Amsterdam University College

Text Mining

Jelke Bloem

Determining an authors age by the written text

Project Report

4 pages (without title page)

Dan Hagen & Luca Wilson Nagel

03.06.2022

Introduction:

Using methods learned in class and getting inspired by a masters thesis in Communication Science we seeked to find out to what extend it would be possible to predict an athors age based on the written text.

Related work:

We initially got the idea from a master's thesis in communication science by Eirik Holbæk of the Norwegian University of Science and Technology written in 2019. The aim of this project was to determine whether the author of a text was a child (defined as below the age of 18) or an adult (defined as being 25 years of age or older). He managed to write a program that could predict wether an author was between the ages 13 and 18 or 20 and 29 with an accuracy score of 77%.

Data set:

The Data Set we used is the Blog Authorship Corpus which can be downloaded from “Kaggle.com” and contains 681288 blog posts from 19320 authors. The blog entries were written in or before 2004 with each writer contributing approximately 35 posts and 7250 words. There is a total of 8240 blogs written by authors ranging from 13 to 17 years of age, 8086 blogs by authors ranging from 23 to 27 years of age and 2994 blogs by authors ranging from 33 to 47 years of age. The csv file is about 700 MB large and can be downloaded from Kaggle..

Methods:

To begin our project we started with a simple baseline model which we used as a baseline to evaluate the performance of the rest of our models. For the baseline model we first prepared the data set using the nltk TweetTokenizer as well as filtering out simple stopwords. After this we used a Countvectorizer and a TFidf Transfromer, both provided by sklearn, and trained a simple Logistic Regression Classifier. To improve upon this initial model we build 3 more models. The first one applied ngrams with range 2 to our data set and as well used a logistic regression classifier. The second one was set up similarly to the baseline with the addition of combining the ages into age groups of size 2 years. The final model we tested was set up like the baseline but with a support vector classifier instead.

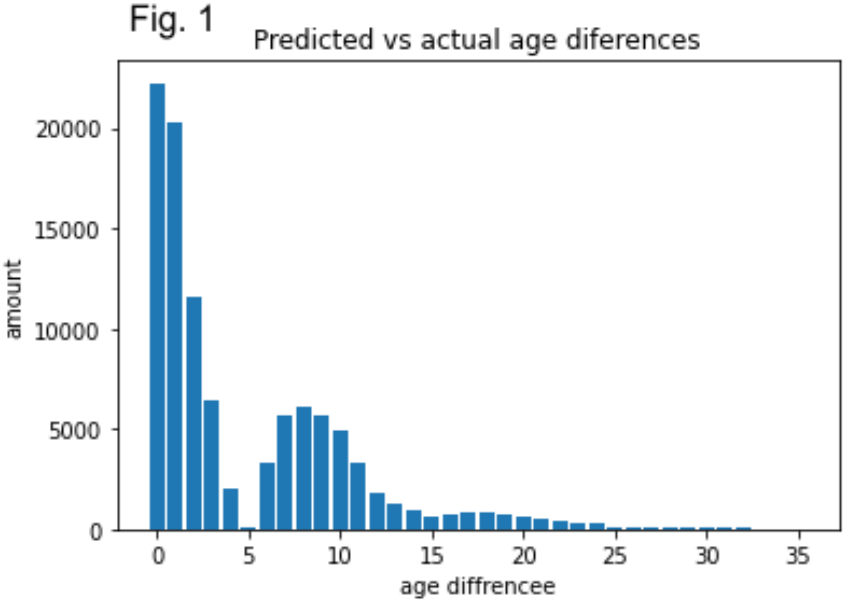
Furthermore to compare our models to the literature as well as to see if they have a higher accuracy than chance we build two more models. One which instead of trying to predict the exact age, we trained to predict if a person was above the age of 18 or under. The model we chose to work as a example of a pure chance model, was a classifier predicting the star signs of an author(with the assumption that star signs and writing style have no noticeable correlation

Results:

Baseline

To evaluate the baseline model we first looked at the accuracy and precision scores. To our surprise this initial model already had an accuracy of around 22-24%. Looking at the precision we saw that some age groups had very low scores. To find out more about where our model failed, we calculated how often the model was off by specific years. THis data can be seen in fig. 1. As visible in the figure most mistakes happened in the 1-2 year difference. This was partly our reason for trying a model with grouped ages. Visible in figure as well is that the biggest mistake we found was 35 years. To further evaluate how good the model performed we calculated the average age our model was off from the actual age of the author. Surprisingly as well this came out to be only 4.7-4.9 years.

Fig. 1



As outlined in the sections of the other models, running them with our completed dataset was not possible due to technical limitations therefore we ran them on a subset of our dataset containing only 50 000 entries. To still accurately compare all models we ran our initial model as well once on this reduced data set. With this reduced dataset this model achieved an accuracy of around 19% with an average age difference between predicted and actual age of around 5.0-5.1 years and a max difference of 35 years.

Logistic Regression + Ngrams:

Implementing ngrams we expected the model to take longer to train. We did however underestimate the severity of this increase. The baseline model took around 1 hour to train. After running this new model for over 6 hours we decided to manually stop the run and try with a reduced data set (50 000 entries). Even with this reduced input size the model still took more than 4 hours to train. To our surprise the model performed slightly worse than the baseline only achieving an accuracy of around 18%.

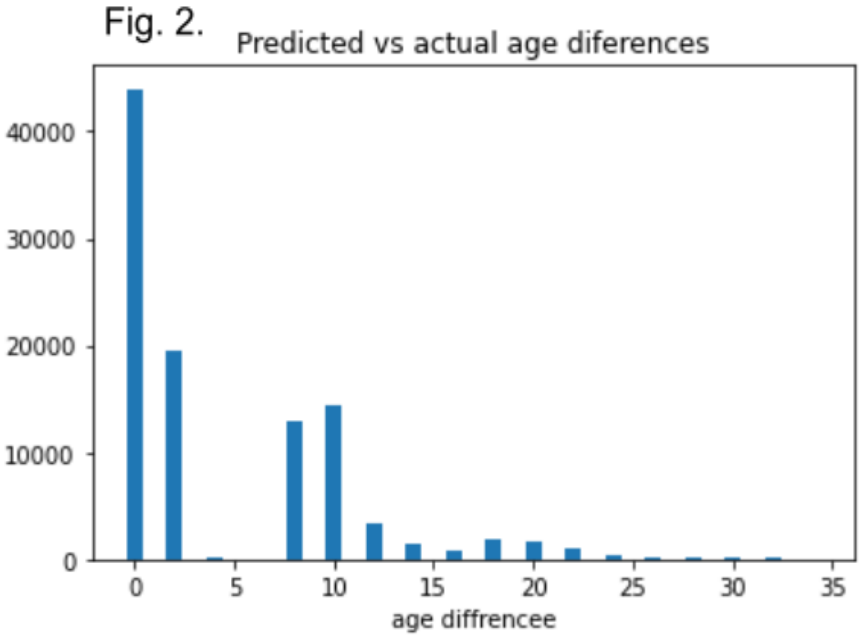
Logistic Regression + combined ages:

We decided to train a model with combined ages of 2 years for 3 reasons. Those were:

- The age number we use does not accurately represent the time difference between people on small scales, 1-2 years, as for example two people aged 15 and 16 could be 364 days apart or just 1. Combining would not completely solve this problem but reduce the influence it has on our model.
- Our main confusion between categories/ages was between ages 1 year apart. Combining ages in groups could prevent these confusions.
- Some age groups, especially older ones, did not have a lot of data. Combining ages would give the classifier more info per category to learn.

Unsurprisingly this model performed better giving us an accuracy of around 42% and an average difference between predicted and actual age of 4.5 years. Building this model and also seeing the accuracy we expected a lower average difference. Receiving this difference and also plotting the difference distribution graph we concluded that the increase in accuracy mostly came from the decrease in categories and not actually from a better trained model.

Fig. 2.



Support Vector Classifier:

Running the SVC we ran into a similar problem as with the Logistic Regression + ngrams model. Running the model overnight was not long enough to actually complete it with our current computers. Therefore we choose to adapt the same reduced dataset method. Running it with 50 000 entries still ended up taking over 5 hours. The model achieved an accuracy of around 21% with a age difference of 4.9 and a max error of 33 years.

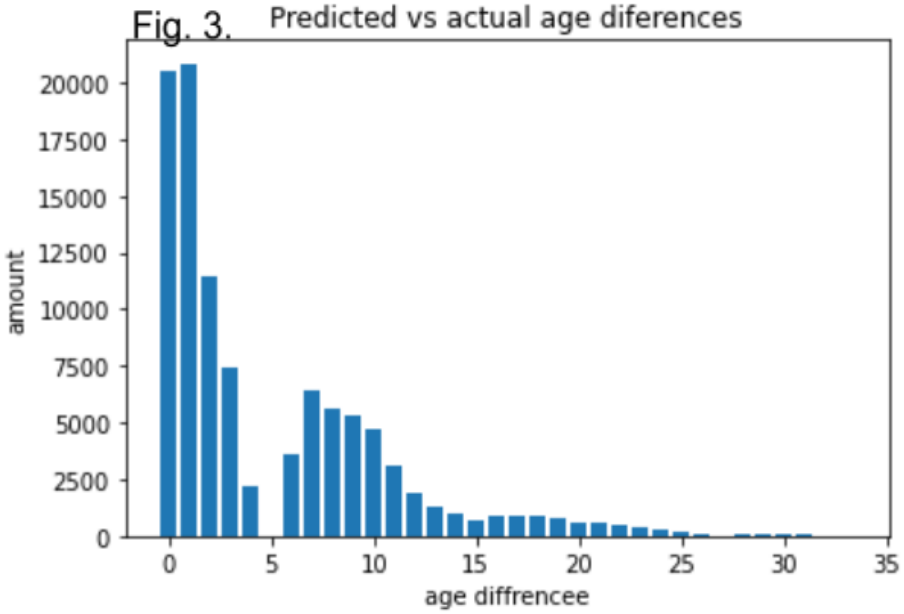


Fig. 3.

Initially we also planned on running this classifier in combination with the ngrams, but based on our previous tests we concluded that this would also not be possible with the current resources available to us.

Comparison models:

star signs:

The star sign model as expected only achieved an accuracy of 17% (not directly comparable as only 12 categories instead of 26). Looking at the average difference as well as at the difference distribution(Fig.4.) it is visible that this model predicts the sign based only on the distribution of the signs provided in the dataset.

U18 vs 18+:

To see how well our simple model holds up to the literature we grouped all ages above 18 and below 18 and trained a simple model(same as baseline) with these two categories. Our model achieved a accuracy of 81% which we found quite surprising as thai is already very accurate with a unoptimized and simple model.

Discussion and conclusion:

Going into this project we expected to be limited by our current resources. We did, however, underestimate how much this would limit us in our progress. Initially we planned on testing a variety of different classifiers, as well as optimizing them through the use of GridSearch or RandomizedSearch. Due to the long waiting times, even with a reduced dataset, we were, however, forced to limit ourself to mostly logistic regression. While we weren't able to fully test the capabilities of the SVC, we did show that it performed better on a smaller data set than the logistic regression classifier. This indicates that with a larger data set SVC would also perform better. Based on this we think combining SVC , the combined age groups and optimization could provide a quite powerful and accurate model. While unsurprisingly comparing our models with the star sign model, we did find that age does indeed change one's writing style and that this is not limited to the earlier years of one's life.