university of groningen

faculty of arts

The Influence of DNA on Writing Style

Authorship Verification with Identical Twins

K. A. Bruijn

Bachelor thesis
Information Science
Kai Bruijn
s3204766
June 6, 2019

ABSTRACT

Authorship verification is the task of identifying whether or not two or more documents are written by the same person. Authorship verification is based on an author's writing style, distinguishing one person from another. This can be useful in detecting plagiarism or in criminology.

Although many previous work has focused on authorship verification, no previous work exists on authorship verification with shared DNA. This is where my research differs from other papers.

In this thesis I researched if it is harder to distinguish between authors when they have genetic resemblance. I do this by examining texts written by identical twins and check if it influences the accuracy scores of a machine learning based classifier. I also take a look behind the scenes on how certain the classifier is. This classifier is built using the programming language Python. The data were collected from a personal circle, meaning that my family had to write several texts on a specific topic. To prove the hypothesis that DNA influences writing style, multiple scores were compared from different research groups.

The first research group consists randomly selected authors, not sharing DNA or growing up in the same environment. The second research group consist of an identical twin, while the third research group consists of the family of the identical twin.

The results show that there is a significant difference in AUC-c@1 scores between several researched groups. The classifier also shows that it is less certain about classifications concerning identical twins. It remains unclear what the cause of this deviation is but it can be concluded that it is hard to tell identical twins apart based on their writing style. There is still room for improvement as described in the discussion section, leaving the possibility of future work.

CONTENTS

PREFACE

It has been a great pleasure writing my bachelor thesis to be able to graduate as information scientist. The past few months have been quite important, which makes the satisfaction of finishing this thesis even greater. It all started with the idea of my supervisor, Ms Nissim, who came up with analyzing authorship verification for identical twins. Of course she thought about my twin brother and me, as it is not very common to supervise the bachelor thesis of identical twins, and we both wanted to write our thesis on authorship attribution. We started with a meeting on how we could implement this in our theses. Ms Nissim provided us with a classifier for which I am very grateful. By then it was time to figure out how Rune and I would still be able to write different theses, considering that we were going to research the same subject with the same classifier. Eventually, we decided that Rune was focusing on texts written in English, while I was focusing on texts written in Dutch, from a personal perspective.

I would therefore like to take this opportunity to thank Ms Nissim for the inspiration and all the help that came with it. This is also my chance to promote my twin brother's thesis on the same subject, but focusing on an alternative approach. He researched columns written by identical twins who did not grow up in the same environment. The texts are written in English, leading to different results. For a more detailed description, I recommend reading his thesis.

I never regretted choosing to study information science as I was always interested in the subjects that we covered. I can add the subject of this thesis to that list, which makes it easier and more fun to write. That is why I am left with a satisfactory feeling and a smile on my face when writing the last words of this research. Again, many thanks to my supervisor and if this topic interests you, do read the thesis "Authorship Identification with Shared DNA", written by Rune Bruijn.

Just to prove my hypothesis in a not so scientific manner, after finishing our theses, Rune and I read each other's papers. Although it is hard to statistically prove that they are very similar, reading both theses will do so. If you enjoy reading this thesis, have a look at his one as well, and see for yourself that our hypotheses are well based on personal experiences.

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 also profile authors (Argamon et al., 2009). 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 verification with shared DNA. Authorship attribution is used to distinguish authors based on their writing style. It can tell if a text is written by a certain person. This could be of great importance in practical applications such as detecting plagiarism but also in criminology. Imagine Donald Trump getting a threatening letter from one of his staff members, not knowing who. Authorship attribution might be able to determine who will probably have to go to the employment agency to find a different job. In this thesis, the focus lies on authorship verification, which answers the question whether a document is written by a certain author, not having a finite set of possible authors, as is the case with authorship identification.

Quite some research has been done on authorship verification, how to improve accuracy, its applications and such (Calle-Martín and Miranda-García, 2012), however, not much research has been done on the impact of DNA on writing style and therefore the impact on authorship verification. From personal experience, I can tell that identical twins have a very similar writing style. Whenever my twin brother and I had to write essays, they would be very similar, sometimes even leading to suspection of plagiarism and I can imagine that many other twins have the same feeling. This leads to the hypothesis that sharing DNA leads to a similar writing style. This could have an impact on the accuracy of a classifier and could lead to new insights about authorship verification combined with genetics. In the example of the president of the United States, if the two cooks in the White House are identical twins, would a classifier based on a machine learning algorithm be able to tell the difference between them?

This thesis will try to answer that question, as this topic has not gotten much attention in previous papers. The research question that this thesis tries to answer is: *how much does DNA influence writing style and thus authorship verification?* This will be researched by answering the question how much shared DNA influences writing style. The structure of the research is as follows. Chapter 2 describes previous works that are of interest for this thesis. As stated above, not much research has been done on authorship verification with shared DNA, but authorship verification itself is a field that many authors have studied. Chapter 3 will cover the data and material used for authorship verification. This is followed by chapter 4, which will explain the method used and chapter 5 will show the results, complemented with a discussion. Additionally, the conclusion can be found in chapter 6.

2 BACKGROUND

The impact of DNA on authorship verification has not yet been researched. There is no related work available that describes or analyzes the impact or influence of DNA on a person's writing style. By the time of finishing this thesis, there will be two papers on this topic. Of course this paper itself, but also the related work of my twin brother, who researched the same subject. The difference between our theses is that this thesis focuses on personal environment. Nevertheless, many authors researched authorship attribution and authorship verification, which is the main part of this thesis and is important to discuss.

A very relevant paper for my thesis is the one written by Hurlimann et al. because they used the same classifier for their research: GLAD: Groningen Lightweight Authorship Detection. I have to give them lots of credits for building a classifier that positioned the system among the top PAN at CLEF 2015 competition shared task participants. This is a competition on authorship identification. For the purpose of this research, the classifier was only minimally changed. The approach, and the reason for using their classifier will be further explained in the method section. The same goes for data and material, for which there are clear reasons to use pre-existing work. This will be explained in the data and material section.

Recently, there was an interesting paper published by Halvani et al. (2018) on unary and binary classifiers for authorship verification. They researched serious implications regarding prerequisites, evaluability, and applicability of underlying classification models. It answers the question whether or not two documents were written by the same author. This is very similar to my research question where I will test if a machine learning algorithm can distinguish between two authors who share DNA. In their paper, Halvani et al. made several approaches to most accurately apply authorship verification on self-compiled corpora. This includes unary, binary, intrinsic and extrinsic approaches. Following their conclusions, they advise to consider unary classifiers as they outperformed binary classifiers. For my thesis however, I stuck with binary classification because it performed better in the PAN competition on authorship identification because of the balanced data. From the intrinsic versus extrinsic approach even more useful conclusions can be drawn. Hurlimann et al. also analyzed these options, so this will be further discussed in the method section.

Another useful paper to consider is the one written by Calle-Martín and Miranda-García (2012). They observe that with the advent of computers and the increasing availability of machine-readable texts authorship attribution is growing. As stated before, authorship attribution is not exactly the same as authorship verification but there is much similarity between the two and much can be learnt from both subjects. They also compared several papers on stylometry, or a person's writing style, and authorship attribution, providing a proper summary of multiple approaches. It gave me some inspiration for my thesis and this is actually the only paper in which twins are mentioned, unfortunately only to state that it is difficult to tell them apart physically and make a comparison to similar writing styles of authors. They do not mention the hypothesis that identical twins could have a similar writing style, or the effect of this hypothesis. In fact, they more or less state that despite the significant resemblance of identical twins, people are characterized by their individuality, which is writing style as well. This is exactly the opposite of the hypothesis in this thesis. The article however is a nice bridge towards my thesis on authorship verification with shared DNA.

Like I said before, these papers are interesting to read and useful to consider when writing a thesis on authorship verification. However, none of the papers seriously consider the possible impact of DNA on authorship verification. This is where my thesis differs from currently available related work and could lead to new insights in this field of study.

3 DATA AND MATERIAL

3.1 COLLECTION

For this thesis, a specific data set is needed. First, it is essential to have randomly selected, written text of which the author is known. This is needed to train the classifier but also to compare scores on regular people to scores on people who share DNA. This means that training data and test data are needed. It does not matter much which data set this is, so I chose the publicly available data set from the PAN competition (Stamatatos et al., 2015). This includes Dutch, English, Greek and Spanish texts written by several authors. There are multiple reasons to use this data set for this research. To begin with, it is the same data set as the one that the original classifier was built on. It is therefore easy to run, there is training and test data available, it is guaranteed to be applicable and the accuracy is known at 0.62 on the test data, positioning the system among the top participants. Secondly, the size of the texts is roughly the same as the size of the texts that a different research group had to write. This will be further explained later on. I chose to use the Dutch texts only, which is easy to understand when considering the facts explained in the next paragraph.

Of course the next data set required is one containing texts written by people who share DNA. In this thesis I researched my identical twin and myself as this group. As we are Dutch native speakers, and the hypothesis is that DNA influences our writing style, it seems logical to stick to Dutch texts. This is because a native language is closer to DNA than a language learnt at the age of around ten. English texts would be more influenced by our environment instead of DNA. That is why the data set contains Dutch texts, written by us in the past. For the purpose of the next group and more data and therefore more certainty in the results, we also wrote three different texts on specific topics.

Comparing the accuracy of the classifier on the regular data to the accuracy on the data from identical twins will show whether or not the system finds it more difficult to distinguish between authors with shared DNA. But it could very well be that the environment in which the identical twins grew up plays a vital part. To research this, I introduced a third research group: my family. My two older brothers grew up in roughly the same environment, just like me twin brother and me. However they do not share as much DNA as we do. My mother and father of course do not share DNA as they come from different families. For authorship verification research, this family is quite interesting. Because of the lack of availability of texts, I gave my family the assignment to write three different texts, with the length of at least one hundred words, on specific topics, including my twin brother and me. This means that group 3 also contains data from the identical twins. This third group complements the other two groups resulting in three different research groups of which the results can be examined.

3.2 ANNOTATION

For the first data set, containing multiple folders with regular texts, a text file is available called `truth.txt` in which the labels are given. This means whether or not two or more texts are written by the same person, given for every folder in the root directory.

For the texts written by identical twins, it is of course known to me who wrote the texts. The answers are kept in a truth file to make sure that it is possible to find out who wrote what.

For the last data set, containing texts written by people who grew up in roughly the same environment, a truth file was also made to be able to check the author of a certain text. Then, all the texts from the second and third research group were added to a different test set, leaving three different test sets. No texts were added to the training data because the research question wants to find out if a classifier struggles to tell the difference between people who share DNA. The training data should therefore not necessarily contain texts written by identical twins.

3.3 PROCESSING

The texts from the original data set were already applicable to the classifier because the classifier was built on this data set. It did not need any processing. The texts from the other two research groups had to be transformed into utf-8 encoding. This is because texts were sent using Word, but use of punctuation was calculated using different characters, for example apostrophes. Additionally to changing the encoding, the new texts had to be tokenized. It was done by a simple Python program that replaces characters and uses the `word_tokenize` function provided in the NLTK package. Some examples of the text files on the same subject, written by research group 3 are shown below.

```
New York is de hoofdstad van Amerika .  Het is een wereldstad met
miljoenen inwoners .  New York is een van de grootste steden van de
wereld .  In New York staat ook de Big Apple .  Verder is New York de
stad waar het Word Trade Centre ( WTC ) zit .  Vroeger stonden daar ook
de Twin Towers .  Dit waren twee hoge gebouwen waar op 11 september twee
vliegtuigen ingevlogen zijn .  Dit was in 2001 .  De vliegtuigen waren
gekaapt en het was waarschijnlijk een terroristische aanslag .  Verder
staat in New York nog het vrijheidsbeeld .  Ik wil nog wel een keer naar
New York op vakantie .
```

```
New York City is een stad in de staat New York in de United States of
de Verenigde Staten van Amerika .  Ik ben er nog nooit geweest , zou
er wel een keer naartoe willen .  Dan zou ik wel in een rustige periode
gaan want ik heb eigenlijk een hekel aan toeristen .  Dan zou ik al
die toeristen-highlights doen en het lijkt me ook wel leuk om naar Wall
Street te gaan .  Het is denk ik net leuk voor een weekje , niet langer .
Dan ga ik een donut halen en een take-away koffie en daarmee over straat
lopen .  Daarna een vette burger naar binnen drukken .
```

```
New York City is een grote stad in het oosten van de Verenigde staten
.  Het ligt aan de kust en is bij iedereen wel bekend .  Er worden veel
films en series opgenomen in New York omdat het zo ’ n bekende stad is .
Ik ben er nog nooit geweest , maar het lijkt me erg leuk om er een keer
naartoe te gaan , ondanks dat het daar waarschijnlijk erg duur is .  Het
is gewoon een stad die je een keer moet hebben bezocht .  Dat de stad
zo ver weg is , helpt ook niet .  Ik zou in New York een dikke hamburger
bestellen en de typische toeristenactiviteiten doen .
```

4 METHOD

4.1 BUILDING A CLASSIFIER

For this research I used the pre-existing and publicly available classifier GLAD: Groningen Lightweight Authorship Detection. A classifier was built based on binary class SVM. For one-class classification, Support Vector Machines (SVM) have been shown to perform very well (Manevitz and Yousef, 2001; Yu, 2003), but when the data distribution is in the ratio of 1:3.5, binary classification outperforms one-class classification. As the task in this research is to find if multiple texts are written by the same author, this problem resembles binary class classification. The instances can be classified as written by the same author, thus class 'Yes', or written by different authors, thus class 'No'. Using the data set from the PAN competition, the data were equally distributed which lead to the choice of a binary class Support Vector Machine classifier. In order to maximize speed and simplicity, they do not introduce any negative examples by means of external documents, thus adhering to an intrinsic approach (Hurlimann et al., 2015). The classifier was built using the programming language Python and, among others, the NLTK and Scikit-Learn libraries. The classifier scores 0.62 on the original Dutch data, details will be elaborated on in the evaluation section.

4.1.1 Training

Training the classifier was done using the available training data from the PAN 15 competition. These data are publicly available and are equally distributed instances of texts written by the same author and texts written by different authors. There are, for Dutch, one hundred directories containing several texts, whether or not written by the same author. In the root directory, there is a file in which the truth can be found. There are no texts written by identical twins in the test data in order to research if the classifier will struggle more with texts written by people who share DNA. Having the data available and, as stated before, already processed, the training data is ready for usage.

4.1.2 Feature Selection

The next step is feature selection to be able to reach the highest accuracy. There are many features to consider when building a classifier for authorship verification. Hurlimann et al. analyzed n-gram features, token features, sentence features, entropy features, visual features, compression features and (morpho)syntactic features. From this analysis, they compiled the best feature selection for Dutch, English, Greek and Spanish. For this thesis, only the Dutch feature selection is of interest, which leads to the choice to use n-gram features, visual features and token features. This is implemented in the classifier by recognizing the language and selecting the best combination of features for all four languages. More details on feature importance are described in their paper, for now it is most relevant to use the optimal feature selection described above.

4.1.3 Testing

This is where the new part of this research shows up. In my thesis I want to find out whether or not it is more difficult for a classifier to distinguish between authors who share DNA. The classifier that I am using is able to show probability scores on the existing test data from the PAN competition. Additionally to these data, I have collected texts written by identical twins and people who grew up in roughly the same environment and added these texts to other test sets. The achieved scores will then be able to show the differences between the researched groups.

4.2 EVALUATION

To be able to answer the research question *how much does DNA influence writing style and thus authorship verification?*, a comparison between the three different research groups is necessary to be able to show differences in scores between groups. For every group, there will be an evaluation, showing the certainty, or probability of the classifier for the assignment to a specific class and the AUC-c@1 score will be calculated, a common score used to evaluate binary classifications. This score multiplies AUC by c@1 where AUC is the area under the curve, meaning correct predictions and c@1 is the probability, meaning how certain the classifier is. The results of these scores are able to show any deviations in research groups. Of course the AUC-c@1 score will be the most important one but probability scores of the classifications might give some information from behind the scenes. Hence, we will also dive into these scores to be able to evaluate the outcome. The baseline of AUC-c@1 is 0.25 as AUC is 0.5 for random guessing, as well as c@1, but the accuracy score of 0.62 on the regular, Dutch data is the score that I will use to compare the results on texts written by people who share DNA. If the other research groups score differently, there is something worth discussing.

With regular authors being research group 1, identical twins being research group 2 and people who grew up in the same environment being research group 3, a comparison between the three groups will be able to answer the research question. If the system has a lower score for research group 2 compared to research group 1, and there is no difference in scores between research group 1 and research group 3, the hypothesis that DNA influences writing style might be correct. The obtained results will be discussed in the next chapter.

A final evaluation manner will be discussed in the next chapter, which will be explained there as well. This evaluation tells us more about the probabilities of classifications of problem instances.

I make my code and data publicly available at `https://github.com/kaibruijn/Thesis_DNA_Authorship_Verification`.

5 RESULTS AND DISCUSSION

5.1 RESULTS

Following the method described in the previous chapter, I obtained results for every research group and they somewhat confirm the hypothesis. First, I ran the classifier on the original data from the PAN competition (table 1). Some numbers do not add up because some problems were unanswered, i.e. not classified. Every research group has met the conditions for optimality of a binary classifier, a ratio less than 1:3.5 and randomly assigned problem instances, problem instances being a classification of either 'Yes' or 'No'.

Table 1: Regular authors

Problems	Correct	Incorrect	AUC-c@1
165	120	41	**0.60**

As can be seen in the table above, the classifier works quite well and has no deviations from the results obtained by Hurlimann et al. (2015). Then, the classifier was run on the data of identical twins, containing 16 texts divided into 7 problem instances.

Table 2: Identical twins

Problems	Correct	Incorrect	AUC-c@1
7	4	3	**0.14**

Obviously, the score goes down a lot, scoring less than 25 percent of the original score. What also jumps to the eye is the reduction in problem instances, because of the use of not so many data. I therefore wanted to rule out that the drop of the score was somehow caused by the reduction of instances, so I randomly selected a similar amount of data from the original test data. The results does not deviate from the original high score.

The two tables above show that the classifier has more difficulties distinguishing between authors who share DNA than for randomly selected authors. To examine whether this is genetically determined or whether the environment plays a part in this drop, data from people who grew up in the same environment were researched as well, leading to the following results.

Table 3: Shared environment

Problems	Correct	Incorrect	AUC-c@1
8	2	6	**0.03**

Needless to say, the imperative thing to note here is that the AUC-c@1 score is ridiculously low at 0.03. The classifier can easily be seen as rubbish when fitted on these data. As a reminder, the baseline for the AUC-c@1 score for this problem is 0.25. This surprising result will need further investigation to be able to completely understand what happened. The conclusion that we can draw from this table is that it is not necessarily the genetical part that the program is struggling with.

To check if there is something wrong with the data obtained from people who grew up in roughly the same environment, I ran the classifier again, now on a different distribution of data. Every person in this research group wrote three texts on a specific topic. I put the texts from every author in one directory, only switching the `unknown.txt` from the identical twins, meaning that only their classification should be 'No'. This leads to similar results as the research done by Hurlimann et al. (2015) (table 4). However, the data do not meet the 1:3.5 ratio and this could not really be called random distribution. This still leaves room for discussion of the results.

Table 4: Shared environment changing twins only

Problems	Correct	Incorrect	AUC-c@1
6	4	1	**0.78**

Besides the quantitative analysis described above, the classifier also assigns certainty scores to each classification instance. Taking a look at these scores might give more insights to the results of the classifier and therefore the differences between research groups. The classifier assigns a score of certainty to each instance as a float between 0 and 1. When the score is 1, it means that the classifier was absolutely certain of its classification that the documents were written by the same autor. When the score is 0, it means that the classifier was absolutely certain of its classification that the documents were **not** written by the same author. In table 1, it is shown that the classifier sometimes does not assign an instance to one of the two classes. This is the case when the certainty score is 0.5, meaning that the classifier has no clue whether it should be 'Yes' or 'No'. This tells us that if these scores are closer to 0.5 for research group 2, the classifier was not so sure about the distinction between authors when they have shared DNA. These results were analyzed because the original analysis lead to some uncertainties in results. Looking at the absolute difference of the obtained scores to 0.5 will provide insights to this matter. The average of the scores was calculated. For research group 1, a similar amount of data as the other two research groups was used in order to equal chances of outliers. The lower these difference scores are, the lower the certainty of the classifier is. The results are shown in table 5.

Table 5: Average absolute difference

Research group 1	Research group 2	Research group 3
0.211	**0.078**	**0.234**

The results from the table above clearly indicate that the texts written by identical twins are more similar than texts written by random authors or authors who grew up in roughly the same environment. This starts to prove the hypothesis that genetics influence writing style, but it should not be discarded that the accuracy scores for research group three were really low. Therefore, the results should be interpreted carefully.

5.2 DISCUSSION

Despite a promising start from the results in tables 1 and 2, the results from table 3 oblige to add a footnote to these results. This table reduces the trustworthiness of the results from the previous ones. There are some possible explanations for the surprising result of the third research group described in the subsections below. The results of table 5 are interesting, though the interpretation of the results depend a bit on the AUC-c@1 score mentioned in the tables before.

5.2.1 Subjects

The third group had to write texts on specific subjects as assigned by me. As my supervisor advised me, writing on extremely different subjects might lead to a unintentionally big difference in texts. However, it could be that writing texts on the same subject has influenced the classifier even more. It would be better to have pre-existing data as no author would then be biased by the assignment. Unfortunately, these were not available and it was therefore not possible for my research.

5.2.2 Number of Texts in Directory

The main difference in the approach which lead to different scores in table 3 and 4 is the number of texts in a problem instance. In table 4, the classifier could compare three texts instead of two, which probably caused the higher score. However, in the original data from PAN, there was a similar distribution as in table 3.

5.2.3 Coincidence and Other Factors

Perhaps it is very difficult to find out what caused the drop in table 3. It might just be coincidence, considering the fact that there was not a large amount of data to work with. It could also be that my family just so happens to have a strong bond and therefore a strong similarity in writing style. As stated before, pre-existing data would be better but not possible for this thesis.

To conclude the discussion, more data are needed to ensure the trustworthiness of the results. It would also be much better to use pre-existing data instead of data that were more or less created. I leave these findings for future work as it was unfortunately not possible for this research.

6 CONCLUSION

In this thesis, I researched whether genetics can be of influence for authorship verification. I conducted this research in a personal circle, leaving a different approach for my fellow student and twin brother, who researched a similar issue. I started with collecting data from my family and pre-existing data from my identical twin brother and myself. I then compared the scores for three different research groups to be able to prove my hypothesis that genetics are of influence for authorship verification.

To conclude my thesis, the answer to the research question *how much does DNA influence writing style and thus authorship verification?* is still somewhat unclear. There are some promising results in the tables that confirm the hypothesis, but the role of the environment is still vague. Some possible explanations are given in the section after the results. Future research, using more and pre-existing data, might give us better insights in this field of study.

Using the amount of data available, tables 1 through 5 show the results of comparisons between different research groups, leading to some doubting but definitely interesting results.

From this thesis and its results, it can be concluded that the classifier struggles more with data obtained from identical twins than data obtained from regular authors. The accuracy scores were lower, and the classifier was less certain of its classifications on the data obtained from people who share DNA. The influence of the environment first seemed of a significant proportion, but taking a look behind the scenes showed otherwise. This also leaves room for future work.

Possible research could be done on identical twins who did not really grow up in the same environment, though this is very rare. For now, one of the first papers on authorship verification with shared DNA has been written. Whether it is genetically determined or influenced by our environment remains somewhat unclear, but the classifier was having a hard time and some excitement may arise after this research on the influence of DNA on authorship verification.

BIBLIOGRAPHY

Argamon, S., M. Koppel, J. W. Pennebaker, and J. Schler (2009). Automatically profiling the author of an anonymous text. *Commun. ACM 52*(2), 119–123.

Calle-Martín, J. and A. Miranda-García (2012). Stylometry and authorship attribution: Introduction to the special issue. *English Studies 93*(3), 251–258.

Halvani, O., C. Winter, and L. Graner (2018). Unary and binary classification approaches and their implications for authorship verification. *arXiv preprint arXiv:1901.00399*.

Hurlimann, M., B. Weck, E. van den Berg, S. Suster, and M. Nissim (2015). Glad: Groningen lightweight authorship detection.

Manevitz, L. M. and M. Yousef (2001). One-class svms for document classification. *Journal of machine Learning research 2*(Dec), 139–154.

Stamatatos, E., M. Potthast, F. Rangel, P. Rosso, and B. Stein (2015). Overview of the pan/clef 2015 evaluation lab. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 518–538. Springer.

Yu, H. (2003). Svmc: single-class classification with support vector machines. In *INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE*, Volume 18, pp. 567–574. Citeseer.