Detecting Income Level of Dutch Twitter Users using Stylometric Features

Léon Melein, S2580861

October 28, 2016

Abstract

Income prediction is a relatively undiscovered aspect of author profiling. Early research on English-speaking Twitter users linked to occupations and average incomes has been promising, but there is no comparable research for Dutch speakers yet. We want to find out to what extent profiling can predict the income level of Dutch-speaking Twitter users. We do so by applying the methodology as Flekova et al. (2016) at these users. Dutch-speaking users will be linked to their occupation and the average income for that occupation to establish ground truth and which, if any, stylometric features correlate significantly with income.

1 Introduction and background

As we rely more and more on the internet and its applications in our daily lives, the opportunities for author profiling continue to grow. Social media services, and the text-based communications they facilitate, provide an ever-growing corpus of texts, which are linked to their author. Furthermore, these services provide additional self-disclosed information about the authors like education, occupation and relationships. If we use this data the right way, we can perform research that simply wasn't possible before or on a much bigger scale than was possible before (Sloan et al., 2015).

An aspect of author profiling is income prediction. A potential use for these estimations of income is in customer relations. Just as the rest of our life, our contact with companies and their customer services move online. People want to be helped quickly and without giving a lot of information in advance. They expect companies to "know" them. Companies want to know their customers so they don't have to ask for a lot of information and can help quickly. Income can be a key factor in this, because it can help in determining which products or services a customer probably already has and in which it is likely interested.

An example: a cable company customer with a relatively low income probably has a low tier service package. He or she will only be interested in cheap(er) options or options that provide significant extra value, in comparison to their additional cost. Knowing this, offering an expensive movie channel package to this customer does not make a lot of sense. Such a package is nice to have, but isn't really of much value - unless the customer is a true movie addict.

There has been some research on income prediction using author profiling on English language Twitter messages. Flekova et al. (2016) looked at both age and income. By codifying stylistic variation into a large number of features, they tried to find a viable writing style-based predictor for age and income. Using this method, they found that readability metrics like the Flesch Reading Ease and the usage of pronouns correlate more strongly with income than age, among other things. The data set used in this study was created by Preotiuc-Pietro et al. (2015). They labeled users with self-disclosed occupational titles extracted from their biography and the income for the extracted occupation in the United Kingdom, even if the user did not live there. Other studies use a second information source to get the occupation of a user Li et al. (2014) used this method to get the current employer of Twitter users. They cross-linked Twitter profiles with Google Plus profiles. From the Google Plus profile they gathered the current employer of a Twitter user. The link between both networks was formed by a link to the Twitter profile that users had added themselves on their Google Plus profiles.

Despite the existing research into English-language Twitter users and their respective incomes, there is currently no comparable research for Dutch-language Twitter users. I will therefore focus my research on the following question: to what extent is it possible to accurately predict the income level of a Dutch Twitter user and which stylometric features are predictive?

2 General approach

2.1 Overview

My research consists of two mayor parts. First, I will look into which features are predictive of income level. Afterwards, I will look into actually predicting the income level of a user. For both parts, I will need a corpus of Twitter messages, for which the author and their income is known. I will discuss the way I intend to build this corpus in the next section.

To find predictive features I will look at the features that Flekova et al. (2016) used in their research. These can be divided into four groups: surface, style, syntax and readability. The details of these features will be discussed later in this proposal. After finding good predictors, I will proceed with the second part of my research. I will look at the correlations between the individual features and income categories and use the features that turn out to be predictive of income level for the next part of my research.

As we are looking at income levels, the task ahead of us is a classification task. Flekova et al. (2016) used regression in their research, as they were able to treat income as a continuous variable because the UK Office for National Statistics provides an average income for each and every occupation, whereas Statistics Netherlands (CBS) does not. We therefore classify our users with the best income data available for their occupation. I will discuss this in more detail later on in this section.

I will use a machine learning algorithm to build a classifier that uses the predictive features. With that classifier, I will find out to what extent I can accurately predict the income level of Dutch Twitter users. I intend to use a number of classification algorithms in order to find the one that provides us with the best results. Because scikit-learn, a popular machine learning package for Python, contains a lot of different algorithms and makes it easy to switch between algorithms, this shouldn't bring about a significant increase in workload.

Afterwards, I will evaluate the performance of the classifier. Details of the expected output and evaluation of the results follows in section 3 of the proposal.

2.2 Data collection and usage

As discussed earlier, the primary data set for my research will be a corpus of Twitter messages, categorized on the income level of its author. I will extract the users from the Dutch-language twitter2 corpus of the University of Groningen. By using the tweet2tab tool to extract the id, username, real name and biography of the author for each tweet, I will create a file with Dutch-speaking Twitter users that we can use for further processing.

For each extracted user, we try to find his or her occupation. We use distant supervision to find these occupations. Using an existing list of occupational titles from Statistics Netherlands (2014a) we look for an occupation in each user's biography. When we find a occupation, we find its class in the International Standard Classification of Occupations (International Labor Office, 2013). The details of these occupational classes will be discussed in the following subsection. After we found the occupational class, we use it to look up the average yearly income for this class in an existing list from Statistics Netherlands (2014b).

Afterwards, we are left with users labeled with their occupation and the average yearly income for the occupational class they fall into. Because we've collected our labels automatically it's necessary to check them for correctness. I will probably select a limited number of user profiles from the large collection for each income level, for which I am certain the label is right.

After the users are selected, I will divide them into income classes. The exact classes are to be determined. Flekova et al. (2016) used two groups at some point in their research: $\leq$ £25.000 and $\geq$ £35.000). Statistics Netherlands use a six-way classification in their regular research: €0 - €10.000, €10.000 - €20.000, €20.000 - €30.000, €30.000 - €40.000, €40.000 - €50.000 and €50.000 or more. I will explore both classifications and use them both in my research, if possible.

Finally, when the users are divided in their respective income levels, their 500 latest Dutch posts are retrieved from Twitter using the REST API, retweets excluded. If a user has less than 500 Dutch self written tweets, he or she will be discarded. The result of our data collection effort is corpus which is divided into directories for each income level. Each directory contains text files, which hold the posts for each user belonging to the income level.

The amount of users in our corpus depends strongly on the successful labeling of our data. The exact amount of tweets I will use to extract users will depend on our yields in labeling. My first intention is to use a month's worth of tweets to extract user.

The corpus will be used to find stylometric features that are predictive of users income levels. The extraction of features from the corpus will be discussed later on. After the discovery of such features, I will use the corpus to perform multifold cross validation of the classifier. The exact amount of validations is to be determined at a later time.

2.3 Labels, features and additional data

Statistics Netherlands (2014a) Statistics Netherlands (2014b)

2.4 Data pre-processing and usage

3 Expected output and evaluation

3.1 Expected output

3.2 Evaluation

Cohen (1988)

4 Literature

References

Cohen, J. (1988). Statistical power analysis for the behavioural sciences. hillside. *NJ: Lawrence Earlbaum Associates*.

Flekova, L., Preoţiuc-Pietro, D., and Ungar, L. (2016). 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)*, pages 313–319, Berlin, Germany. Association for Computational Linguistics.

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

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

Preotiuc-Pietro, D., Lampos, V., and Aletras, N. (2015). 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)*, pages 1754–1764, Beijing, China. Association for Computational Linguistics.

Sloan, L., Morgan, J., Burnap, P., and Williams, M. (2015). 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.