INDIAN INSTITUTE OF TECHNOLOGY MADRAS

CS6370 – Natural Language Processing Project Proposal Email classification using Co-EM

Aravind Sankar (CS11B033)
Sriram V (CS11B058)

October 25, 2014

1 MOTIVATION

Our chosen topic to work on is Text Classification. This is a very broad area which has been well researched in the case of supervised learning of classifiers. We want to focus on classification using limited labeled data, i.e semi-supervised learning. The classification of email into different types is interesting problem where we face the problem of scarcity of labeled data. Of the many approaches to semi-supervised learning, we came across [1] which introduces the idea of Co-EM, which is hybrid method combining Co-Training and Expectation Maximization (EM). Existing techniques for Email Classification have tried EM and Co-Training algorithm separately as shown here [2]. We feel that combining these techniques using Co-EM would yield better results for classification. To the best of our knowledge, Co-EM has not been used to solve email classification till now. The reason why we want to use Co-EM is —- . Our work will also involve trying out various classifiers and identifying the best suited one. This is a very essential task, as the learning problem has a very good chance of having a class imbalance.

2 DATASETS

3 BASELINE APPROACHES

The baselines with which we'd like to compare our approach would be the one described in [1] along with a few other approaches. These are :

1.

4 EVALUATION METRICS

REFERENCES

[1] Analyzing the Effectiveness and Applicability of Co-training. Kamal Nigam and Rayid Ghani - CIKM '00 Proceedings of the ninth international conference on Information and knowledge management.

[2] Email classification with co-training. Svetlana Kiritchenko and Stan Matwin - CASCON '01 Proceedings of the 2001 conference of the Centre for Advanced Studies on Collaborative research