Assignment 4 Report

Max Neerken - s1081717
Paul Verhoeven - s1086755

November 18, 2024

1 Feature design

In this section we go over the extracted features (175) from the texts and briefly talk about the motivation. Many features were inspired by Zheng et al. (2006), who described a framework with many features and feature groups.

1.1 Number of characters

First of all we count the number of characters in a text, this feature was also used to normalize features 2 and 3. Some author for example may have many words/tokens or sentences but can ultimately use very few characters as a writing style/word choice.

1.2 Number of sentences

The second feature is the number of sentences proportional to the length of the text (counted by chararacters). We again expect this to be a very stylistic and author-specific feature.

1.3 Number of tokens

The third feature was a simple token count also proportional to the length of the text as we again expected wordy texts to be an indication of a specific writing style an author might have.

1.4 Number of Number of words without vowels

The fourth feature was a count of words without vowels, whether typo or not these words are rare and could possibly be attributed to an author.

1.5 Special characters

We wrote a function that counted the number of special characters and digits given a text. We considered the following characters special: , . / ? : ! % & ( ) + - = _ " ' \ @ # $ ‘ ˜ { } [ ] < > | ˆand the sum of digits spanning number 0 through 9. This accounted for 31 of the 175 features. These features were selected as we expected them to be very indicative of a certain author.

1.6 Continuous punctuation characters

The continuous punctuation such as '...' or '!!!' can be an indication of writing style. A story-telling text could use it to accentuate a character yelling!!! While a tweet may want to indicate sarcasm or disappointment... Which is an important feature to have for authors.

1.7 Contraction count

Contractions can of course also be very style-specific, while some authors **do not** use contractions, some authors simply **don't** care for such things and thus we added it to our wash-list of features.

1.8 Words in all capital letters

The reasoning behind counting the all-cap words is quite similar to the continuous punctuation of characters. It can be very indicative of an author trying to get a point across in their own unique way. ITS WHY WE ADDED IT AS A FEATURE.

1.9 Emoticon count

Some authors that write informal texts might have used text-based-emoticons which could be a dead giveaway of their hidden identity. The emoticons we searched for are as follows: :), :(, :D, :P, :o, :/, >:(, ^_^, T_T, :-), :-(, :-D, :-P, :-o, :-/, >:), ^^, :'(.

1.10 Happy emoticon count

To go into further depth we considered happy emoticons! Everyone knows happy people use happy emoticons, so this makes identifying them trivial. We considered the following emoticons to make us happy: :), :D, :-),:-D, (^_^), (^o^), (:, (^.^), :], c:, ^^

1.11 sentence first-letter lower case

starting sentences in lower case, can indicate a person is influent in english, or is sloppy with their spelling.

1.12 Part of speech tags

The number of different part of speech words such as nouns and verbs proportional to all words. This provides insight in the type of words an author uses.

1.13 Letter frequency

The number of letters proportional to all characters. This informs us about the the quantity of alphabetic characters in a text compared to non-alphabetic characters such as symbols and numbers.

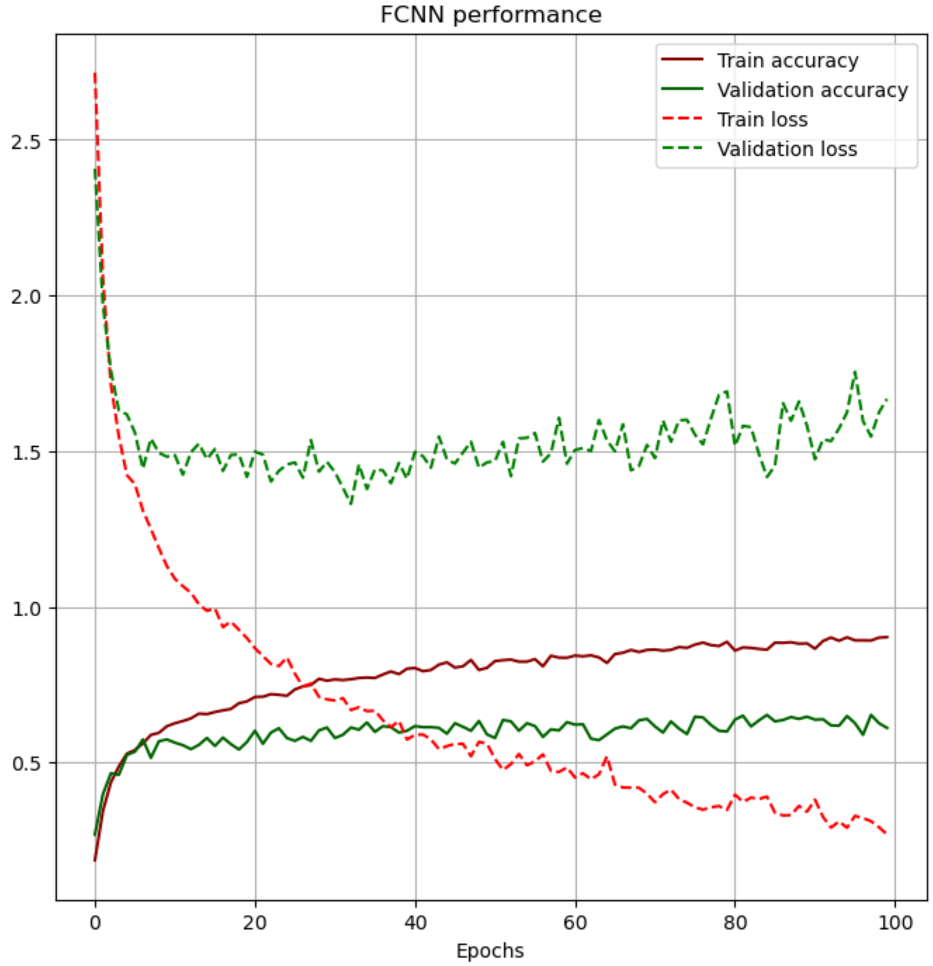

Figure 1: Train and validation loss and accuracy over epochs. We can see that the network does not improve from about epoch 30.

1.14 Function word frequency

Function words such as determiners or conjunctions add context to sentences. Some people use them often, whilst others use them sparsely. The function words were inspired by Zheng et al. (2006)

1.15 Lower case 'i' frequency

Spelling I in lower case, can indicate a person is influent in english, or is sloppy with their spelling.

2 Evaluation of classifier

For our classifier we trained a fully connected neural network. We trained this neural network until convergence as can be seen in figure 1. We can see that while the train accuracy is still trending upwards our validation accuracy flat-lines and its respective loss increases which a clear indication of over-fitting the model. At best we reach approximately 60% accuracy on the validation set with our model.

Recall	0.67
Precision	0.71
F-Score	0.66

Table 1: Result metrics on test set

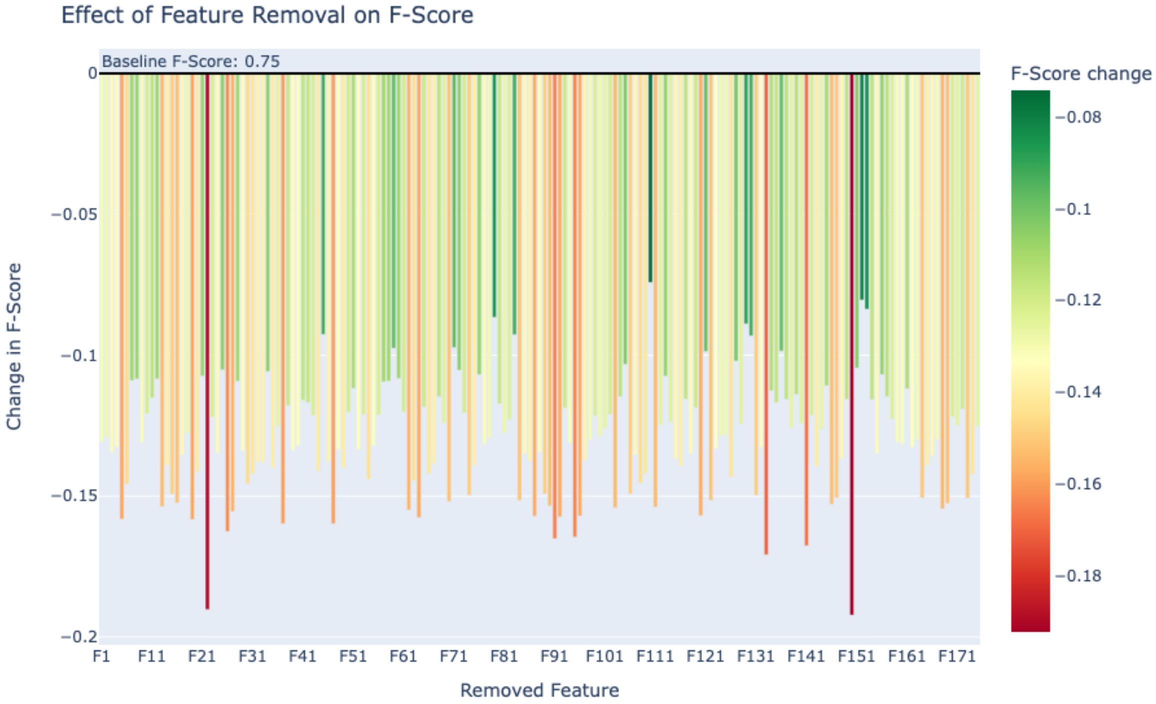

Figure 2: The difference in F-Scores when removing different features. All features improve the F-Score of the model.

The performance on the test set (table 1) shows that while the model performs significantly better than random guessing, is not reliable enough for professional use.

To investigate the effectiveness of our extracted features, we conducted ablation analysis. The model was trained and evaluated multiple times, each time leaving out 1 feature. This allows us to get an indication of how important the features are for the model performance. The results can be seen in figure 2

We can see that some features had a bigger impact than others. However, all features contribute to the F-Score. None of them increased the F-Score. This indicates that leaving out any features, would not increase the effectiveness of the model. Furthermore, as the model is not that accurate, it would not make sense to drop features in order to make the model more efficient if that will hurt the effectiveness.

3 Reflection

Firstly the fact that our model only reaches about 60% accuracy (and its validation loss is also not great) feels more like an educated guess as opposed to a model that has a high certainty. We have tried to implement a different classifier namely a binary classification neural network that would train and predict that given a text does it belong to a certain author yes or no. However, some authors were under-represented whilst others were over-represented and we could already see issues forming with the training. Additionally we would have to train 20 networks, 1 for each author, which would have been more computationally intensive.

1. Did not look at leaving out a combination of features 2. Only evaluated the f-score once, which could lead to fluctuating results due to the initial weights of the model. 3.

When provided with some more time, we could have looked at more features to represent the texts of different authors, maybe zooming in on some texts to come up with dataset specific features.

We would also have changed the ablation analysis to use a mean of F-Scores over multiple runs. Now in the ablation analysis the model is trained and the F-Score is measured. However, this causes the results to vary between runs because of the way the weights are initialized. By averaging over multiple runs we solve this problem.

We also got the insight that from a linguistic point of view authorship attribution is hard because of things like:

- Stylistic Similarity
  - Between Authors; authors who share similar cultural, educational, or linguistic backgrounds may have indistinguishable writing styles.
  - Mimicry: An author intentionally mimicking another's style to obscure their identity or create confusion complicates attribution.
- Stylistic Variability of an Author
  - Across Genres: An author's style can vary between writing genres (e.g., fiction vs. academic papers).
  - Over Time: Writing style evolves due to age, experience, or changing contexts.
  - Different Contexts: Informal communication differs significantly from formal writing.

It is difficult to come up with meaningful features

- Make sure that your report answers these questions:
  - What problems did you encounter? Do you see systematic patterns in the mistakes that are made by your best performing feature set (failure analysis)?
  - What problems could you solve with 10 hours extra time?
  - What problems do you think are real challenges in authorship attribution?
- Feel free to add other issues you want to reflect upon.

4 References

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the American society for information science and technology, 57(3), 378-393.