Predicting Income Level of Dutch Twitter Users using Stylometric Features

Léon Melein, S2580861

November 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 **?** 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 stylometric features correlate significantly with income. Afterwards, we use these features to build a classifier, in order to find an answer to our research question.

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 (**?**).

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 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 posts. **?** 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 **?**. 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. **?** used this method to get the current employer of Twitter users. They linked Twitter profiles with Google Plus profiles. From the Google Plus profile they gathered the current employer of the 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 speaking Twitter users and their respective incomes, there is currently no comparable research for Dutch speaking 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 major parts. First, I will look into which features are predictive of income level. Afterwards, I will look into actually predicting the income level of users. For both parts, I will need a corpus of Twitter posts, for which the author and their income is known. I will discuss the way I intend to build this corpus in the next subsection.

To find predictive features I will look at the features that **?** 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 on in this proposal. 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 I work with income levels, I will use a classification algorithm in order to predict users' income level. **?** used regression in their research. 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 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 returns 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 the next section.

2.2 Data collection and usage

As discussed earlier, the primary data set for my research will be a corpus of Twitter posts, 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 provided tweet2tab tool to extract the id, username, real name and biography of each post's author, 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 their occupation. We use distant supervision to find these occupations. Using an existing list of occupational titles from **?** we look for an occupation in each user's biography. This list also holds the corresponding class from the International Standard Classification of Occupations (ISCO) (**?**). With that class, we can look up the average yearly income in an existing list from **?**. The details of these occupational classes will be discussed in the following subsection.

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 their respective income classes. The exact classes are to be determined. **?** 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 into 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 posts, it will be discarded. The result of our data collection effort is a 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 labelling of our data. The exact amount of posts I will use to extract users will therefore depend on the yield of the labelling process. My first intention is to use a month's worth of posts to extract users.

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 develop a classifier and perform a multifold cross validation. The exact amount of validations is to be determined at a later time.

2.3 Labels, features and additional data

As already discussed in the subsection on data collection, each user included in our corpus is labelled with their average yearly income. To derive this income we use a three step process. First, we determine a user's occupation and the corresponding ISCO class. Second, we look up the average income for that class. As the average incomes in our data set are hourly, we need to multiply them with the average amount of worked hours per year to calculate the average yearly income. According to the **?** the average worked hours per year in the Netherlands are 1677. Finally, we label the user with its average yearly income.

We rely on two additional data sources for our labelling process. The occupational titles and their respective classes were derived from a file with occupational classes and example occupations provided by **?**. The average yearly income per occupational class is derived from another file created by **?**.

The ISCO classification contains four levels of classes, starting with nine very general classes and terminating in hundreds of very specific classes. The average hourly wages are only available for the first and second level of the classification. We use the second level to link our occupations to, as this provides the most specific income data available.

The features that will be used in this research will be the same as those used in **?**, given that can be extracted from Dutch texts. I will discuss them per class.

The first group of features is surface. This group includes measures like average post length (in words and characters), average length of words used in the posts and the type-token ratio. These features should be relatively straightforward to implement.

The second group is readability. This group includes measures like the Automatic Readability Index, the Flesch Reading Ease metric and the SMOG grade. These metrics were specifically built for English. At the time I am not sure if all metrics have Dutch counterparts. I will look for them, as these group of metrics performed best in the study by **?**.

The third group is syntax. This involves calculating the ratio of all parts-of-speech in a user's posts. In order to count the parts of speech, we need a parser that provides us with part-of-speech tags.

The final group is style. This includes the amount of Named Entities in a user's posts, the ratio of standalone numbers and the ratio of hapax legomena. As with the first group, these features should be relatively straightforward to implement.

2.4 Data pre-processing

As described in the previous subsection, all features are derived from our corpus. In order to calculate most features the posts need to be tokenized and non-word features like hashtags and hyperlinks must be removed. Furthermore, the posts must be parsed to get the parts-of-speech. I will use the Alpino parser to get the needed part-of-speech tags.

3 Expected output and evaluation

3.1 Expected output

This research will have two outputs, one intermediate and one final. The intermediate output will be formed by the significance and correlation coefficient of our features. We use these outcomes to determine which features are predictive and should be used by our classifier. The features will be selected on significance and correlation strength.

The final output will be that of our classifier. The most important results will be precision and recall, as we need those to calculate the F1-score, their harmonic mean. These will be gathered by performing a multifold cross validation. We can use that, together with accuracy, to compare our classifier with our baseline. Furthermore, a confusion matrix and the support of each class will be included in order to interpret the results.

3.2 Evaluation

In order to evaluate the classifier's performance a baseline is needed. As there is no established baseline yet, we will create a baseline by classifying all users as the most frequent class in our corpus. We will then compare the classifier to the new baseline in order to evaluate its performance and provide a full answer to our research question.

4 Literature

References

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.