COMP30027 - Project 2: Blog Authorship Classifier

Anonymous

1. Introduction

The purpose of this project was to build and critically analyse one or more supervised machine learning systems, with the goal of identifying the age of blog authors. This report aims to briefly outline the task and related literature, summarise the development process for the machine learning system, and evaluate the system's performance on the development documents.

I will then discuss the behavior and properties of the machine learning methods, and perform error analysis on them.

2. Background

2.1. Project Task

I was provided with a heavily-altered subset of the Blog Authorship Corpus, described in "Effects of Age and Gender on Blogging" (Schler, Koppel, Argamon, & Pennebaker, 2006). The goal was to automatically identify each of the blog author's ages. The machine learning system was to assign the authors to one of four age classes: 14-16 years, 24-26 years, 34-36 years, or 44-46 years.

For the purposes of this project, a machine learning system was defined as a combination of the following:

- data representation
- choice of learner
- choice of hyperparameters for the learner
- a mechanism for building the model ("training") on the given training data
- a mechanism for using the resulting model ("classifier") to make predictions on the development/test data.

2.2. Related Literature

Author age identification is a sub-problem of the extensively-researched text authorship and text classification problems. Author attribution supported by statistical methods pre-date machine learning, commencing in the 19th century (Stamatatos, 2009). One of the primary uses for author identification is plagiarism detection (de Vel, 2000). Another common application is forensic analysis for criminal cases (de Vel, 2000).

Modern methods make use of stylometric features: lexical, character, syntactic, semantic and application-specific (Stamatatos, 2009). Attribution methods utilise either profile-based approaches or instance-based appraoches (Stamatatos, 2009).

3. Machine Learning System

3.1 Development Process

Initially, I built a baseline authorship classifier by running the Multinomial Naïve Bayes and Linear Regression learners on the provided "Top 10" data subsets. Based on the results, there was evidence of a skew towards the younger age class, particularly after closely inspecting the pre-selected word tokens, which were largely "blog words" (neologisms such as *lol*, *haha*, *ur*) (Schler et al, 2006).

Accordingly, for the next iteration of the machine learner, I decided to process the raw data files instead. There are two distinguishing features in this dataset: style-related and content-related (Schler et al, 2006). Since Schler and colleagues (2006) found content-related features to be a better predictor of age, I decided to narrow the datasets down to my chosen parameters of user ID and blog post text. I pre-processed the text fields by removing punctuation and transforming to lower case. I also removed any entries that didn't align with the four age classes. I ran chi-square statistics for $n=8$ most correlated unigrams and bigrams. Numerous non-English words were present. I ascertained these words were blog pseudonyms; therefore, I consolidated each of the blog posts and grouped the dataset by author, to ensure there was no over-representation by any individual author.

I used cross-validation to test the accuracy of four different learners: Random Forest Classifier, Linear Support Vector Classification, Logistic Regression, and Multinomial Naïve Bayes (Table 1).

Accuracy	Machine Learner
0.504357	Random Forest Classifier
0.646037	Linear Support Vector

	Classification
0.666037	Multinomial Naïve Bayes
0.670171	Logistic Regression

Table 1: Accuracy of the Tested Machine Learners (Cross-Validation)

Since Logistic Regression and Multinomial Naïve Bayes yielded the highest accuracy, they were the learners I selected for fitting the development data and test data. Finally, I chose the top 16 most correlated unigrams and bigrams for each age group as the parameters to tokenize the blog posts.

4. Results

4.1. Behaviour & Error Analysis

As shown in Table 2, the accuracy of the machine learners was lower when testing on the development dataset.

Accuracy	**Machine Learner**
0.496453	Multinomial Naïve Bayes
0.561170	Logistic Regression

Table 2: Accuracy of the Selected Machine Learners (Development Data)

After pre-processing the development dataset, its size significantly reduced – from 45,332 to 1497. Furthermore, the remaining data was heavily skewed towards the 14-16 and 24-26 age groups (Figure 1).

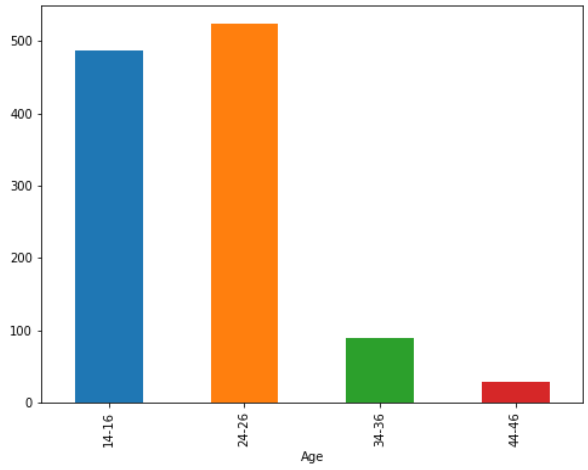

Figure 1: Age Class Distribution for Development Data

The Logistic Regression confusion matrix revealed one main error class: the majority of errors were from predicting the 24-26 age group, when it should have been 14-16, whereas with the Multinomial Naïve Bayes confusion matrix, the error class was reversed (see Figures 2 and 3). I checked the list of tokens selected and noticed significant overlap between both groups' unigrams, leading me to hypothesise this to be the cause of the low accuracy rate. I tested my hypothesis by adjusting the parameters and dropping the unigrams, focusing solely on the bigrams; however, this resulted in lower accuracy, increasing the bias towards the 14-16 group.

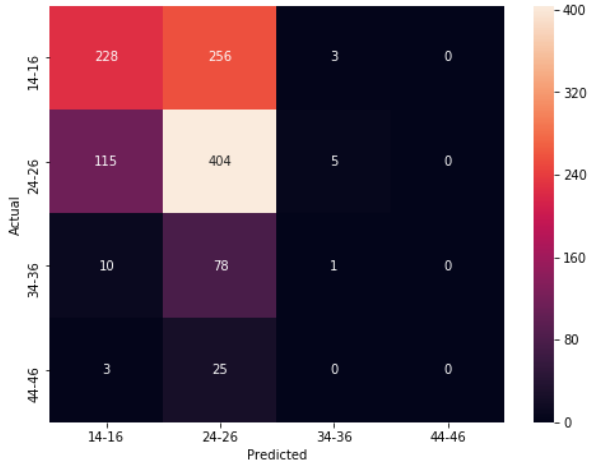

Figure 2: Logistic Regression Confusion Matrix

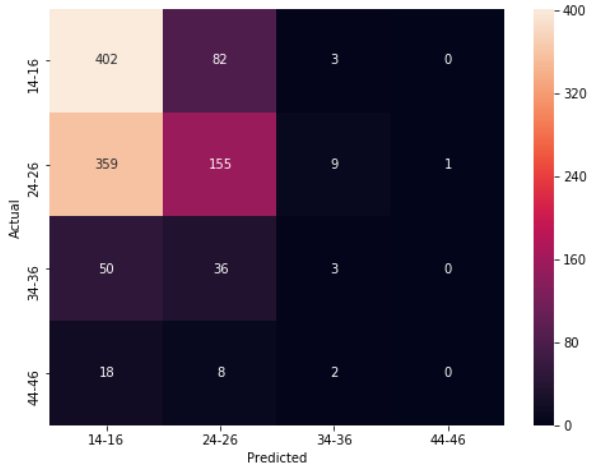

Figure 3: Multinomial Naïve Bayes Confusion Matrix

The significant accuracy difference from the cross-validation hold-out test leads me to believe that the development data size and skewed distribution are the primary cause of the lower accuracy scores (see Figure 4).

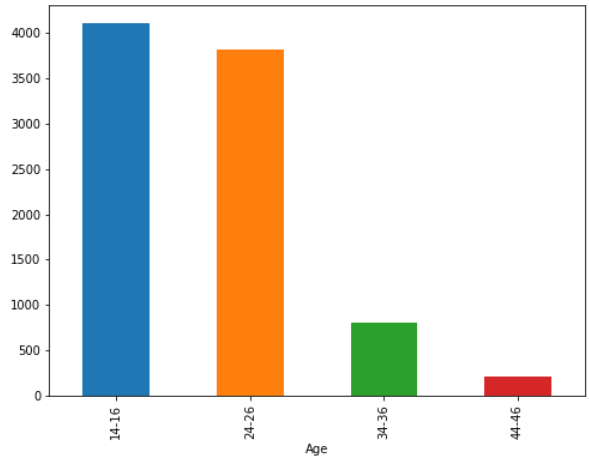

Figure 4: Age Class Distribution for Training Data

5. Summary

The aim of this project was to build and critically analyse one or more supervised machine learning systems that identify the age of blog authors. After building the baseline authorship classifier, I benchmarked four different machine learners and chose Multinomial Naïve Bayes and Logistic Regression, with accuracy scores of 49.64% and 56.11%, respectively. I attempted various text manipulation methods to test my hypotheses regarding the low accuracy; however, the accuracy either decreased or remained largely the same. I believe the small dataset size and uneven distribution are the primary cause of the inaccurate learners; therefore, a much larger development dataset is required for significant accuracy improvements.

References

de Vel, O. (2000). Mining e-mail authorship. *Proc. Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining.* Salisbury.

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. (2006). Effects of Age and Gender on Blogging. *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs.* Stanford.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology, 60*(3), 538-556.