INDIAN INSTITUTE OF TECHNOLOGY MADRAS

# CS6370 – Natural Language Processing
# Project Proposal
# Email Classification Using Co-EM

Aravind Sankar (CS11B033)
Sriram V (CS11B058)

October 26, 2014

## 1 MOTIVATION

The broad area of Text Classification has been a topic of research for a very long time now, and has been well researched especialy in the case of supervised learning of classifiers. Our focus is therefore directed towards tackling the problem of classification of limited labeled data, otherwise referred to as semi-supervised learning.

Although text classification is a problem that has been looked into right from the 1960s, email classification is highly relevant in today's fast-paced world, where one does not have the time to sift through vast amounts of mail but would rather like information clustered together, so that one may disregard the irrelevant ones such as promotional offers, and focus attention on the important ones at hand. Triaging has gained a lot of focus in recent times, with GMail launching personalized mail classification, and with other services such as Mailbox and Boxer focussing on improving user efficiency. Thus the problem is of immense interest in recent times, and improving performance and accuracy is of prime concern.

The classification of email into different categories is technically an interesting problem as well, since there is a scarcity of labelled data. While there are many approaches to semi-supervised learning, the idea of Co-EM (as explained in [1]) is a unique one, as it is a hybrid method that combines Co-Training and Expectation Maximization (EM), and has been shown to outperform Co-Training. Previous efforts (such as [2]) in the area of Email Classification have involved either EM or Co-Training alone, but to the best of our knowledge, combining the two methods has hitherto been unexplored.

Apart from merely exploring a new semi-supervised learning technique, we also intend to delve deeper into the specific classifiers that could be used in combination with these techniques as there is a very good chance of the learning problem having a class imbalance. [4] talks about utilizing SVMs instead of Bayesian methods for Co-EM, which is an interesting idea that we would like to look into. We also plan to explore other aspects, such as extending [2] to handle multi-class classification.

## 2 DATASETS

Email datasets with category labels appear to be scarce, primarily due to privacy concerns. There appears to be some sort of categorization of mails in the Enron Corpus[1], and this has been used for email classification

[1] http://www.cs.cmu.edu/~enron/

in [6]. Further details about this corpus are also covered in [5]. Apart from this dataset, we also plan to look at Spambase Dataset[2] (a 2-class classification problem) and the Reuters Newsgroup dataset [3] (a multi-class classification problem), should we not be able to obtain well defined classes in the Enron Corpus.

## 3 BASELINE APPROACHES

Our baseline techniques include the ones described in [1], apart from a few other approaches, namely:

1. Expectation Maximization
2. Co-Training with Naive Bayes
3. Co-Training with SVM
4. Co-Training with Random Labelling

## 4 EVALUATION METRICS

We plan to evaluate our proposed technique on the same metrics as [1], as well as some others such as:

1. Graphical visualisations of Accuracy vs. number of iterations
2. Absolute difference in accuracy between the 1st and the 50th iterations
3. Precision, Accuracy and Recall measures

## 5 FUTURE WORK

[3] introduces Co-EMT, a multi-view algorithm which combines semi-supervised and active-learning, to handle classification problems with incompatible, correlated views. This technique could be applied to the problem of email classification, and its performance could be compared with the proposed as well as the exisiting semi-supervised learning methods. This paper makes use of Naive Bayes as the underlying algorithm, but SVMs too could also be explored as an alternative classifier.

## REFERENCES

[1] Analyzing the Effectiveness and Applicability of Co-training. Kamal Nigam, Rayid Ghani. *In Proceedings of the Ninth International Conference on Information and Knowledge Management, CIKM 2000.*

[2] Email classification with co-training. Svetlana Kiritchenko, Stan Matwin. *In Proceedings of the 2001 conference of the Centre for Advanced Studies on Collaborative Research, CASCON 2001.*

[3] Active + Semi-Supervised Learning = Robust Multi-View Learning. Muslea, Ion, Steven Minton, Craig A. Knoblock. *In Proceedings of the International Conference on Machine Learning, 2002.*

[4] Co-EM Support Vector Learning. Ulf Brefeld, Tobias Scheffer. *In Proceedings of the International Conference on Machine Learning, 2004.*

[5] Introducing the Enron Corpus. Klimt, Bryan, and Yiming Yang. *In CEAS, 2004.*

[6] The enron corpus: A new dataset for email classification research. Klimt, Bryan, and Yiming Yang. *In Machine learning: ECML. Springer Berlin Heidelberg, 2004*

[2] `http://archive.ics.uci.edu/ml/datasets/Spambase`
[3] `http://www.cs.umb.edu/~smimarog/textmining/datasets/`