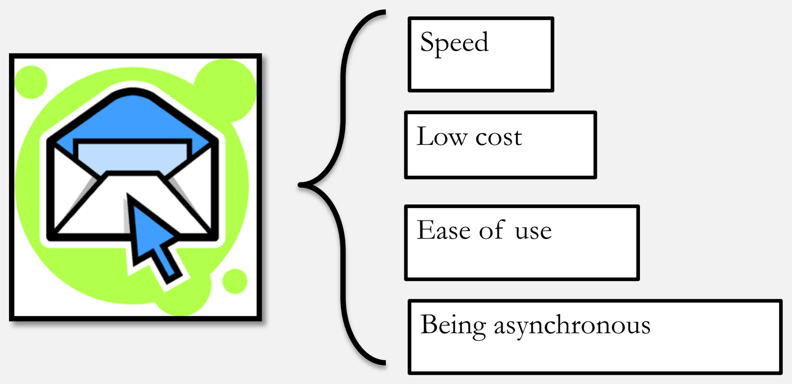
Reply Prediction of Email messages using Interaction- and Content-based features

Abtin Zohrabi and Ali A. Ghorbani

Introduction

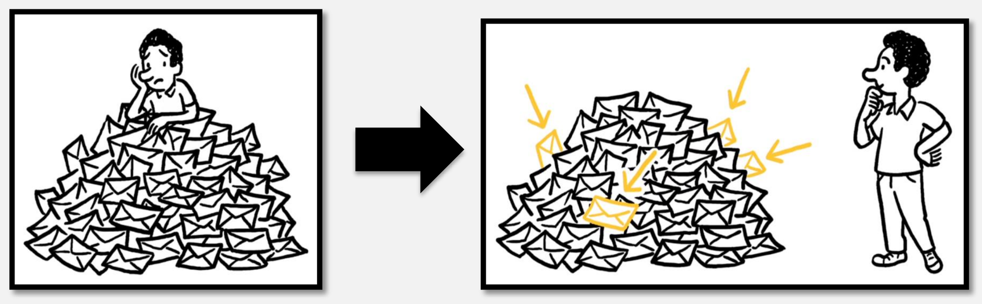
- As a subdomain of Text Mining research area, **Email Mining** can be defined as knowledge discovery on textual email data.
- Main features of Email include:



- Some distinct characteristics of emails comparing to any other unstructured text data:
  1. Additional information exists in headers of email in addition to its content (which is a plain text)
  2. it is significantly shorter comparing to texts from news or blogs, as a result a subset of Text Mining techniques will not produce the expected accurate results in email data.
  3. There is also a high probability of spelling and grammar mistakes in email bodies.
  4. it is almost impossible to have access to a public dataset for experiments, due to privacy and ethical issues.

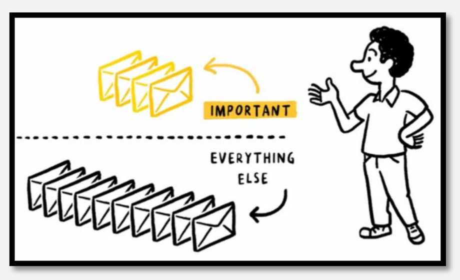
Motivation

- There's no explicitly defined "level of interest" or "level of importance" integrated into email systems, therefore users have to spend valuable time to deal with a large volume of emails.



* Taken from "Google Priority Inbox" ad.

- The problem with all of those tools is that they are not facilitated with deep text mining and machine learning techniques and also there is less intention towards extracting useful and descriptive features out of emails textual contents and previous user interactions.
- The idea behind this assumption was that the probability of an email being replied or not is not only dependent on previous user interactions but the "language" reflecting in the message body.



Purpose

- It is believed that every email with a high probability of being replied can be considered as "important" for a specific user. Therefore this abstraction has been implemented into a real-world application: Email reply prediction, not only to have a better understanding of what is "importance", but also to be able to propose a process and analyze the results in terms of accuracy.

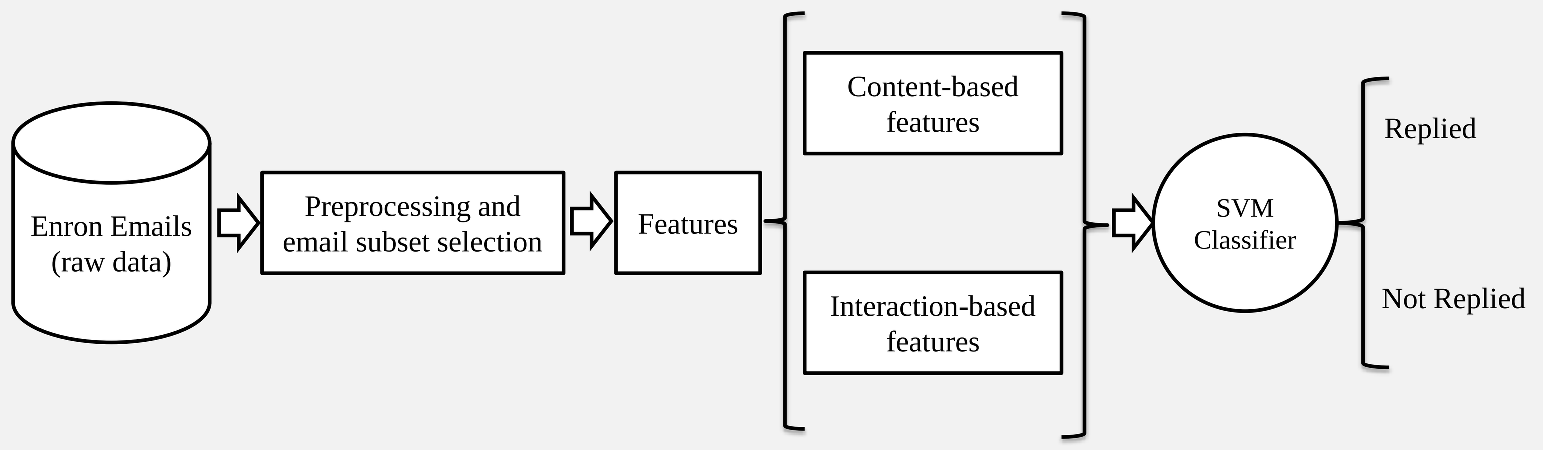
Reply Prediction → Classifying the emails into replied or not-replied.

Interaction-based Features → History of sent, received, and replied emails between users. (previous communication patterns)

Content-based Features → We used computational linguistics methods on content of an email. One score will be calculated, as a measure for the intensity of request words in the body of email.

Problem definition

- Through a process, two types of features have been extracted out of user interactions and also content of an email. Then an SVM model is learned to classify incoming emails as "Replied" or "Not Replied".



Features

- Interaction-based features

Feature 1 → $indegree\ (Sdr(e))$

Feature 2 → $outdegree\ (Sdr(e))$

Feature 3 → $totalDegree$

Feature 4 → $\frac{|RepliedBy(Sdr(e))|}{|Sent(Sdr(e))|}$

Feature 5 → $\frac{|RepliedTo(Sdr(e))|}{|Received(Sdr(e))|}$

Feature 6 → $\frac{|Replied\ (Sdr(e) \rightarrow u_r)|}{|Emails(u_r \rightarrow Sdr(e))|}$

Feature 7 → $|Emails(Sdr(e) \rightarrow u_r)|$

Feature 8 → $|Replied(Sdr(e) \rightarrow u_r)|$

Feature 9 → $\frac{|Replied\ (u_r \rightarrow Sdr(e))|}{|Emails(Sdr(e) \rightarrow u_r)|}$

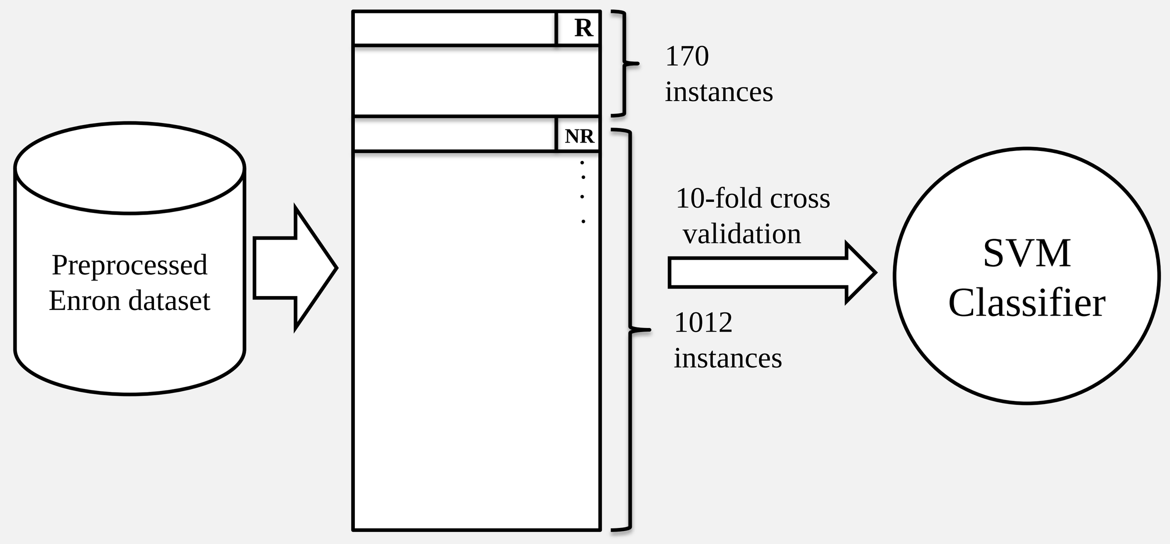
- Content-based bag of words

Searle Keywords	Request, send, deliver, please
Neighboring Question	?
Modal	May I, May you, can you, can I, shall I…
Sentences begins with WH Questions	What, which, who, why, when,…
Plan Phrases	I am going to, I am planning to,…

Experiments and Results

- For each email a vector has been made that includes all of the values of proposed features:

$for\ each\ Email_{i,j} \rightarrow [F_1, F_2, F_3, F_4, F_5, F_6, F_7, F_8, F_9, S, Class]$



- Two classifiers are trained and tested: one with all features (SVM_{I+C}) and the other without content-based features (SVM_I).
- We also applied 10-corss fold validation for better estimation of how accurately our model will perform in practice.

Classifiers	Correctly Classified	Incorrectly Classified	Average Accuracy
SVM_{I+C}	960	127	88.3 %
SVM_I	945	142	86.9 %

Conclusion

- By using content-based analysis besides those user interaction histories, the accuracy of the SVM classifier is enhanced.
- The SVM classifier reports 88.3% of accuracy using both feature sets and 86.9% with only interaction-based features. Our SVM classifier reported improvements by adding content-based features.
- We can use more Interaction-based features, extracted from social network of email users. Measures like Between-ness, PageRank measure, or Clustering coefficients may have some additional contribution to our classification task.

2013 Faculty of Computer Science Research Expo. April 11-12, Wu Conference Centre, University of New Brunswick