

# COMP 348 : Ass 1 Part 1, Report

Ilun Ahn (41404505)

March 21, 2008

## 1. Introduction

Many organizations are interested in finding out the writers of particular texts. Such task is important for marketers and governments who want to find out demographic data through texts. Currently, most of this work is done by requesting users to fill their information when they subscribe to a website. Another way to do this is identifying demographic characteristics from a lot of texts in the Internet. In this paper, we will handle the latter way.

## 2. System Description

This section will be filled after AdAge's algorithm is clearly defined.

## 3. Experiment and Result

As data we have classifications for 1000 blogs by AdAge, along with their actual classes.

We choose a baseline as 0.742 because if

we classify all posts as most frequent category, written by young generation, we can get accuracy 0.742. (I can't sure that I properly understand most-frequent-category baseline) By comparing actual classification data and classification by AdAge, we get 0.77 for the accuracy rate.

Here, we will use z-Test of proportion. This method is used to test proportions. Z value is calculated along this formula:  $z = (p - p_0) / \sqrt{p_0 * (1 - p_0) / n}$ , where p is actual accuracy and  $p_0$  is baseline we set.

Z-statistic is 2.0237 when we consider actual accuracy rate and baseline. According to the Z-statistics table, this means just 4% of other accuracy value which can be generated is farther from baseline than our accuracy. This can be considered statistically significant.

## 4. Conclusion

Comparing actual classification accuracy from given data and our baseline accuracy in Z-test, we find that the difference between the two proportion is statistically significant.