# Authorship Attribution

## Assignment 4 - Text and Multimedia Mining

Feida Wei
s1086990

Vinícius R.M. Schmidt
s1123702

November 19, 2024

## 1 Introduction

This report goes over the process of building a classifier capable of performing authorship attribution. The dataset used for this task was sampled from the PAN2020 dataset[1], and consists of fan-fiction pieces written by 20 different authors. In the following sections we will briefly discuss the choice of features used to train the classifier, as well as the impact of each feature group on the performance of the classifier.

## 2 Feature selection

For selecting the features, we took inspiration from the work of Solorio et al. (2013), who based most of their work on common features for authorship attribution. These features were proven quite useful for this sort of task.

### 2.1 Counts over lexical characteristics

This group consists of the following features:

- **Sentence count**: Sequence of characters separated by ".", ";", ":", "!" or "?"
- **Token count**: Sequence of characters with no spaces between them
- **Words without vowels count**
- **Uppercase words count**: Words consisting of two or more letters, all uppercase. This excludes words like "I", which are probably very frequent across all texts.

These features were included because they are quite easy to extract and they can already reveal quite a bit of information about the author. For instance, if two texts have the same character count, but one has a significantly lower token count, it probably uses longer and more complex words than one with a higher token count. The same applies for sentence count, which indicates longer or shorter sentences.

### 2.2 Aggregated character-level features

This group of features includes counts of different types of characters, e.g. alphabetical, numerical, punctuation. The features belonging to this group that we included in our model were:

- **Character count**: The length of the text itself
- **Punctuation marks count**: This feature counts the total frequency of the characters ".", ",", ";", ":", "!" and "?" in the text.
- **Continuous punctuation marks count**: This feature counts any sequence of two or more of ".", "!" or "?". Combinations of characters are allowed, e.g. "!?".

[1] https://pan.webis.de/clef20/pan20-web/author-identification.html

The reason for choosing the aforementioned features specifically is because punctuation tells something about sentence length. Some authors write longer sentences and thus use fewer punctuation characters, whereas others write shorter sentences leading to more frequent use of punctuation. In addition, punctuation can be used to express the intensity of a feeling in a dialog, e.g. an author uses "???" frequently to express intrigue of a character. This is also a question of preference that may vary per author.

### 2.3 Syntactic features

These group includes the following features:

- **POS-tags count**: In total, 36 POS-tags from the Penn Treebank project (Marcus et al., 1993) were individually counted. For a complete list of the tags used, see Appendix A
- **Function words count**: Based on Solorio et al.'s work, we counted 150 different function words. The full list of the words used can be found in Appendix B

These features were chosen because authors tend to unconsciously use the same sentence structure in their texts. This also applies to function words, which seem irrelevant on the surface, but their usage can be quite unique per author.

## 3 Classifier

### 3.1 Ablation analysis

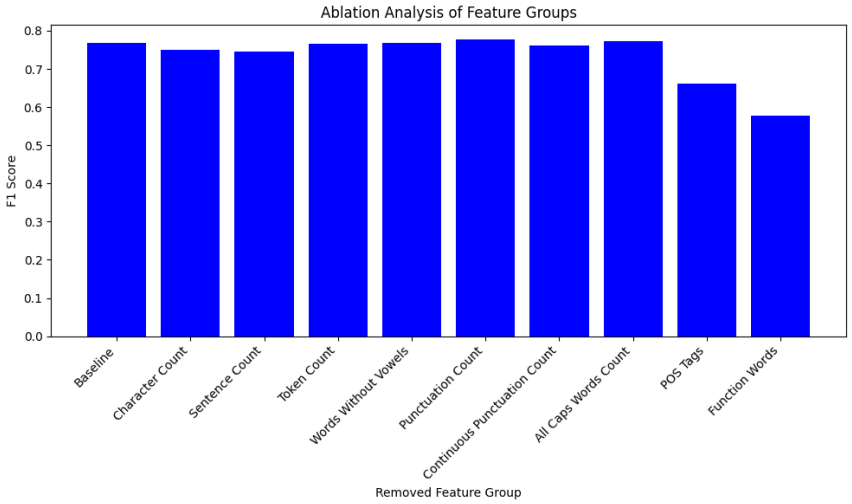


Figure 1: Ablation analysis of the SVC classifier by F1 score on the development set. Each column represents the performance of the classifier when leaving out each of the features mentioned in Section 2.

For our classifier, we decided to use a Support Vector Classifier (SVC)[2] from the scikit-learn library (Buitinck et al., 2013). As we can see in Figure 1, the classifier achieved a pretty good F1 score of 0.769 on the development set when using all of the features. However, we can see that most of the features do not contribute significantly to this result, since the F1 score stays around the same value when excluding lexical and character features, so applying feature selection would definitely be useful in this case. The most significant changes come from POS-tags and function words counts. We believe the reason for this to be two-fold: firstly, these groups comprise the vast majority of the features (36 and 150 respectively), secondly, these features tend to tell more about someone's writing style because people unconsciously use the same writing structures in their texts.

Since POS-tags and function words were the most useful features, we decided to run our classifier again on the development set, using only them to see how well it would perform. It achieved a F1 score of 0.713, which is already pretty good, since it surpasses the 0.7 cutoff point. When it comes to the test set, our classifier achieved a F1 score of 0.734 when using all of the feature and 0.674 when using only POS-tags and function words. We believe that if we had more time to run experiments,

[2]https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

we could find more suitable combinations of POS-tags and function word features, combined with other lexical features, to achieve a F1 score of at least 0.7.

### 3.2 Confusion matrix

In order to see what mistakes the classifier was making, we plotted its predictions of the development set on a confusion matrix (Figure 2).

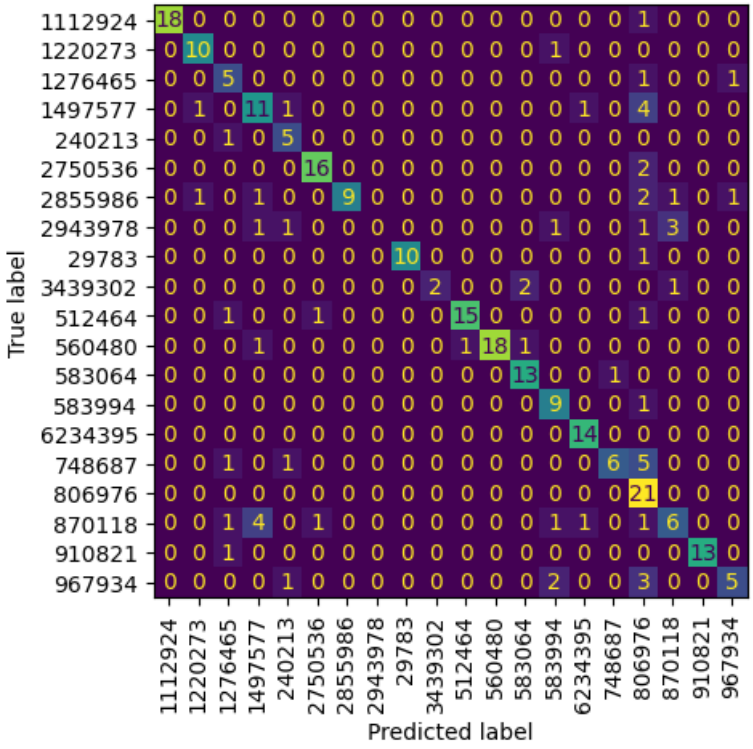


Figure 2: Confusion matrix of the classifier results on the

As we can see, most of the authors are quite easy to recognize, with only one or two instances being misclassified. However, if we look at author "2943978", we can see that none of the texts were classified correctly. Upon closer inspection of the training set, we can see that this author in particular appears significantly less than other authors, which could explain why the classifier performs so poorly on this author's texts. Specifically, there are 51 texts by this author in the training set, whereas other authors all have 71 or more. Also, we can see that in general there is a class imbalance in the dataset, which is an impactful aspect on any machine learning task. Thus, dealing with this imbalance could help improve the model's performance in addition to feature selection, which was mentioned previously.

## References


Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). API design for machine learning software: Experiences from the scikit-learn project. *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108–122.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank (J. Hirschberg, Ed.). *Computational Linguistics*, *19*(2), 313–330. https://aclanthology.org/J93-2004

Solorio, T., Hasan, R., & Mizan, M. (2013, June). A case study of sockpuppet detection in Wikipedia. In C. Danescu-Niculescu-Mizil, A. Farzindar, M. Gamon, D. Inkpen, & M. Nagarajan (Eds.), *Proceedings of the workshop on language analysis in social media* (pp. 59–68). Association for Computational Linguistics. https://aclanthology.org/W13-1107

# Appendix

## A POS-tags

| *Tag* | *Description* |
|---|---|
| CC | Coordinating conjunction |
| CD | Cardinal number |
| DT | Determiner |
| EX | Existential there |
| FW | Foreign word |
| IN | Preposition or subordinating conjunction |
| JJ | Adjective |
| JJR | Adjective, comparative |
| JJS | Adjective, superlative |
| LS | List item marker |
| MD | Modal |
| NN | Noun, singular or mass |
| NNS | Noun, plural |
| NNP | Proper noun, singular |
| NNPS | Proper noun, plural |
| PDT | Predeterminer |
| POS | Possessive ending |
| PRP | Personal pronoun |
| PRP$ | Possessive pronoun |
| RB | Adverb |
| RBR | Adverb, comparative |
| RBS | Adverb, superlative |
| RP | Particle |
| SYM | Symbol |
| TO | to |
| UH | Interjection |
| VB | Verb, base form |
| VBD | Verb, past tense |
| VBG | Verb, gerund or present participle |
| VBN | Verb, past participle |
| VBP | Verb, non-3rd person singular present |
| VBZ | Verb, 3rd person singular present |
| WDT | Wh-determiner |
| WP | Wh-pronoun |
| WP$ | Possessive wh-pronoun |
| WRB | Wh-adverb |

Table 1: Overview of the POS-tags (Marcus et al., 1993) features used for our classifier.

## B Function words list

a, between, in, nor, some, upon, about, both, including, nothing, somebody, us, above, but, inside, of, someone, used, after, by, into, off, something, via, all, can, is, on, such, we, although, cos, it, once, than, what, am, do, its, one, that, whatever, among, down, latter, onto, the, when, an, each, less, opposite, their, where, and, either, like, or, them, whether, another, enough, little, our, these, which, any, every, lots, outside, they, while, anybody, everybody, many, over, this, who, anyone, everyone, me, own, those, whoever, anything, everything, more, past, though, whom, are, few, most, per, through, whose, around, following, much, plenty, till, will, as, for, must, plus, to, with, at, from, my, regarding, toward, within, be, have, near, same, towards, without, because, he, need, several, under, worth, before, her, neither, she, unless, would, behind, him, no, should, unlike, yes, below, i, nobody, since, until, you, beside, if, none, so, up, your