university of groningen / faculty of arts

Genetic Influence on Writing Style
Authorship Discrimination and DNA

Kai Bruijn

Master thesis
Information Science
K.A. Bruijn
s3204766
June 8, 2020

ABSTRACT

Although authorship attribution itself is a much researched field, previous work has had no focus on the influence of genetics on the possibility of being able to distinguish between authors. This study shows that genetics have a very limited influence on writing style. Based on personal experience and biological intuition, I hypothesised that DNA has an influence on a person's writing style and tried to answer the question to what extent this is true.

The paper shows that humans disagree on writing style similarity and perform bad on authorship discrimination, not being able to tell whether or not two texts are written by the same person. Hence, research was done by creating a data set of texts written by twins, both identical and fraternal, as well as their siblings and testing an existing, robust machine learning model by Hürlimann et al. (2015) on 130 instances. The model provides probabilities of the texts being written by the same person and was trained on a new data set containing Reddit texts, based on the genre of the twin texts.

The results show that twins write the most similar, followed by siblings and that texts written by two random people show the least similarity. This confirms the hypothesis that DNA influences writing style. However, as twins and siblings have either the same or similar ages, this had to be taken into account as well as gender. After correcting for age and gender, the hypothesis is rejected as the probability scores are the same for all groups.

Also there are no linguistic features specifically influenced by DNA, as confirmed by the feature analysis after correction for age and gender. This concludes that no authorship de-identification is possible for authors with genetic resemblance in a group of random people.

The conclusion shows that authorship discrimination is harder when the two authors have the same gender and especially when their difference in age is less than 10 years.

CONTENTS

PREFACE

The past few months, it has been a great pleasure to write a master thesis on a subject that lies so closely to my heart. I enjoyed researching the subject of writing styles of identical twins as there have been situations where I had to explain why my work was so similar to my twin brother's work, one time even leading to a reduction of points.

This paper is closely related to the bachelor thesis of my brother and I, where we carried out research in the same field. In my bachelor thesis I researched in a personal environment how similar my family writes, because we have such a suitable family to research this topic with four boys, two of them being an identical twin, one of them writing this at this very moment (Bruijn, 2019a). My twin brother, Rune, researched resemblance in writing styles of an identical twin who grew up in different environments, an interesting research as well (Bruijn, 2019b).

I also want to mention his master thesis, again researching authorship attribution and genetics (Bruijn, 2020). If you enjoy reading this paper, you will likely want read his paper as well. We collaborated on the data collection, for which he deserves part of the credits. We worked together on this, smooth as always, so I want to thank him for the cooperation.

Being part of an identical twin and studying Information Science, this thesis is a perfect ending to another great year of studying at the University of Groningen. Unfortunately, this last year was not as I had expected it to be because of the COVID-19 virus. Because of it, a lot of work had to be done from home and I want to thank my supervisor, prof. dr. M. Nissim, especially for supervising me in such an inconvenient and challenging time, as well as for the provision of a research topic that fits so well.

1 INTRODUCTION

A person's writing style is something that has a part in the individuality of a person. It distinguishes one writer from another and can profile authors, i.e. uncovering various characteristics of an author based on stylistic and content-based features (Rangel et al., 2013). The analysis of writing style can also be used for authorship de-identification to identify features which can properly capture an author's writing style (Swain et al., 2017). But the applicability of analyzing writing style does not end here.

An interesting field to discover is authorship attribution, which is where the focus lies in this thesis: authorship discrimination with genetic resemblance. Authorship discrimination is a sub field of authorship attribution used to distinguish authors based on their writing style where the objective is to predict if two or more texts are written by the same person. This could be of great importance in practical applications such as detecting plagiarism but also in criminology.

Building on preliminary research for my bachelor thesis and based on personal experiences, being part of an identical twin who followed the same study, I know that my twin brother and I share many characteristics, one of which is the way we write. I noticed that my twin brother and I have a similar writing style, which leads to my expectation that other twins may have similar writing styles as well. Based on this notion, I hypothesise that genetics have an influence on an individual's writing style, and in this paper try to answer the research question: *"To what extent does DNA influence an author's writing style?"*. After running predictive models, a feature analysis is done to research which linguistic features are influenced by our DNA.

This research could lead to new insights in a yet unexplored field where authorship attribution and genetics are combined, as, to the best of my knowledge, no previous work has done any research on this matter. The research question will be answered by using the best available model for authorship discrimination, tested on a created data set containing texts written by twins and their siblings, while trained on a collected data set with a similar genre.

The model used for obtaining the probability of two texts being written by the same person is GLAD (Hürlimann et al., 2015). This model was re-adapted to fit the needs of this research. I also experimented with state-of-the-art models such as the Dutch Transformer model BERTje (de Vries et al., 2019), as well as clustering, alternative support vector machines and linear regression to try and outperform the model of Hürlimann et al. (2015), however I did not reach better performance than they did.

The structure of this research is as follows. Chapter 2 describes previous work that has focused on similar topics. As mentioned before, no previous work has really focused on this field but on authorship attribution, as well as genetics, much research has been done. Chapter 3 will cover the data and materials used for the research. This includes the created data set of twin texts as well as the training data scraped from Reddit. This will be followed by Chapter 4, explaining the method used and Chapter 5 will show the results, complemented with a discussion. This also includes a correction for the influence of age and gender and an analysis based on leaving features out to compare probability scores. Finally, the conclusions can be found in Chapter 6.

2 BACKGROUND

As no previous work has focused on the combination of authorship attribution and genetics, these two fields will be discussed separately in this section. First, important previous work on authorship attribution will be discussed, followed by a section on genetics. As a non-expert in this field, in the second section conclusions ought to be drawn cautiously.

2.1 AUTHORSHIP ATTRIBUTION

Authorship attribution is the process of identifying the unknown author of a given text, by looking at differences between texts in writing style, for example sentence length and use of punctuation. This can be applied in criminology or plagiarism detection. Authors can be distinguished by using these kind of features, often done in the field of computational linguistics. Authorship attribution has a long history starting from the 19th century where heavily statistical models were used instead of prediction models, and has been getting more attention in the past few decades (Bourne, 1897).

Stamatatos (2009) provided a survey of authorship attribution methods in his paper, where he analyses authorship attribution methods, focuses on computational requirements and discusses evaluation methodologies and criteria for authorship attribution studies. He concludes that significant steps have been made with regard to the applicability of authorship attribution as it provides robust methods, such as support vector machines, able to handle real-world texts with relatively high-accuracy results, if the test set is balanced, by looking at stylometric features such as character n-grams.

Also, there are annual competitions on authorship attribution problems called the PAN[1] authorship attribution shared tasks, to obtain better performance and new insights in the field. The model that I re-adapted was trained for the 2015 edition of PAN (Hürlimann et al., 2015).

In such shared tasks, participants often submit systems that make use of support vector machines, as did Hürlimann et al. (2015), and sometimes neural networks, though the first generally perform better. The best systems usually obtain accuracy scores of around 0.7, where often the problem is approached using a closed set of possible authors. Hürlimann et al. (2015) obtained an accuracy of .73 in their research for a Dutch authorship discrimination problem, positioning themselves amongst the top participants. They trained a binary linear classifier both on the features describing known and unknown documents individually, and on the joint features comparing these two types of documents. The list of feature types includes, among others, character n-grams, the lexical overlap, visual text properties and a compression measure.

As the model of Hürlimann et al. (2015) was built several years ago, it may not be very state-of-the-art compared to other natural language processing tasks, where neural networks and recently Transformer models have taken over the lead position. However, to date there have been no real breakthroughs when neural networks are applied on authorship attribution tasks. In 2019, more than half of the participants who entered the shared task of Kestemont et al. (2019), where the task was to attribute an unknown text to a previously seen candidate author, used

[1] http://pan.webis.de/

support vector machines and in fact none submitted neural networks as their final system.

It is difficult to explain why neural networks perform well in almost all natural language processing tasks except authorship attribution, but the key may be that authorship attribution has such a focus on writing style. Bozkurt et al. (2007) attempt to come up with an explanation using the following observation:

> *"Authorship attribution is a kind of classification problem. But it is different from text classification, because style of writing is also important in authorship attribution as well as text content which is the only factor used in text categorization. Also, with different data (e.g. books, articles), the classifiers and feature sets may behave differently. Also in authorship attribution, the feature set is not deterministic as in text categorization. So, these differences make authorship attribution task more challenging."*

In the same paper they also prove that neural networks perform much worse on their task than other authorship attribution methods, for example support vector machines as they observed success rates of 30 and 90 percent respectively.

They used Turkish newspapers as a data set and a feature set consisting of stylometry, vocabulary diversity, bag of words and frequency of function words. The features specified to stylometry are the number of sentences in an article, number of words in an article, average number of words in a sentence, average word length in an article, vocabulary size of author (word richness), number of periods, number of exclamation marks, number of commas, number of colons, number of semicolons and number of incomplete sentences. This resembles the approach and features of Hürlimann et al. (2015).

The shared task organized by Kestemont et al. (2019) is important to mention for yet another reason as the setting of the task is cross-domain. Accuracy scores of authorship attribution shared tasks are very informative with regard to performance, however it can be much influenced by the nature of the task. If the task has an in-domain setting, it means that the genre of the texts of authors is similar and constant. However, in real-life applications, texts are not necessarily of the same genre, considering for example a threatening letter to a political person. When applied on a cross-domain setting, the results could be much worse and thus highly influenced by the genre of texts and therefore not really applicable in actual tasks.

However, though many of these papers are very interesting or groundbreaking, no papers exist that have focused on the influence of genetics during the process, which is based on the distinctions made, or as Hürlimann et al. (2015) did, the recognition, between writing styles. The only ones to have combined genetics with writing style are Bruijn (2019a) and Bruijn (2019b) who carried out preliminary research for their bachelor thesis. Bruijn (2019a) used Dutch texts of at least one hundred words written by 6 people on 3 specific topics, thus 18 texts. Bruijn (2019b) used English texts of a blogger and her identical twin, also a blogger. He used 122 texts with an average of around 500 characters.

They concluded that indeed authorship attribution might be harder for identical twins, but future work was needed. They also stated that the corpus used was too small and pre-existing texts were preferred to exclude the topical influence.

2.2 GENETICS

The hypothesis that DNA influences writing style is also somewhat supported when put into a biological background perspective. Experts in this field often refer to it as the nature-nurture debate, nature being genetics and nurture referring to environmental factors and its influence on individuals from a psychological perspective Paul (1998). As a non-expert in this research area, some previous work on the nature-nurture debate will briefly be discussed.

It is observed that human psychological traits are largely effected by genetic factors and specifically variation in individuals' DNA (Plomin, 2019). In his book, Plomin claims that people's personalities are more dependent on human genes rather than their environment and claims that the influence of the environment on psychological differences are mostly random, unsystematic and unstable, which means that we cannot do much about them. If people's personalities are more dependent on genetics rather than environment, then so could be an author's writing style.

Using examples of identical twins who grew up in different environments turning out to have similar lives, scientists Clark and Grunstein (2004) show that psychological similarities are influenced by genetics. This is again insinuating confirmation regarding the hypothesis that DNA influences writing style. The same example is used by linguist and psychologist Pinker (2005), joining the nature-nurture debate by publishing his book where he states that that human behaviour is, for a large part, formed by evolutionary psychological adaptations. He refers to identical twins growing up in different environments but sharing very specific qualities.

The observations mentioned above by scientists in the biological field claim that human behaviour is somehow and somewhat dependent on our DNA. Much is still unknown, but support is being found that there is a significant influence by genetics at play. This in its turn supports the hypothesis that certain linguistic aspects could be affected by genetic influences.

This paper differs from previous work in the way that it researches the linguistic influence of genetics on an authorship discrimination task. This means combining two fields where one could be dependent of the other, namely writing style being influenced by DNA. Combining these two fields in a research could lead to new insights in computational linguistics with a focus on authorship attribution, as well as in the biological field with a focus on the nature-nurture debate. To the best of my knowledge, this specific combination of different fields has not been properly researched yet.

3 DATA AND MATERIAL

For this research, I made use of four different data sets which will be discussed in this chapter. First I will explain the two created data sets called Twin-20 and RedSet-20. The Twin-20 data set was created for testing whether DNA influences writing style. As identical twins are genetically the same, this data set could function as a test set for the hypothesis. The RedSet-20 data set was created for training the model as the Twin-20 data set had only three true positive instances, i.e. texts written by the same person, which makes it very difficult to train the model, considering the highly skewed data distribution.

Then the two existing data sets used are discussed, which are PAN-15 Verification and CLIN29. The PAN-15 data set was used because the model that I re-adapted was built on this data set. It also has the correct labels for authorship discrimination, i.e. 'Y' and 'N', equally distributed. This could therefore also be used to train and test the model on an authorship discrimination task. The CLIN29 data set differs from the other three in the fact that it is an authorship profiling data set, containing information about the gender of authors. This was used to obtain information on the influence of gender on writing style.

3.1 CREATED DATA SETS

These data sets were obtained in collaboration with Bruijn (2020). We created two data sets, one obtained through a questionnaire and one by scraping certain web pages. The following sections will explain the process of collecting, annotating and processing these data sets.

3.1.1 Twin-20[1]

The Twin-20 data set was created as a test set to test the hypothesis. It consist of texts written by 71 individuals who either are part of a twin, or a sibling of a twin. Most instances consist of one text only, however 3 authors provided us with a second text. An example text can be found below.

```
"NIEUWE TEKST" Aangezien we als familie graag 'wist-je-datjes' vertellen
over onze broers en zussen, bedacht ik dat ik alvast wat mocht gaan
graven in mijn herinneringen en gedachten.  Het is nog niet zover, maar
we hopen dat Gerrianne ooit nog een heel leuke man tegenkomt met wie
ze wil gaan trouwen.  Daarom alvast een tweetal van deze overheerlijke
weetjes, waarvan we vinden dat iedereen ze mag horen...  Wist je dat
Gerrianne op haar 12e een bril kreeg, maar toen al zulke slechte ogen
had dat de opticien verbaasd was dat ze nog mee kon komen op school.  ...
Gerrianne meer mayonaise eet datn een gemiddelde Nederlander.  Ga maar
uit van een pot van 750mL per twee maanden ..  en nee, die eet ze dan
echt helemaal alleen op.
```

Example text of Twin-20 data set.

[1] https://github.com/kaibruijn/Master_Thesis

Collection

The texts were obtained via a survey, posted on Survio[2]. We sent the link to as many people as we could, with the help of the Dutch Twin Register[3]. The survey asks twins to add in some information about themselves, as well as a written text, preferably pre-existing. They are also asked to upload texts written by their other siblings. The details of the questions can be found in our survey[4]. With 50 uploading participants, we obtained texts of the previously mentioned 71 individuals. The format of the first 5 uploads is shown in Figure 1. The range of characters varies between 45 and 1,654 with an average of 634 characters.

	A	B	C	D	E	F	G	H
1	ID	Survio_ID	Twin_or_S	Type_of_T	Gender	Age	Text	Text2
2	1	3	T	I	F	75	Wij (Betty en ik) zijn	
3	2	3	T	I	F	75	Betty zei over onze ba	
4	3	3	S	-	F	72	Ik ben jaloers, zijnhen	
5	4	4	T	I	F	62	Ben er een van een tv	
6	5	4	T	I	F	62	Fijn dat we een goede	

Figure 1: Twin-20 in Excel.

Annotation

The information about two texts was stored in an Excel database as shown in Figure 1, where information about a combination of two texts can be found. This information was translated as an annotation of either Twin, Sibling, Random or Fraternal. The meaning of these annotations and the data distribution can be found in Table 1.

Table 1: Twin-20: Label Meaning and Distribution.

Label	Definition	#
T	Text of an identical twin + text of the other half of the identical twin.	21
S	Text of any twin + text of the first person's non-twin sibling.	34
R	Text of any twin + text of any non-related person.	71
F	Text of a fraternal twin + text of the other half of the fraternal twin.	4

One instance consists of a set of two texts to be compared. As we have 71 individuals, there are 71 random comparisons made.

Processing

The Twin-20 texts needed only minimal processing. As mentioned before, we stored the texts in an Excel database, where we removed "NIEUWE TEKST", because we asked the participant to add this to the beginning of a text if they were planning to write a new text. This information was stored in the original Excel database. Two texts were written in English but we did not remove them from the data. This data set was only used for testing, so they would not influence the training process or results for other texts. These two texts make up a small experiment when the language for training and testing is different.

[2] https://www.survio.com/nl/
[3] https://tweelingenregister.vu.nl/
[4] https://www.survio.com/survey/d/G7A6Q4CoY9I9U3R4X

3.1.2 RedSet-20[5]

RedSet-20 was created as a training set for predictions of the model. As mentioned before, there are only three true positive cases, i.e. texts written by the same person, in the Twin-20 data set, which is why we created a training set. It consist of texts written on Reddit by 149 individuals. This data set was created based on the genre of the Twin-20 data set, because the model needs true positives in the training set, suggesting an in-domain setting at first. We wanted to have the least influence by topic and therefore decided to use a similar genre for training. An example text can be found below.

```
Ik ben zelf best wel tegen vleesconsumptie (laat staan -promotie),
maar ik zie de tegenspraak niet zo.  Terugbrengen van vleesconsumptie
is slechts één manier om aan het klimaat te werken.  Kennelijk wil
Timmermans het - om welke reden dan ook - op andere manieren doen.  Het
zou trouwens wel fijn zijn als hij een beetje concreet was met wat hij
bedoelt met 'verduurzamen', want het lijkt alsof hij bewust vaag aan het
doen is.
```

Example text of RedSet-20 data set.

Collection

To collect a data set from Reddit we used the PRAW API[6]. We collected Dutch posts on Reddit where the same author had posted another reaction in the same thread, where we used a text length filter to only obtain texts with a range of 200 to 1500 characters, based on the characteristics of the Twin-20 data mentioned in the previous section.

Annotation

The information about two texts was stored in an Excel database, where information about the relation between two texts can be found. This information was translated as an annotation of either Y or N. The meaning of these annotations and the data distribution can be found in Table 2.

Table 2: RedSet-20: Label Meaning and Distribution.

Label	Definition	#
Y	Two texts written by the same author.	149
N	Text written by some author + text written by some other author.	148

Processing

The Reddit texts needed only minimal processing. As mentioned before, we stored the texts in an Excel database, where we replaced links with "LINK". If the full text was only a link, which was the case for two texts, to be able to fit to our model we replaced the link with "LINK link". Also, we replaced a user mention with "MENTION". After this, we filtered out the English texts as they were not applicable for training.

3.2 PRE-EXISTING DATA SETS

Besides the created data sets mentioned above, I also used two pre-existing data sets for this research. One is the data set that the model I used was trained for, the

[5] https://github.com/kaibruijn/Master_Thesis
[6] https://praw.readthedocs.io/en/latest/

other has information about the gender of the authors, used for experiments. These data sets will be explained in more detail in the next subsections.

3.2.1 PAN-15 Verification[7]

The PAN-15 Verification data set is an existing data set consisting of 100 training and 165 test instances with combinations of 2 or more Dutch texts that are either written by the same author or not, created by Stamatatos et al. (2015). It is a modified version of the CLiPS Stylometry Investigation corpus by Verhoeven and Daelemans (2014), which comprises documents from two genres (essays and reviews), written by language students at the University of Antwerp between 2012 and 2014. The outline of the data set is comparable to the RedSet-20 data set as we tried to create a similar data set but with a more fitting genre when compared to our test data. The model that I re-adapted was built on this data set. My experiments in trying to outperform the model of Hürlimann et al. (2015) were first run on this data as it contains GLAD's original training and test set. The label distribution of this data set can be found in table 3.

Table 3: PAN-15 Verfication: Label Distribution.

Data Set	Y	N	Total
PAN-15 Verfication Train	50	50	100
PAN-15 Verfication Test	83	82	165
PAN-15 Verfication	133	132	265

3.2.2 CLIN29[8]

This data set differs from the others in the way that it is an author profiling data set and not an authorship discrimination data set, i.e. there is no data written by the same author. The CLIN29 data set is an existing data set consisting of Twitter, Youtube and news texts where the gender of the author is known (Haagsma et al., 2019). Because of the genre of the Twin-20 data and mainly the text length, I did most experiments with the news genre, containing news articles with 2,138 training and 1,000 test instances as it best resembled the Twin-20 data. The data was used for a shared task called GxG at the University of Groningen where the goal was to predict the right gender to a specific text. I used this data set mostly for experiments with clustering as it fits better to the needs of clustering, with clear gold labels 'F' and 'M', female and male respectively. This data set has less focus on authorship discrimination than the other data sets discussed. I also used the performance on this data set to compare the influence of gender on an authorship attribution task. The label distribution of this data set can be found in table 4.

Table 4: CLIN29: Label Distribution.

Data Set	F	M	Total
CLIN29 Train	916	916	1,823
CLIN29 Test	500	500	1,000
CLIN29	1,416	1,416	2,832

[7] https://https://pan.webis.de/clef15/pan15-web/
[8] https://www.let.rug.nl/clin29/shared_task.php

4 METHOD

In this chapter, the method used for the research will be described. First, the theoretical approach is discussed as authorship discrimination differs from most natural language processing task in the way that it focuses on writing style. This is also previously mentioned in Chapter 2. Then, the framework of comparison is explained, going into detail about how the results for different labels should be interpreted, given the fact that there are so few texts written by the same person in the Twin-20 data set. The model selection is described afterwards, which explains the models that were used for experiments and trying to reach best performance, followed by the method to analyse stylistic features.

4.1 THEORETICAL APPROACH

Whereas usual natural language processing tasks have subjective labels, e.g. sentiment analysis, authorship discrimination has clear binary labels as the texts are either written by the same person or not. For this reason, a human evaluation can provide clear insights on how well humans perform on authorship discrimination.

It is hard to explain in a theoretical context what makes two people write more similar. In fact, together with Bruijn (2020) we show that humans disagree on writing style similarity and do not at all perform well on the authorship discrimination task provided. Although no previous research is published that focuses on human performance on authorship attribution, Luyckx (2011) states that people will know the author of some of the texts in their model, but are unaware of the authors of other texts and are also likely to combine models of genre with models of authorship.

In our human performance experiment, we asked 12 random people, 8 female and 4 male, in the age range of 20 to 57 to rate a combination of two texts on how sure they were that the two documents were written by the same author. This was concluded with a score between 1 and 5, where 1 indicates absolutely not and 5 indicates absolutely certain. We used combinations written by identical twins, siblings, random people and fraternal twins as well as actual documents written by the same author, i.e. true positives. Results for this small-scale research can be found in Chapter 5.

Because of the aforementioned issues, in this research there are two main assumptions made. The first is that if we train our model on a genre-similar data set, this is a proper way for learning true positive and true negative cases. Because we were unable to obtain enough texts written by the same individual, this was the best possible way to obtain true positives. It was noticed that when training the model on a different genre, i.e. PAN-15, the model did not show any expected results which explains the choice of a similar genre.

The second assumption is that if the model, re-adapted from Hürlimann et al. (2015) or any model that reaches better performance on the training data, is confident enough that two documents have the same author, the documents are written with a similar writing style. This way, we can compare the probability scores for different groups, respectively identical twins, siblings, random combinations and fraternal twins and draw conclusions based on the results. An elaboration on both assumptions can be found in the following two sections.

4.2 FRAMEWORK OF COMPARISON

After obtaining the best performing model, which will be described in the next section, it can be tested on the Twin-20 data set. The model is trained on the RedSet-20 data again, but this time on all instances, thus including those previously used for testing. This is based on the assumption that more training data leads to a better performance.

If this model is then tested on the Twin-20 data, it will produce probability scores for every comparison of two documents. These scores can then be compared for every group in the data set and from here the conclusion can be drawn whether people who share DNA have a higher probability of writing similarly or not.

For example, if the average probability outputted by the model for two texts labeled 'T' is higher, say 0.70 than for two texts labeled 'R', say 0.40, this could confirm the hypothesis. The average probability score can give more insights in this matter, however there is an issue that needs to be attended to.

Because the objective of this research is to find if genetics influence writing style, all possible factors need to be excluded as best as possible. Two important factors are the gender and age of authors. Chapter 5 will provide an overview of the impact of these characteristics.

Also, this research wants to find as much information as possible with regard to the influence of genetics on certain linguistic features. If people with genetic resemblance write more similarly, we want to find out which aspects of writing style are most affected by our DNA. The method for obtaining information with regard to this matter is explained in the last section of this chapter.

4.3 MODEL SELECTION

Being able to re-adapt an existing model that positioned itself amongst the top participants in the PAN-15 shared task is a privileged position (Hürlimann et al., 2015). However, as the model is already somewhat outdated, much research has gone into trying to outperform this model. Recently, other natural language processing tasks are taken over by Transformer models and other neural networks, rather than support vector machines (Devlin et al., 2018). To be able to obtain the model with best performance, I experimented which such models, as well as linear regression and clustering, and compared the accuracy scores when trained on half of the training data and tested on the other half. These models will be described in section 4.3.3.

I mentioned before that accuracy scores are not useful when tested on the Twin-20 data set, but the RedSet-20 data set has as many true positives as true negatives, which makes the accuracy score a useful evaluation tool for ranking the best model. In the next subsections, the training and testing of several models will be explained. Note that testing is still not on the Twin-20 data set.

4.3.1 Training

As mentioned before, the model is trained on the created RedSet-20 data set, where there are 149 true cases and 148 false cases. When trained on half of this data and tested on the other half with an equal distribution, accuracy scores can tell which model performs best. The models were also trained and tested on the PAN-15 data, mentioned in Chapter 3, to ensure that models are not too genre-specific, i.e. in-domain oriented.

4.3.2 Testing

The models, trained as mentioned above, are then evaluated on the test data, which is half of the RedSet-20 data set. This allows to obtain accuracy scores to conclude on the performance of the model. Again, they are also tested on the PAN-15 data set, and vice versa. The results for several models can be found in Chapter 5.

4.3.3 Models

After experimenting with Transformer models, clustering, linear regression and support vector machines, the final model remains the slightly re-adapted version of GLAD (Hürlimann et al., 2015). In this subsection, the models will briefly be discussed with an extra elaboration on the final model.

BERTje (de Vries et al., 2019)

The BERTje model reads the train and test set that are given as input. First, the data is transformed into the correct format for the simpletransformers[1] package that I use. For every line in the train and test set, the two texts to be compared are concatenated and labeled accordingly, i.e. 'Y' or 'N', converted to 0 and 1 as needed for the classifier. I observed that the way of concatenating does not influence performance and stuck with three new lines as concatenating characters, because there can be regular new lines in the texts.

The BERTje model, in this package called 'bert-base-dutch-cased' can then be trained and tested and will provide a classification report and confusion matrix. Changing parameters such as learning rate, epochs and batch sizes did not lead to higher performances and did not even outperform the baseline of .50.

Clustering

There are multiple implementations of clustering used for this research. I used scikit-learn's K-Means[2] package as it allows to control the number of classes, which should be 2 for this research, and is most commonly used. All implementations read the train and test set that are given as input and convert the words into a vector and concatenate the two texts as mentioned before.

I experimented with tf-idf and Word2Vec[3], both commonly used as a vectorizer. I also implemented the spaCy[4] module which allows to get the mean vector for an entire sentence. All of the vectorizers are available for Dutch, which is the language used in the data sets. Changing parameters such as maximum iterations and relative tolerance only slightly enhances performance. However, the results vary for every run and often do not outperform the baseline.

A special feature that I implemented in the clustering algorithms is called bleaching (van der Goot et al., 2018). This transforms words into tokens, which puts the focus more on style rather than word meaning. For example punctuation, vowels and the length of words are transformed into characters which provide information. This implementation was provided by Bruijn (2020), altered from van der Goot et al. (2018). This also did not reach high performance.

Linear Regression

This basic machine learning algorithm was implemented with the notion that if twins are given value -1 and random people are given value 1, perhaps siblings will be predicted as 0 as they are somewhat in the middle. The model reads the train

[1] https://github.com/ThilinaRajapakse/simpletransformers/
[2] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
[3] https://radimrehurek.com/gensim/models/word2vec.html
[4] https://spacy.io/models/nl

and test set as input and applies a tf-idf vectorizer after concatenating the two texts, after which it is already set for predicting. I used scikit-learn's LinearRegression[5] for this implementation. Changing the few parameters available did not enhance performance.

Support Vector Machines

Another basic machine learning algorithm that I implemented is scikit-learn's LinearSVC[6]. I also applied the regular SVC and the LinearSVR, but they did not enhance performance. The idea behind these implementations is that the GLAD model also makes use of a support vector machine. A basic implementation would also provide insight into the value of additional features in the model of Hürlimann et al. (2015). The model reads the input texts and concatenates them into one text ready for predicting. It makes use of a tf-idf vectorizer.

GLAD (Hürlimann et al., 2015)

The GLAD model reads the documents from directories in a format as necessary in the PAN-15 shared task mentioned before. This means that there is a different train and test directory, which both contain directories, starting with 'DU001' and counting up for every problem instance. In here, there is exactly one text file called 'unknown.txt' and one or more files 'known01.txt' and counting forth for every other text file. The directory structure is visualized in Figure 2.

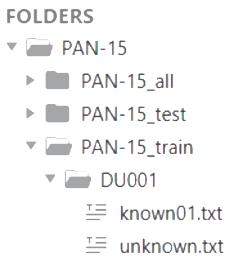

Figure 2: GLAD Directory Structure.

After reading and tokenizing the files and adding features like character n-grams, punctuation similarity, sentence length and a compression measure, a binary support vector machine compares the documents on the joint features and produces a probability score between 0 and 1, where >0.5 and <0.5 mean similar and different writing styles respectively. It then writes the predictions, a probability between 0 and 1, to a file called 'answers.txt'. The model uses an svm with radial basis function kernel and default parameters. A more detailed description of the model and features used can be found in the paper of Hürlimann et al. (2015).

I also implemented linear regression and a linear support vector machine in the model, but this did not enhance performance. Also adding a maximum character limit, changing parameters and using only certain groups of features, which are called combos, had no positive effect on the results. The final model therefore remains the model of Hürlimann et al. (2015).

4.4 FEATURE ANALYSIS

Using the model of Hürlimann et al. (2015) allows for an easily implemented feature analysis because of the code format, as they did in their research as well. They add several features at a time, allowing to easily leave features out. For this research, I

[5] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
[6] https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.htmlsklearn.svm.LinearSVC

chose to leave one feature out at a time instead of using one feature at a time, based on the results shown in Chapter 5. Because the scores are already quite close to 0.5, I suggest that leaving out one feature at a time is more informative than using only one feature. The results for this analysis can also be found in the next chapter.

5 RESULTS AND DISCUSSION

This chapter shows the results of the method used as described in the previous chapter. First, the results of the human evaluation are shown, where the observation is that humans disagree on writing style similarity and perform bad on an authorship discrimination task, suggesting that the task of recognizing if two texts are written by the same person is not quite feasible for people. This is followed by the results and accuracy scores of several models, leading to the selection of GLAD for authorship discrimination (Hürlimann et al., 2015). Several scores after using GLAD are then shown to indicate its performance on PAN-15 and RedSet-20, as well as cross-domain, which is followed by the results of the main research based on the Twin-20 data set and a feature analysis. Lastly, the results are discussed in the final section.

5.1 HUMAN PERFORMANCE

As can be seen in Table 5, humans disagree on writing style similarity very much as the minimum and maximum ranges are very large, indicating complete disagreement for individual instances. In the table, some examples are shown.

Also, the average for the instances is often in the very middle, even for documents written by the same author. In Table 6, the average scores are shown for each label. A full disclosure including the probability assigned by GLAD and individual answers can be found on the GitHub page[1] of this paper. For comparison, the assigned probability of GLAD is shown in the table as well.

Table 5: Human Evaluation: Examples of Individual Instances

Instance #	Label	Minimum	Maximum	Average
1	T	1	4	2.4
2	S	1	4	1.9
7	R	1	4	2.8
11	F	1	5	2.8
21	TP*	1	5	2.8

*TP (true positives) are two documents written by the same author.
For every label, i.e. T (Twin), S (Sibling), R (Random) and F (Fraternal), the first instance to occur was chosen. A score of 1 indicates that the participants think that the texts are not written by the same person, 5 indicates that they are absolutely certain that the two texts are written by the same person.

5.2 MODEL SELECTION

For the authorship discrimination task, as stated before, I have experimented with several models. Five of these models, which were most promising or important to experiment with, are discussed in this section. Table 7 shows the performance of

[1] https://github.com/kaibruijn/Master_Thesis

Table 6: Human Evaluation: Average Scores with GLAD

Label	Minimum	Maximum	Average	GLAD*
T	1	5	2.6	2.4
S	1	5	1.9	3.2
R	1	5	2.7	3.0
F	1	5	2.8	3.6
TP	1	5	2.8	2.7

*Scores of GLAD are mapped from a value between 0 and 1 to a [1,5] scale, then averaged.
In 3 out of 23 cases, the difference in age is more than 10 years, the gender is always the same.

the models based on accuracy while Table 8 shows the size of the training and test sets used.

Table 7: Performance of Models.

Model	PAN-15	RedSet-20	Average
GLAD	**0.73**	**0.68**	**0.71**
Linear Regression	0.60	0.50	0.55
BERTje	0.47	0.47	0.47
Support Vector Machine	0.66	0.50	0.58
Clustering	~0.50	~0.50	~0.50

Clustering was experimented with using tf-idf, Word2Vec, Sentence2Vec and bleaching (van der Goot et al., 2018), where the results have a range of ± 0.1 for every run. The best model is shown above. More detail on this and other models can be found in Section 4.3.3.

Table 8: Size of Train and Test Sets.

Data Set	Training Instances	Test Instances
PAN-15	100	165
RedSet-20	136	117

In this table, it can also be seen that a Transformer model does not work for the described authorship discrimination task. As mentioned in Chapter 1, authorship attribution is difficult sub field of natural language processing for neural networks. The results show that with state-of-the-art models, this is still the case, which may be explained by observing that BERTje, overly simplified, learns a language rather than stylistic patterns.

The main observation of Table 7 remains that GLAD is the best model to work with as it outperforms all other models in both settings.

To better experiment with authorship attribution, especially combined with clustering, I also used the CLIN-27 data set for experiments. Although the experiments are not fully related the main research and clustering did not even outperform the baseline of .50, it is important to obtain the performance of recognizing the gender of an author, as this explains the influence of gender on the concerning genetics.

Hence, it is important to note that the best of many models is a support vector machine reaching 68 percent accuracy on a balanced data set, showing that gender influences writing style to a certain degree. This simple model makes use of scikit-learn's LinearSVC[2] with a tf-idf vectorizer and a regularization parameter of 1. This model positions itself second place with regard to the GxG shared task, as can be

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

seen in the paper of Haagsma et al. (2019) where the best participant reached an accuracy score of .69.

5.3 GLAD RESULTS

As Table 7 shows, GLAD remain the best model by far. It also shows that the RedSet-20 data, which is used for training before testing on the Twin-20 data, is much more difficult than the PAN-15 data. Because only GLAD outperforms the baseline, and quite significantly in both tasks, this suggests that Hürlimann et al. (2015) have built a model that is robust, or at least more robust than the other models mentioned before. However, this is still in an in-domain setting, which is easier for a machine learning based model. To test the robustness of GLAD, the model was also run in a cross-domain setting using PAN-15 and Redset-20 for training and testing respectively, and vice versa. The results can be found in Table 9, where the model seems to be performing quite well in a cross-domain setting.

Table 9: GLAD Accuracy and Robustness.

Train Set	Test Set	Accuracy
PAN-15	PAN-15	0.73
RedSet-20	RedSet-20	0.68
PAN-15*	RedSet-20*	0.66
RedSet-20*	PAN-15*	0.69

*Full data set.

5.4 RESULTS TWIN-20

Table 9 shows that GLAD seems to be quite robust, which is necessary for the framework of comparison used in the Twin-20 data set. As there are only 3 true positive cases, accuracy scores can be misleading in this data set given the majority class baseline of .98 for a binary classification, so we need to know that the model performs well enough. To be able to compare all different groups, i.e. labels T, S, R and F, the probability scores of the model are compared. Table 10 shows the results of this comparison, where no correction for gender and age have taken place.

Table 10: Twin-20 Probability Scores.

Train Set	T	S	R	F
RedSet-20	0.63	0.59	0.49	0.86*

*As there are only 4 instances of fraternal twins, this score can be considered as an outlier.

The scores show that indeed there is a deviation for the different labels T, S, R and F. However, although this seems promising, the influence of age and gender has to be taken into account. Figure 3 shows the average scores for the labels after corrections made for the characteristics mentioned.

In this figure, it can be seen that having a similar age, i.e. a difference in age of maximally 10 years, has a big influence on the probability scores assigned by the model. Also, gender influences the results. After both corrections, the average of all labels is the same with a probability of .63, except for fraternal twins. However, it needs to be noted that there are only 4 cases of this class, indicating that this score is probably exaggerated.

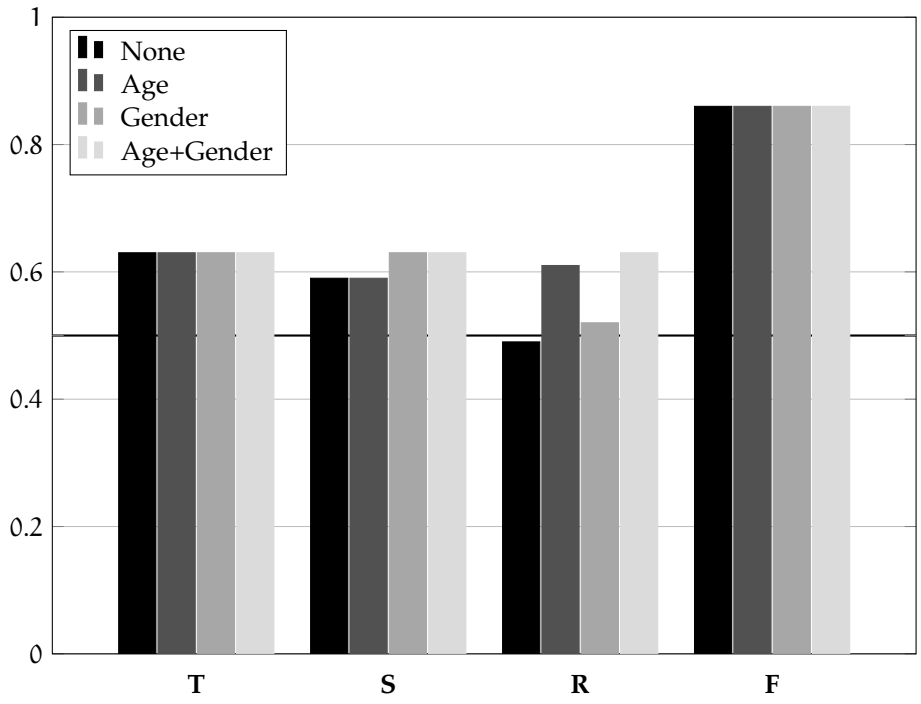

Figure 3: Twin-20 Probability Scores with Corrections.

5.4.1 Feature Analysis

Figure 4 shows the influence of leaving out certain features built by Hürlimann et al. (2015). However, the results in Figure 3 already indicate not much difference after corrections for age and gender, which is confirmed by the feature analysis. There is no label specifically influenced by leaving a feature out. This indicates that there probably is no difference in linguistic features between twins, siblings or random people, e.g. it is not the case that twins have a more resembling punctuation use than random people. It could also be the case that there are enough other features to recover for the feature that is left out.

The analysis does show that leaving 5-gram similarity out leads to the biggest decline in probability scores. This makes sense as the model uses words as input for the classification, and if 5 words following each other are the same, then it looks very similar. However, the impact is still only marginal, again indicating that other features recover for the feature that is left out.

An overview of all features can be found on the GitHub page of this paper[3].

5.4.2 Discussion

Although this is not confirming the hypothesis that writing style is influenced by genetics, it makes sense that gender and especially age influence writing style. Many will recognize that women write differently than men and even more that elderly have a different style for writing than adolescents.

This observation is also supported by Schler et al. (2006) who analyzed a corpus of tens of thousands of blogs, incorporating close to 300 milllion words. They observed significant differences in writing style and content between male and female bloggers as well as among authors of different ages. They even managed to exploit these differences to determine an unknown author's age and gender on the basis of the blog.

[3] https://github.com/kaibruijn/Master_Thesis

The results show some promising insights into this matter, however as this was not the focus of this paper, it is only mildly researched. Hence, these insights should be concluded carefully, but are supported by previous work.

It is also important to make a note on the robustness of the model built by Hürlimann et al. (2015). Although Table 9 shows that GLAD performs well on both data sets with different genres, the results of testing GLAD on the Twin-20 data when trained on the PAN-15 data are very different from the results when trained on the created RedSet-20 data. Table 11 shows that all scores drop to .5, including the high-scoring label 'F' when GLAD is trained on PAN-15. This of course questions the robustness of GLAD and therefore the results of the research.

Table 11: Twin-20 Probability Scores.

Train Set	T	S	R	F
PAN-15	0.46	0.48	0.46	0.52

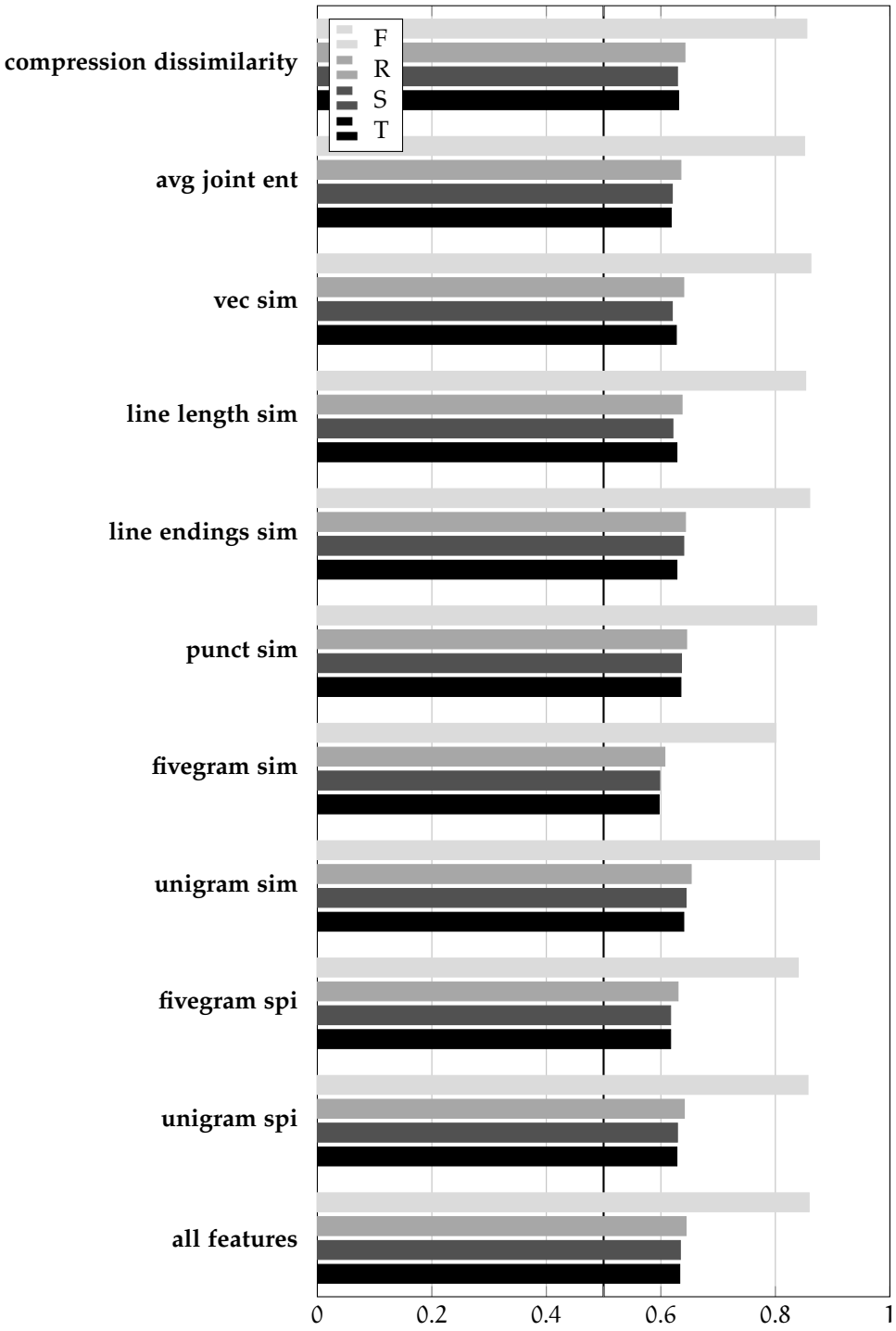

Figure 4: Feature Influence.

6 CONCLUSION

The results of Chapter 5 show to what extent an author's writing style is affected by genetic influences. Building on the observations made in this chapter, several conclusions are drawn with regard to the influence of DNA on writing style. As can be seen in Table 3, the gender and range of age of two different authors has an impact on the probability of similar writing styles. Especially when the difference in age is less than 10 years, the probability as predicted by the model goes up significantly, exceeding the .5 threshold of uncertainty of the model. Gender has a more limited influence on scores, but still has an influence, which is confirmed by the GxG shared task results outperforming the baseline.

However, although the two previously mentioned characteristics influence an author's writing style, the answer to the research question *"To what extent does DNA influence an author's writing style?"* mentioned in Chapter 1 remains that DNA has a very limited impact. After correcting for age and gender, the probability scores for all labels, except the underrepresented fraternal twins, are equal. This means that sharing fifty percent of DNA, or even one hundred percent, does not increase this probability score. The main conclusion therefore is that DNA does not influence writing style, except that it determines a person's gender.

Because of this conclusion, it comes as no surprise that there are no linguistic features specifically influenced by DNA. This was done after correcting for age and gender. Figure 4 supports this statement and shows no deviations indicating differences between groups.

Another important observation made during this research is that humans disagree on writing style similarity and perform bad on authorship discrimination. Tables 5 and 6 shows that there is much disagreement and that true positives are not recognized. This makes it even harder for computational linguists to build a model that is robust and reaches high performance.

Yet another conclusion, supporting the paper of Bozkurt et al. (2007), is that the state-of-the-art natural language processing model BERTje does not perform well on an authorship attribution task at all, not even outperforming the baseline of 0.50. Also, clustering algorithms fail to meet the expectations after many experiments with tf-idf and Word2Vec. Even after applying bleaching and therefore looking at linguistic patterns, it fails to significantly outperform the baseline.

An important notion with regard to the credibility of this research is the conclusion on the robustness of the GLAD model that was used. Table 9 shows that GLAD seems quite robust after its accuracy scores for PAN-15 in-domain, RedSet-20 in-domain as well as cross-domain, whereas linear regression and support vector machines only perform well on a specific type of genre. However, Table 11 questions the robustness of GLAD showing different results for the Twin-20 data after training on a different data set. Having a robust model for predicting probability scores is imperative, because there are very limited true positive cases, which is why future work could focus on testing GLAD on different data sets or building a model that performs well on different genres.

Building on the questioning of GLAD's robustness, it has to noted that the model only classified one of the three true positives correctly. As stated before, there is too little data to report accuracy scores, but if this would happen it would be .33. Because of this, future work might also focus on expanding the data set, or changing the setting of data collection, with true positives. Unfortunately, this is quite a difficult task as there is a need for many twins willing to write or upload multiple texts to be able to create such a data set. Lastly, they could also include

adopted siblings to research the influence of growing up in the same environment, having totally different genetic characteristics.

BIBLIOGRAPHY

Bourne, E. G. (1897). The authorship of the federalist. *The American Historical Review 2*(3), 443–460.

Bozkurt, I. N., O. Baglioglu, and E. Uyar (2007). Authorship attribution. In *2007 22nd international symposium on computer and information sciences*, pp. 1–5. IEEE.

Bruijn, K. (2019a). The influence of dna on writing style.

Bruijn, R. (2019b). Authorship identification with shared dna.

Bruijn, R. (2020). Authorship verification with shared dna.

Clark, W. R. and M. Grunstein (2004). *Are we hardwired?: The role of genes in human behavior*. Oxford University Press on Demand.

de Vries, W., A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim (2019). Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Haagsma, H., T. Kreutz, M. Medvedeva, W. Daelemans, and M. Nissim (2019). Overview of the cross-genre gender prediction shared task on dutch at clin29.

Hürlimann, M., B. Weck, E. van den Berg, S. Suster, and M. Nissim (2015). Glad: Groningen lightweight authorship detection. In *CLEF (Working Notes)*.

Kestemont, M., E. Stamatatos, E. Manjavacas, W. Daelemans, M. Potthast, and B. Stein (2019). Overview of the cross-domain authorship attribution task at {PAN} 2019. In *Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019*, pp. 1–15.

Luyckx, K. (2011). *Scalability issues in authorship attribution*. ASP/VUBPRESS/UPA.

Paul, D. B. (1998). *The politics of heredity: Essays on eugenics, biomedicine, and the nature-nurture debate*. SUNY press.

Pinker, S. (2005). *The blank slate*. Southern Utah University.

Plomin, R. (2019). *Blueprint: How DNA makes us who we are*. Mit Press.

Rangel, F., P. Rosso, M. Koppel, E. Stamatatos, and G. Inches (2013). Overview of the author profiling task at pan 2013. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, pp. 352–365. CELCT.

Schler, J., M. Koppel, S. Argamon, and J. W. Pennebaker (2006). Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, Volume 6, pp. 199–205.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology 60*(3), 538–556.

Stamatatos, E., W. D. amd Ben Verhoeven, P. Juola, A. López-López, M. Potthast, and B. Stein (2015, September). Overview of the Author Identification Task at PAN 2015. In L. Cappellato, N. Ferro, G. Jones, and E. San Juan (Eds.), *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. CEUR-WS.org.

Swain, S., G. Mishra, and C. Sindhu (2017). Recent approaches on authorship attribution techniques—an overview. In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, Volume 1, pp. 557–566. IEEE.

van der Goot, R., N. Ljubešić, I. Matroos, M. Nissim, and B. Plank (2018). Bleaching text: Abstract features for cross-lingual gender prediction. *arXiv preprint arXiv:1805.03122*.

Verhoeven, B. and W. Daelemans (2014). Clips stylometry investigation (csi) corpus: a dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *LREC 2014-NINTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION*, pp. 3081–3085.