COMP30027 Report: Age Classification from Text Documents

Anonymous

May 9, 2018

1 Introduction

This report tackles the problem of text classification, specifically age classification of blog posts. Utilizing a modified data set from the paper¡citation¿, over two hundred thousand samples of blog posts were analyzed and processed allowing for features to be extracted and utilized. Machine learning models were then trained with this data set and classification predictions made.3 learners were then chosen to fit the data. This report aims to explain classifier behaviour in the context of a Naive Bayes classifier;to explore/explain its predictions through error analysis and interpretation.

2 Data

The dataset used is an altered form as described in Schler, Jonathan, Moshe Koppel, Shlomo Argamon, and James Pennebaker (2006) Effects of Age and Gender on Blogging. The dataset has been split into a training portion and a development portion. The training portion contains 276414 instances of blog posts from 8950 unique authors. The models were trained on the training portion and evaluated on the development portion.

2.1 Preprocessing

The dataset contained metadata describing the author, this was discarded however as the testing data did not contain this information. The only information needed was the target age and text content of each blog. The training text content was abundant with statistical noise that first and foremost needed to cleaned. Firstly, instances were stripped of extra whitespace and all characters changed to lower case to avoid repeating case sensitive

features. Next, using a Bag of Words model, each document was transformed into a feature vector that contained word counts for each word in the document. This was then transformed into a matrix of inverse document term frequencies by dividing each count by the total number of words in the document. This is done to avoid discrepancies between long and short documents, as longer documents may have a higher average count of values than short documents. This step was streamlined by one call to scikit learns TF-IDF Vectoriser. This vectoriser compiled the above steps while removing stop words and commonly occuring words (words above a max frequency of 50%)

3 Results

Below are final results and are the result of applying trial and error to hyper parameters and processing methodologies

Model	Accuracy (%)
Multinomial Naive Bayes	58.846%
Logistic Regression	61.991%
Perceptron	62.057%

4 Interpretation

4.1 Naive Bayes

Applying our cleaned feature extracted representation of our data set to our first learner yields underwhelming results. The first pass at training this model yields an accuracy of around 50%, only slightly better than simply predicting the majority class. This prompts an investigation into what properties of the data is causing the NB learner to predict so poorly. Below is a confusion matrix for a an initial simple Naive Bayes model.

TP	"14-16"	"24-26"	"34-36"	"44-46"	"?"	Accuracy Per Label(%)
"14-16"	5640	7798	1	2	0	58.01
"24-26"	729	16568	0	1	0	95.77
"34-36"	21	2563	0	0	0	0.00
"44-46"	8	542	1	0	0	0.00
"?"	1519	9933	5	1	0	0.00

An overwhelming majority of instances are being classified as either 14-15 or 24-25 while zero correct predictions are made for classes 34-36 and 44-46. Taking a look into the training set, we find that our data is largely unbalanced and contains a large majority of training instances of class 14-16 and 24-26 meaning our learner becomes heavily biased towards these instances.

Secondly we find that our testing instances contain ages not in the range of our prior classes. No matter how good our model gets, if it sees instances of classes it hasnt been trained on, it will always misclassify these them. Applying this finding to our mode, we realize we need a way to classify ages not within the range of seen ages. One can hypothesize that an unseen classes will arise from unfamiliar words and features associated with the blogs corpus. This implies our learner will predict this unknown instance with a low probability of confidence. Thus we create a hyperparameter called a Confidence Threshold that defines the cutoff point where our model should predict an all encompassing ? label for ages not in the range seen in training. Experimenting with this cutoff point leads us to an optimal value of 0.4 for this Naive Bayes classifier. By increasing this value we move too many values from correct positions where the model was confident into an incorrect position.

Lastly the third issue lies within how we treat an instance. Since there are multiple blog posts per unique author, our model sometimes gets a specific authors age correct but also sees another instance by this same author but miss-classifies it. In order to remedy this issue it is necessary to classify authors instead of individual blog posts, this can be done since an authors age is consistent for every blog post they write. In other words there is a one to one mapping from unique author to age. Applying this solution, our learner now predicts an age for a blog post and adds that prediction to a count of predictions for that unique author. To finalize an age prediction for an author we simply iterate over each prediction and pick the class that is predicted the most frequently. Below is an example of this mechanism.

Taking this idea further, if we group each blog post instance by author and simply concatenate the text together we can achieve better performance on our NB model. This effectively causes each instance to be independent from each other, this decreases auto-correlation and improves the models ability to learn distinct relationships between class labels and instance fea-

tures. By applying these modifications to our baseline model we achieve an accuracy of 58.846%.

Figure 2: Final Confusion Matrix for Naive Bayes Classifier

TP	"14-16"	"24-26"	"34-36"	"44-46"	"?"	Accuracy Per Label(%)
"14-16"	404	110	0	0	4	77.99
"24-26"	42	476	3	0	3	90.84
"34-36"	1	84	2	0	2	2.25
"44-46"	8	25	2	0	0	0.00
"?"	81	253	3	0	1	0.29

5 Conclusion

This low accuracy score can possibly be attested to the nature of the data we feed it. Inherently with the Bag of Words model our NB learner only cares about the presence and frequency of words in a document. While of course this allows our NB model to understand a set of vocabulary used by different age cohorts, it has no knowledge of sentence structure and writing styles. The lack of meaning and context behind text content severely handicapped our model, making it difficult to correctly make predictions where more nuance is required. For future exploration of age classification in a blog context, models that take into account sentence structure and word embeddings may be able to classify a data set like this better.

6 Works Cited

Schler, Jonathan, Moshe Koppel, Shlomo Argamon, and James Pennebaker (2006) Effects of Age and Gender on Blogging. In Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs. Stanford, USA