university of groningen / faculty of arts

Authorship verification with shared DNA

the influence of genetic and situational factors on writing style

R.K. Bruijn

Master thesis
Information Science
Rune Bruijn
s3204758
April 22, 2020

ABSTRACT

Authorship verification is the task of verifying whether two texts were written by the same person or not and is a hot topic amongst data scientists. The author of a given text can be identified based on writing style. But what about the influence of DNA? Does DNA have an influence on writing style? Do other situational factors like gender and age have an influence on writing style? In this research, I dove into these problems.

This research on the influence of genetic and situational factors on writing style is done by using authorship verification models that predict from two texts whether they were written by the same person or not. The main authorship verification model in this research is the Groningen Lightweight Authorship Detection model (Hürlimann et al., 2015). This model reads two texts and outputs a score between 0 and 1, indicating how likely the model thinks the two texts were written by the same person.

Other models, like a LinearSVR model and a bleached version of it, are also tried in this research for the authorship verification task. Although the bleached version performed slightly better than the regular LinearSVR model, the GLAD model outperformed both of them significantly.

In this research, the texts for the authorship verification task are from several data sets. The first data set is the PAN (2015) data set, where the GLAD model was built on. Secondly, a new data set is generated in this research with texts from Reddit. Finally, a data set with texts of twins and siblings was made, in order to research the influence of DNA on writing style. All models and data that is used in this research is available on GitHub .

During this research, the GLAD model has proven to work for both in-domain and cross-domain classification tasks for the PAN data, Reddit data and twin data. The model reached in-domain classification scores of up to 0.76 accuracy and cross-domain accuracy scores of up to 0.69, significantly outperforming the baselines of 0.50. Running the GLAD model on the twin data indicated that people with shared DNA do write more similar to each other than people who do not have shared DNA.

Deeper feature analysis showed that punctuation similarity and vector similarity are textual features which influence the writing style of twins and siblings. However, the situational factors gender and age neglect these textual features.

Finally, a human evaluation was done during this research, where contestants had to give a score between 1 and 5 to pairs of texts, indicating whether they thought two texts were written by the same person or not. The results clearly indicated that humans cannot perform an authorship verification task between two texts.

CONTENTS

PREFACE

At this moment, you are reading the Master thesis "Authorship verification with shared DNA, the influence of genetic and situational factors on writing style", written by me. My name is Rune Bruijn and I am currently a Master Information Science student at the University of Groningen, where I finish this study by writing this thesis.

For the past six months, I have worked on this Master thesis about the influence of genetic and situational factors on writing style by looking at authorship verification in combination with shared DNA. This is a subject of interest to me, as I am an identical twin myself. My twin brother Kai and I think that we write very similar to each other and thus wanted to do research about the influence of DNA on writing style. This is why Kai, who also studies Information Science, and I both decided to our Master thesis about this subject.

During our Bachelor Information Science, we both wrote a Bachelor thesis about this same topic. The main conclusion from these Bachelor theses was that twins do write somewhat more similar to each other than others. However, as there were some limitations to the theses due to a small amount of data, we decided to do our Master thesis about the same subject, but with more data, deeper feature analysis and some other improvements.

I would like to thank my supervisor prof. dr. M. (Malvina) Nissim a lot for her help during the past few months. She helped me a lot by initially coming up with the idea of doing research about writing style and DNA and also provided me with updates of new approaches that I could use for my research. Furthermore, in 2015, she and her colleagues made a model that takes two texts as input and then outputs a score indicating how likely it is that the two texts were written by the same person (Hürlimann et al., 2015). This model was made for the PAN (2015) task, where it performed really well and was essential during my Master research, so I would like to thank her for that as well.

Next to my supervisor, I would also like to thank the Dutch Twin Registry for posting a survey for twins and their siblings that Kai and I made on their Facebook page. They usually do not post anything on behalf of other people, but they were willing to make an exception for us as they were very excited about our ideas. Above that, I would really like to thank all the people that took the time to fill in our surveys, as I could not have done this research without them.

Finally, I would like to recommend reading the Master thesis of Kai, who did a similar research for his Master thesis. His thesis is called "The Genetical Influence on Writing Style, Authorship Discrimination and DNA" and has a few approaches that are different than the approaches I used during my research. So, if you are interested in this topic, please have a look at his thesis as well.

But first, please have a look at my thesis. I hope you enjoy reading it!

Rune Bruijn

1 INTRODUCTION

Authorship verification is the task to decide whether a given text was written by a certain author or not, as mentioned by Stamatatos (2009). More precisely, given a number of sample documents of an author A and a document allegedly written by A, the task is to decide whether the author of the latter document is truly A or not (Halvani et al., 2016). It is a hot topic amongst data scientists and can be used for example in forensics, to decide whether a threat note was written by a particular person or not.

Authorship verification is based on writing style. Factors like punctuation use can be an indication for an authorship verification model that a certain text is or is not written by a certain person. But what happens when two people with shared DNA, who might write very similar to each other write two texts? Will the model still be able to see the differences in writing style or will the model treat two texts of people who write very similar as written by the same person? During this Master thesis, I will do research on this area.

In this research, I will investigate whether DNA has an influence on writing style, by investigating whether twins write more similar to each other than siblings or random people do. Even though from personal experience I think twins write very similar to each other, it has not been proven yet, as no research has been done on the writing style of twins yet. That is why I am going to investigate this in this research. In order to do research on this topic, the following main research question will answered:

- *1) What is the influence of genetic and situational factors on writing style?*

In order to answer this main research question, several sub research questions should be addressed. The first sub research question is to investigate whether identical twins write more similar to each other than to other people: *1.1) Do identical twins with one hundred percent shared DNA have a more similar writing style than people who do not have shared DNA?* The second sub research question is based on textual features of writing style to see which features are important for shared DNA: *1.2) Which textual features of writing style are most important for shared DNA?* The last sub research question is about situational factors, which should be considered in order to make conclusions about DNA: *1.3) What is the influence of situational factors like gender and age on writing style?*

To answer these research questions, I will input pairs of texts of twins, siblings and random people in several authorship verification models. These models are trained on a data set where it is annotated whether the texts were written by the same person or not. This trained model will than be run on the pairs of texts of the same person, identical twins, nonidentical twins, siblings and random people where the model outputs a value indicating how likely it is that the texts were written by the same person. One very important and solid authorship verification model that already exists is the Groningen Lightweight Authorship Detection model (Hürlimann et al., 2015). Among some self-made models, like a LinearSVR model, this model will be used during my research.

As mentioned above, no research has been done on the subject of writing style of twins in combination with an authorship verification model. This is also what makes it even more interesting for me and thus is a great motivation why I chose this subject for my Master thesis. Another motivation for this subject is the fact that I am an identical twin myself and I think I write almost identical to my twin brother. This resulted into both of us getting zero points on an assignment that we did not

cooperate on during our Bachelor, due to the fact that the two assignments looked too much alike. This is also a big motivation for me to do research on the writing style of twins with shared DNA.

This thesis is structured as follows. In Chapter 2 I will describe previous work that is important for this research. Chapter 3 describes the data that I used for this thesis. This chapter is split up in two parts, where the first part describes the already existing data and the second part describes how I created data sets myself for this research. After that, Chapter 4 describes what methods I used during this research in order to answer the research questions. In Chapter 5, the results of this research are shown. Chapter 6 contains the conclusions of this research with the answers to the research questions, talks about the limitations of this research and presents some improvements for future work.

2 BACKGROUND

In this chapter, relevant works and definitions will be mentioned and explained. The background of this research can be divided into three sections, namely authorship attribution, writing style and genetics influence. These sections will now be elaborated on.

2.1 AUTHORSHIP ATTRIBUTION

Authorship attribution is a task in computational linguistics which focuses on trying to identify the author of a given text, based on the writing style of that particular author. Two sub fields of authorship attribution are authorship verification and authorship identification, both of which will now be elaborated on.

One sub field of authorship attribution is authorship identification. According to Zheng et al. (2006), authorship identification determines the likelihood of a piece of writing to be produced by a particular author by examining other writings by that author.

The second sub field of authorship attribution is authorship verification, which lies closely to authorship identification, but is slightly different. For authorship verification, the task is to decide whether a given text was written by a certain author or not (Stamatatos, 2009).

The slight difference between authorship identification and authorship verification can thus be defined as follows. While for authorship identification a model has to identify out of a set of authors, which author wrote a specific document, a model for authorship verification only has to verify whether a specific document was written by a specific author or not. In this research, the focus will lie on authorship verification, as models will be used that have to predict whether two texts were written by the same author or not.

For this research, some state-of-the-art models like BERTje will be used, which are improving every day. The main reason that the state-of-the-art of authorship attribution is improving so fast, is the open availability and accessibility of shared tasks. This is also mentioned by Stamatatos (2009), who states that shared tasks can push research forward and offer a good understanding of the state-of-the-art of the field. The best-known state-of-the-art shared task, which is widely known among data scientists, is the PAN shared task, with sub tasks on author identification, author verification and author obfuscation.

In 2015, Hürlimann et al. (2015) participated in one of these PAN shared tasks. This was an authorship verification task where the task was to predict for an unknown text and one up to five known texts whether the unknown text was written by the same person or not. They cited the overview of the authorship verification task at PAN (2015) as follows: "In the authorship verification task as set in the PAN competition, a system is given a collection of problem sets containing one or more known documents written by author A_k and a new, unknown document written by author A_u, and is then required to determine whether $A_k = A_u$ without access to a closed set of alternatives. In this form, the task is generally interpreted as a one-class classification problem, in the sense that the negative class is not homogeneously represented and systems are based on recognition of a given class rather than discrimination among classes. This is akin to outlier or novelty detection and different from standard authorship attribution problems, where a system must choose among a set of candidate authors in a more standard, multi-class text

categorisation fashion" (Hürlimann et al., 2015). The task was split in several languages, including Dutch. In their research, they used a Support Vector Machine with several features, including character n-grams, the lexical overlap, visual text properties and a compression measure. They obtained competitive results that outperformed the baseline and positioned their system among the top PAN shared task participants. The model made by Hürlimann et al. (2015) will be used during this research.

The evaluation of authorship verification is often done by measuring the accuracy. This is easy to calculate as the prediction whether the pairs of text are written by the same person or not is either correct or incorrect. However, it is also interesting to research how well humans perform in this task in order to compare an authorship verification model with human performance. Rexha et al. (2018) researched how well humans performed in two authorship tasks. They conducted two studies, where both studies confirmed that this authorship verification task is quite challenging for humans. This is why a human evaluation is also added in this thesis.

2.2 WRITING STYLE

The models that are used for authorship verification tasks look at the writing style in order to predict whether a text was written by a specific author or not. Hence, the writing style determines what the model predicts. However, writing style can be influenced by several factors.

One factor that influences writing style is the genre or topic of a text (Baayen et al., 2002). For example, if there are four texts, out of which two texts are fiction and two texts are non-fiction, the texts with the same genre might be more similar to each other. The same idea applies for texts with a similar topic.

Another factor that influences writing style is the domain of the text. For example, texts written on Facebook tend to be quite different than texts written on Twitter, where there is a lower limit of characters that can be used. This way, users might tend to use much more abbreviations on Twitter.

This is why their is a difference between in-domain authorship attribution and cross-domain authorship attribution. With in-domain authorship attribution the training set has the same domain as the test set, while with cross-domain authorship attribution, the domains are different. Overdorf and Greenstadt (2016) determines that state-of-the-art methods in stylometry do not perform as well in cross-domain situations as they do in in-domain tasks, to to the influence of domain on writing style as mentioned above. Hence, in this research, the models that will be used for authorship verification will be tested both on in-domain classification tasks and cross-domain classification tasks.

2.3 GENETICS INFLUENCE

From above, it can be concluded that a lot of research has been done on authorship verification, by looking at writing style, where situational factors like genre and domain influence the writing style of an individual. However, while it is proven that DNA influences someone's behaviour, as Plomin (2019) states that the influence of DNA on behaviour can be seen in a lot of actions from humans, genetics have not yet proven that this is also a factor which influences writing style. The influence of DNA on writing style can be researched by looking at texts from twins with an authorship verification model.

The only two researches that have been done on the influence of DNA on writing style are two Bachelor theses. During the BSc. Information Science, I wrote

a Bachelor thesis on this topic (Bruijn, 2019b). In this research, the influence of DNA on writing style was investigated using the Groningen Lightweight Authorship Detection (GLAD) model of Hürlimann et al. (2015) on columns written by two identical twin sisters. While the conclusions of this research were that there is a small correlation between DNA and writing style, the data set used in this research was too small, the texts used during this research were only a few sentences and no deep research has been done on what features are relevant. Also, the data set in this research was in English, while the GLAD model is more accurate for Dutch data.

Similar to the previous study, is the research done by Bruijn (2019a). In his research, Bruijn investigated whether genetics can be of influence for authorship verification. As data set, he used texts written by himself, his own family and his twin brother. While his research had some promising results on the correlation between DNA and writing style, the results were also not very solid due to the same small data set issue mentioned above.

In this research, several new data sets are created, in order to improve on the limitations of the researches of Bruijn (2019b) and Bruijn (2019a). The first newly created data set consists texts and some meta data of twins and siblings, obtained via an online survey. The other data set that was newly created contains pairs of texts from Reddit, where some pairs are written by the same user and some are written by two different individuals.

3 DATA

This data section is divided into two parts. The first part is about data that I used during my research, that already existed. The second part is about data sets that I create myself, together with my twin brother. This data set contains texts from twins and siblings who filled in a survey that we made ourselves. Also, we created a Reddit data set that contains pairs of texts written by the same person and pairs of texts written by two different individuals.

3.1 EXISTING DATA

For this research, a data set that contains pairs of texts annotated whether they were written by the same person or not was required. This data is needed for the training phase of the authorship verification model that will be used to investigate whether twins write more similar to each other than to random people or not. The data set that is used for this part is the PAN 2015 data set.

3.1.1 PAN 2015 Data Set

A data set that is used during my research, which already existed, is the PAN 2015 data set. This is a data set that was used during the PAN at CLEF 2015 shared task (PAN, 2015). The Groningen Lightweight Authorship Detection model, which will be used during this research and was made by Hürlimann et al. (2015), was built for this particular shared task.

Collection

Luckily for this research, the PAN 2015 data set is freely available and downloadable on their website. The complete data set consists of multiple data sets in several languages. Available languages are Dutch, English, Greek and Spanish. As the data of the self-collected data set from the twin survey in Section 3.2.1 is in Dutch, I only need the Dutch part of the complete PAN 2015 data set.

The Dutch data set is split up into a training set and a test set, which have a similar structure. The structure of the data sets is as followed. In the train or test folder, there are multiple folders called 'DU001', 'DU002', 'DU003', and so on. Inside each folder there is a text file called 'unknown.txt'. In the same folder, there are between one and five known texts files, called 'known01.txt', 'known02.txt', 'known03.txt', and so on.

In total, the training set consists of 100 folders for Dutch where each text contains about a few hundred up to a few thousand words. The test set consists of 165 folders, where the length of the texts is similar to that of the training set.

Annotation

In the train or test folder mentioned in the previous section, there is also a file called 'truth.txt', which has the gold labels stating whether in the Dutch folders the unknown text is written by the same person as the known texts or not. As the PAN data set is already annotated this way, no further manual annotation is needed.

Processing

As the PAN data set is already tokenized, no further tokenization or other pre-processing of the data is necessary and thus the data set is ready for this research.

3.2 SELF-COLLECTED DATA

During this research, some self-collection of data was required. As there exists no data set of texts written by twins, this data set had to be made. Also, during this research, a Reddit data set, which contains pairs of Reddit posts written by the same person and pairs of Reddit posts that are written by two seperate individuals, had to be made. The self-collected data sets are available on GitHub. I will now elaborate on both these data sets.

3.2.1 Twin Survey Data Set

In order to do research about the influence of DNA on writing style by looking at the writing style of twins and siblings, texts written by them is required. Figure 1 shows what the twin data set looks like.

Figure 1: Twin data set.

	B	C	D	E	F	G	H	I
1	**Relation**	**LongTexts**	**Gender**	**ID1**	**ID2**	**Value**	**Text1**	**Text2**
2	I	N	FF	1	2	1	Wij (Betty en ik) zi	Betty zei over (
3	I	N	FF	4	5	1	Ben er een van een	Fijn dat we een
4	N	Y	FF	7	8	0	Zoals de geschieder	Dit condoleanc
28	S	N	FF	1	3	0	Wij (Betty en ik) zi	Ik ben jaloers, a
64	R	N	FF	1	4	-1	Wij (Betty en ik) zi	Ben er een van
65	R	N	FF	2	4	-1	Betty zei over onze	Ben er een van

Collection

As no data set of texts written by twins exists yet, this data set had to be made during this research. I did this together with my twin brother Kai, because he has done research on the same topic and thus needed the same data set. In order to make the data set as large as possible, we decided to cooperate on this task.

To obtain texts written by twins and their siblings, we made an online survey. In this survey, we ask for the gender, the age, and written texts of twins and their siblings. To get as many responses as possible, we sent the survey to all of our friends and teachers and asked them to send the survey to any twins that they know. Also, we contacted the Dutch Twin Registry, who posted the online survey on their Facebook page.

In total, we received 26 pairs of texts from twins, out of which 21 pairs are from identical twins and the remaining 5 pairs are from nonidentical twins. Furthermore, we obtained 36 pairs of texts from siblings. The length of all texts varies between 45 and 1,654 characters with an average of 634 characters.

Annotation

For every pair of texts, we manually annotated the relation between the two texts. The four possible relations between the texts are: identical twin (I), nonidentical twin (N), sibling (S) and random (R). In this case, the random relation is simply a text written by someone in combination with a text written another individual who filled in the survey and has no shared DNA with the first individual.

In the twin survey, we also asked the gender and age of every person who filled in a text. In this research, this annotation is used to rule out the situational factors in order to make conclusions about the influence of DNA on writing style.

Processing

For the twin survey data, some processing had to be done before it can be used in the authorship verification models. Firstly, as mentioned in the previous section, pairs of texts had to be made with the relations between the texts annotated with it. To do this, we had to manually scrape the texts from a PDF file and copy it into an Excel file, including the annotations. As the texts were not tokenized yet, this also had to be done before it could be used in the authorship verification models.

3.2.2 Reddit Data Set

During my research, I found that the PAN data set is too different from the twin survey data to obtain solid results and conclusions when first training it on the PAN data set and then running it on the twin survey data. That is why I needed a new data set, consisting of pairs of texts written by the same person and pairs of texts not written by the same person, that has a more similar structure as the twin survey data than the PAN data set. Figure 2 shows an example of what the Reddit data set looks like.

Figure 2: Reddit data set.

	A	B	C	D	E	F
1	**Value**	**User1**	**User2**	**Text1**	**Text2**	
2	1	wafflewaldo	wafflewaldo	Marktwerking ? Geniaal . Wa	Ik ben zelf best wel te	
3	1	Netherviking	Netherviking	> Dit bewustzijn is hard nodi	> Wat dacht je van ge	
73	-1	Slabananananana	Willie1982	Vlees (indirect de boeren) Ik	Dit is wel mooi , maa	
74	-1	Bixbeat	MalleBeer	Wat een raar apparaat is de	De ontwikkelingen die	
75	-1	centerofdickity	Shrexpert	Bizar om te weten dat het m	Hetzelfde , uiteraard	

Collection

With the help of the PRAW module, we created a Python script that scrapes posts from Reddit into an Excel file. We collected data from popular Dutch Reddit pages and scraped two texts that were written by the same person and two texts that were written by two others.

In total, we scraped 253 pairs of texts, out of which 131 pairs of texts were written by the same person and 122 pairs were not. The training set consists of 136 pairs of texts and the test set consists of 117 pairs of texts.

Annotation

The Python script that scrapes the posts from Reddit into an Excel file checks whether the pair of two texts is written by the same person and then annotates a value of 1 if they were written by the same person and a value of -1 if not. As we do not have any other information on the users that posted the texts, no additional annotation is provided.

Processing

Some pre-processing was necessary for the Reddit data set as it was scraped from the website. First of all, the scraped texts had to be tokenized. Secondly, we replaced the use of links in a post with the word 'LINK'. After that, we replaced a mention to someone, e.g. '@user1', with the word 'MENTION'. Finally, pairs of texts were completely or partly English texts, which are removed from the data set as the authorship verification model only uses Dutch texts in this research.

4 METHOD

In this chapter, the methodology behind this research is explained. First, the general theoretical approach is introduced. After that, an explanation of which classification models are used in this research is provided. Also, feature analysis of the authorship verification task will be mentioned. Finally, the human evaluation part of this research will be introduced.

4.1 APPROACH

In this research, the influence of genetic and situational factors on writing style is investigated. This will be done by using multiple authorship verification models in combination with pairs of texts from twins and siblings. These classification models will be explained in the next section. Each classification model takes pairs of text as input and outputs a score, indicating how likely it is that the pairs of texts were written by the same person, according to the model.

During this research, one vs. one authorship verification will be performed between pairs of texts from twins, siblings and random people. These texts are obtained via an online survey, as mentioned in Chapter 3. When looking at a text of Twin 1 (T1), a text of Twin 2 (T2), a text of a sibling (S) and a text of someone else who filled in the survey (R), a comparison of the scores of the six combinations that are shown in Table 1 can be made. In this table, the percentage of shared DNA within the combination is shown, as this is necessary in order to make conclusions about the influence of DNA.

In total, there are four relations possible for the six combinations in Table 1. The abbreviations for the four relations are as follows: identical twins (I), nonidentical twins (N), siblings (S) and random (R).

The outputted scores of the authorship verification models of all possible combinations will be compared, after which it will be verified whether for example the scores between two identical twins are higher than the scores of nonidentical twins, siblings or random people. This might give an indication whether DNA has influence on writing style or not, as a score close to 1 indicates that the model thinks it is written by the same person and thus indicates a similar writing style, while a score close to 0 indicates that the model thinks it is not written by the same person and thus indicates a different writing style.

Before running an authorship verification model on the combinations of Table 1, these models have to be trained first. The model will not be trained on any data of the twin survey for two reasons. The first reason is that there simply is not enough data to do this. The second reason is that I want to want to check which of the four

Table 1: All possible combinations of texts.

Combination	% of shared DNA	Relation
T1 - T2	100 (I) or 50 (N)	I or N
T1 - S	50	S
T1 - R	0	R
T2 - S	50	S
T2 - R	0	R
S - R	0	R

relations writes most similar. To check this, a model should be used that is trained to see whether two texts were written by the exact same person. This is why the model will be trained on the freely available and downloadable Dutch data set from the PAN (2015) shared task.

4.2 CLASSIFICATION MODELS

For the approach explained in the previous section, authorship verification models are needed. There are several models that can be used, which will be explained now.

4.2.1 GLAD Model

The most important model and probably the most robust one is the Groningen Lightweight Authorship Detection (GLAD) model (Hürlimann et al., 2015). This is an already existing model and thus not self-made. The description of the model is as follows. The system is implemented using Python's scikit-learn machine learning library as well as the Natural Language Toolkit (NLTK). They used an SVM with default parameter settings in all final models, with an implementation based on libsvm. The model predicts the output based on several visual features.

The first visual feature is that the model looks at the punctuation of the text. It registers the frequency of question marks, exclamation marks, comma's etc. Secondly, the model looks at line endings of a text. It checks the frequency of full stops, comma's question marks, etc. Thirdly, the model checks the letter case. It calculates the ratio of uppercase characters to lowercase characters and the proportion of uppercase characters. Also, the model takes line length into account. Here, it looks at the following features: sentences per line, words per line and proportion of blank lines. Finally, the model looks at the block size of a text. This is split up in two parts. The first part is the number of lines per text block and the second part is the number of characters per text block.

Apart from all the visual features, the model also works with N-grams features, token features, sentence features, entropy features, compression features and (morpho)syntactic features.

4.2.2 LinearSVR Model

Next to the GLAD model mentioned above, a Linear Support Vector Regression (LinearSVR) model will also be used. Specifically, a regular LinearSVR model from sklearn with a TF-IDF vectorizer will be used during this research, as this is easy to implement.

4.2.3 Bleached LinearSVR Model

In addition to the LinearSVR Model mentioned in Section 4.2.2, a bleached version of the LinearSVR model will be created. Models that bleach text, i.e. transforming lexical strings into more abstract features, has shown to perform better than lexical models, as mentioned in van der Goot et al. (2018). This is why there will also be made a LinearSVR model that uses bleaching during this research. There are several alternative textual representations used within bleaching. For this research, the following bleaching methods will be used: PunctC, Shape, Vowel-Consonant and Length. The meanings of these alternative textual representations, further explained in van der Goot et al. (2018), are as follows.

- **PunctC:** Merges all consecutive alphanumeric characters to one 'W' and leaves all other characters as they are (C for conservative).
- **Shape:** Transforms uppercase characters to 'U', lowercase characters to 'L', digits to 'D' and all other characters to 'X'. Repetitions of transformed characters are condensed to a maximum of 2 for greater generalization.
- **Vowel-Consonant:** To approximate vowels, while being able to generalize over (IndoEuropean) languages, we convert any of the 'aeiou' characters to 'V', other alphabetic character to 'C', and all other characters to 'O'.
- **Length:** Number of characters (prefixed by 0 to avoid collision with another transformation).

4.2.4 BERTje Model

To include the state-of-the-art work of language models, this research will also use the BERTje model of de Vries et al. (2019). This is a monolingual Dutch BERT model. As de Vries et al. mentions, the transformer-based pre-trained language model BERT has helped to improve state-of-the-art performance on many natural language processing (NLP) tasks. Because the texts from the surveys in the twin data set are in Dutch, BERTje will be used to see whether it also works as an authorship verification model.

4.2.5 Other Models

Next to the four main authorship verification models mentioned above, a few basic scikit-learn models will be tried, namely a Linear Regression model, a Logistic Regression model and a Support Vector Machine model. As these are basic implementations and similar to the LinearSVR model mentioned above, I will not elaborate on these models. Also, if it turns out that there are multiple models that perform very well, an ensemble model of these models will can be made.

4.3 FEATURE ANALYSIS

Some deeper feature analysis will be done by either adding or removing a feature and investigating what the influence is on the scores outputted by the models. The feature analysis is split up in three parts: textual features, age and gender.

4.3.1 Textual Features

For the textual feature analysis, one by one a textual feature will be added within the GLAD model and the influence on the results will be investigated. This will indicate which features are more important for the model and thus are of influence of the writing style of twins and siblings. Examples of the textual features that will be analyzed are punctuation similarity and line endings similarity.

4.3.2 Age

In order to make conclusions about the influence of DNA on writing style, the situational factor of age should be accounted for. This is why this will be included in the feature analysis by checking if the overall results are similar to the results of people with roughly the same age.

4.3.3 Gender

Similar to the age factor mentioned above, the gender of a person should be taken into account. This will also be included in the feature analysis, by checking if the overall results are similar to the results of people with the same gender.

4.4 HUMAN EVALUATION

Next to the usual evaluation by looking at the accuracy of the models, a human evaluation part will be implemented. For this part, it will be investigated how well humans perform in an authorship verification task on the survey data. There will be a questionnaire with five random pairs of each one of the four relations in Table 1, summing up to a total of twenty pairs of texts. Next to the relations I, N, S and R, there are also three pairs of texts that were actually written by the same person, as three people filled in the survey twice. These are also added to the questionnaire for the human evaluation part. This will give a better view on how well the models perform, compared to humans.

In the questionnaire, people are given twenty three pairs of texts, where they have to give a score between 1 and 5 to each pair of texts, indicating how likely they think it is written by the same person.

5 RESULTS

In this chapter, the results of this research are shown. First, the results on each authorship verification model will be addressed, focusing on the best models. After that, the results of the best models on the twin survey data are summarized. Then, the results of the feature analysis will be shown. Finally, the results on the human evaluation part will be explained.

5.1 GLAD

The GLAD model (Hürlimann et al., 2015) performed best on all authorship verification tasks, both in-domain and cross-domain, independent of which training set was used in combination with which test set. The scores of all models on both in-domain classification and cross-domain classification are summarized in Table 2 . I will now elaborate on both the in-domain and cross-domain results of this model.

5.1.1 In-domain classification with GLAD

Looking at Table 2, the highest accuracy score of the GLAD model was obtained when training it on the PAN training set and testing it on the PAN test set, with an accuracy score of 0.76. This makes sense as the model was built for this task. Also, the model performed well for the in-domain classification task on the Reddit data, with an accuracy score of 0.68. Both these accuracy's are the highest in-domain classification scores of all models.

5.1.2 Cross-domain classification with GLAD

As cross-domain classification is often harder than in-domain classification, it does not come as a surprise that most of the cross-domain classification scores are somewhat lower than the in-domain classification scores. However, the GLAD model got some solid results on cross-domain classification, with even a higher score for the cross-domain classification Reddit-PAN than the in-domain classification Reddit-Reddit, shown in Table 2.

5.2 SVR

In this section, the scores of the LinearSVR model from sklearn with a TF-IDF vectorizer are elaborated on.

5.2.1 In-domain classification with SVR

As you can see in Table 2, the SVR model works fine for the in-domain classification task on the PAN data. However, the acccuracy score for the in-domain classification task on the Reddit data is only just above the baseline of 0.50, with a score of 0.53.

Table 2: Accuracy scores of all models for in-domain and cross-domain classification.

Model	Training set	Test set	Accuracy
GLAD	PAN	PAN	0.76
	Reddit	Reddit	0.68
	PAN	Reddit	0.66
	Reddit	PAN	0.69
SVR	PAN	PAN	0.70
	Reddit	Reddit	0.53
	PAN	Reddit	0.51
	Reddit	PAN	0.51
Bleached SVR	PAN	PAN	0.71
	Reddit	Reddit	0.55
	PAN	Reddit	0.50
	Reddit	PAN	0.52
BERTje	PAN	PAN	0.47
	Reddit	Reddit	0.47
	PAN	Reddit	0.49
	Reddit	PAN	0.47

5.2.2 Cross-domain classification with SVR

Looking at the cross-domain classification task with the SVR model, we can conclude that the model does not perform good at all, with an accuracy score of 0.51.

5.3 BLEACHED SVR

In this section, I will elaborate on the perfomance of both in-domain and cross-domain classification of the bleached version of the SVR model mentioned in Section 5.2.

5.3.1 In-domain classification with Bleached SVR

The bleached SVR model performs quite similar to the regular SVR model on the in-domain classification task. Actually, the bleached model performs even slightly better, with a solid accuracy of 0.71 on the PAN data and a score of 0.55 on the Reddit data, slightly outperforming the baseline on the Reddit data and significantly outperforming the baseline on the PAN data.

5.3.2 Cross-domain classification with Bleached SVR

The cross-domain classification scores with the bleached SVR model are similar to the scores of the cross-domain classification scores with the regular SVR model and thus do not significantly outperform the baseline.

5.4 BERTJE

As you can see in Table 2, the BERTje model does not work in this authorship verification task. Neither in the in-domain classification task nor in the cross-domain classification task does it outperform the baseline of 0.50.

Table 3: Results of the best model, tested on the twin data.

Model	Training set	I	N	S	R
GLAD	PAN	0.47	0.52	0.49	0.50
	Reddit	0.62	0.86	0.59	0.51

5.5 OTHER MODELS

During the research, it turned out the the other models (Linear Regression model, Logistic Regression model, Support Vector Machine and an ensemble) did not come close to the scores of the models mentioned above. That is why there will be no elaboration on the performance of these models.

5.6 RESULTS ON TWIN DATA

In Table 3, the results of the GLAD model, trained on the PAN data and the Reddit data and tested on the twin survey data, are summarized. Only the GLAD model showed to perform quite well on these data sets, even for cross-domain classification, which is why only this model is shown. Looking at Table 3, you can see that the GLAD model, trained on the Reddit data, shows some significant differences in predicted scores between the relations of identical twins (I), nonidentical twins (N), siblings (S) and random people (R). The highest score, 0.86 for nonidentical twins, is probably an outlier as there are only four instances of this relation.

The scores from the table indicate that people with shared DNA write more similar to each other than people who do not have shared DNA. Neglecting the scores for nonidentical twins due to the low number of instances, the scores for identical twins, having one hundred percent shared DNA, is the highest. The second highest score is the score for siblings, having fifty percent shared DNA. The score for people with no shared DNA is the lowest, with a score of 0.51.

The GLAD model does not show the same results when trained on the PAN data. A possible explanation for this would be that the texts in the PAN data set are too different from the twin survey data set, while the Reddit data set is probably somewhere in between.

5.7 FEATURE ANALYSIS

As the GLAD model outperformed all other models on all tasks, the feature analysis is performed on this model. The feature analysis is split up in textual features, age feature and gender feature.

5.7.1 Textual Features

The GLAD model uses a lot of textual features. The influence of each one of these features are researched by adding them one by one to the model, before training it on the Reddit data and then testing it on the twin survey data. The results of this feature analysis is shown in Table 4.

In this table, it is shown that two features indicate similar differences as the complete model in scores for the four relations, whereas the scores of the other features do not differ significantly. These two features are punct_sim and vec_sim. punct_sim is the similarity in punctuation use and vec_sim is the vector similarity. It seems that these two textual features are most important for the differences in the scores for the four relations and thus most influential for writing style differences.

Table 4: Textual features GLAD.

Feature	I	N	S	R
all features	0.62	0.85	0.58	0.50
punct_sim	0.58	0.59	0.52	0.51
line_endings_sim	0.57	0.57	0.56	0.56
linelength_sim	0.44	0.36	0.43	0.45
lettercase_diff	0.55	0.56	0.51	0.56
textblock_diff	0.50	0.49	0.51	0.53
vec_sim	0.61	0.74	0.57	0.53

5.7.2 Gender

From the results of the GLAD model, trained on the Reddit data and tested on the twin data in Table 3, it can be concluded that people with shared DNA write more similar to each other than people who do not have shared DNA. However, the gender of the two people who wrote the texts can be of influence.

In Figure 3, you can see that if the situational factor of gender is accounted for, i.e. only considering the scores for pairs of texts where the authors have the same gender, the scores of the four relations become more close to each other, except for the nonidentical twins, but as there are only four instances of this relation, this might be an outlier.

5.7.3 Age

Another situational factor that can be of influence for the scores of the GLAD model, trained on the Reddit data and tested on the twin data in Table 3 is age.

In Figure 3, you can see that if the situational factor of age is accounted for, the scores of the four relations are very similar, except for the nonidentical twins, but again this might be an outlier. In this figure, the accounting for factor age is done by checking whether the age differs for more than 10 years.

Looking at Figure 3, the scores of each relation when both the gender and age factors are corrected are similar. This indicates that these two situational factors are the most influencing factors, neglecting the influence of the textual features mentioned in Section 5.7.1.

Figure 3: Age and gender factors correction.

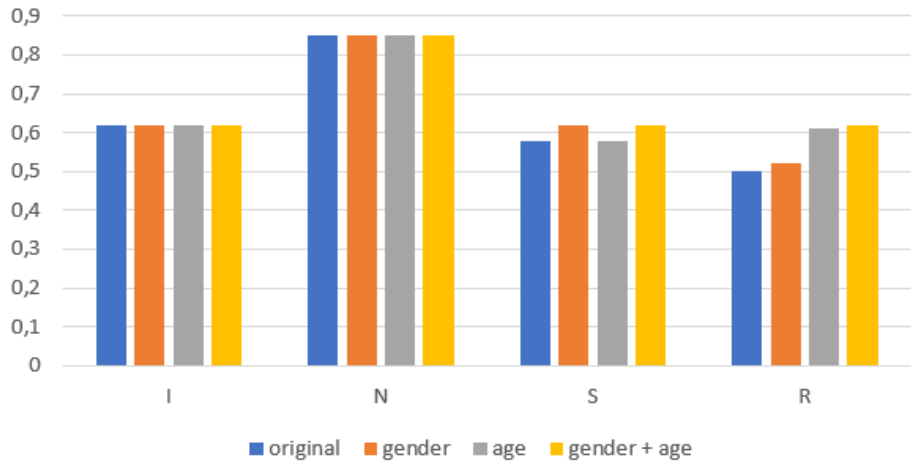

Table 5: Overview of the results on the human evaluation questionnaire.

Question	Genders	Ages	Relation	Human 1	Human 6	Human 13
2	F-F	30-33	S	1	1	1
3	M-M	30-30	I	3	3	1
14	F-F	75-62	R	5	1	1
18	F-F	29-29	N	3	1	1
22	F-F	27-27	X	5	2	1

Table 6: Summary of the human evaluation and GLAD model.

Question	Genders	Ages	Relation	Human avg	GLAD	GLAD (1-5)
2	F-F	30-33	S	1.85	0.72	4
3	M-M	30-30	I	2.62	0.45	3
14	F-F	75-62	R	2.69	0.29	2
18	F-F	29-29	N	2.54	0.86	5
22	F-F	27-27	X	3.31	0.80	5

5.8 HUMAN EVALUATION

For the human evaluation part, a questionnaire was filled in by 13 people, who had to give a score between 1 and 5 to pairs of texts, indicating whether they thought they were written by the same person or not. Table 5 gives an overview of what the results of this questionnaire look like, where some random results are shown. An example for every relation of all five relations is shown: same person (X), identical twin (I), nonidentical twin (N), sibling (S) and random (R). Looking at this table, it can be concluded that humans do not agree much when giving the scores to the pairs of texts. Also, they do not give significantly different scores to different relations.

In Tabel 6, the average score of the humans is compared to the scores of the GLAD model for some random example question pairs. In this table, the column GLAD contains the scores outputted by the GLAD model and the column GLAD (1-5) are the outputted scores transformed to a score between 1 and 5. In this table, you can see that the model is better at distinguishing between different relations than humans.

6 CONCLUSION AND DISCUSSION

6.1 CONCLUSION

In this research, the influence of genetic and situational factors on writing style was researched, out of which several conclusions can be drawn.

First of all, the Groningen Lightweight Authorship Detection (GLAD) model of Hürlimann et al. (2015) turns out to be a robust system, which can deal with both in-domain and cross-domain authorship verification tasks. It significantly outperformed other systems like a LinearSVR model, where the bleached version of this system worked slightly better than the regular version.

Secondly, the GLAD model showed that two authors with shared DNA tend to write more similar to each other than people who do not have shared DNA. Especially twins write quite similar to each other. The most important textual features that are of influence for this phenomenon are punctuation similarity and vector similarity. However, when situational factors like gender and age are taken into account, this influence seems to be negligible.

Thirdly, from the human evaluation part, it can be concluded that humans can not perform an authorship verification task between a pair of two texts written by the same person, identical twins, nonidentical twins or random people.

Also, from the performance of the GLAD model on the PAN data, the Reddit data and the twin data, it can be concluded that the PAN data and the twin data are too different for the model to be trained on the PAN data, before tested on the twin data. The Reddit data set seems to lie somewhere in between the other two data sets.

In conclusion, the answer to the research question of this research, *"What is the influence of genetic and situational factors on writing style?"*, is as follows. While people with shared DNA write more similar to each other than people who do not have shared DNA, the textual features like punctuation similarity that showed to be of influence here can be neglected when the situational factors gender and age are taken into account. Hence, the situational factor that are investigated in this research are of more influence than the DNA factor for writing style.

6.2 LIMITATIONS

Even though I tried to make my research as firm and solid as possible, there are some limitations in this work and thus some improvements to be made in future work. Most of these limitations are due to the fact that I did not have enough time or resources during my research.

First of all, twin data set is not big enough to make one hundred percent solid conclusions out of it. For example, there are only four pairs of texts from nonidentical twins. It would be interesting to see if the conclusions of this research are still similar to the conclusions of a research that uses much more data from twins and siblings.

Secondly, the relations that I used during this research are not perfect. In order to research the influence of shared DNA, it would be better to have identical twins that grew up separately and adopted people who grew up in the exact same environment. This is because identical twins have one hundred percent shared DNA, but often also grow up in the same environment, which may also influence writing

style. The same idea applies for adopted people who grew up in the exact same environment, as they would be perfect for this research, having no shared DNA at all, but they do grow up in the same environment. However, I did not have the time or resources to find any of these people, let alone enough to make a solid data set.

Finally, while the Groningen Lightweight Authorship Detection (GLAD) model of Hürlimann et al. (2015) turned out to be a solid and robust system, it still is not perfect, as the accuracy lies around seventy percent. The conclusions would be more solid if a more accurate system is used for the same research, but this system does not exist (yet).

6.3 FUTURE WORK

As mentioned in the previous section, there are some limitations to this research, due to a lack of time and resources. This results into some recommendations for future work.

For a more solid conclusion, I would recommend bigger data sets for future research on this topic. Also, if it is feasible, I would recommend to add texts of adopted people and twins who grew up separately in the data sets. Finally, an authorship verification model that has a higher accuracy score is recommended for future research in this area, if it is possible.

BIBLIOGRAPHY

Baayen, H., H. van Halteren, A. Neijt, and F. Tweedie (2002). An experiment in authorship attribution. In *6th JADT*, pp. 29–37.

Bruijn, K. (2019a). The influence of dna on writing style. authorship verification with identical twins.

Bruijn, R. (2019b). Authorship identification with shared dna. how dna influences writing style.

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*.

Halvani, O., C. Winter, and A. Pflug (2016). Authorship verification for different languages, genres and topics. *Digital Investigation 16*, S33–S43.

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

Nederhoed, P. (2010). *Helder rapporteren: Een handleiding voor het opzetten en schrijven van rapporten, scripties, nota's en artikelen*. Bohn Stafleu van Loghum.

Overdorf, R. and R. Greenstadt (2016). Blogs, twitter feeds, and reddit comments: Cross-domain authorship attribution. *Proceedings on Privacy Enhancing Technologies 2016*(3), 155–171.

PAN (2015). Pan at clef 2015. https://pan.webis.de/clef15/pan15-web/author-identification.html.

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

Rexha, A., M. Kröll, H. Ziak, and R. Kern (2018). Authorship identification of documents with high content similarity. *Scientometrics 115*(1), 223–237.

Rudman, J. (1997). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities 31*(4), 351–365.

Sari, Y., M. Stevenson, and A. Vlachos (2018). Topic or style? exploring the most useful features for authorship attribution. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 343–353.

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

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*.

Zheng, R., J. Li, H. Chen, and Z. Huang (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.