A Survey of Document Classification Problem Using Modified Naïve Bayes and E-M Algorithm*

Michael Peven
Johns Hopkins University
100 Bleecker Street, 27A
New York, NY 10012, USA
mpeven1@jhu.edu

Christian Reotutar
Johns Hopkins University
3700 N. Charles St.
Baltimore MD, USA
creotut1@jhu.edu

Abstract

Document classification is an important machine learning and natural language processing topic that has applications in spam filters and summarization techniques. In this typically multi-class problem, the Naïve Bayes classifier is a popular model with results rivaling that of the Support Vector Machine classifier with shorter training time. In this survey, we seek to find improvements in the Naïve Bayes classifier such that accuracy is improved and compare this to another classifier used in the problem space.

1 Introduction

Here we give a brief overview and definition the document classification background

1.1 Problems in document classification

Here we talk about problems in this problem space and how it's applied.

1.2 Summarization

Here we talk about our specific domain, summarization, and how document classification relates to this.

*This document has been adapted from the instructions for earlier ACL and NAACL proceedings, including those for NAACL-HLT-09 by Joakim Nivre and Noah Smith, for ACL-05 by Hwee Tou Ng and Kemal Oflazer, for ACL-02 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats. Those versions were written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence*.

2 Data Description

Here we talk about Twitter data and how it relates to our problem

3 Machine Learning Techniques

Here give a broad overview of what we're comparing, what techniques we used and why

3.1 Naïve Bayes

Here we give background into the classifier

3.2 Modified Naïve Bayes

Here we give background into the classifier + word2vec library

3.3 Other Classifier

Here we give background into the classifier

4 Methods

4.1 Data Description

Here we talk about the filtering done on the data, the importance of feature vector formation and attribute picking, and the problems we came upon (e.g. language, more than one hash tag).

4.2 Code for Naïve Bayes

4.3 Code for Modified Naïve Bayes

4.4 Code for Other Classifier

5 Results

Here we give a broad overview of our results

5.1 Comparisons

Lots of tables and graphs

5.2 Explanations

It is important that we relate our results to the classifiers we used and other ML topics.

5.3 How we can improve

Any improvements for the classifiers we did not have time to implement

6 Conclusion

7 Comparison to Proposal

- A list of what we did differently from proposal

Acknowledgments

We would like to acknowledge Ilya Shpitser for giving feedback on our project and offering suggestions. Ilya for the skeleton for the code.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.