Computational Intelligence, Communications, and Business Analytics (CICBA 2018)
(27 – 28 July, 2018)

Understanding Email Interactivity and Predicting User Response to email

Soumyadeep Roy*, Nibir Pal, Kousik Dasgupta, Binay Gupta
Indian Institute of Technology Kharagpur
soumyadeep.roy9@iitkgp.ac.in

Technical Session 2.6 Paper Id. 101

Disclaimer

The use of general descriptive names, registered names, trademarks, service marks, etc. in this presentation does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The authors and the editors are safe to assume that the advice and information in this presentation are believed to be true and accurate at the date of presentation. Neither the organisers nor the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Paper Outline

1. Motivation
2. Related Works
3. Proposed Method
4. Results
5. Conclusion
6. Future Work
7. References

Motivation

- Email overload problem and increasing volume of email traffic, as email is essential for personal and work-related use
- Understanding factors responsible for determining user replying behavior help improve organizational productivity and developing targeted strategies for customers
- Email expectation and breakdown points are influenced by the recipient, urgency of topic, timeshift between sender and recipient. They also vary based on content and recipient type
- Most existing studies work on small data samples or surveys on employees of a particular organization

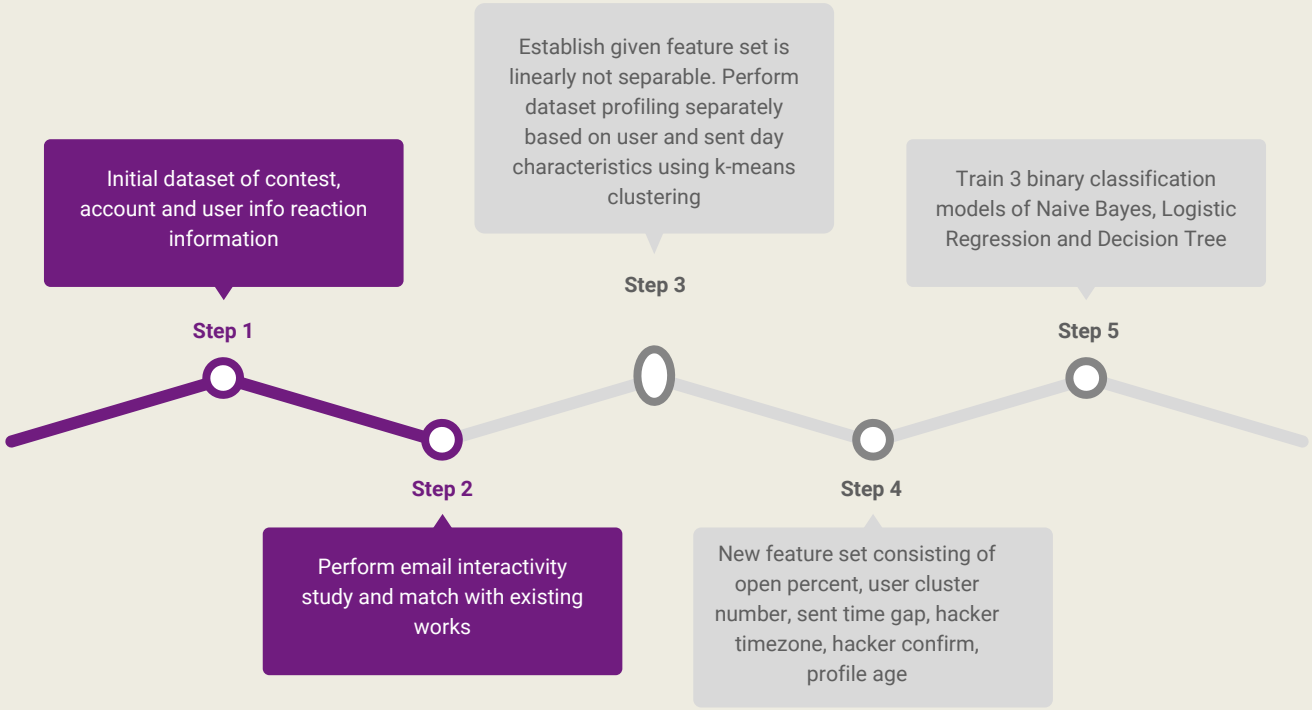
Related Works

- Survey information and interviews are unable to verify whether the user perceptions are reflected during work
- Some predict message importance while others directly predict user actions like read(open), reply, delete, delete-without-read
- Email content and metadata, historical interactions, temporal features, stage and history of conversation, email load, day of week, time message is received, user demographics, are used for determining time and length of reply
- Focus on quantitative measures of overload and its corresponding effect on users. Explore impact of frequency of checking emails on subjective well-being

Problem Formulation & Proposal

- Address the email overload problem, i.e, inability to address all incoming mails, leading to decline of usual effectiveness
- Understand the pattern of interaction between the email and the recipients using dataset profiling based on k-means clustering on user and sent day characteristics
- Understand factors determining user replying behavior and predict the email recipient's response, specifically whether he/she opens the email

Proposed Method



Methodology Workflow

Preprocessing

- Remove observations with missing mail category values
- Address the multi-collinearity problem, by removing one out of each feature pair having high correlation values
- Remove mail id and user id features
- Use one-hot-encoding to transform categorical variables for utilising the Logistic Regression classifier
- For the Email interactivity study, only use data points of users belonging to the timezone 18000

Dataset Profiling

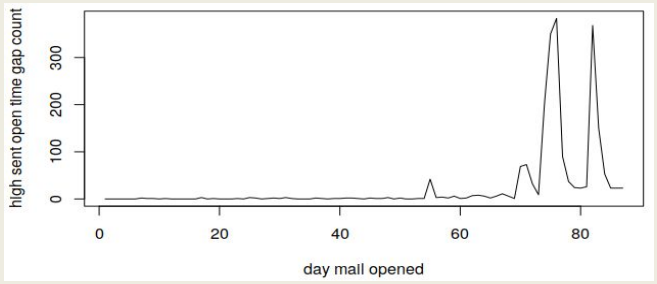
- Perform feature-engineering where we create new features
  - total mid open, total high open, max open time gap
- Assign each user to its most recent data
- Perform k-means clustering of user and sent day properties and obtain optimal value of “k”
- Add features
  - age of recipient in weeks, open percent
  - user activity related to clicked and unsubscribed actions

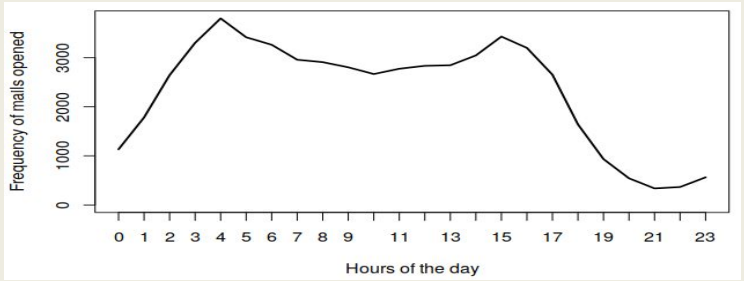
Feature Importance

- Features ranked on information gain associated when the new feature set is fed into a Decision Tree classifier
- Generate boxplots comparing value distribution of all mails opened vs those not opened

Results

- Observe similar patterns to Jackson(2003) in volume of email sent per hours of the day
- Varied trend between percent of positive user response per day the emails are opened and volume of mails after high threshold of time gap
- Observe sudden spikes in volume of email sent





Jackson, T.W., Dawson, R., Wilson, D.: Understanding email interaction increases organizational productivity, pp. 80-84, Commun. ACM 2003

Results

- Establish that original feature space is not linearly separable
- Perform user profiling and sent day profiling and obtain final set of features: open percent, user cluster number, sent time gap, hacker timezone, hacker confirm, user profile age in weeks
- Opened percent most significant

Approach	Accuracy	Precision	Recall	F1
Logistic Regression	**0.7941**	**0.6895**	0.5052	0.5832
Decision Tree	0.7588	0.5605	**0.7139**	**0.6279**
Naives Bayes	0.7588	0.6	0.6529	0.6253

Conclusion

- Observe similar patterns to Jackson, 2003 in our detailed Email Interactivity study
- Propose novel feature selection methodology based on user and volume of emails sent by sender, using k-means clustering
- Develop three classification models using Decision Tree, Logistic Regression and Naive Bayes
- Decision Tree classifier perform best with F1 score of **0.6279**. The most significant feature being opened percent, the fraction of emails opened by the user in the past

Future Work

- Predict other user action-types like unsubscribe, delete, delete-without-read, reply
- Use clustering algorithms like k-nearest neighbor, DBSCAN and OPTICS, which automatically learns the number of clusters
- Linear and rbf SVM took very long to train. Overcome by simple random sampling measures or advanced subsetting measures based on active learning

References

1.Ai, Q., Dumais, S..T., Craswell, N., Liebling, D.: Characterizing email search using large-scale behavioral logs and surveys, pp. 1511-1520, WWW 2017

2. Dabbish, L.A., Kraut, R.E., Fussell, S., Kiesler, S.: Understanding email use: Predicting action on a message, pp. 691-700, CHI 2005

3. Jackson, T.W., Burgess, A., Edwards, J.: A simple approach to improving email communication, pp. 107-109, Commun. ACM 2006

4. Jackson, T.W., Dawson, R., Wilson, D.: Understanding email interaction increases organizational productivity, pp. 80-84, Commun. ACM 2003

References

5. Kooti, F. et.al. : Evolution of conversations in the age of email overload, pp. 603-613, WWW 2015

6. Tyler, J.R., Tang, J.C.: When can I expect an email response? a study of rhythms in email usage, pp. 239-258, ECSCW 2003

7. Yang, L., Dumais, S.T., Bennett, P.N., Awadallah, A.H.: Characterizing and predicting enterprise email reply behavior, pp. 235-244, SIGIR 2017

Thank you