Spam Prediction Using NLP

PRESENTED BY: SHREYANS SHAH(200460185)

Introduction

- In today's business world, email and messaging is a primary source of communication Spam emails have become a major issue as a result of the increased use of electronic communication.

- Apart from being annoying, spam emails and messages can also pose a security threat to computer system.

- Some of the most common types of spam emails and messages that pose a significant threat to security include fraudulent e-mails, identity theft, hacking, viruses, and malware, among other things.

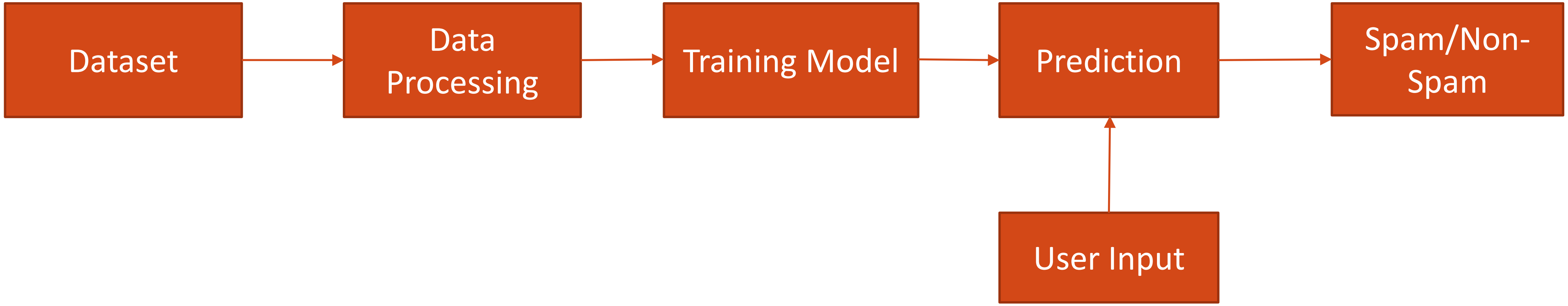
- As a result, the development of an autonomous system that would identify spam messages by identifying their pattern is the most important thing to do.

- In this project I'm using Kaggle dataset of spam and non-spam messages which will use to train the model, which will subsequently able to predict whether user messages are spam or non-spam.

Natural Language Processing

- Natural language processing aspires to create programs that can interpret and react to text or voice input, and then answer with text or speech of their own, in a manner similar to how people do it themselves.
- The most effective spam detection tools make advantage of natural language processing's text classification skills to analyze emails for terminology that is often associated with spam or phishing.
- Python have Natural Language Tool Kit (NLTK) Library to process Natural Language data.

System Flow Chart



DataSet

- ➢Kaggle Dataset : Spam Email Detation
- ➢Two Files: Sms_Train.csv , Sms_Test.csv

Train Dataset	Test Dataset
Spam Messages: 126	76
Non Spam Message:847	49

Sr.NO	Message_Body	Lable
1	30% discount for you	Spam
2	Complete work asap.	Non-Spam

Data Processing

- For Text processing it is necessary to pre process the data for better results
- Data Processing consist of 4 Process:

1) Punctuation Removal
2) Stop Words Removal
3) Lemmatization
4) Stemming

Stop Words

- Stop Words are the common words that is in every text input or sentences.
- Some example of Stop Words : A , An , The , You , Your
- These all words are unnecessary for training models as they are not helpful for text analysing
- Nltk have have default list of words that are common and using nltk methods we can remove this words.

Stemming And Lemmatization

- Stemming is a process that is used to remove affix from the words.
- Example: Eats become eat
- NLTK have : Porter , Lancaster, Regex , SnowBall Stemmers.

- Lemmatization is same like stemming but major difference is output of this will be the Original form of the word.
- Example : Believes become Believe

Training Model

Model	Accuracy
Linear SVC	96%
Multinomial Naive Bias	88%
Random Forest	81%
K-Nearest Neighbor	80%
Decision Tree	80%

Conclusion And Future Work

- This project is useful to detect spam messages.
- User can input the message for prediction
- For Future , I would like to focus on making user interface which will automatically store detected spam messages into training dataset for betterment of the accuracy of the model