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 our daily lives rely more and more on the internet and its applications, 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 textual data in the right way, we can perform research that simply wasn't possible before or on a much bigger scale than 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 user 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, income or both. 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 crosslink with the original data set. Li et al. (2014) uses 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

2.2 Data collection and use

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.

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*.

Statistics Netherlands (2014b). *Uurlonen van werknemers naar beroepsgroep, 2012*.