university of
groningen

faculty of arts

Detecting Income Level of Dutch Twitter Users using Stylometric Features

Léon Melein

Bachelor thesis - Concept version
Information Science
Léon Melein
S2580861
December 18, 2016

ABSTRACT

CONTENTS

PREFACE

To be added to the final version of my thesis.

1 | INTRODUCTION

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).

2 | BACKGROUND

Research into author profiling on Twitter is relatively new, especially in the field of income prediction. All prior work has focused exclusively on English-speaking Twitter users.

The most recent study was performed by Flekova et al. (2016). Their goal was to find a viable writing style-based predictor for age and income. For each dimension, a different data set was used. Regarding income, a data set created by Preotiuc-Pietro et al. (2015) was used. It contains almost six thousand Twitter users labeled with their occupation. The researches used the occupations to label the users with the mean UK yearly income for their occupation. This was done regardless of the location of the users involved (e.g., an English-speaking user from Amsterdam would also be assigned a UK income). As the exact income for every occupation was known, the machine learning task was framed as regression. Flekova et al. codified stylistic variation into a large number of features, which were grouped into four categories: surface, readability, syntax and style. Using ten-fold cross validation they discovered that readability metrics like the *Flesch Reading Ease* metric and the relative use of pronouns correlated with income over age ($r_{income} = .297, r_{Flesch} = .315; p < 0.001$). They concluded that the differences in style can be used to "tailor the style of a document without altering the topic to suite either age or income individually".

The data set used in Flekova et al. (2016) was created during an earlier study by Preotiuc-Pietro et al. (2015). They used the corpus to classify users according to their occupational class. They used the occupational titles and classes from UK Standard Occupational Classification (SOC) (Office for National Statistics, 2010) for their research. The SOC is a hierarchical classification of occupations. It has four levels: starting with nine very general classes and terminating in hundreds of very specific classes. Each layer is indicated with a different number of digits. The coarsest level is indicated with one digit and the finest level with 4 digitals (e.g., class 1: 'managers, directors and senior officials' and class 1116: 'elected officials and representatives', respectively). The classification is based on the International Standard Classification of Occupations (International Labor Office, 2013). For each occupation they used the Twitter REST API to find at most 200 users for each occupation. The accumulated users were divided into the three-digit groups they belong to. Users that were companies, had no description or had a contradicting description were removed from the collection by hand. Furthermore, three-digit groups with less than 45 users were discarded. The final collection contained 5191 users divided into 55 three-digit groups.

Preotiuc-Pietro et al. (2015) mention two papers that describe different labeling strategies in their related work section. Both use distant supervision, like Preotiuc-Pietro et al., but in a different manner.

The first is a paper by Li et al. (2014) which tries to label users with the name of their employer, among other things. In this study, Twitter profiles were cross-linked with profiles on a different social networking site, Google Plus. To ensure the both profiles are interrelated they looked at the friend connections on both sites and made sure that there was a large enough intersection between them. The employer name was extracted from the Google

Plus profile and used to label the Twitter user. Using this method they were able to collect 7208 users with a known employer. This strategy relies heavily on profiles from networking sites other than Twitter, like Google Plus or LinkedIn. Without unfettered access or a large number of cross-linkable profiles on both platforms, the strategy cannot be applied.

The second strategy only relies solely on the occupational titles from the SOC and profiles on Twitter. Sloan et al. (2015) labeled users with class in the National Statistics - Socio-Economic Classification (NS-SEC), a classification closely related and interoperable with the SOC. They extracted users from a feed provided by the Twitter API, which constitutes a 1 percent representative sample of public tweets. For each user, they looked for an self-disclosed occupational title in the biography line. The titles they used for detection were gathered from the SOC. After finding an occupation, the user would be labeled with their NS-SEC class by looking up the SOC class for their occupation and then looking up the corresponding NS-SEC class. In case a user mentioned multiple occupations, the authors hypothesized that the first one would be most important and therefore be used. Using this method they were able to label 32032 users with their NS-SEC class. A random survey of 1000 users resulted in an accuracy of 57,8 percent. The authors mention several caveats of this method. Hobbies and former occupations may be falsely detected as current occupations. Commonly occurring phrases like "Doctor Who fan" may also result in false matches.

This second strategy depends a lot less on outside data. It is therefore easier to implement and use in further research. Specifically for this thesis, a corpus of Dutch tweets and their authors available already at the university and data on occupational titles and incomes is available from the Dutch government bureau Statistics Netherlands. It therefore makes sense to use this strategy for our data collection and annotation efforts. The implementation details of this strategy follows in the next chapter.

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. 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 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 in-house *tweet2tab* tool to extract the user ID, username, real name and biography line for each user from the corpus. 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 multiplied 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, high (above 34.500 euros) and low (below 34.500 euros). Afterwards, 1500 users were randomly selected from each group and their tweets were gathered using the Twitter API. Retweets and non-Dutch tweets (as explained in the next section) 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 as mentioned in the previous subsection. 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-trivially 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 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.*

Flekova, L., D. Preoţiuc-Pietro, and L. Ungar (2016, August). Exploring stylistic variation with age and income on twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Berlin, Germany, pp. 313–319. Association for Computational Linguistics.

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

Li, J., A. Ritter, and E. Hovy (2014, June). Weakly supervised user profile extraction from twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland, pp. 165–174. Association for Computational Linguistics.

Office for National Statistics (2010). *The Standard Occupational Classification (SOC) 2010 Vol 1: Structure and Descriptions of Unit Groups.* Palgrave Macmillan.

Preotiuc-Pietro, D., V. Lampos, and N. Aletras (2015, July). An analysis of the user occupational class through twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China, pp. 1754–1764. Association for Computational Linguistics.

Sloan, L., J. Morgan, P. Burnap, and M. Williams (2015, 03). Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. *PLoS ONE 10*(3), 1–20.

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.