MACHINE DRIVEN PREDICTIONS OF THE SOCIO-ECONOMIC STATUS OF TWITTER USERS

Fons Mentink (s1010247)

Master thesis Business Information Technology
Track: Business Analytics
Faculty of Electrical Engineering, Mathematics
and Computer Science (EEMCS)

UNIVERSITY OF TWENTE.

Deloitte.

Machine driven predictions of the socio-economic status of Twitter users.

Author:
Fons Mentink (s1010247)
fmentink@deloitte.nl

Supervisors University of Twente:
Dr. Robin Aly
Dr. Chintan Amrit

Supervisor Deloitte:
Dr. Stephan Peters

October 2015-June 2016. Amsterdam, the Netherlands.

FOREWORD

Dear reader,

For the past 8 months, I've enjoyed working on this thesis. Even though I was completely new to the subjects of machine learning, socio-economic statuses and ethical frameworks, I found it to be a good fit with my interests. Though challenging and frustrating at times, this project has taught me more about analytics, being a responsible data scientist, writing and programming than I'd ever imagined.

I couldn't possibly have gotten to the point where I am if it weren't for the outstanding supervision from both the University and Deloitte. I owe a lot of thanks for the time and effort that were invested in me and my thesis. It made this study understandable, valuable and fun.

As this project comes to an end, I've started with a new challenge. I will start my career at Deloitte Consulting, where I will be a part of the Analytics & Information Management team.

Thanks go out to my family, friends and my dear girlfriend for all the support.

Fons Mentink

Utrecht, June 2016

ABSTRACT

Over the past few years, there has been a rise in the number of studies to infer the latent attributes of Twitter users. A large portion of these studies examines the possibilities of inferring the gender and age of users, while some take on more challenging characteristics such as occupation, income or political affiliation. Our study attempts to infer the Socio-Economic Status of the Twitter users, based on the occupation and education of the users.

Most of these studies use the same approach, where a sample of users with known labels is created and used to train classification algorithms. We use the same approach, but create the labels based on LinkedIn profiles that match the Twitter users in our sample. Before the data collection process was started, we employed an ethical framework to make the intended values of the research clear and document and discuss the value-tradeoffs.

The data was collected and stored on restricted servers, after which the assignment of labels began. This was done according to internationally used classification schemes for the education, level of education and occupation of the users in our "train/validation set". After analyzing the collected data, we find that it dataset is skewed. A comparison with data from the Dutch Statistics about the Dutch population shows an overrepresentation of the higher educated and those with jobs in the professional services in our dataset. We create four different groups of features: Shallow User features, Language Use features, Description features and Text features.

We employ two different approaches to classifying the users in our dataset. The first, named the 'individual' approach, determines the best performing classifier per featuregroup and consequently combines these via a soft-voting ensemble method. The second, named the 'combined' approach, calculates the performance scores for all possible combinations of classifiers and their respective ensemble (also via soft-voting). For this study, we use classifiers with their default settings. We make use of Logistic Regression, Support Vector Machines, Naive Bayes and Random Forest algorithms.

We find that the 'combined' approach outperforms the 'individual' approach by a little, and it outperforms the dummy classifiers. Peak performance is achieved for the occupation predictions at the highest level of the scheme with a score of 0.6224. The skewed dataset does create a consistent error in the predictions and therefore we set up another round of experiments. Here we test the performance of the classifiers per featuregroup with an increasing amount of possible classes. We see a decline in performance scores but promising results as we reach a peak of 0.7569 for education level prediction.

Further studies may focus on improving the dataset in size and/or distribution over classes, inclusion of more data (such as the network of users) and tuning the parameters of the classification algorithms.

CONTENTS

CHAPTER 1. MOTIVATION

1.1 MOTIVATION

Historically, statistic agencies have been analyzing (inter)national statistics from varying perspectives. These range from macro-economic indices, to economic growth and household incomes and spending. As the data for these studies is not always readily available, statistic agencies have to go to lengths to acquire data. This typically includes surveys on a large scale. Online Social Networks (OSN) such as Twitter and Facebook have been a valuable new source of data. Their main advantages being the easier, faster and cheaper creation of more extensive datasets.

Twitter has a lot of potential for the creation of insightful statistics. The platform enables users to post tweets, micro-messages of up to 140 characters, which will be shown to their followers. The users tend to speak their mind and this has proven to be a valuable source of information for many organizations. The unfiltered opinions of customers, reviews of products, values and beliefs of users, are often expressed explicitly in tweets.

However, studies about the Twitter users remain just that; about the Twitter users. Some studies shed light on demographics of the Twitter users in terms of age, gender and location, but it is not clear what the Socio Economic Status of these users is. This study aims to bring more light to this question.

1.2 PROBLEM STATEMENT

By knowing the SES of Twitter users, a better understanding of the OSN can be established. For our study, we examine Dutch Twitter users. Unfortunately, studying individual users by hand is not an option due to the vast volumes, as there are an estimated 2.5 million Dutch Twitter users (drs. Neil van der Veer, 2016). In order to overcome this issue, we can make use of machine learning techniques. With these techniques we can create automated predictions based on a sample of Dutch speaking Twitter users. The sample requires extensive processing in order to be used for this study. With the use of this set we can train the algorithms to classify users into different categories. The more examples (i.e. users) are in our training set, the more opportunities there are to learn for the algorithms. The creation of the right dataset and the training of machine learning algorithms are the main objectives of this project.

1.3 RESEARCH QUESTIONS

In order to achieve the objectives of this research, we aim to answer the following research question:

How can we make machine driven predictions of the socio-economic status of individual Dutch Twitter *users?*

As this is question requires an extensive answer, the following sub questions are used in order to establish an answer to the main research question:

SQ1. What data do we need to analyze and predict socio economic classes?
Answered in Chapter 3
SQ2. What can we conclude when we compare our dataset to national statistics?
Answered in Chapter 3
SQ3. With which features can we predict a socio economic status for Twitter users?
Answered in Chapter 4

SQ4. What algorithms should we use, and how do we measure their performance?
Answered in Chapter 5

SQ5. What are the outcomes of the experiments and what do these outcomes mean?
Answered in Chapter 5

1.4 APPROACH TO ACHIEVE THE OBJECTIVE

The study starts with a literature review into related studies. Then we start with the creation and processing of our own dataset, based on a national Twitter archive. The dataset will be used to train the algorithms by looking at several characteristics (features) of the users. The algorithms learn by associating the scores on these features with certain labels. More users in the dataset provide more learning opportunities for the algorithms. Therefore, we strive for a dataset of substantial size (over 750 users). After training the algorithms, we validate their performance on a separate part of this training set. Based on the performance we make our conclusions.

1.5 SCOPE OF THIS STUDY

This population is limited to the study of Dutch Twitter users, the classification of the SES is limited to the field of education, level of education and occupation. The income, which makes the third pillar of SES, is kept out of the scope, as it cannot be collected without contacting the subjects and it largely depends on the occupation. Furthermore, the study focuses on the application of machine learning algorithms, not the development of these. The studied data comprises the tweet text, some variables concerning the behavior and the profile description text. The technical perspective is focused on model performance, but not on a computational efficiency perspective. Also, hardware related questions are left out of scope.

1.6 RELEVANCE

This study is one in the active field of information retrieval. Studies like these are positioned on a crossroads of multiple disciplines. There are sociolinguistic interests, with closely related studies into opinion mining per social class and studies about voting intention differences (Preotiuc-Pietro, Lampos, & Aletras, 2015). There are studies that examine the inferring of hidden user characteristics (Al Zamal, Liu, & Ruths, 2012; Filho, Borges, Almeida, & Pappa, 2014; Volkova, Bachrach, Armstrong, & Sharma, 2015). The results of these studies are not only of interest to statistics agencies around the world, but also relevant for researchers in the social science domain.

There are also strong commercial interests for applications based on the underlying technology, such as targeted advertising, improved user experience, personalized recommendations of users to follow or user posts to read and the possibility of extracting authoritative users (Pennacchiotti & Popescu, 2011).

1.7 STRUCTURE OF THE REPORT

This report shows the steps that were taken in order to predict the socio economic classes of Dutch twitter users. After a briefing on the used definitions and techniques in Chapter 2, we cover the dataset creation and the handling of the data in Chapter 3. In Chapter 4 we describe the processes of predicting the Socio Economic Status of the Twitter users, followed by the required experiments and their respective outcomes Chapter 5. Chapter 6, the final chapter of this study, concludes the work with a brief examination of the outcomes compared to the national statistics.

CHAPTER 2. BACKGROUND

As the classification of Socio Economic Status (SES) of Dutch Twitter users is no easy task, it is important to establish clear definitions and gain understanding of the used techniques. This chapter starts with an explanation of the used definition of SES and how to determine this, how users are classified and a brief description of the used technologies.

2.1 SOCIO ECONOMIC STATUS

The concept of the SES classification in research goes back to the start of the 20th century (Chapman & Sims, 1925). Based on an older survey with questions such as: “Do you have a telephone?”, “How many years did your father go to school?” and “Do you work at some regular job out of school hours?” the socio economic statuses of the people were determined. In today’s times, there is a commonly accepted understanding is that SES is based on three important factors, being cultural, social and material capital (Bourdieu, 2011; Coleman, 1990). Similarly, it is the common accepted understanding that their main influencers are education, occupation and wealth (Berkel-van Schaik & Tax, 1990; White, 1982; Winkleby, Jatulis, Frank, & Fortmann, 1992).

2.1.1 Cultural capital and education

Cultural capital is comprised of skills, capacities and knowledge. The main influencer on these fields is education. The amount of education one receives positively correlates with the socio-economic status, as shown by a meta-analysis of over a hundred studies in this field (White, 1982). In this study, education is defined by two dimensions: the field of education and the level at which the education was followed.

2.1.2 Social Capital and occupation

Social capital is the social network combined with the status and power of the people in that network. The occupational status is an important influencer for the social capital of individuals. Most connections with regard to status and power are made through professional contacts (Winkleby et al., 1992). In this study we examine the user’s occupational field.

2.1.3 Material Capital and income

Income influences the material capital. However, income is somewhat more disputed, as many state that ‘wealth’ is a more relevant measure. Both the wealth and income are unverifiable with passive data collection techniques, therefore it is kept out of scope for this project.

2.1.4 Classification

The users in our dataset are classified using education and occupation schemes. This will enable the learning algorithms to understand which users have (near) similar jobs for example. As the scope of this research is limited to the Dutch demographic, there is a need for a national applicable classification method. Statistics Netherlands (CBS) has developed the ‘*Standaard Onderwijs Indeling 2006’* (*SOI-2006*). This classification scheme is based on the level and field of the education. Furthermore, *SOI* comes with the added advantage of being directly linked to the ‘*International Standard Classification of Education’* (*ISCED-11*) developed by UNESCO. Also used by the CBS, is the ‘*International Standard Classification of Occupations’* (*ISCO-08*) developed by the *International Labor Office*. Like the *ISCED-11*, the *ISC0-08* method is linked into many different national standards around the world. It is a four-level hierarchically structured classification that ends in a total of 436 unit groups. As an example, the classification of a University Teacher is shown in Figure 1.

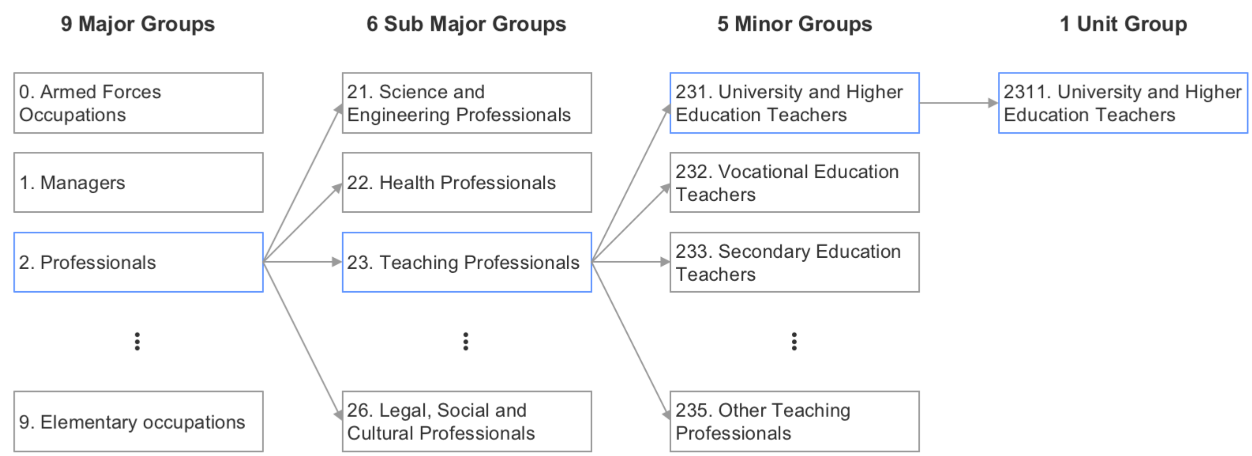

Figure 1 ISCO-08 classification for University and Higher Education Teachers

2.2 MACHINE LEARNING AND CLASSIFICATION

Machine learning is the study of algorithms that can be trained to do a particular task without being explicitly programmed for that task. The two main paradigms of machine learning are supervised and unsupervised learning. The first makes use of labeled examples and tries to calculate the relationships between the examples and their labels, while the latter is concerned with discovering patterns in unlabeled data. This study makes use of supervised learning algorithms.

2.2.1 Supervised learning algorithms

There are various algorithms that can be used in supervised learning settings. The choice of the model depends on the task at hand. In this study, the task concerns classification. We run the algorithm to determine what occupation, education and education-level '*label*' should be given to the users. These predictions are based on their characteristics, also known as '*features*'. The model is able to predict these labels by learning from examples that we provide. We subsequently test the performance of the model on a subset of our data, of which we have predetermined the actual labels. In this study we make use of Logistic Regression, Naive Bayes, Support Vector Machines and Random Forests. In order to give some insights into the workings of such models, the theory behind the Support Vector Machine (SVM) is explained below.

2.2.2 Support Vector Machine

The SVM algorithm is a popular machine learning algorithm, with applications in image analysis, biomedical data analysis and text analysis. In classification tasks, the SVM determines where to 'draw the line' between separate groups. As can be seen in Figure 2, in a setting with two classes and two features this 'decision line' can represented by a line. The instances in the dataset are plotted on two dimensions (X_1 and X_2) as dots. The colors of these dots represent the label these belong to. The vector that separates these groups is represented by the continuous line. The parallel vectors that cross the closest-by instances from either class are represented by the dotted lines. These vectors are called the support vectors.

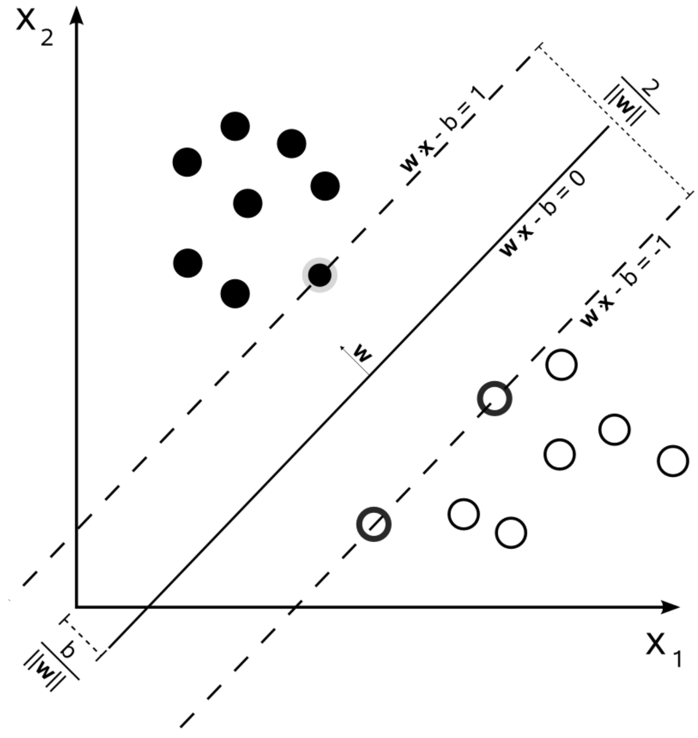

Figure 2 - The Support Vector Machine in a two-dimensional plane trained on data with two classes

The goal of the SVM is to find the settings for the vector where the support vectors lay as far as possible from the main vector. This ideal situation is shown in Figure 2. This is equal to the optimization of the following formula:

$$\left[\frac{1}{2}\sum_{i=1}^{n}\max\left(0,1-y_i(\vec{\omega}\cdot\vec{x_i}-b)\right)\right]+\ \lambda||\vec{\omega}||^2$$

However, in real world applications the data may not be so easily separated. With the use of more advanced mathematics however, we can change the main vector to take other shapes than a straight line, such as curves, or circles. The applications generally have more than two dimensions, so the line changes into a multidimensional hyperplane. There may be more than two classes, which is treated for by constructing multiple hyperplanes. In case the data overlaps, the hyperplane may place instances of different classes on the same side of the plane. These errors can be taken into account by adding a penalty clause in the function.

2.3 RELATED WORK

There are studies about the inferring of user characteristics from all kinds of sources, such as email, forum- or blog posts and those based on OSN. An overview of related work that uses Twitter as a source can be seen in Table 1. The work done by (Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011) concerned the possibilities of studying the Twitter population with predictive models. The labels are based on human annotations of the Twitter profiles. They study the gender and ethnicity of users by examining (last) names. They find that the U.S. based users are predominantly male and discover over and under sampling of ethnical groups. They acknowledge the shortcomings of relying solely on Twitter data, as their labels not validated with the use of a second or different source.

Other work, performed by (Rao, Yarowsky, Shreevats, & Gupta, 2010) focusses on predicting the gender, age, origin and the political orientation of English Twitter users. They used socio-linguistic features and word (combination) matching to help determine differences between male and females, below and over 30 year olds, Northern Indians and Southern Indians and Democrats and Republicans. Predictions were made with the use of an SVM algorithm. They made use of an ensemble technique, called a 'stacked' model, which is essentially a classifier that creates predictions based on the predictions made by the lower level classifiers. This stacked model outperformed basic models both on gender and age prediction.

A study performed by (Sloan, Morgan, Burnap, & Williams, 2015) shows the possibilities of examining the user profile descriptions. Their classification is also based on the ISCO-08 scheme. They examine the performance of algorithms deriving the occupation based on matching the profile description with jobs from the SOC2010 list (which is linked to the ISCO-08 scheme). However, they do not include predictive models in their paper. They classify people in the National Statistics-Socio-Economic-Classification scheme, which has 8 classes based on the occupation of people. They find that their algorithm predicts many users to be in the second class, containing 'Lower managerial, administrative and professional occupations', due to the fact that the job names may overlap with hobbies or interests. This also goes to show that there is a need for an external label.

Work done by (Preotiuc-Pietro et al., 2015) concerns the prediction of income based on Twitter profiles. The income is based on the occupation of the user, which is predicted by the model. The predictions are made with SVM, Gaussian Processes and Logistic Regression models, based on a variety of features, ranging from the number of followers, to the number of tweets displaying anger. They find a list of features that show observable patterns in their relationships to the income of the user. Their work shows the feasibility of our study.

Table 1 - Related work

Subject	Information source	Targets	Algorithms
(Rao et al., 2010)	English tweets	Age, gender, regional origin and political orientation	Lexical feature based and socio-linguistic based models
(Pennacchiotti & Popescu, 2011)	English tweets	Political affiliation, brand affection and ethnicity	Hybrid of text-based, behavior and community based classification
(Mislove et al., 2011)	English tweets (by self-reported US residents)	Gender, location and ethnicity	Text-based classification
(Ikeda, Hattori, Ono, Asoh, & Higashino, 2013)	Japanese tweets	Age, gender, area, hobby, occupation and marital status	Hybrid of text-based and community based
(Nguyen, Gravel, Trieschnigg, & Meder, 2013)	Dutch tweets	Age	Text-based classification
(Siswanto & Khodra, 2013)	Indonesian tweets	Age and occupational status	Text-based classification
(Preotiuc-Pietro et al., 2015)	English Tweets	Occupation	Text-based and user interaction based classification
(Preoţiuc-Pietro, Volkova, Lampos, Bachrach, & Aletras, 2015)	English Tweets	Income	Text-based, user interaction emotional-expression based classification
(Sloan et al., 2015)	English tweets	Age, occupation and social class	Text-based detection (no predictions)

CHAPTER 3. DATASET CREATION

For this study, existing datasets could not be used because datasets containing Personally Identifiable Information (PII) are generally inaccessible by third parties. Besides the restricted access, the demands that this study placed on the dataset rendered pre-existing datasets insufficient. For example, many datasets did not contain information about both the occupation and education of users, or did not use an external source to verify this. The use of an external source can help defeat the challenges of unreliable, out-of-date, incomplete and inaccurate data (Mislove et al., 2011). Therefore, the choice was made to construct a new dataset. Because the dataset would contain PII, we apply a framework in order to scrutinize the ethical value trade-offs that we make, in order to keep from poor decisions made in other studies.

3.1 PRIVACY AND ETHICS

The dataset required for this research contains PII, information which can be used (in combination with other information) to identify a single person, or individual in context. This study is positioned on an intersection of social networks, big data and classification studies. These three aspects require a sense of awareness for the researchers and caution in their approach to ethically responsible research. In this chapter, these concerns are examined and the researchers approach to these concerns is explained.

3.1.1 Concerns

The use of social network data does not go undisputed. While one may wonder why publicly available data is subject to discussion, there are strong arguments to take a step back and consider what data is included in the study and why. For social network data, it is important to realize that 'publicly available' often depends on the person who looks at the data (Zimmer, 2010). For example, one user can see pictures of friends-of-friends on Facebook, while someone who is multiple connections away from these users may not. Another important realization concerns the power imbalance between the researcher and the subject of the study (Boyd, 2011). Researchers have tools and access, while the subjects generally don't. The users place content in spaces that are highly context-sensitive. Often the users are not aware of the researchers that are studying their data, let alone that their data is saved and used in studies years later. The researchers are not in the imagined audience of most users. The last concern that should be raised when using social network data, is the event of a stolen or leaked dataset. It is important to realize that harmful use is not the issue at hand, but the dignity of the user. The focus on dignity recognizes that one does not need to be a personal victim of hacking, or have tangible harm take place, in order for there to be concerns over the privacy of one's personal information (Miller, 2007).

The field of big data studies concerns studies that analyze data with large volumes, high veracity, high variety, high velocity or a combination of these. Regardless of the size of a data set, it is subject to limitation and bias. A large number of studies aims at the collection of as much data as possible to eliminate doubt about validity and bias. However, this is not automatically the case and it can lead to a more complex nature of the data, limiting the researchers understanding of the data. The use of combined (online) sources can lead to a magnification of data errors, gaps and increased unreliability, severely limiting the interpretability of the data. Interpretation is at the center of data analysis. Without those biases and limitations being understood and outlined, misinterpretation is the result (Boyd, 2011).

Based on their approach, classification studies can be divided into two main groups. Approaches based on assumptions of relative uniformity among individuals within a given functional unit (Lawrence, Lorsch, &

Garrison, 1967), and approaches based on communities of practice, which are networks of individuals within or across functional units (which may be grouped together based on commonality of interests, practices and personal associations)(Brown & Duguid, 1991). The use of these two perspectives has enabled researchers to enrich their understanding of the relationships between social actors and technology in varying organizational contexts (Berente, Gal, & Hansen, 2008). However, limiting the classification techniques to functional groups, or communities of practice brings along an ethical risk, as in both of these schemes there are members whose interests, values, or identification align with these neglected issues may be inadvertently marginalized by the research approach.

3.1.2 Solutions

The identified concerns listed above, call for proper actions to make sure that these issues are handled correctly. In order to do so, a framework developed by Dr. A. van Wynsberghe, is used. This framework (Wynsberghe, Been, & Keulen, 2013) can be used in cooperation with an ethicist. In this case Dr. A. van Wynsberghe conducted the required interviews. As an example, the first interview concerned the discussion of the objectives of the project, and how the researcher should approach value trade-offs. The project is subjected to careful consideration and discussion from a normative ethics viewpoint of the ethical value trade-offs. The goal of using this framework is to give insight into value trajectories over the progression of the research, and to improve the design of the tool.

3.1.3 Application of the framework

For this study, the guidelines as proposed by (Wynsberghe et al., 2013) for value analysis have been used. These guidelines helped the researchers recognize and make the intended values of this research explicit. In short, there are five guidelines to help with the value analysis:

1. Make explicit the key actors: direct and indirect subjects, researchers etc.
2. What is the context and what does privacy mean in this context? (Location and data content).
3. Type and method of data collection (passive vs active).
4. Intended use of info and amount of info collected.
5. Value Analysis: making the explicit and scrutinizing intended values of the researchers.

1. Make explicit the key actors: direct and indirect subjects, researchers etc.

For this study, the users of the system will be the researchers at the University of Twente. They are the only ones with access to the data during this research. In later stages, there may be other researchers who use the data for (additional) studies. The subjects are the users in our '*Train/Validation set*'. The indirect subjects are their connections, whose data is also collected, but those users were not subjected to any manual inspection.

2. What is the context and what does privacy mean in this context? (Location and data content).

The data is sampled from a limited national archive of the Twitter online Social Network, and supplemented with specific data from the Twitter API. Twitter is intended as an open network, on which the idea is to share messages with the world (http://twitter.com/tos). The messages are actively stored in the Library of Congress (http://www.loc.gov/today/pr/2013/files/twitter_report_2013jan.pdf) and are available for everyone. The users generally know that their messages are openly visible.

The data has been enriched with information about the (presumably same) users found on the business-oriented Social Networking Site LinkedIn. LinkedIn lets users set the amount of information shown to users who are outside their network, users who are one-step-away in their network and users who are in their network. When connections look at a profile, the user is notified about this. This leads to a higher perceived idea of privacy on this platform than what can be expected of Twitter users.

3. Type and method of data collection (passive vs active).

The data collection for this study has been performed in a passive way. This means that users who are in the dataset are not aware of the fact that they are in the dataset. The source for our dataset saves nearly all of the Twitter data (that is indicated to be Dutch). This is a passive manner of collecting data. The data that was added to the source data was also collected passively, users were not notified that their friend and follower lists were collected. Users that are not in our core sample but are in the network, are also collected in a passive way. When the product of this research is applied to users that are not in our sample, the data collection will also happen passively. So there is no expected active collection in the future.

4. Intended use of info and amount of info collected.

For this research, there are different uses for the data that is collected. One use concerns the creation of features, a sort of user characteristics, ranging from the number of tweets, to measurements of sentiment and word use (as can be seen in 4.2 Features). Another use concerns the constitution of 'labels', or target values that we want the algorithms to predict. Subsequently, the combination of features and labels is used to train an algorithm to learn the relationships between the scores and the occupations, educations and educational level of users.

For the creation of features, the Twitter information of users is used. This concerns a list of a user's tweets in the time period of November 2014 – October 2015, in their original JSON format (an example can be found in Appendix B – Tweet Example). These lists of the users are read by the script and translated into measureable and understandable figures. We received a sample of 5011 users, 989 of which are actively used, and part of the '*Train/Validation set*'.

For the creation of the labels, the users in the '*Train/Validation set*' were subject to manual inspection by the author, as they were matched with LinkedIn profiles. For this step, the URL of their LinkedIn profile was saved, together with the provided descriptions from their most recently completed education and their occupation in the time period. Also, annotations were made of their gender based on profile picture and/or name, the year in which these educations were completed and added was at which institution (based on the school logo) if this was not provided in text. Based on this information the users were manually classified into classes of occupation, education and level of education

During the data acquisition of this study, more data was collected than eventually used. The original scope of the project was broader and had to be scaled down, therefore rendering the network data unnecessary. Also, some of the annotations were not used in this study, such as the gender and year of completion of the study. However, the data was still saved, as it may be used in future studies to achieve the objectives this study first set out with.

The features derived from the Twitter information and the labels derived from the LinkedIn information, are then used to train the algorithms. The scores on the features and the corresponding labels contain relationships that the algorithm tries to learn and simulate. This process happened multiple times for different sets of users.

5. Value Analysis: making the explicit and scrutinizing intended values of the researchers.
The main value for the researchers is to create a better understanding of the Twitter users. The application of the algorithm concerns other values. Applications can help statistics agencies in their analysis of online social networks, allow researchers to perform more extensive analysis on important issues for segments of the population based on education or occupation and may provide relevant insights for social scientists.

The realization of the values listed in the previous paragraph may also lead to the jeopardizing of other values. There are several red flags that arise when studies classify people based on passively collected social media data on such a scale. The use of passively collected social media data rises concerns over privacy values. These kinds of value trade-offs were monitored closely as the dataset collection progressed, and are shown below:

- *Concerns for the use of Twitter data*

As can be read in the terms of service of Twitter, the platforms intent is to share messages between its users. Users of Twitter are generally aware that their tweets are stored in both the Library of Congress. This makes the creation of this dataset acceptable in the opinion of the researchers. When looking at this issue from a consequentialism perspective, there is no reason to keep from using this data, as it is a relatively small sample, a sufficiently random sample and already outdated. Therefore the sample is not of interest to parties who can make their own larger, specified and contemporary datasets. Also, users that make up the '*train/test set*' were not stored separately. Those users are simply selected from the larger random sample when it is needed. The files that contain the information required to create these selections are also not stored in the same location.

- *Concerns for the use of LinkedIn data*

The perceived privacy on LinkedIn is different compared to that on Twitter. This is due to the fact that you must approve people before they can follow you and see all of your information, you get a notification about who viewed your profile and being unable to view full profiles of people that you are not connected with. However, it is no secret that LinkedIn sells the information to jobs agencies, headhunters and corporate recruiters. Since 2010, all the privacy settings are set to public by default (Manzanares-Lopez, Muñoz-Gea, & Malgosa-Sanahuja, 2014). From a consequentialism perspective, there is no significant harm done by collecting this data as it is less complete than the publicly available information. It is also stored in a separate location, so there is no way to link the data to the individuals or the corresponding Twitter accounts.

This expected level of privacy was taken into account when the decision to use the LinkedIn data was made. By visiting the profiles from an anonymous browser window, only the information that the users decided to share publically was examined. Also, in order to minimize the chance of including innocent bystanders, this entire process was performed manually. This means that each user from the eligible set was examined briefly and compared to search results on the LinkedIn website. A match was determined on similar characteristics such as name, profile picture, profile descriptions and location. Only if there was sufficient overlapping information between the profiles, the LinkedIn profile was visited and its information saved.

- *Concerns regarding the combination of this data*

As datasets containing either Twitter users or LinkedIn users are relatively easy to create, the main value for third parties lies in the combination of these datasets. The authors made the concisions decision to keep the data separated throughout the project. The Twitter data remained on a closed server, unfiltered and in its original shape, while the data from LinkedIn was stored on separate files only accessible to one of the researchers. Also, the decision was made that the dataset will be not be released publicly, but may be used for other academic studies.

As the study progressed, some of the annotations were no longer needed (gender, year of finishing education). The information from these annotations was still saved, as it could be used in future studies that focus on the same dataset. This could eliminate the creation of extra datasets in the future. Therefore, the researchers deemed this to be the right thing to do.

- *Concluding on this matter*

In order to answer the question of whether or not we are allowed to access and save the information for our own research purposes, we examine the incentives that we as researchers and our work have. Ultimately, the incentive of this study is to create a better understanding of users on online social networks. We do this by using efficient techniques and only data that is open to the public. By keeping the data from the different sources separated and on restricted servers the possibility of the data ending up in studies with fraudulent incentives in minimized. Even if the data were to be used in a fraudulent way, there would be no harm done. The only data that is created for this study concerns multiple manual annotations in the shape of numbers ranging from one to ten. Therefore we can justify the value trade-off identified earlier in this process, meaning that this study falls within ethical limits.

3.2 ACQUISITION OF ORIGINAL DATASET

The construction of this dataset started with a sample from a Dutch Twitter archive. This archive, named '*Twiqs.nl*' (Twiqs) is constructed by the Dutch eScience Center and hosted on the surfSARA servers for a consortium of Dutch educational institutions. It can be used to search in roughly 40% of the Dutch tweets from 2010 onwards, which amounts to over three billion tweets. Their website employs a search function, which returns only user- and tweet ids. These ids can be examined with the use of the Twitter API. For selected studies, Twiqs.nl is willing to share the data that these ids represent.

As per request, we received the tweets from users in the time window of 1 November 2014 till 31 October 2015 (i.e. one full year). The only restrictions that we placed on these users is that these would be 'active' and not 'overly active' in the aforementioned time window. This means that only users with 100 to 1000 tweets could be selected. This left us with 5011 active users that were indicated to be tweeting in Dutch, dubbed the 'original set'.

However, as Twiqs.nl indicates on their website, the archive does not contain just Dutch tweets. This is due to the three ways Twiqs uses to include a user: based on (rough) geolocation, detected language and Dutch Celebrities. The original set reflected those criteria, containing non-Dutch users, organizations and celebrities. Other issues with this dataset are the amount of anonymous users and users are still in school.

3.3 DATASET FILTERING AND CLEANING

In order to overcome the issues with the dataset, the data was cleaned down in three steps (a graphical representation can be seen in Figure 3. These steps were performed manually to ensure high performance and close monitoring of the filtering in practice. The used protocols can be found in Appendix A – Manual Classification Protocols. The repetitive nature of the annotation work proved to be an error prone process. The step-wise approach did prevent some issues to be resolved, but it does not completely eliminate the possibility of these errors to exist.

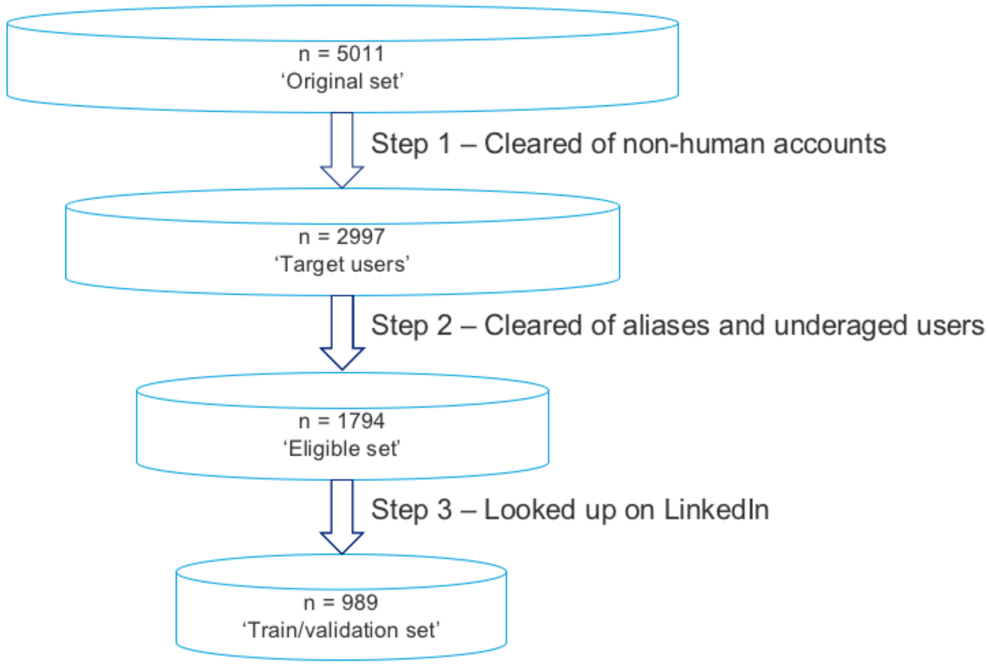

Figure 3 - Dataset refinement funnel

3.3.1 Target users

In our target set, we want to keep users who appear to be individuals tweeting in the Dutch language. So in the first step, the dataset is cleared from non-Dutch speakers, organizations and users tweeting 'spammy' (or automatically generated) messages. This step concerns a brief examination of the ten most recent tweets by users to establish the language used, whether the account was only showing automated messages and a brief examination of username and description to see if it concerned an organization or a human. This resulted in 2997 users in the 'target users' set.

3.3.2 Eligible set

In the second step, the users without a full first and last name were eliminated, as the use of an alias displays the desire to stay anonymous by the user. If the users showed any signs of still attending high school or an age below 18, they were also filtered from the dataset. This was done by examining the name and tweets of the user. This set, dubbed the 'Eligible set' was used as input for scripts that started calling the Twitter API for the lists of friends, followers and mentions of these users. Preferably, this would be done for only the users in our train/validation set, but as the API would only provide the current friend, follower and mention lists, this means that delaying this step would have caused a growing discrepancy between the friends, followers and mentions at the time of tweeting and the time of measuring.

3.3.3 Train/validation set

The '*Eligible set'* contains Dutch speaking users with an estimated age over 18 and a full name. The users in the '*Eligible set'* were looked up on LinkedIn to acquire their self-reported occupation and education. From these 1794 users in this set, 989 could be located on LinkedIn and be assigned at least either their education level, field or occupation on some level. The annotation took place over the course of three weeks, by one annotator. If a user had a name that was listed several times on LinkedIn, the correct user was selected based on profile picture, matching (elements of) descriptions and location. The selection of users of which relevant information could be determined, is named the *'Train/validation set'*.

3.4 DATA ANNOTATION

As can be seen in the information provided by Twiqs.nl has been extended with annotated data and data from the Twitter API. The data that was gathered via the Twitter API consists of lists of friends, followers and mentions, the user information of these profiles and their tweets in the same time window.

The annotated data differs from what can be seen in other studies. It is common to have manually annotated data and use this as a label for the classifier, but it is often based on the Twitter profile itself. The inclusion of data from another source (in this case LinkedIn) has enabled us to study new cases, as previous research has discarded users who did not list their profession in their description. With the approach used in this research, these are still eligible to be used as training examples for the algorithm. Previous studies have acknowledged the shortcomings of blindly accepting the user's self–reported data, but argued that this is a common problem in offline methods as well. The use of LinkedIn data is our attempt to avoid these shortcomings. The LinkedIn profiles created by the users are visible to everyone who is in their professional network, which means that all the data on LinkedIn is under some form of social scrutiny. Therefore, the information provided on LinkedIn is held to a higher truth standard than data provided on Twitter (Manzanares-Lopez et al., 2014).

The information that was gathered via the LinkedIn annotation sessions was limited to the URL of the public profile, the textual description made by the user regarding the highest level of finished education and the textual description of their occupation in the timeframe of November 2014 – October 2015. The education and occupation were subsequently translated into classes of the SOI and ISCO-08 classification schemes as mentioned in Chapter 2. Translating the self-reported occupation and education in to the lowest of the schemes was not always possible and were then limited to a higher level of the classification scheme. Due to the time constraints of this study, the network data is not used for any analysis in this study, but may be used in the future. The data that was used in this study can be seen in Figure 4 in the navy-blue (middle) section.

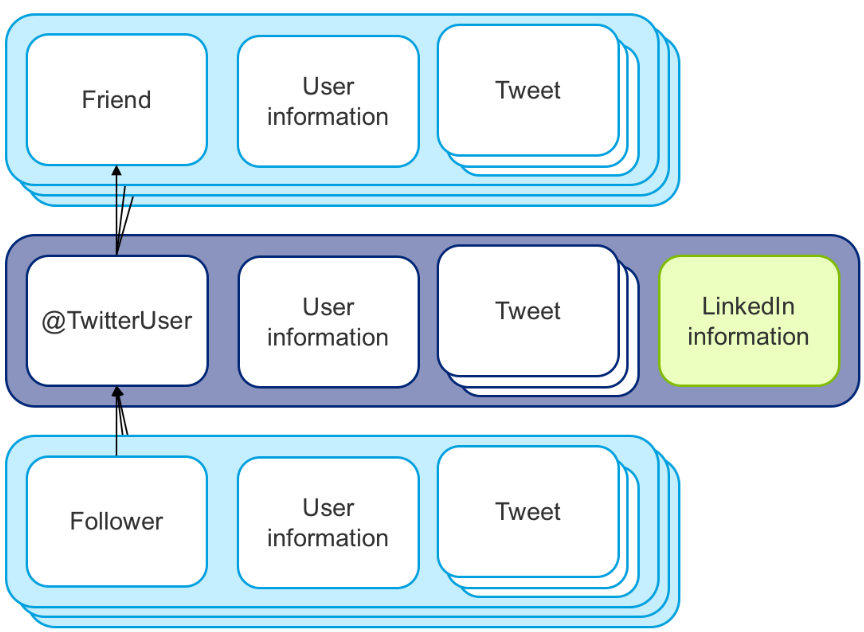

Figure 4 - Dataset overview. Navy-blue data was supplied by Twiqs.nl, light-blue data was gathered via the API and green data is manually annotated

3.5 DATASET CHARACTERISTICS

The users that are in the *'Original set'*, were randomly selected. The users in the *'Train/validation set'* are active Twitter users, tweeting in Dutch, not sending out 'spammy' messages, over 18 years old and could be identified on LinkedIn. The effects of these selection criteria need to be examined by exploring the data that was selected to be in the *'Train/validation set'*.

3.5.1 Characteristics of the users in the 'Train/validation set'

By examining the general characteristics of the users in the *'Train/validation set'*, we gain insights into the randomness of the sample that will be used to train our algorithms. The following figures provide insights in the size and content of the data. As shown in Figure 5, the average number of friends is a little higher than the amount of followers. About twenty users follow more than 2000 accounts. The maximum was a little over 3600. There were two major outliers for the amount of followers, as there was one account with over 45.000 followers, and one with more than a 100.000 followers.

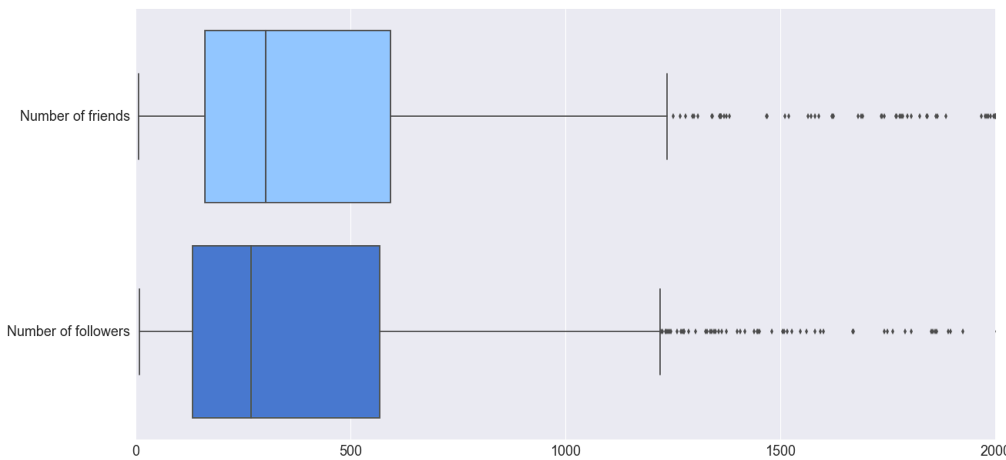

Figure 5 - Number of friends and followers per user

During the manual annotation process, the gender of the users was also determined. The can be seen in Figure 6. The percentages are 62% for males and 38% for females. This is in line with other studies that discovered a higher level of male users (Mislove, 2011).

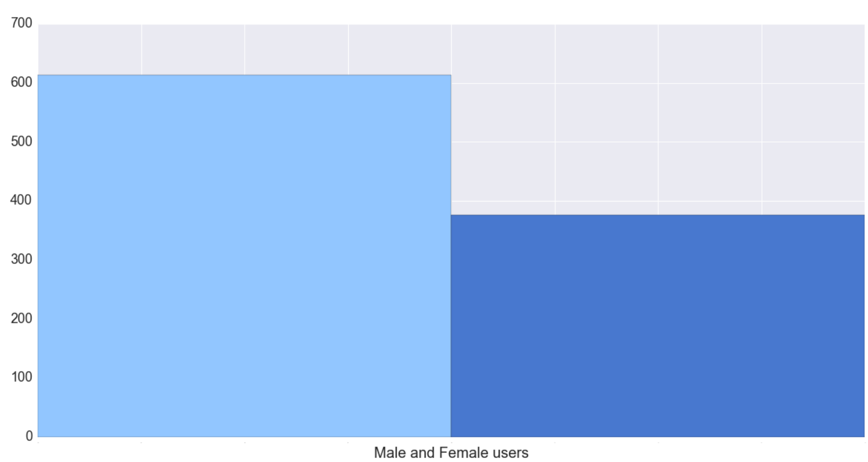

Figure 6 - Amounts of male and female users

If we take a look at the contents of the collected data, we can see what the most frequently used words are in the tweets and the profile descriptions. These show some expected differences, as the words in the profile descriptions (Figure 7Figure 7 - 100 Most frequently used words in profile descriptions) seem more descriptive of the user, than the words that users employ in the tweets (Figure 8). For example, we find the Dutch words for mother ('moeder') and father ('vader') in Figure 7. Also, we can see highly valuable information for the discovery of occupations in the profile descriptions, as some occupations are apparently very frequent ('consultant', 'journalist' and 'manager').

Figure 7 - 100 Most frequently used words in profile descriptions

Figure 8 - 100 Most frequently used words in tweets

3.5.2 Labels of the users in the 'Train/validation set'

When we examine the labels that are annotated for this dataset, we see that the dataset contains users from all main categories on educational field, educational level and occupational field (except for armed forces). However, is must be noted that the distribution occupation is rather skewed, as can be seen in Table 2. Top ten educations and occupations can be seen in Table 3.

Table 2 - Dataset details

Educational field (Edu1)	Amount	%	Educational level	Amount	%	Occupational field (Occ1)	Amount	%
Basic education	28	3%	PHD, PostDocs	23	2%	Armed forces	0	0%
Teaching	33	3%	Masters, MBA's	235	24%	Managers	244	25%
Humanistic	250	25%	Bachelors	429	43%	Professionals	535	54%
Economics	187	19%	Community College	95	10%	Technicians and associate professionals	114	12%
Jurisdiction	84	8%	High school	8	1%	Clerical support workers	30	3%
Mathematics	35	4%	Lower Education	0	0%	Service and sales workers	27	3%
Technology	52	5%				Skilled agricultural, forestry and fishery workers	7	1%
Agrarian	10	1%				Craft and related trades workers	12	1%
Health	73	7%				Plant and machine operators, and assemblers	5	1%
Hospitality	27	3%				Elementary occupations	3	0%
Unclear	210	21%	Unclear	199	20%	Unclear	12	1%
Total	989		Total	989		Total	989	

Table 3 - Top 10 Categories of Education and Jobs

Categories of education (Edu4)	Amount	%	Occupation (Occ4)	Amount	%
(Dutch) Law	39	5,01	Journalists	54	5,53
Business Economics	37	4,75	Public relations professionals	41	4,11
Journalism	32	4,11	Sales and marketing managers	37	3,79
Communication	31	3,98	Senior government officials	36	3,61
Marketing, Commercial Economics	30	3,85	Advertising and marketing professionals	34	3,41
Public Administration	20	2,57	Management and organization analysts	30	3,01
Management studies	16	2,05	Authors and related writers	29	2,91
Human Resources	16	2,05	Policy administration professionals	21	2,11
Study of Education	16	2,05	Announcers on radio, television and other media	21	2,11
Economics	15	1,93	Social work and counselling professionals	19	1,91

Because we used the standard classification schemes, we can compare the labels of the users in our dataset with the statistics about the Dutch population. However, the field of education could not be established for the Dutch population from the figures of Statistics Netherlands. Their reports are focused on annual enrollment and graduation of educational programs and these do not show the amount of people who followed an education in the field of Law for example. The level of the education could be determined though, as there is data available on the highest level of education per person, as can be seen in Figure 9. We did have to merge the 'PhD' and 'Masters' classes, as these numbers were not separately available. Overall, we see a strong overrepresentation of the higher educated. This is in with our expectations, as our sample is based on users who are also active on LinkedIn. The latter used for professional networking, which is valued more among higher positions.

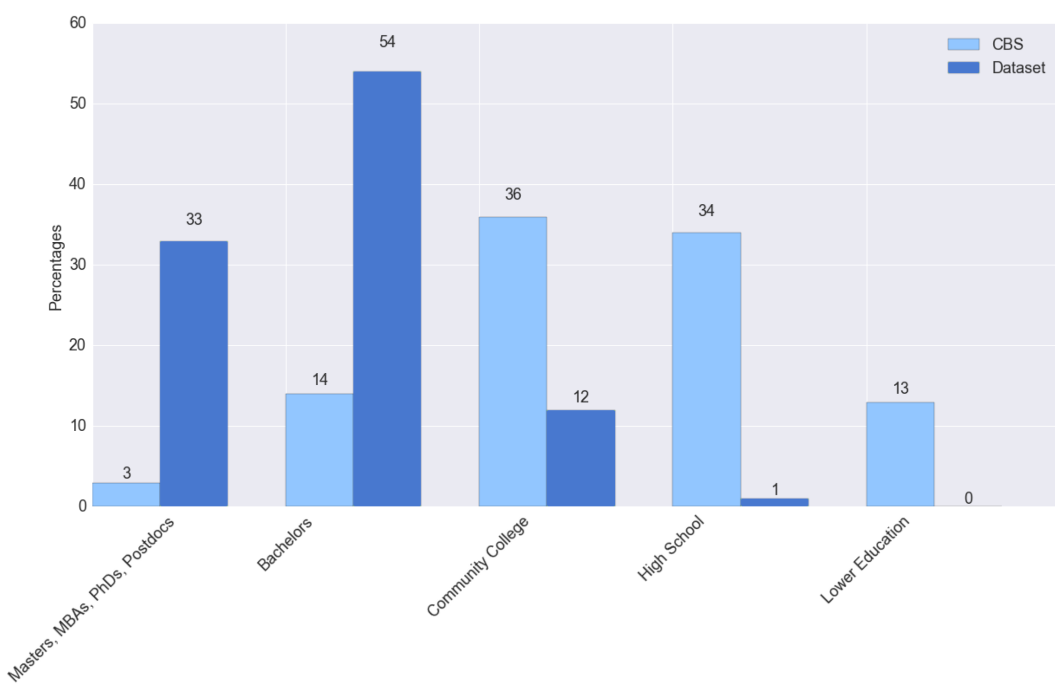

Figure 9 - Side by side comparison of the Dutch population and the users in our dataset for the level of Education

The classification scheme for the occupation of the users also allows for comparisons with the Dutch population, as shown in Figure 10. In this case, we also see a strong overrepresentations that may be caused by the inclusion of LinkedIn as an external source. Professionals are those who require professional networks for their careers, while people working as machine operators do not. The same goes for people in the Clerical support workers, Service and sales, Agricultural, Trades and Elementary classes.

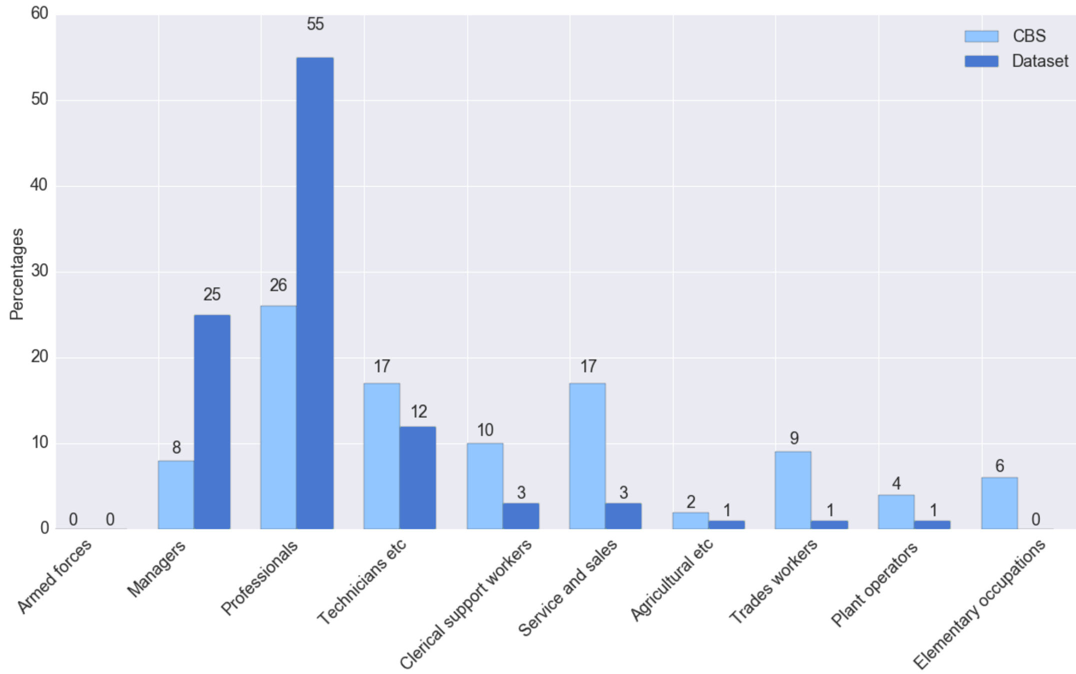

Figure 10 - Side by side comparison of the Dutch population and the users in our dataset for the Occupational class

CHAPTER 4. SOCIO-ECONOMIC STATUS PREDICTION

In order to extract meaningful insights from the data, the filtered data from the '*Train/test* set' has to be preprocessed before it can be used by machine learning algorithms. After the preprocessing, a set of features is extracted from the data. This Chapter describes these preprocessing steps, feature extraction and the subsequently combination into one larger algorithm.

4.1 PREPROCESSING

There are two kinds of data used in this project; structured data, and unstructured data. The structured data does not require any preprocessing, as its values are directly translatable into understandable metrics. For unstructured data, this is not case. The unstructured data in this project are the tweets of the users and their profile descriptions. The unstructured data requires preprocessing, in this study this concerns stemming, stop word removal and tokenization.

4.1.1 Tokenization

Fundamentally, computers do not understand language, but mathematics. This means that the text needs to be transformed to a numerical format, that the computer can use and we can translate into insights as the analysis is done. The values in this format will be used to calculate word occurrence and other aspects that affect the word importance. The texts are still one complete string at this point. We cut these strings into separate 'words' with the use of tokenization. The tokenization algorithm uses regular expressions to understand how to handle emoticons, hashtags, URLs and substrings comprised of numbers. After tokenization the texts are no longer strings, but lists of words.

4.1.2 Stemming

Different styles of writing and spelling words may lead to slightly different versions of the same word. In our analysis we want these words to be considered as the same. That's why we make use of a stemming algorithm. It is based on Porters Snowball Stemming algorithm, for which a separate Dutch branch is available. Stemming a sentence reduces the words in this sentence to their stem. Note that these stems may appear distorted to humans, to a computer model this makes no difference. The accuracy of the predictions will improve as long as this mapping is applied consistently (Porter, 1980). As an example: '*twijfels'*, '*twijfeling'* become '*twijfel'*.

4.1.3 Stop word removal

There are many words that are not directly relevant for our analysis. The use of 'stop words' merely add noise the use of actually interesting words for our analysis. These words do not contain any relevant information or are omnipresent. Due to the high level of English words used on Twitter, both Dutch and English stop words were removed. Examples are: '*het*, '*die*, '*al'*, '*omdat',* '*nog*', 'the', 'your', etc. But in order to be able to handle negated sentences, words such as '*niet'*, '*geen'*, '*not'*, '*none'* were kept in the sentences. Also, English stop words that are part of the Dutch language were kept in (i.e. '*haven'*, '*been'*).

4.2 FEATURES

In this research, several sorts of features are created. These are grouped into four main categories: 'user shallow features', 'language use features', 'description features' and 'text features'. A full overview can be seen in Table 4. The shallow user features and the language use features are based on structured data. The extraction of the content features requires data preprocessing as described in the previous section.

Table 4 - Feature table

id	Feature Group
	Shallow user features
F1	Number of Tweets
F2	Number of Followers
F3	Number of Friends
F4	Follower/Friend ratio
F5	Number of Times listed
F6	Amount of favorites
F7	Average number of tweets per day
F8	Average amount of retweets per tweet
F9	Percentage of tweets that are retweets
F10	Percentage of tweets that are direct replies (start with @)
	Language Use features
F11	Average number of mentions per tweet
F12	Average number of links per tweet
F13	Average number of hashtags per tweet
F14	Percentage of tweets with positive sentiment
F15	Percentage of tweets with neutral sentiment
F16	Percentage of tweets with negative sentiment
F17	Subjectivity of tweets
	Description features
F18	Bag-of-Word based on the profile description
	Text features
F19	Bag-of-Words based on the tweets

4.2.1 Extraction of shallow user features

The extracting of these features requires the analysis of the Twitter data provided by Twiqs.nl. Each tweet also includes the information about the user and its profile (for an example of a tweet and its full information, see Appendix B – Tweet Example). Features F1, F2, F3, F5 and F6 could be directly derived found in the data. The features of F4, F7 and F8 were determined with some basic operations on the structured data. The final two features in this group, F9 and F10, were determined by examining the first few characters of each tweet.

4.2.2 Extraction of language use features

The extraction of the language use features required some basic operations for F11, F12 and F13, as this information was available in the structured data. The F14 score is calculated by looking at the amount of hash tagged characters in a tweet if the tweet contains a hashtag Features F14-F18 were calculated with the help of the CLiPS pattern-nl sentiment analysis plugin. The plugin analyses the provided sentence and returns polarity and subjectivity values for each word (combination) that it recognizes (De Smedt & Daelemans, 2012). These scores are combined into a polarity and subjectivity score per tweet. The tweet is marked positive if the polarity score is higher than 0.2 and negative if the polarity score is below -0.2. Anything in between is marked as neutral.

4.2.3 Extraction of description features

The profile descriptions of Twitter users can contain a lot of information. In order to include this information in this analysis, we create a Bag-of-Words for these descriptions. In short, this means that we create a row vector with dimensions of (1, 5000). This means that we can represent 5000 words with columns. For reach profile description that is added to the bag-of-words, we determine how often the words are present. If the total amount of different words exceeds 5000, only the 5000 most frequent are used. With the use of this table and the labels, the algorithm can learn how the choice of words in the profile description correlates to the occupation, educational level and field of education of the user. An example of this approach is shown in Table 5.

Table 5 - Bag of Words approach for profile descriptions

Word vector:	**1**	**2**	**3**	**4**	**5**	**6**	**…**	**4999**	**5000**
Word that is represented	best	day	example	firm	simple	show		this	works
Example profile descriptions									
'This is a simple example to show how this works'	0	0	1	0	1	1	…	2	1
'Every day is my favorite day!'	0	2	0	0	0	0	…	0	0
'Trying my best, working at the coolest firm in the world!'	1	0	0	1	0	0	…	0	0

4.2.4 Extraction of text features

In order to analyze the discussed topics by users for F19, the contents of the tweets are analyzed. The tweets are grouped together per user and translated to numbers with the Bag-of-Words approach. This approach enables the algorithm to see what words are used by journalists, or users with an IT education, or users with a relatively high level of education.

4.3 COHERENT MODEL

We use the groups of features mentioned in the previous paragraph to create our predictions. In order to combine these groups of features, we take two approaches. The first approach is the '*Individual*' approach, in which we choose the optimal performing algorithm per feature group. An example of this approach can be seen in Figure 11. We train the classifier on 90% of the dataset, and test it on the remaining 10%. Based on the performance scores we can determine per label which classifier provides the best predictions per group of features. As an example: the best performance for Occ1 on the Shallow User Features may be achieved by a Support Vector Machine, while the other feature groups may require Naive Bayes, Logistic Regression or Random Forest models. These predictions are subsequently combined with an unweighted mean and its performance is measured in the same way as the second approach.

The second approach, '*Combined'* approach is based on the prediction probabilities. The algorithms provide probabilities that a user belongs to a certain class. The sum of all the predictions for one user is equal to one, as each user is presumed to belong to one of these classes. With the use of the four groups of features, we create four different tables with probabilities. An aggregate table is build, based on the unweighted average of these four tables. These calculations are performed for ten percent of the *'Train/validation set'* population. The other ninety percent is used to train the algorithm. A graphical representation for the occupational class at the ISCO-08 minor group (3 digits) is given in Figure 12.

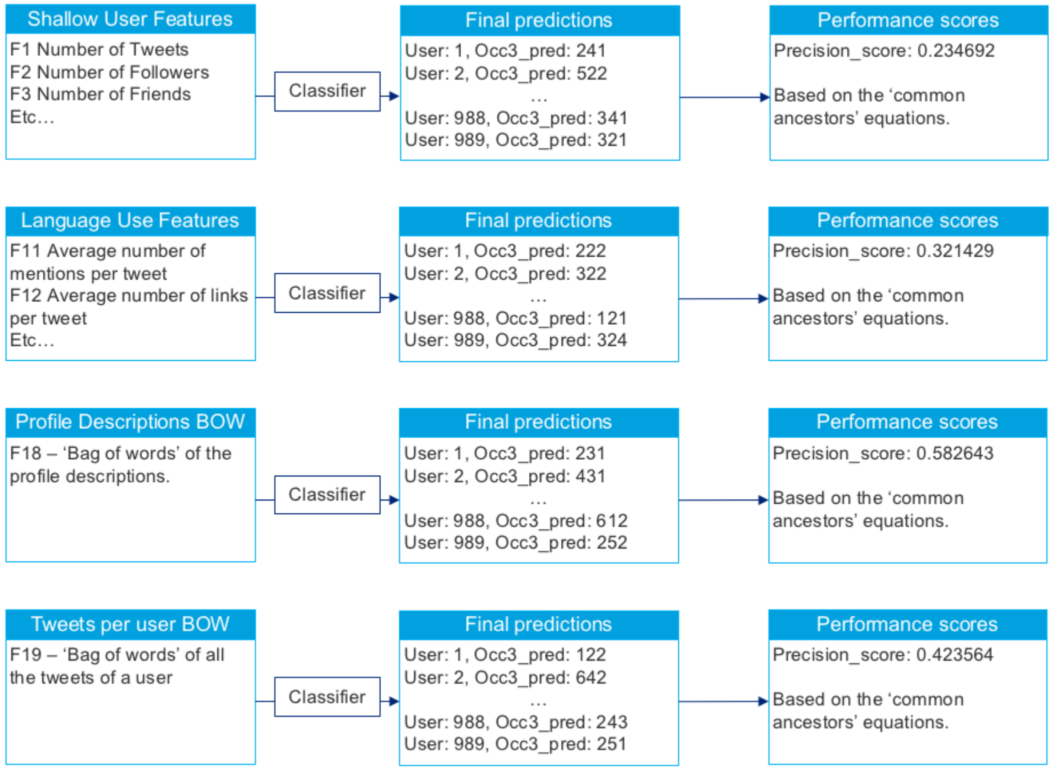

Figure 11 - Representation of the 'Individual' approach, with example values for the ISCO-08 Occupation3 label

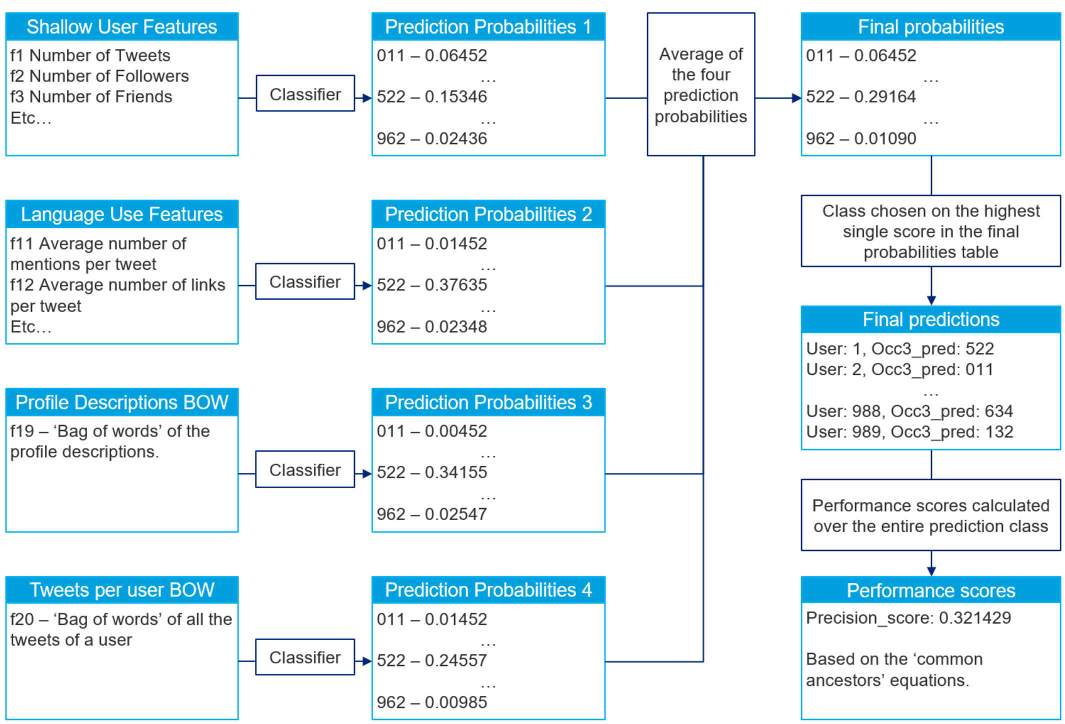

Figure 12 - Representation of the 'Combined' approach, with example values for the ISCO-08 Occupation3 label

CHAPTER 5. EXPERIMENTS

In order to test the setups described in the previous chapter, a set of experiments is designed to test the performance of the model. Based on the experiments conducted in this chapter, we can see how well the model performs and which configuration performs best. This chapter describes the set-up, the performance metrics, expected outcomes, the results and provides a discussion.

5.1 SETUP

All the code for this project is written in Python 2.7, a high-level dynamic programming language. The data is stored on a restricted server of the University of Twente. The scripts are designed to be run on the same location as the data, eliminating the need to copy the data. As stated in 4.3 Coherent model, we use two approaches. Therefore, we created two separate scripts. Both start with the creation of the feature tables. These are placed in Python Pandas DataFrames (McKinney, 2010), which allow for easy handling of data. While building these tables every time we run the script is inefficient, it is in line with the ethical guidelines we imposed on ourselves, to create no derived versions of the data.

For the '*Individual'* approach, each feature table is split up in randomly chosen test and train part. The same goes for the respective labels. Subsequently, different classifiers are fitted and asked to predict. With the use of the Sci-kit learn module for python (Pedregosa et al., 2011), we are able to run Logistic Regression, Support Vector Machine, Naive Bayes and Random Forest models after creating the train and test sections. The predictions are then scored and based on these scores the best performing setup per label is determined. This setup is then created, merging the prediction probabilities created by the classifiers by taking the mean. These are then examined as if they are the final predictions, and are subsequently scored. Merging the prediction probabilities by taking the mean is also known as an ensemble method with soft-voting.

The '*Combined'* approach continues after the feature table creation by running all possible combinations of classifiers. As there are four feature tables, which are subject to one of four classifiers, which need to be ran for nine labels, this amounts to a total of 2304 configurations ($4^4 \cdot 9$). We combine the prediction probabilities of each classifier by taking the mean of these four scores and subsequently check the performance of the configuration.

Besides these approaches, we test the performance of the model versus the performance of three *'dummy classifiers'*. The first dummy generates predictions with uniform probability. The second dummy generates predictions in the same proportions as the labels of the test set. The third dummy classifier picks the most common value in the labels of the test set, and predicts that for all the instances.

5.2 PERFORMANCE

As every experiment will end with a prediction for users in the test set, it is important to determine how the performance of this set of predictions is measured. When working with a classifier, there are four archetypes of outcomes to the predictions that the model. These types are related to the predicted value and the actual value of the instance that it classifies. An example of these can be seen in Figure 13. If the predicted and actual class are the same, the instances are dubbed a True Positive (TP). If an instance is labelled as A, while it should have been B (in Figure 13 this is E_{AB}), it is called a False Positive (FP), or type 1 error for A. If an instance is labelled as B, while it should have been A (E_{BA}), it is named a False Negative

(FN) or type 2 error for A. There are also instances labelled as True Negatives (TN), which is equal to the sum of all columns and rows, excluding the class's column and row.

		PREDICTED		
		A	B	C
ACTUAL	A	TP_A	E_{AB}	E_{AC}
	B	E_{BA}	TP_B	E_{BC}
	C	E_{CA}	E_{CB}	TP_C

Figure 13 - Outcome examples for multiclass classification

Based on these TP, TN, FP, and FN numbers, performance measures can be made. For this study, we measure accuracy, precision and recall. Accuracy represents the total correct answers relative to the total set. Precision concerns the percentage of 'positive predictions' were correct, while recall is the amount of positives actually labelled as positive. Precision and recall differ based on the amount of type 1 or type 2 errors. In this study, there is no type of error that is preferable over the other. Therefore, the accuracy is the performance metric of choice. However, accuracy is not fit for all the labels in this study, as we also require performance indicators that incorporate the hierarchical nature of the classification task that is being studied in this research (Sokolova & Lapalme, 2009).

The measures that are specifically designed for multiclass hierarchical problems are able to value errors differently, based on the level where the mistake is made. For example, the prediction that a user belongs to occupation class *2352 - Special needs teachers*, while it should be occupation *2351 - Education methods specialists*, is not as bad as guessing occupation *4221 - Travel consultants and clerks*. There are measures for recall and precision. These measures combined can be used to calculate the F-Measure (also known as F1 score or F-Score), which combines the precision and recall scores to measure the test's accuracy.

As indicated by (Costa, Lorena, Carvalho, & Freitas, 2007) the precision and recall can be calculated via the ancestors of the predicted classes. The problem with descendant measurements is that it assumes that the predicted class is either a subclass or a superclass of the true class. Therefore, the choice was made to work with common ancestors via the following equations (Kiritchenko, Famili, Matwin, & Nock, 2006). In this study, predictions (C_p) are always made on the same 'level' as the label (C_t), resulting in identical scores for recall and precision (and F-Measure if β is left at its default value of 1). For all predictions made in this study where we encounter a hierarchical structure, we use the Precision hP as the performance indicator.

$$hP = \frac{|\, Ancestor(C_p) \cap Ancestor(C_t) \,|}{|\, Ancestor(C_p) \,|} \qquad hR = \frac{|\, Ancestor(C_p) \cap Ancestor(C_t) \,|}{|\, Ancestor(C_t) \,|}$$

$$F - measure = \frac{(\beta^2 + 1) * hP * hR}{\beta^2 * hP + hR}$$

5.3 RESULTS AND DISCUSSION

As can be seen in Table 6, we collected performance scores for the two main approaches and the dummy classifiers. Overall, the '*combined'* model showed the best performance figures. The dummy classifiers don't show interesting performance, except for the Most-Frequent model. However, it is important to keep in mind that the dataset is heavily skewed and therefore brings an advantage to this dummy classifier. Also, the scores show an expected gradual decline as more levels are added to the classification schemes (with an exception for Education-3). The '*individual*' model also outperforms the dummies, albeit by a smaller margin. When looking at the configurations that performed best, there are little similarities or consistencies to be found.

Table 6 – Full model performance scores

Label: / Performance measure:	Education-1 Accuracy	Education-2 Precision	Education-3 Precision	Education-4 Precision	Education Level Accuracy	Occupation-1 Accuracy	Occupation-2 Precision	Occupation-3 Precision	Occupation-4 Precision
'Individual' Approach	0.3421	0.1429	0.2697	0.1438	0.5949	0.5918	0.3980	0.3401	0.2934
Features:									
Shallow User	Logistic Reg.	Ran. Forest	Logistic Reg.	Ran. Forest	Logistic Reg.	Logistic Reg.	Logistic Reg.	SVM	SVM
Language Use	Logistic Reg.	Ran. Forest	Logistic Reg.	Naive Bayes	Logistic Reg.	Logistic Reg.	Logistic Reg.	SVM	SVM
Description	Logistic Reg.	Ran. Forest	Logistic Reg.	Ran. Forest	Logistic Reg.	Logistic Reg.	Naive Bayes	SVM	SVM
Text	Logistic Reg.	Ran. Forest	Ran. Forest	Ran. Forest	Logistic Reg.	Logistic Reg.	Logistic Reg.	SVM	SVM
'Combined' Approach	**0.3684**	**0.2532**	**0.3224**	**0.1575**	**0.6076**	**0.6224**	**0.4388**	**0.3878**	**0.2934**
Features:									
Shallow User	Logistic Reg.	Ran. Forest	Logistic Reg.	SVM	Ran. Forest	Ran. Forest	Ran. Forest	Ran. Forest	SVM
Language Use	SVM.	Naive Bayes	Logistic Reg.	Ran. Forest	SVM	Ran. Forest	Logistic Reg.	SVM.	SVM
Description	Naive Bayes	SVM	Ran. Forest	Ran. Forest	Logistic Reg.	Naive Bayes	Ran. Forest	Logistic Reg.	SVM
Text	Ran. Forest	SVM	SVM	Ran. Forest	Ran. Forest	Ran. Forest	Logistic Reg.	Ran. Forest	SVM
Dummy's:									
D- Uniform	0.0921	0.1039	0.1009	0.0240	0.2532	0.1122	0.1633	0.2007	0.1250
D- Stratified	0.2368	0.0844	0.0877	0.0788	0.4177	0.0486	0.2500	0.1905	0.1888
D - Most Frequent	0.3026	0.1818	0.1886	0.1438	0.5316	0.5510	0.3673	0.3129	0.2551

When examining the actual predictions of the algorithms, we immediately noticed that even the best performing algorithm (Occupation 1, setup: Random Forest on Shallow User features, Random Forest on Language Use features, Naive Bayes on Description features and Random Forest on Text analytics) with an accuracy score of 0.6224 has big issues dealing with the skewed data, as can be seen in Table 8. The type 1 error predictions for class 2 (users incorrectly identified as professionals) are the main issue. Also, the test set ended up with only 7 of the 9 classes (see also Table 2). When looking at a less skewed label set (Educatio-1, setup: Logistic Regression on Shallow User features, Support Vector Machine on Language Use features, Naive Bayes on Description features and Random Forest on Text analytics) with an accuracy score of 0.3684 in Table 7, we see similar issues concerning the type 1 errors for the most frequent class and missing one class. The type 1 errors may be due to the fact that there are not sufficiently distinguishing features in the current setup. The missing classes are a direct effect of the skewed dataset, as some classes contain less than 10 users and are therefore not in the test set.

Table 7 - Confusion Matrix for the Combined approach for Education-1, setup: Random Forest, Support Vector Machine, Logistic Regression, Random Forest

Predicted: Actual:	1	2	3	4	5	6	7	8	9
1	0	3	0	0	0	0	0	0	0
2	0	25	1	0	0	0	0	0	0
3	0	13	2	0	0	0	0	0	0
4	0	6	2	0	0	0	0	0	0
5	0	2	0	0	0	0	0	0	0
6	0	5	2	0	0	0	0	0	0
7	0	0	2	0	0	0	0	0	0
8	0	9	2	0	0	0	0	0	0
9	0	2	0	0	0	0	0	0	0

Table 8 - Confusion Matrix for the Combined approach for Occupation1, setup: Random Forest, Random Forest, Naive Bayes, Random Forest

Predicted: Actual:	1	2	3	4	5	6	7
1	4	15	0	0	0	0	0
2	1	57	0	0	0	0	0
3	0	15	0	0	0	0	0
4	0	3	0	0	0	0	0
5	0	1	0	0	0	0	0
6	0	1	0	0	0	0	0
7	0	1	0	0	0	0	0

5.4 ADDITIONAL ANALYSIS

As the previous experiments have shown, the skewed data heavily influences the predictions in a negative way. Therefore we designed an extra set of experiments. In this set we examine the performance of different classifiers as we add more classes to the prediction task. We start by having the model predict the label for only two classes. In the case of occupation, this would be predicting if a user belongs to the '*Professionals'* class (54% of the dataset) or to the class 'other'. Then we add another class, '*Managers*' (25%), followed by *'Technicians and associate professionals'* (12%). These experiments are conducted for the same classifiers (Logistic Regression, Support Vector Machines, Naive Bayes and Random Forest) and the three main labels (Occupation 1, Education 1 and Education-level). We use ten-fold cross validation with a 90% training and 10% testing ratio. The results of these experiments can be seen in Tables 9-11.

Table 9 - Classifier performance based on accuracy on single feature groups with increasing difficulty for Education-1

Education 1

Classes: 2 — Humanistic, other

Featuregroup	**Logistic Regression**	**SVM**	**Naive Bayes**	**Random Forrest**
F1 - Shallow User Features	0,2810	0,6048	0,6996	0,7569
F2 - Language Use Features	0,2810	0,6048	0,6889	0,7222
F3 - Description Features	0,2810	0,6048	0,6889	0,7190
F4 - Text Features	0,2810	0,6048	0,6889	0,6558

Classes: 3 — Humanistic, Economics, other

Featuregroup	**Logistic Regression**	**SVM**	**Naive Bayes**	**Random Forrest**
F1 - Shallow User Features	0,3738	0,4246	0,4159	0,4683
F2 - Language Use Features	0,3738	0,4246	0,3266	0,3435
F3 - Description Features	0,3738	0,4246	0,3640	0,4603
F4 - Text Features	0,3738	0,4246	0,4681	0,4510

Classes: 4 — Humanistic, Economics, Jurisdiction, other

Featuregroup	**Logistic Regression**	**SVM**	**Naive Bayes**	**Random Forrest**
F1 - Shallow User Features	0,2814	0,2714	0,3586	0,3307
F2 - Language Use Features	0,2814	0,2714	0,3150	0,3032
F3 - Description Features	0,2814	0,2714	0,3318	0,3792
F4 - Text Features	0,2814	0,2714	0,4239	0,3300

Table 10 – Classifier performance based on accuracy on single feature groups with increasing difficulty for Education level

Education - Level

Classes: 2 Bachelors, other

Featuregroup	**Logistic Regression**	**SVM**	**Naive Bayes**	**Random Forrest**
F1 - Shallow User Features	0,5325	0,5125	0,5536	0,5099
F2 - Language Use Features	0,5325	0,5125	0,5129	0,5474
F3 - Description Features	0,5325	0,5125	0,5133	0,5450
F4 - Text Features	0,5325	0,5125	0,5615	0,4825

Classes: 3 Bachelors, Masters, other

Featuregroup	**Logistic Regression**	**SVM**	**Naive Bayes**	**Random Forrest**
F1 - Shallow User Features	0,5119	0,3198	0,4348	0,5341
F2 - Language Use Features	0,5659	0,3198	0,2251	0,4541
F3 - Description Features	0,5659	0,3198	0,5156	0,4344
F4 - Text Features	0,5659	0,3198	0,3819	0,4801

Classes: 4 Bachelors, Masters, Community College, other

Featuregroup	**Logistic Regression**	**SVM**	**Naive Bayes**	**Random Forrest**
F1 - Shallow User Features	0,5399	0,3716	0,2031	0,4826
F2 - Language Use Features	0,5399	0,3716	0,0864	0,4637
F3 - Description Features	0,5399	0,3716	0,5182	0,4627
F4 - Text Features	0,5399	0,3716	0,3574	0,4981

Table 11 – Classifier performance based on accuracy on single feature groups with increasing difficulty for Occupation-1

Occupation 1

Classes: 2 Professionals, other

Featuregroup	**Logistic Regression**	**SVM**	**Naive Bayes**	**Random Forrest**
F1 - Shallow User Features	0,6236	0,5794	0,3684	0,6232
F2 - Language Use Features	0,6236	0,5794	0,3823	0,5723
F3 - Description Features	0,6236	0,5794	0,4443	0,5836
F4 - Text Features	0,6236	0,5794	0,4667	0,5912

Classes: 3 Professionals, Managers, other

Featuregroup	**Logistic Regression**	**SVM**	**Naive Bayes**	**Random Forrest**
F1 - Shallow User Features	0,5736	0,3945	0,1926	0,6236
F2 - Language Use Features	0,6236	0,3945	0,2308	0,5491
F3 - Description Features	0,6236	0,3945	0,4734	0,5745
F4 - Text Features	0,6236	0,3945	0,4175	0,6134

Classes: 4 Professionals, Managers, Technicians, other

Featuregroup	**Logistic Regression**	**SVM**	**Naive Bayes**	**Random Forrest**
F1 - Shallow User Features	0,6324	0,3602	0,1265	0,5791
F2 - Language Use Features	0,6324	0,3602	0,1947	0,5301
F3 - Description Features	0,6324	0,3602	0,4825	0,5817
F4 - Text Features	0,6324	0,3602	0,3867	0,6099

There are some striking results. For all the tables, we see (almost) identical scores for the Logistic Regression and SVM predictions across different feature groups. This is due to the fact that the logistic regression classifier predicted the same class for all the users in the test set. The SVM also showed similar scores across different feature groups, but did include all possible classes in its predictions, as the two confusion matrices below show:

Table 12 – Confusion Matrix of SVM classifier based on Language Use features for Occupation 1 predictions with three classes.

Predicted: **Actual**:	Managers	Technicians	Professionals
Managers	1	13	6
Technicians	4	7	6
Professionals	12	18	31

Table 13 – Confusion Matrix of SVM classifier based on Profile Description features for Occupation 1 predictions with three classes.

Predicted: **Actual**:	Managers	Technicians	Professionals
Managers	1	13	6
Technicians	4	7	6
Professionals	12	18	31

The confusion matrices show the exact same prediction pattern, while the classifier was trained on a whole different set of features. This behavior is displayed by both the SVM and the logistic regression classifiers. The Naive Bayes and Random Forest show more diversity in between predictions when they are trained on other sets of features. The scores of the Naive Bayes and the Random Forest classifier do show an expected decrease in performance as de task gets harder. Overall, the Random Forest classifier does perform more steadily than the Naive Bayes, which shows extreme peaks and dips in its performance, for example the 0.0864 score for Education – Level with four classes based on the Language Use features. Overall, the Random Forest classifier shows the most promising results.

In general, the results show the need for more advanced parameter tuning. In this study, the classifiers' parameters have been left at their default values. Logistic Regression was given a C value of 100, SVM was given a C value of 100000 and gamma of $1e^{-16}$, Naive Bayes was left as is and the Random Forest was given 25 trees. There are promising results when examining the Education of users, but these can be further improved when the classifiers are optimized for the task at hand.

CHAPTER 6. CONCLUSION AND FUTURE WORK

In this study, we've shown that we are able to create machine drive predictions of elements of the socio-economic status of individual Dutch Twitter users. After scrutinizing the ethical value trade-offs that are embedded in the dataset creation process, we created our own dataset and meticulously processed it in order to create a set of useful research subjects. We used the 'standard' Twitter data and the labels that we manually gathered from LinkedIn in order to create a dataset that has an external 'ground truth', providing a more certain label.

Features were created from structured and unstructured data, which were analyzed by the out-of-the box classifiers 'Logistic Regression', 'Support Vector Machine', 'Naive Bayes' and 'Random Forest'. The 'Combined' approach beat the 'Individual' approach by an average of 3.72 percentage points. When comparing the performance against that of the best dummy classifier, we find that the 'Combined' approach scored an average of 6.85 percentage points above the Most Frequent Dummy classifier.

We also examined the performance of the classifiers on classification tasks with only 2, 3 or 4 classes. This allowed us to see that the Shallow User features and Text Features yield the best results for Education 1. For Education-level the scores show no clear winner, but for Occupation 1 the best scores are also provided by Shallow User features and Text features. Overall, the Random Forest classifier shows the most promising results, with a peak performance of 0.7569 for the Education classification based on Shallow User features. These scores justify further studies with tuned algorithms and more data.

As any other study, this research is subject to limitations and bias. The main issue for this study is the manual annotation by one researcher, which could be heavily biased. Future studies based on the same dataset can focus on validating the manual annotations, extending the dataset, including the network data in the analysis, the tuning of the algorithms and the creation of additional features.

APPENDIX A – MANUAL CLASSIFICATION PROTOCOLS

STEP 1 – FILTERING OF ORGANIZATIONS, NON-DUTCH AND SPAM ACCOUNTS

Annotator was prompted with following information:

***** USER DESCRIPTION: *****
Naam: Fons Mentink
Account: fmentink
Omschrijving: Stagair bij Deloitte Technology Consulting, Studeert Business Analytics. Views are my own.

***** TWEET SAMPLES *****

1. Zekerheid op een schaal van 1 tot 10, met onder andere de opties blauw-verbaasd en geel-chill.
2. The most important perspective on the #panamapapers. Do we want media cherry picking or open data?
3. Jack Garratt timmert aan de weg in de VS, zie zijn top performance bij de @colbertlateshow!
4. Food for thought voor alle data-scientists (in spe)!
5. Google weet waar je foto is gemaakt door hem te vergelijken met 90 miljoen andere! http://www.theverge.com/2016/2/25/11112594/google-new-deep-learning-image-location-planet … Kan jij dat beter?
6. Mooi, nu zal er in ieder geval een debat over volgen. Trump staat alvast achter.
7. Loon naar werken voor deze toffe Twentse startup! #SciSports
8. Iedere week meer plezier van de #SpotifyDiscoverWeeklyPlaylist. Erg benieuwd naar achterliggende algoritmes!
9. Terug op de UT, op bezoek bij het #datacamp van de #UT en het #CBS, erg toffe onderzoeken en inzichten!
10. Het #DutchStudentInvestmentFund zoekt nieuwe bestuurders! Kijk snel op http://www.dsif.nl voor meer informatie! #innovatie #studenten

This is an example based on the profile of the author. The tweets in this selection contain no references to personal accounts, while the tweets that were presented to the annotator were simple the 10 most recent and therefore may have.

Used protocol:

1. If 9 or 10 tweets were in a different language than Dutch, user was marked as ‘other’.
2. If 9 or 10 of the tweets showed a very similar structure (i.e. automated messages about uploading something on Youtube, or facebook for example), user was marked as ‘other’.
3. If the *Naam, Account* or *Omschrijving* field revealed signs of the account being about an organization/controlled by multiple persons, the user was marked as ‘other’.

Step 2 – Filtering of underage users and aliases

Annotator was prompted with following information:

***** USER DESCRIPTION: *****
Naam: Fons Mentink
Account: fmentink
Omschrijving: Stagair bij Deloitte Technology Consulting, Studeert Business Analytics. Views are my own.

***** TWEET SAMPLES *****
Zekerheid op een schaal van 1 tot 10, met onder andere de opties blauw-verbaasd en geel-chill.
The most important perspective on the #panamapapers. Do we want media cherry picking or open data?
Jack Garratt timmert aan de weg in de VS, zie zijn top performance bij de @colbertlateshow!
Food for thought voor alle data-scientists (in spe)!
Google weet waar je foto is gemaakt door hem te vergelijken met 90 miljoen andere! http://www.theverge.com/2016/2/25/11112594/google-new-deep-learning-image-location-planet …
Kan jij dat beter?
Mooi, nu zal er in ieder geval een debat over volgen. Trump staat alvast achter.
Loon naar werken voor deze toffe Twentse startup! #SciSports
Iedere week meer plezier van de #SpotifyDiscoverWeeklyPlaylist. Erg benieuwd naar achterliggende algoritmes!
Terug op de UT, op bezoek bij het #datacamp van de #UT en het #CBS, erg toffe onderzoeken en inzichten!
Het #DutchStudentInvestmentFund zoekt nieuwe bestuurders! Kijk snel op http://www.dsif.nl voor meer informatie! #innovatie #studenten

This is an example based on the profile of the author. The tweets in this selection contain no references to personal accounts, while the tweets that were presented to the annotator were simple the 10 most recent and therefore may have.

Used protocol:

1. User was tested to the criteria of the previous step.
2. If the users *Naam*, *Account* and *Omschrijving* didn't provide a full first and/or last name they were marked as '*alias*'.
3. If the user listed their age and this was below 18, the user was marked as '*underage'*
4. If the user tweeted about high school, or celebrating birthdays under 18 they were marked as '*underage'*.

Step 3 – Finding users on LinkedIn

Annotator was prompted with following information:

***** USER DESCRIPTION: ***** Naam: Fons Mentink Account: fmentink Omschrijving: Stagair bij Deloitte Technology Consulting, Studeert Business Analytics. Views are my own.

This is an example based on the profile of the author.

Used protocol:

1. User was tested to the criteria of the previous steps.
2. In a Google Chrome Incognito Mode browser window, the name was used to search for the LinkedIn page.
3. The results were examined based on the content of the *Omschrijving*.
4. If there was no clear match, the location of the LinkedIn user was compared to the *Omschrijving*.
5. If there was no clear match, the user was looked for on Twitter. A profile picture comparison was performed.
6. If there was no clear match, the user was marked as '*unfindable*'.
7. If there was a matching profile; the following information was stored:

,	0
0	user_id: 297981243
1	gender: m
2	li_url: https://nl.linkedin.com/in/fonsmentink
3	edu: BSc University Twente Industrial Engineering & Management
4	year: 2014
5	occ: student

7.0 – user_id was stored automatically.
7.1 – gender was annotated, but never used in project.
7.2 – li_url: copied string URL of the matching profile
7.3 – edu: copied string from the matching profile. Highest level, completed, education.
7.4 – year: the year in which the education was finished. Never used in project.
7.5 – occ: the current occupation (copied). Students were given the '*student'* annotation.

Step 4 – From annotation to classification

Annotator was prompted with following information:

***** USER DESCRIPTION: *****
user_id: 297981243
Edu: BSc University Twente Industrial Engineering & Management
Occ: student

This is an example based on the profile of the author.

Used protocols:

For field of education:

The information that was copied from the LinkedIn Profile should be sufficient to provide the annotator with the ability to place the subject at some level in the SOI scheme. The relevant excerpt can be seen below:

…
2 Humaniora, sociale wetenschappen, communicatie en kunst
3 Economie, commercieel, management en administratie
    …
    35 Administratie, secretarieel
    37 Economie, commercieel, management en administratie met differentiatie
        371 Economie met wiskunde, natuurwetenschappen/ techniek
            3711 Econometrie
            3712 Bedrijfstechniek, technische bedrijfsvoering
            3713 Actuariaat
        372 Economie met andere differentiatie
        …
4 Juridisch, bestuurlijk, openbare orde en veiligheid
…

For education level:

The same prompt was used to place user into one of the following categories, based on the same SOI:

Nummer	Name	Kinds of education
0	Onduidelijk	Onduidelijk
1	Hoger onderwijs, derde fase	PhD, PostDocs
2	Hoger onderwijs, tweede fase	Masters, officiersopleidingen, MBA 's
3	Hoger onderwijs, eerste fase	Universitaire Bachelors, HBO Bachelors, POST HBO, post MBO-4
4	Secundair onderwijs, tweede fase	MBO-3, MBO-4, Havo/VWO Bovenbouw
5	Secundair onderwijs, tweede fase	VMBO-b/t/g/k, MAVO, Onderbouw Havo/VWO

The levels 6 and 7 are not included as these represent (pre)primary education, which is not applicable to the users that were still in the dataset at this point.

For occupation:

The information that was copied from the LinkedIn Profile should be sufficient to provide the annotator with the ability to place the subject at some level in the ISCO-08 scheme. As an example, here's the relevant excerpt for a system administrator:

```
…
1 Managers
2 Professionals
        …
        24 Business and administration professionals
        25 Information and communications technology professionals
                251 Software and applications developers and analysts
                        2521 Database designers and administrators
                        2522 Systems administrators
                        2523 Computer network professionals
                252 Database and network professionals
                …
        26 Legal, social and cultural professionals
        …
3 Technicians and associate professionals
…
```

The Dutch version of the ISCO-08 was used to manually assign users to classes.

APPENDIX B – TWEET EXAMPLE


```
JSON
Object
    index : 0
    0
        Object
            author
                Object
                    contributors_enabled
                    default_profile
                    description : Stagair bij Deloitte Technology Consulting, Studeert Business Analytics. Views are my own.
                    favourites_count : 18
                    follow_request_sent
                    followers_count : 88
                    following
                    friends_count : 289
                    has_extended_profile
                    id_str : 297981243
                    is_translator
                    listed_count : 2
                    location : Utrecht, The Netherlands
                    notifications
                    profile_background_color : C0DEED
                    profile_background_image_url_https : https://abs.twimg.com/images/themes/theme1/bg.png
                    profile_banner_url : https://pbs.twimg.com/profile_banners/297981243/1432655187
                    profile_image_url_https : https://pbs.twimg.com/profile_images/480341629903990784/I96Z-RYW_normal.png
                    profile_sidebar_border_color : C0DEED
                    profile_text_color : 333333
                    protected
                    statuses_count : 89
                    url
                    verified
            coordinates
                null
            entities
                Object
                    hashtags
                        Array
                            Object
                            Object
                                indices
                                    Array
                                        117
                                        127
                                text : innovatie
                            Object
                    symbols
                    user_mentions
                    urls
            favorite_count : 0
            geo
                null
            id_str : 660053542191939584
            in_reply_to_status_id
                null
            in_reply_to_user_id
                null
            is_quote_status
                False
            place
                null
            retweet_count : 0
            source : Twitter Web Client
            text : Het #DutchStudentInvestmentFund zoekt nieuwe bestuurders! Kijk snel op https://t.co/5dpg16DQpy voor meer informatie! #innovatie #studenten
            user
                Object
```

BIBLIOGRAPHY

Al Zamal, F., Liu, W., & Ruths, D. (2012). *Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors.* Paper presented at the ICWSM.

Berente, N., Gal, U., & Hansen, S. (2008). The Ethics of Social Stratification and the IT User. *Ethics, 1*, 1-2008.

Berkel-van Schaik, A. B., & Tax, L. C. M. M. (1990). *Naar een standaardoperationalisatie van sociaal-economische status voor epidemiologisch en sociaal-medisch onderzoek: rapport op basis van de werkzaamheden van de Subcommissie Sociaal-economische Status van de Programmacommissie Sociaal-economische Gezondheidsverschillen*: Ministerie van Welzijn, Volksgezondheid en Cultuur.

Bourdieu, P. (2011). The forms of capital.(1986). *Cultural theory: An anthology*, 81-93.

Boyd, D. a. C., Kate. (2011). Six provocations for big data. *Proceedings - A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*.

Brown, J. S., & Duguid, P. (1991). Organizational learning and communities-of-practice: Toward a unified view of working, learning, and innovation. *Organization science, 2*(1), 40-57.

Chapman, J. C., & Sims, V. M. (1925). The quantitative measurement of certain aspects of socio-economic status. *Journal of Educational Psychology, 16*(6), 380-390. doi:10.1037/h0075372

Coleman, J. S. (1990). *Equality and achievement in education*.

Costa, E., Lorena, A., Carvalho, A., & Freitas, A. (2007). *A review of performance evaluation measures for hierarchical classifiers.* Paper presented at the Evaluation Methods for Machine Learning II: papers from the AAAI-2007 Workshop.

De Smedt, T., & Daelemans, W. (2012). Pattern for python. *The Journal of Machine Learning Research, 13*(1), 2063-2067.

drs. Neil van der Veer, R. S. B., Isabelle van der Meer MSc. (2016). National Social Media Onderzoek 2016. *Newcom Research & Consultancy*.

Filho, R. M., Borges, G. R., Almeida, J. M., & Pappa, G. L. (2014). *Inferring user social class in online social networks.* Paper presented at the Proceedings of the 8th Workshop on Social Network Mining and Analysis, SNAKDD 2014.

Ikeda, K., Hattori, G., Ono, C., Asoh, H., & Higashino, T. (2013). Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems, 51*, 35-47. doi:10.1016/j.knosys.2013.06.020

Kiritchenko, S., Famili, F., Matwin, S., & Nock, R. (2006). Learning and evaluation in the presence of class hierarchies: Application to text categorization.

Lawrence, P. R., Lorsch, J. W., & Garrison, J. S. (1967). *Organization and environment: Managing differentiation and integration*: Division of Research, Graduate School of Business Administration, Harvard University Boston, MA.

Manzanares-Lopez, P., Muñoz-Gea, J. P., & Malgosa-Sanahuja, J. (2014). Analysis of linkedin privacy settings: Are they sufficient, insufficient or just unknown? *WEBIST 2014 - Proceedings of the 10th International Conference on Web Information Systems and Technologies, 1*, 285-293.

Miller, J. M. (2007). Dignity as a New Framework, Replacing the Right to Privacy. *Thomas Jefferson Law Review, 30*(1).

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. *ICWSM, 11*, 5th.

Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T. (2013). *" How Old Do You Think I Am?"; A Study of Language and Age in Twitter.* Paper presented at the Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research, 12*, 2825-2830.

Pennacchiotti, M., & Popescu, A. M. (2011). *Democrats, republicans and starbucks afficionados: User classification in twitter.* Paper presented at the Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130-137.

Preotiuc-Pietro, D., Lampos, V., & Aletras, N. (2015). An analysis of the user occupational class through Twitter content. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 1754-1764.

Preoţiuc-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., & Aletras, N. (2015). Studying User Income through Language, Behaviour and Affect in Social Media. *PloS one, 10*(9), e0138717.

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). *Classifying latent user attributes in twitter.* Paper presented at the Proceedings of the 2nd international workshop on Search and mining user-generated contents.

Siswanto, E., & Khodra, M. L. (2013). *Predicting latent attributes of Twitter user by employing lexical features.* Paper presented at the Proceedings - 2013 International Conference on Information Technology and Electrical Engineering: "Intelligent and Green Technologies for Sustainable Development", ICITEE 2013.

Sloan, L., Morgan, J., Burnap, P., & Williams, M. (2015). Who tweets? Deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PloS one, 10*(3), e0115545.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427-437.

Volkova, S., Bachrach, Y., Armstrong, M., & Sharma, V. (2015). *Inferring Latent User Properties from Texts Published in Social Media.* Paper presented at the AAAI.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91*(3), 461-481. doi:10.1037/0033-2909.91.3.461

Winkleby, M. A., Jatulis, D. E., Frank, E., & Fortmann, S. P. (1992). Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. *American journal of public health, 82*(6), 816-820.

Wynsberghe, A., Been, H., & Keulen, M. (2013). To use or not to use: guidelines for researchers using data from online social networking sites.

Zimmer, M. (2010). "But the data is already public": on the ethics of research in Facebook. *Ethics and information technology, 12*(4), 313-325.