SPAM CLASSIFIER

MISBAH UL HAQUE / misbahul8@gmail.com / +91 8376069521

CONTENT

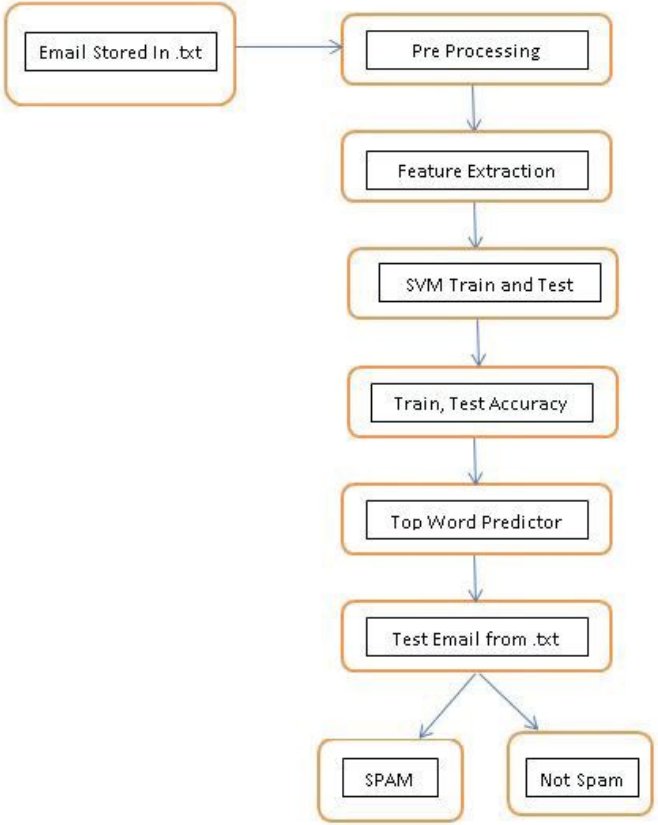
- PROBLEM STATEMENT
- MODEL INCLUDED
- METHODOLOGY
- CONCLUSION

PROBLEM STATEMENT

Spamming is one of the major attacks that accumulate the large number of compromised machines by sending unwanted messages, viruses and phishing through emails. We have chosen this project because nowadays there are lot of people trying to fool you just by sending you fake emails like you have won 1000 dollars, this much amount is deposited in your account once you open this link then they will track you and try to hack your information. Sometimes relevant emails are considered as spam emails

- Unwanted email irritating Internet consumers.
- Critical email messages are missed and/or delayed.
- Consumers change ISP's all the time looking for consistent email delivery.
- Loss of Internet performance and bandwidth. 2
- Millions of compromised computers.
- Billions of dollars lost worldwide.
- Identity Theft.
- Increase in Worms and Trojan Horses.
- Spam can crash mail servers and fill up hard drives.

MODEL INCLUDED



METHODOLOGY

A given message is spam or not! We can do this by using a simple, yet powerful theorem from probability theory called **Bayes Theorem.** It is mathematically expressed as

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)},$$

where A and B are events and $P(B) \neq 0$.

- $P(A)$ and $P(B)$ are the probabilities of observing A and B without regard to each other.
- $P(A \mid B)$, a conditional probability, is the probability of observing event A given that B is true.
- $P(B \mid A)$ is the probability of observing event B given that A is true.

With this we use two additional concepts:

- BOW-Bag of Words
- TF-IDF -Term Frequency-Inverse Document Frequency

We use these two methods in the implementation of the projected Spam classifier
They work as following:
Bow- In Bag of words model we find the 'term frequency', i.e. number of occurrences of each word in the dataset.

$$P(w) = \frac{Total\ number\ of\ occurrences\ of\ w\ in\ dataset}{Total\ number\ of\ words\ in\ dataset}$$

$$P(w|spam) = \frac{Total\ number\ of\ occurrences\ of\ w\ in\ spam\ messages}{Total\ number\ of\ words\ in\ spam\ messages}$$

In addition to Term Frequency we compute Inverse document frequency.

$$IDF(w) = \log \frac{Total\ number\ of\ messages}{Total\ number\ of\ messages\ containing\ w}$$

In this model each word has a score, which is TF(w)*IDF(w). Probability of each word is counted as:

$$P(w|spam) = \frac{TF(w|spam) * IDF(w)}{\sum_{\forall\ words\ x \in train\ dataset} TF(x|spam) * IDF(x)}$$

In addition to this, if the word 'w' is not present in the tran dataset, the TF(w)=0
Additive smoothing can be done, In additive smoothing we add a number alpha to the numerator and add alpha times number of classes over which the probability is found in the denominator.

$$P(w|spam) = \frac{TF(w|spam) * IDF(w) + \alpha}{\sum_{\forall\ words\ x \in train\ dataset} TF(x) * IDF(x) + \alpha \sum_{\forall\ words\ x \in spam\ in\ train\ dataset} 1}$$

This is done so that the least probability of any word now should be a finite number. Addition in the denominator is to make the resultant sum of all the probabilities of words in the spam emails as 1.

CONCLUSIONS

In this assignment the proposed a method to filter out spam messages
This was accomplished using Baye's Theorem and methods of bow and tf-idf.
An accuracy of 93% was accomplished to filter out ham and spam messages from the data set