ML Project Presentation

Project Category: Email Spam Prediction

Roll No.: 1901049

Name: Ayushman Singh Chauhan

Institute Name: Indian Institute of Information Technology Guwahati

Incoming Mails
Spam Filter
Spam mail to a holding area (or to the trash)
Non-spam mail to your inbox

Outline

- Introduction
- Objectives
- LITERATURE Survey / REVIEW
- Data understanding / Description
- NAÏVE BAYS CLASSIFIER
- Results
- Conclusion
- Novelty

Introduction / Problem definition

- Spam e-mails can be not only annoying but also dangerous to consumers.

- ▯ Unwanted e-mails irritating internet connection
- ▯Critical e-mail message are missed and / or delayed.
- ▯Millions of compromised computers
- ▯Billions of dollars lost worldwide
- ▯Identity theft
- ▯Spam can crash mail servers and fill up hard drives

Objective

- The objective of identification of Spam e-mails are :
- To give knowledge to the user about the fake e-mails and relevant e-mails
- To classify that mail spam or not.

Spam e-mails can be defined as :

1. Anonymity
2. Mass Mailings
3. Unsolicited:

Spam e-mail are message randomly sent to multiple addressees by all sorts of groups, but mostly lazy advertisers and criminals who wish to lead you to phishing sites.

LITERATURE Survey / REVIEW

I consulted from articles below that are all after year 2019

- Paper-1: [Email classification via intention-based segmentation] (https://sci-hub.mksa.top/10.23919/eecsi50503.2020.9251306)
- Paper-2: [Feature Extraction aligned Email Classification based on Imperative Sentence Selection through Deep Learning] (https://iecscience.org/uploads/jpapers/202108/rvRUTTmGYo1MiJ86KpRHOop8SqW8HWawp2xvrkn0.pdf)
- Paper-3: [Email Classification Research Trends: Review and Open Issues] (https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7921698)
- Paper-4: [Classification Of Power Relations Based On Email Exchange](https://sci-hub.mksa.top/10.1109/gucon48875.2020.9231072)

Data understanding / Description

- Initial Dataset: Enron-Email-Dataset

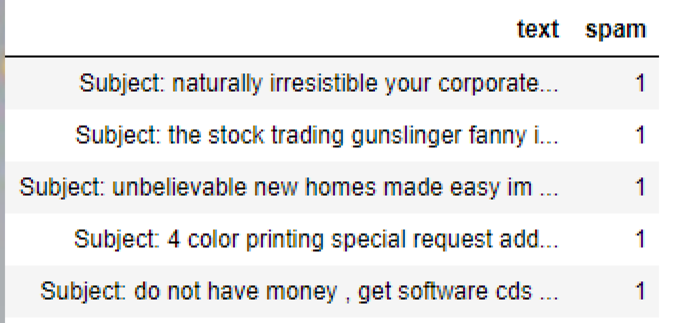
- Enron Email Dataset downloaded from : https://www.cs.cmu.edu/~enron/. And it is the May 7, 2015 Version of dataset.This email set is a gzipped tar file of emails stored in directories.
- This generates the following directory structure:
- maildir
- - $userName subdirectories for each user
- - $folderName subdirectories per user
- -mail messages in folder or additional subfolders This directory structure contains over 500,000 small mail files without attachments. These files all have the following layout:
- Message-ID: 31335512.1075861110528.JavaMail.evans@thyme

Date: Wed, 2 Jan 2002 09:26:29 -0800 (PST)

From: sender@test.com

To: rec1@test.com,rec2@test.com Cc: Bcc: Subject: Kelly Webb ... Message Body Some headers like To, Cc and Bcc or Subject can also be multiline values.

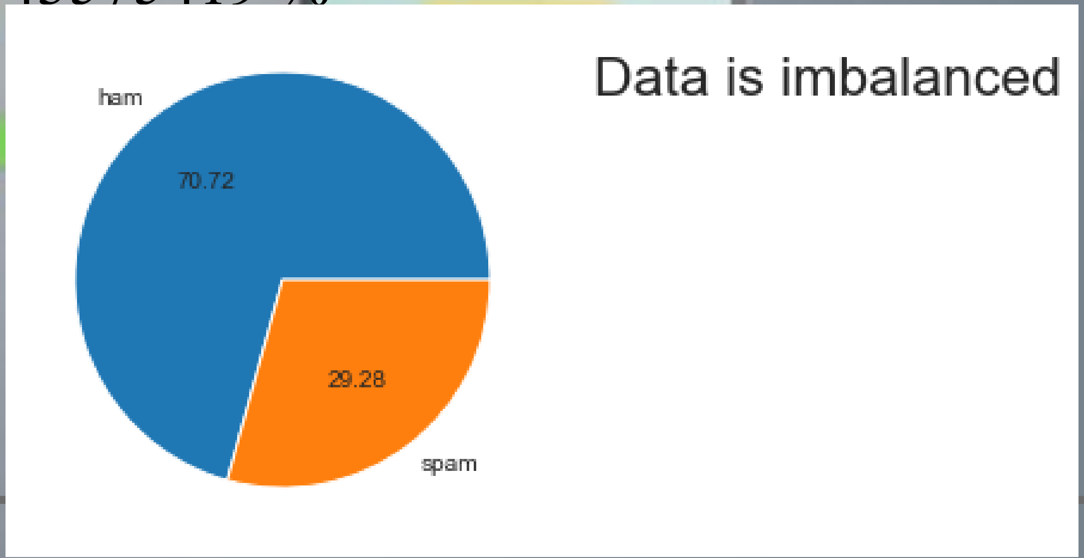
The new dataset contains two columns. The descriptive feature consists of text. The target feature consists of two classes ham and spam, the column name is spam. The classes are labeled for each document in the data set and represent our target feature with a binary string-type alphabet of {ham; spam}. Classes are further mapped to integer 0 (ham) and 1 (spam).



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

Spam email percentage in the dataset = 29.88268156424581 %

Ham email percentage in the dataset = 70.11731843575419 %



Data is imbalanced

Best Resust by : NAÏVE BAYS CLASSIFIER

NAÏVE BAYS CLASSIFIER SIMPLE PROBABILISTIC CLASSIFIER THAT CALCULATES A SET OF PROBABILITIES BY COUNTING THE FREQUENCY AND COMBINATION OF VALUES IN A GIVEN DATASET.

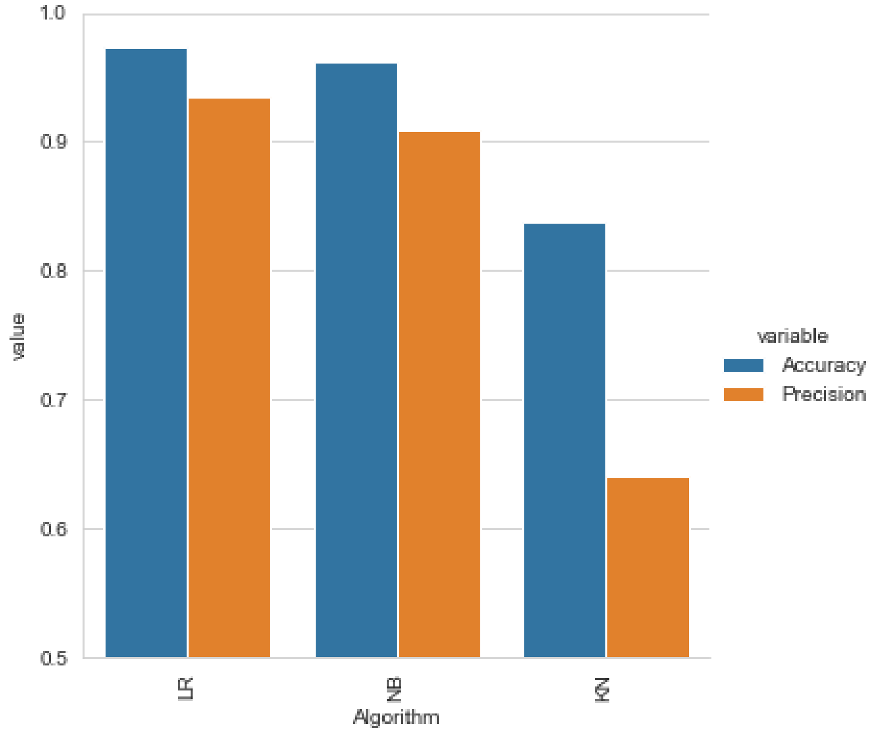
REPRESENT AS A VECTOR OF FEATURE VALUES.

IT IS VERY USEFUL TO CLASSIFY THE E-MAILS PROPERLY

THE PRECISION AND RECALL OF THIS METHOD IS KNOWN TO BE VERY EFFECTIVE

Results (in brief) and analysis of results

- Results (in brief) is :



- and analysis of results is in my excel file [Ayushman_1901049_Project_Result_Analysis.xlsx]..

CONCLUSION

- We are able to classify the emails as spam or non-spam. With high number of emails lots if people using the system it will be difficult to handle all possible mails as our project deals with only limited amount of corpus.

Novelty

- ● My project is Heroku Deployed
- ● Launch my Project (https://email-spam-cassifier.herokuapp.com/)

Thank You