Term Project Proposal

Advance Algorithm and Complexity

Efficient Spam Filtration

Created By :

Surbhi Sharma (2020H1030148H)

Himanshu Gupta (2020H1030121H)

Guided By :

Apurba Das

Abstract :

A spam-detector algorithm must find a way to filter out spam while and at the same time avoid flagging authentic messages that users want to see in their inbox. Sometimes, static rules can help.

Some spam emails include too many BCC recipients, very short body text, and all caps subjects. Likewise, some sender domains and email addresses can be associated with spam. But for the most part, spam detection mainly relies on analysing the content of the message.

Current Algorithm :

Currently the most famous algorithm which is used in Spam detection is Naive Bayes Algorithm in backend which uses supervised learning model Technique. For this we have to provide the training set, that includes spam and non-spam mails, to the machine learning model and let it find the relevant patterns that separate the two different categories.

For instance, every time you flag an email as spam in your Gmail account, you're providing Google with training data for its machine learning algorithms.

The Naïve Bayes Algorithm tokenize the mail content and for each word, it calculates the probability of occurrence in Spam / Non-Spam folder. In case the probability of occurrence of a word in Spam folder is higher than the probability of occurrence of same word in Non-spam folder, the new email is moved to Spam, else, it is moved to Non-spam folder / Inbox.

Issues in Existing Approach :

1. Bulk emails from some random senders
   - Due to indirect linking of Distribution lists (Transitive dependency), the users receive bulk emails from same sender on different subjects (as shown in image below).
2. Spam Filter not efficient
   - Email from same sender are not filtered efficiently. Some are in spam and others in Inbox.
3. Emails from unknown senders are not recognized

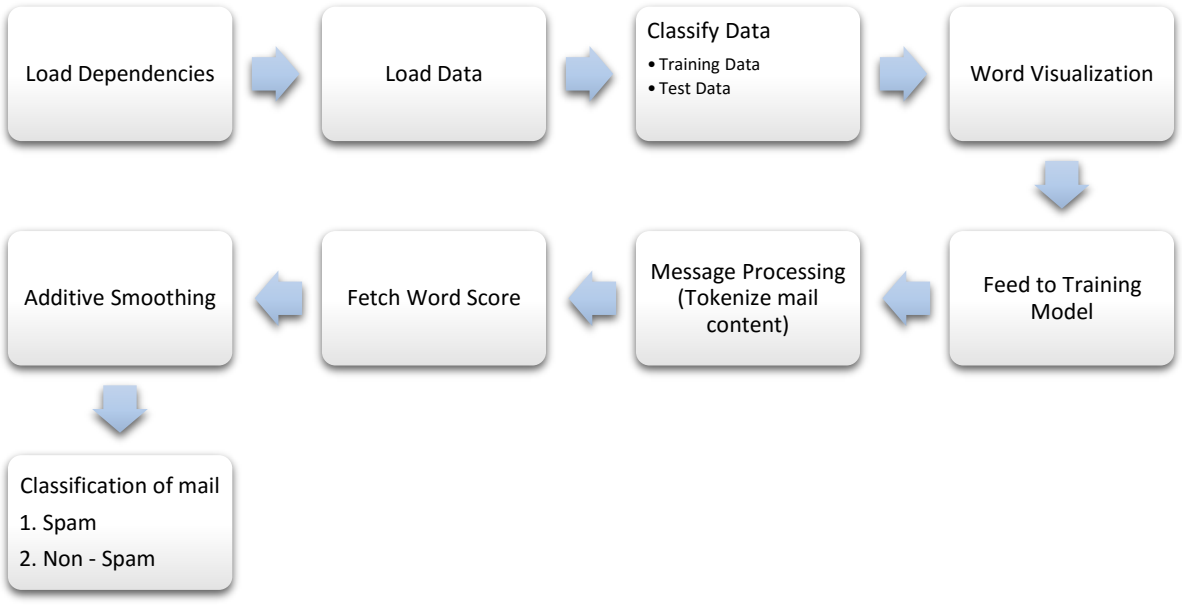
- Email from unknown sender are not considered as spam and are available in Inbox.

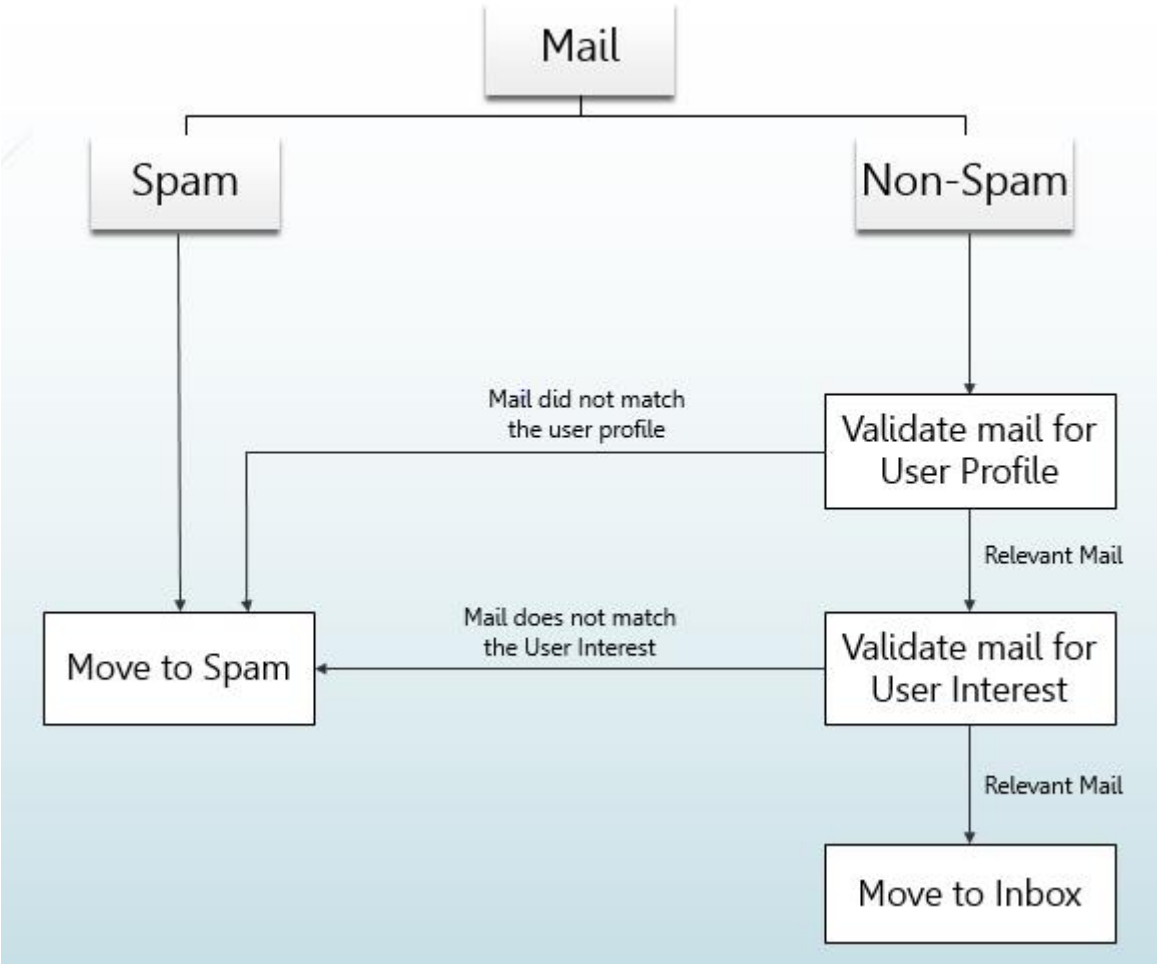
4. Lists un-subscribed long ago still sends bulk emails
   - The list(s) unsubscribed long ago still sends emails and are available in Inbox.
   - For Example: Sender 'Nykaa.com' was unsubscribed few days back, but still user received email showing some offer.
5. Mails are not categorized on basis of User profile
   - Users getting mails which are not useful to them
   - For Example : Women University sending admission mails to every user (irrespective of their gender / other eligibility criteria)

Suggested Solution / Approach :

Our solution would work on the below -

1. Increasing the efficiency of spam filtration using Naïve Bayes Algorithm.
2. Categorization of mails on basis of User profile and User interest to enhance user experience.
3. Detection of mails that have some data / content which may be used to make a fraud.
4. Update the algorithm on the basis on usefulness / user profile (this is a part of Algorithm maintenance)





References :

1. http://www.ijceronline.com/papers/Vol8_issue6/Version-3/D0806032632.pdf
2. http://www.ajer.org/papers/v2(10)/F02106373.pdf
3. https://towardsdatascience.com/clustering-concepts-algorithms-and-applications-f512a949549a
4. https://bdtechtalks.com/2020/11/30/machine-learning-spam-detection/amp/