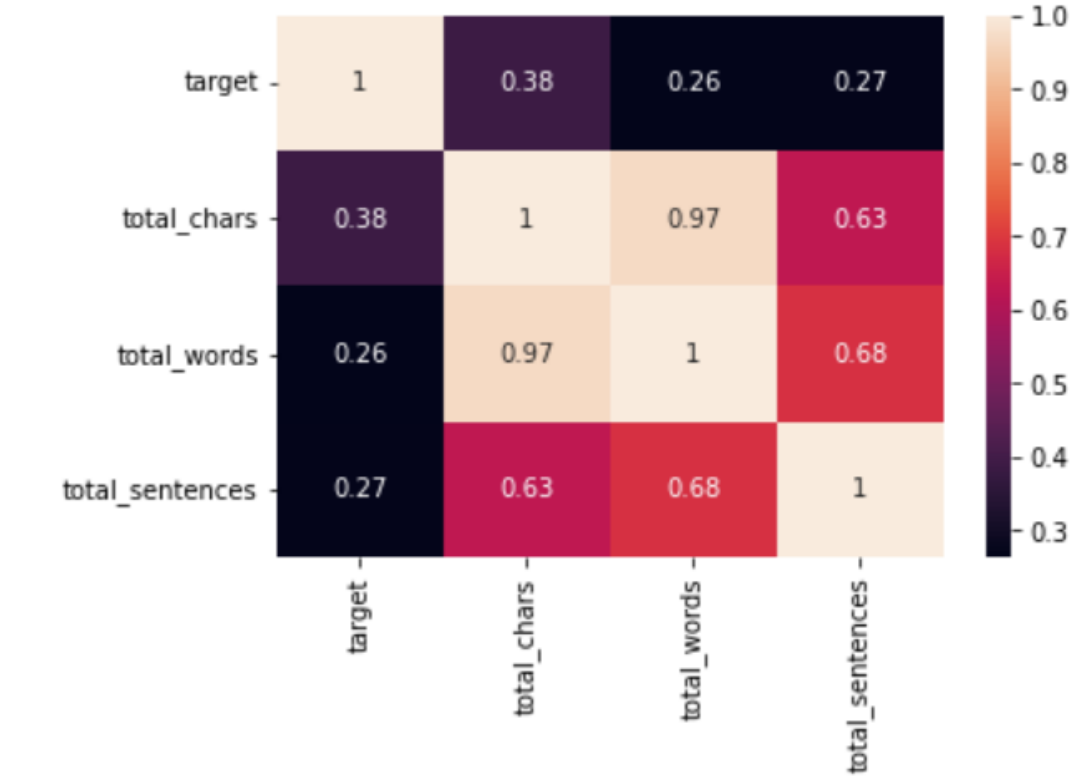
After this we have clean the null values in the data as they don't have any significance to the data modeling or may diminish the performance of the model and we even made sure we don't have any null values. The next step is removing duplicate values in the data, even duplicates values may diminish the performance of the model. This cleaning makes the data clean and can be used for modeling. The next step is labeling data based on ham and spam or we can say it as transformation of data. Below is the pictorial representation of the transformed data after data cleaning, in the transformed data we have approximately 87 %of ham data and the rest classified as spam data.



Exploratory Data Analysis (EDA):

Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. It is a good practice to understand the data first and try to gather as many insights as possible from it. EDA is all about making sense of data in hand, before getting them dirty with it. In the Exploratory data analysis of the data, we will find the number of words, number of tokens and number of sentences for each message from the data. As to get the detailed analysis of the data describe function is

used, as a part detailed analysis, count, medium, minimum, maximum, mean etc. details of tokens, words and sentences are extracted. The heat map represents the data visualization technique that shows magnitude of a phenomenon as color in two dimensions.



Data Preprocessing:

Machines don't understand free text, image, or video data as it is, they understand 1s and 0s. So it probably won't be good enough if we put on a slideshow of all our images and expect our machine learning model to get trained just by that. Data preprocessing is the process of transforming raw data into an understandable format. In other words, the features of the data can now be easily interpreted by the algorithm. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms.

A feature is an individual measurable property or characteristic of a phenomenon being observed.

Categorical: Features whose values are taken from a defined set of values. For instance, days in a week: {Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday} is a category because its value is always taken from this set. Another example could be the Boolean set: {True, False}

Numerical: Features whose values are continuous or integer valued. They are represented by numbers and possess most of the properties of numbers. For instance, number of steps you walk in a day, or the speed at which you are driving your car at.

In the data preprocessing he given sentence will be made into token by removing punctuation and all tokens are in small case. Once the required tokens are generated the stop words present in the English language will be used and these words are removed from the sentences, as these words can less or no information , so they are. removed from the sentences for improving the performance of the model.

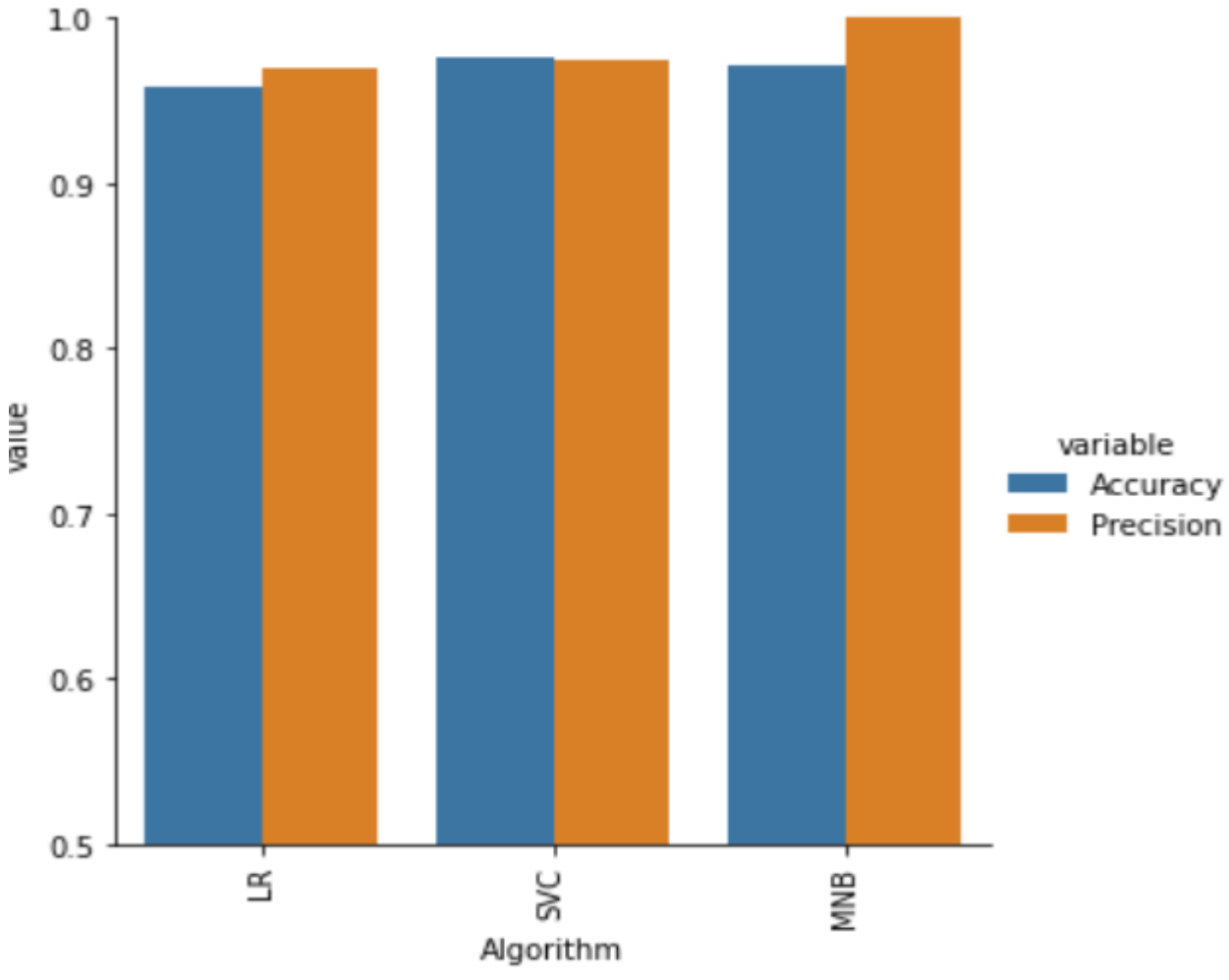
	target	text	total_chars	total_words	total_sentences	transformed_text
0	0	Go until jurong point, crazy.. Available only ...	111	24	2	go jurong point crazi avail bugi n great world...
1	0	Ok lar... Joking wif u oni...	29	8	2	ok lar joke wif u oni
2	1	Free entry in 2 a wkly comp to win FA Cup fina...	155	37	2	free entri 2 wkli comp win fa cup final tkt 21...
3	0	U dun say so early hor... U c already then say...	49	13	1	u dun say earli hor u c alreadi say
4	0	Nah I don't think he goes to usf, he lives aro...	61	15	1	nah think goe usf live around though

Model Building:

Machine learning models are powerful tools used to perform vital tasks and solve complex problems efficiently and effectively. An exponential increase in data across the modern world means organizations from a range of sectors are ready to deploy machine learning models. These models have a huge range of uses, whether machine learning in finance proactively monitoring bank transfers for signs of fraud, or machine learning in healthcare. The process of building a machine learning model is often complex, driven by specialists in data science. But an understanding of the process is important as machine learning is adopted by more and more organizations. This guide explores the basics of building a machine learning model, breaking the process up into six steps.

We are using more than one model in the project which will provide more insights into finding the most reliable model to work in real life. The models are MultinomialNB, Logistic Regression And SVC. For each model data is divided into train and test data. The train data is used to create each model and test data used to test the trained model. Over, that model's accuracy is calculated to find the performance and to know about reliability of the model.

	Algorithm	Accuracy	Precision
2	LR	0.958414	0.970297
1	SVC	0.975822	0.974790
0	MNB	0.970986	1.000000



Spam Filtering API

This is a spam filtering API to detect whether the given message is spam or not. We found the good performing algorithms such as multinomial Naive Bayes, Logistic Regression, Support vector classifier to compare the accuracy and precision and to find the better algorithm to detect the content whether it is spam or not. Later, we observed that Multinomial Naive Bayes gave 97% accuracy and 1.0 precision for this data sets. Hence, we can say clearly for the spam message containing words in this data set. it detects 97% of them. Therefore, we selected Multinomial Naive Bayes algorithm to predict it.

This API consists of only two requests:

Get request.

This is like homepage request for this API, and it shows the welcoming message as below in Json format.

Response*:*

["Welcome to spam Filtering API, please provide your content in post request body message attribute to determine whether it is spam or ham"]

Post Request:

This is the request where user sends his message or context to check whether it is spam or not. The user must sends his message in the Json format and in the message attribute and in the post request body. as below:

Request body:

{"message”: “Congratulations!! you have won 1000 call on this number to get your prize"}

Response:

{"Response": "Spam Message"}

Setup

- You should have pip and python installed in your system and python interpreter should be selected based on your virtual environment or whatever the version you have installed it.
- Visual studio or Py charm or any other text editors should be installed.
- Postman should be installed to test this application.

Execution:

1. You need to install all things in the requirements file:
   cd to project directory
   pip install -r requirements.txt
2. You need to download punkt and stop words package from nltk:
   Go to terminal
   python
   import nltk
   nltk.download('punkt')
   nltk.download('stopwords')
   exit()
3. Run this command to start the uvicorn server in terminal:
   uvicorn api:app –reload

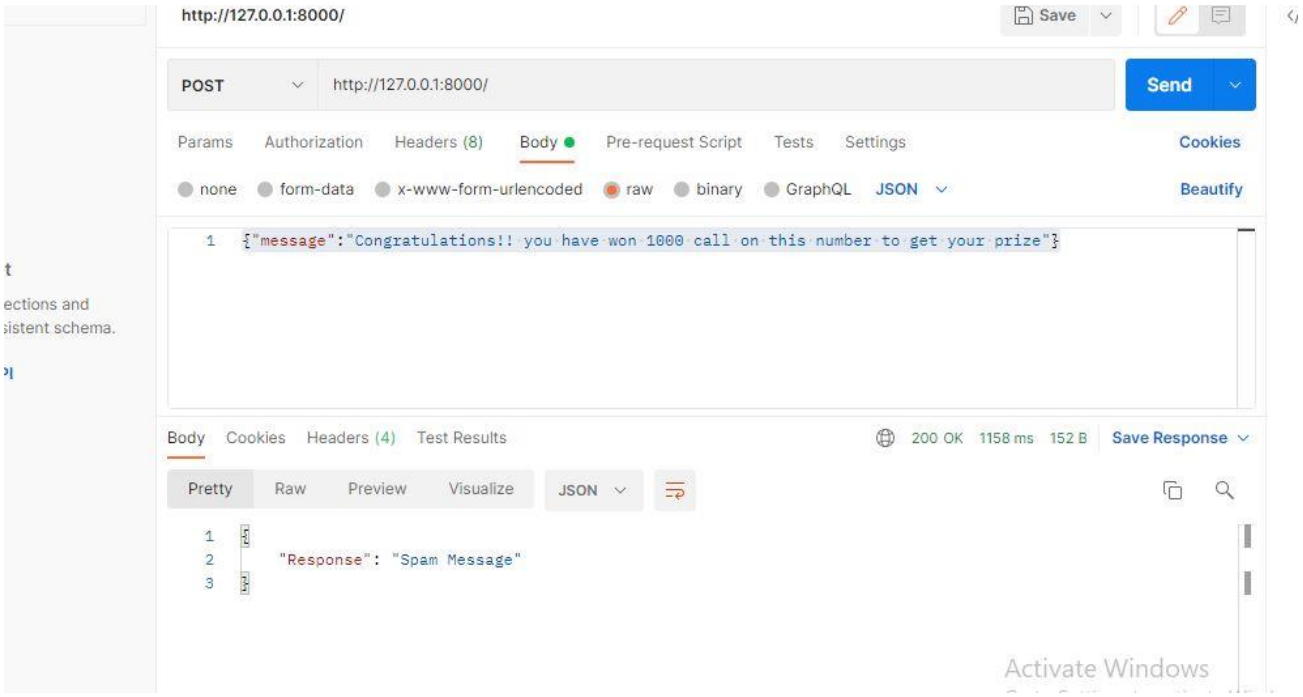
Output:

Spam message:

{"message":"Congratulations!! you have won 1000 call on this number to get your prize"}

Response:

{ "Response": "Spam Message" }



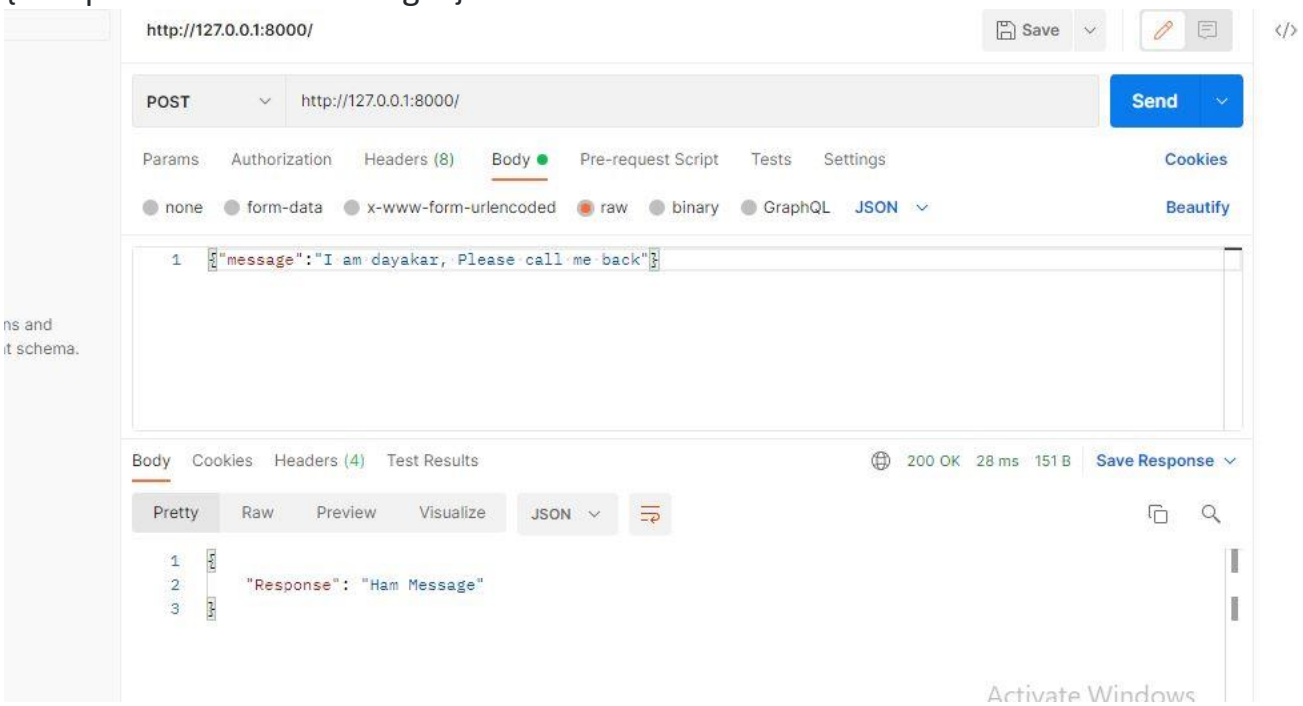
Ham message:

Request*:*

{"message":"I am dayakar, please call me back"}

Response:

{"Response": "Ham Message"}



Examples of messages:

Spam messages:

congratulations you won 1000 dollars call on this number to get your prize.

You could be entitled up to $5000 in compensation from mis-sold PPI on a credit card or Loan. Please reply PPI for in for STOP to opt out.

A Loan for $3000 is approved for you if you receive this sms. 1 min verification & cash in 1 hr at ww.[redacted].co.uk to opt out reply stop

Ham Messages:

Did you see the match ? It was insane.
I am Dayakar, please call me back.
Do you want to go to football match ?

Conclusion and Future Work:

Thus, we have developed a spam filtering API model using small data set and few machine learning algorithms. However, API which can detects all spam content can be built using a very big data set and many machine learning algorithms. That API can be used to integrate into any application like Gmail, Outlook, SMS, WhatsApp, and Discord. Additional features like redirecting and blocking spam messages can be included.

References:

1. https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset
2. https://fastapi.tiangolo.com/tutorial/
3. https://www.kaggle.com/code/adamschroeder/countvectorizer-tfidfvectorizer-predict-comments/notebook
4. https://www.sciencedirect.com/science/article/pii/S2405844018353404