university of groningen / faculty of arts

Detecting Income Level of Dutch Twitter Users using Stylometric Features

Léon Melein

Bachelor thesis
Information Science
Léon Melein
S2580861
December 5, 2016

ABSTRACT

This is what Kenneth K. Landes says about abstracts: "The abstract is of utmost importance for it is read by 10 to 500 times more people than hear or read the entire article. It should not be a mere recital of the subjects covered, replete with such expressions as *is discussed* and *is described*. It should be a condensation and concentration of the essential qualities of the paper.". In other words, it should comprise the goal of the thesis, describe the problems, the method used to solve them, the obtained results, and what consequences the results have (conclusions). Try to fit the abstract on one page (remember: it is abstract, not an essay).

CONTENTS

PREFACE

To be written at a later time.

1 INTRODUCTION

The introduction is the first chapter of your thesis. Every reader needs an introduction to the topic of your thesis (it doesn't need to be called "Introduction"). It is not always easy to write a coherent and comprehensible introduction: it requires practice. It sometimes makes sense to write the other chapters first, but this also has the risk that many things are taken for granted for the reader. The introduction provides that stage for the topic of the thesis, and usually has three key parts (**?**):

1. Which research questions the thesis is providing answers to (or trying to);
2. Why answers are given to these questions (motivation);
3. How the answers are given (structure).

It is not necessary to write a very long introduction. Try to make it snappy with a catchy opening that arouses interest of the reader – often the first sentence decides whether a reader is to read your report or not. There are many ways of doing this (**?**).

The rest of the thesis usually has a fixed structure. First of all more about the theoretical background is given in a separate chapter (unless this is very brief – in which case it could be part of the introduction). Then a description of the data material is provided: this could be done in a separate chapter or part of the method chapter. The method chapter gives a detailed overview of the used approach is given, and usually this chapter is called *Method*, *Methodology*, *Method and Tools*, or something similar.

2 BACKGROUND

I am currently writing up the background section of my thesis, but I could not finish it before I had to hand in this concept.
It will be included in the next concept version.

3 DATA AND MATERIAL

3.1 COLLECTION

The primary data set for this research is a corpus of Dutch Twitter users with their 500 latest tweets, categorized on income class. The tweets are saved as a text file per user, with a new post on each line. All users in a particular class are grouped in a folder. As there was no suitable data set available off the shelf, a new corpus was created. As a starting point, the University of Groningen (UG) twitter2 corpus was used to gather user profiles. The twitter2 corpus contains all Dutch tweets provided by Twitter's Streaming API, which constitutes a 1 percent representative sample of all public messages posted on Twitter.

In order to gather user profiles, we used all tweets from september 1th till september 5th, 2016. For each tweet in the corpus, we looked up its user by using the UG's inhouse *tweet2tab* tool to extract the user ID, username, real name and biography line for each user from the corpus. Furthermore, the user's biography line was used to find an occupational title. That title was then linked to an occupational class and consequently the average hourly income for that occupational class. The average hourly income was then muliplied by the average number of worked hours in the Netherlands to compute the average yearly income. All users with a known occupation were labeled with their average yearly income. This resulted in a collection of 36113 users with known occupations and incomes.

After removing no longer existing accounts, private accounts and accounts with less than 1000 tweets, 21862 users were still available. These users were divided into two income classes. Afterwards, 1500 users were randomly selected from each group and their tweets were gathered using the Twitter API. Retweets and non-Dutch tweets were left out of the collection. Users with less than 500 Dutch, self-written tweets were discarded. From the remaining users, 1000 users were randomly selected per class for further use in our research. More details about the processing of the users for use in the collection will follow in the subsection on *processing*.

3.2 ANNOTATION

In order to divide the users into income classes, they need to be annotated with their average yearly income. Distant supervision is used to find the average yearly income of a user. We look for an occupational title of a user in the user's biography line. In case a user has multiple occupations, we use the first one we can find. With the found title title, we look up the user's occupational class and the average hourly income for that class. We then label the user with the average yearly income by multiplying the average hourly income with the average total hours worked in The Netherlands.

There are three additional data sources needed in order to make our annotation process work. First, we use a list of occupational titles and their respective classes from Statistics Netherlands (2014a) to look up the occu-

pation of a user. These classes correspond with classes in the International Standard Classification of Occupations (International Labor Office, 2013). As this file was never meant for machinal consumption, the file had to be modified. All titles formed one long string, which had to be split in order to get the individual titles per class. Furthermore, the titles contained a lot of shorthand notations, e.g. "assistent-, coach" for the similar occupations "coach" and "assistent-coach" and slashes for synonyms like "typist / tekstverwerker". These were removed by hand as there was no suitable way to do this correctly in an automated manner. Second, we use a list of occupational classes and their respective average incomes from Statistics Netherlands (2014b) to look up the average hourly income for a particular class. Finally, to derive the average yearly income we need to know the average worked hours per year in The Netherlands. According to the European Observatory of Working Life (2015) the average Dutch worker makes 1677 hours a year.

To evaluate the performance of our distant supervision method a random survey of 100 users per class was taken. Their labels were manually checked in a two class setting: a low class with incomes below the average national income of 34.500 euros per year and a high class with incomes above that. The accuracy over the whole group of 200 users was 68,5 percent, with 59 percent in the low class and 78 percent in the high class. In 33 cases, the labels were wrong because the account was simply not used by a person but by a company. As there is no surefire way to distinguish between human and non-human users, we disregard these cases. The overall accuracy without these cases is 82 percent. The last few cases that were wrongly labeled consist of hobbies or voluntary work labelled as an occupation (15 cases), occupational titles that got bodged during the transformation of the needed file (11 cases) and four miscellaneous cases, which consist of users describing their former occupation in a non-trivally detectable way.

3.3 PROCESSING

Concerning the users, after extracting users with a known occupation from the twitter2 corpus a number of processing steps are involved to bring the entire group of 36113 users to a manageable number of suitable users. First, users that no longer exist, have their profile set as private or have less than 1000 tweets are removed to ensure we can get enough data per user for our research. The users are checked by using the Lookup API of Twitter, which allows us to check users in batches of 100. The remaining users are saved in a Python dictionary and written to disk with the built-in Pickle library. Afterwards, the remaining users are divided in the chosen income classes. From each class 1500 users are randomly selected for further processing by using the random.choice function of the NumPy Python library. For these users, their latest 1000 tweets are collected. Retweets and non-Dutch tweets were discarded. The language classification of each tweet was performed by the *langid* Python library. Users with less than 500 suitable tweets were left out of the data set. For all remaining users the tweets are written to a text file per user per class. From the remaining set of users, 1000 were randomly selected per group to be used in our research.

After the processing of the users is completed, the tweets are prepared for further use. URL's, hashtags and usernames are removed and the tweets are tokenized so that relevant features can be derived from them. The pro-

cessing relies mainly on the *TweetTokenizer* included in the NLTK Python library, a popular library for natural language processing in Python.

4 METHOD

To be written at a later time.

5 RESULTS AND DISCUSSION

To be written at a later time.

6 CONCLUSION

To be written at a later time.

BIBLIOGRAPHY

European Observatory of Working Life (2015). *Developments in collectively agreed working time 2014*.

International Labor Office (2013). *International Standard Classification of Occupations 2008 (ISCO-08)*. International Labor Office.

Statistics Netherlands (2014a). *Codelijsten ISCO-08*. Retrieved from https://www.cbs.nl/nl-nl/onze-diensten/methoden/classificaties/onderwijs

Statistics Netherlands (2014b). *Uurlonen van werknemers naar beroepsgroep, 2012*. Retrieved from https://www.cbs.nl/nl-nl/maatwerk/2014/15/uurlonen-van-werknemers-naar-beroepsgroep-2012.