Method {#chap:meth}

As previously discussed, there are a number of different ways to appraoch authorship attribution tasks. Specifically, we experiment with both authorhip identification and authorship verification tasks, using
different methods. These methods can be broadly broken into

- Statistical methods, in which we compare texts using (unsupervised) statistical methods.
- Support Vector Machine (SVM) methods, in which we use SVMs to discriminate between authorship styles.
- Neural Network approaches, in which we experiment with various formulations of AA tasks and various models, including Siamese Networks and generative language modelling.

In the rest of this chapter, we explain the set up for various experiments covering these three methods. Due to practical and theoretical limitations, we did not experiment with all methods on all tasks. Nor did we use every dataset for every experiment. In !REF, we present an overview of which methods we tried on which datasets.

Basic statistical approaches

Previously, we discussed the distinction between *intrinsic* and *extrinsic* approaches to authorship attribution. Recall that the former relies only on comparing the target texts (e.g. a known and an unknown text) to each other, while the latter compares the target texts to a corpus of external texts.

In this section, we describe our experiments involving *extrinsic* approaches. Supervised machine learning has become the go-to method for many text classification tasks, but the resulting models are often seen as a "black box". Data goes in and predictions come out, and it can be difficult to justify or explain these predictions. Here, we explain how we can meaningfully analyse texts using some basic statistical approaches, including correlation techniques, on a number of different features. For these techniques, we experimented only with authorship verification tasks.

Authorship verification tasks

To decide if two texts are written by the same author, we can compare various features in each text to some large corpus of (preferably similar) texts. If, for example, both texts have (when compared to the corpus) a higher-than-average sentence length, or a lower-than-average adjective frequency, then we have some small piece of evidence that the two texts were written by the same author. If the two texts relate to the corpus similarly over a variety of different features, then we have more evidence that the two texts share an author.

We experiment with two ways of comparing our target texts to a corpus. Firstly, we use a simple count-based approach based on supporting and opposing points for the hypothesis that the texts share an author. Secondly we use a correlation analysis, to see if the two texts diverge from the corpus in a similar fashion.

We use lexical and syntactic features for these experiments, including sentence length, word length, various readability scores, parts-of-speech (PoS) tag relative frequency, and frequency of the most-common function words. Because we use more features here than in our later experiments, we carry out these experiments only on the smaller English datasets, C50 and the English parts of PAN.

Count-based statistical experiments

First, we use a simple count-based approach. If two texts, compared to the corpus, either:

a) Both have a feature that is higher-than-average (e.g. both texts have a higher-than-average adjective frequency), or

b) Both have a feature that is lower-than-average (e.g. both texts have an average sentence length that is lower than the average of the corpus)

then we take this to be a point supporting the hypothesis that the texts share an author.

If there is a feature such that one text is higher-than-average while the other is lower-than-average (e.g. Text 1 uses more commas than the corpus-average while Text 2 uses fewer commas than the corpus-average), then we take this to be an opposing point.

We then classify texts into same-author or different-author classes based on whether there are more supporting points than opposing ones.

Correlation based statistical experiments

A more refined way is to look do a correlation test between the two sets of scores. If one text has a much higher than average sentence length, and the other has only a slightly higher than average sentence length,
then this should be less indicative of same-author than if the two texts both have a much higher than average sentence length.

If two texts arewritten by the same author, we expect the exact way that each diverges from the reference corpus to correlate quite strongly. If the texts are written by different authors, we expect a weaker correlation.

Support vector machine approaches

Support Vector Machines (SVMs) have proven to be efficient and powerful for a wide-variety of text-classifciation tasks. For all of our multi-class SVM experiments, we use a one-against-all strategy. That is, if we have 50 candidate authors, we train 50 SVMs, and each one is trained to discriminate in a binary fashion between its author and all other authors.

Because SVMs are fast to train and scale well to large datasets, we present SVM-based experiments for all of our datasets.

Authorship identification tasks

For the identification task, we can simply train SVMs on the authors knowns texts, and attempt to predict the author of unknown texts from the pool of candidate authors. While this is perhaps the most-common set up for authorship attribution tasks, it is also the least interesting. Generally, good results are only feasible for a small pool (fewer than 100) candidate authors, and practically there are very few cases where we want to know which of small, well-defined set of authors wrote a particular text. Nonetheless, we use this set up to experiment with different numbers of authors and different features, predominately word and character n-grams represented as Term Frequency - Inverse Document Frequency (TF-IDF) vectors.

Authorship verification tasks

Normally for text classification tasks, the goal is to apply a label to each text. For authorship verification tasks, we need to say something meaningful about the relation between a pair of texts. A naive approach might be to concatenate the two texts and then attempt to assign a "yes" or "no" label. However, we don't want to learn rules such as "if either text contains the word *discombobulation* then the two texts are likely to be written by the same author" which is unfortunately the kind of rule that a traditional classifier would learn if we trained it on concatenated pairs of texts.

Therefore, to learn something meaningful about text pairs, we need to look at the similarity between the two texts. Often, the cosine distance between the vector representations of each text is used for such tasks. However, two different authors often write texts that have a strong cosine similarity (for example, two authors writing about the same topic), and conversely a single author will often produce two texts that appear very dissimilar by the cosine metric (for example, if the author writes on two different topics).

By subtracting TF-IDF vectors of each text, we get a feature set that ismore meaningful. The SVM can learn which word- and character n-grams are indicate authorship, and can learn rules such as "if the count of the word "the" is similar in both texts, then it is likely that they are written by the same author".

Neural network approaches

Neural Networks (NNs), especially Recurrent Neural Networks (RNNs) with Long Short-Term Memory (RNN-LSTMs) or with Gated Recurrent Units (RNN-GRUs) have been the poster-child of many Natural Language Processing (NLP) advancements in the last few years. As discussed before, they have failed so far to become state-of-the-art for Authorship Attribution tasks. This is for several reasons, including:

- RNNs are difficult to train and not yet fully understood
- RNNs usually require much larger amounts of data than existing AA datasets
- The feedback cycle for RNN classifiers can be much longer than alternative approaches such as SVMs. For example, for one of our experiments FIXME REF, the SVM took about three minutes to train and evaluate, while the RNN model took over 14 hours.

Our experiments with Neural Network classification can be broken into three main categories:

- **Discriminative Text Classification**: Similarly to our SVM identification experiments, we model the authorship identification problem as a traditional multi-class classification task.

Instead of TF-IDF vectors, we use word embeddings, and instead of the SVM classifier, we use a neural network model.
- **Generative, author-specific language modelling**: Following recent advances in neural character-based language modelling, it is possible to generate text in the style of a specific author (for example, Karpathy FIXME REF). By training language models to mimic the writing style of specific authors, we can classify unknown texts by seeing which author-specific language model best models that text.
- **Siamese neural networks**: We use a neural network that contains two sets of identical weights and learns a custom distance function between pairs of inputs. This fits our verification task very well.

Authorship identification tasks

Discriminative classifiers

TODO or remove

Generative language modelling

If we can generate text in the style of a specific author, we can also recognise it. Previous work has shown that generative classifiers can be more effective than discriminative ones, even for discriminative tasks. For these experiments, we analyse various ways to create personalized language models, including different architectures, and using transfer learning.

Authorship verification tasks

Discriminative classifiers

Generative language modelling

Because identification tasks can often be reformulated as verification tasks, and vice-versa, we also attempt some of the verification tasks using the generative personalized models described above. For verification, we train personalized language models for each *known* text. Then we compute the cross entropy for each unknown text, for each model. If the unknown text has a lower cross entropy than more than half of the personalized models, that provides an indication that the two texts are by the same author.

Siamese networks

TODO