Detecting Income Level of Dutch Twitter Users using Stylometric Features

Léon Melein, S2580861

October 3, 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 by applying the same methodology on these users. Dutch-speaking users will be linked to their occupation and the average income for that occupation to establish ground truth. Afterwards, we will find out which, if any, stylometric features correlate significantly with income.

1 Introduction and background

(i) Include here your motivation to do the research you propose. (main area, subarea, problem statement, most important references)
(ii) End this section with your research question. (your research question, clearly specified)

- Main area: author profiling
- Sub area: profiling on social media
- Problem statement:
  - Customer relations are more and more moving online. People want to be helped fast and without giving a lot of information, they expect companies to "know" them.
  - Companies want to know their customers so they don't need to ask a lot of information and can provide relevant information and offers.
  - Income can be a key factor because it can help in determining which products a customer already has and in which it is likely interested.
  - An example: a cable company customer with a low income probably has a low tier service package and will only be interested in cheap(er) options or options with significant extra value. Knowing this, offering them an expensive movie channel bundle - something which is nice to have, but isn't really of much value - doesn't really make sense.
- Most important references
  - Flekova et al. (2016)
- Research question
  - To what extent is it possible to accurately predict the income level of Dutch Twitter Users, taking solely stylometric features into account?

2 General approach

(i) Specify the approach you adopt in order to execute your research (i.e., to answer your research question) successfully.
(ii) Describe your data collection, the corpus of input data (amount and type(s) of data), the way you will obtain it, and how you intend to use it.
(iii) Describe your labels and data representation/features (amount and label(s), type of data features). Also describe any additional annotations or data resources you may use.
(iv) Describe the way you intend to use your data as explicitly as possible, including the way you intend to pre-process your data.

- Approach
  - Collect Dutch Twitter profiles including tweets, link them to their occupation and label them with the respective average income

  - Extract features from tweets using software
  - Train a machine learning model using some of the extracted feature vectors as a training set

    Question: do I want to turn this into a classification or regression task? Predicting exact incomes might yield inaccurate results, the authors of Flekova et al. (2016) chose to turn it into a classification task. It might be wise to follow that choice and adhere to the categories the Dutch Bureau of Statistics (CBS) uses in income classification (yearly income: 0-10.000, 10-20.000, 20-30.000, 30-40.000, 50.000 euros or more).
  - Test the model with multi-fold validation

- Data collection and use
  - The authors of the paper used an English language corpus created by Preoiuc-Pietro et al. (2015). I have to look into the full details of the collection method in that paper in order to evaluate if that method can be used for our purposes.
  - Another option proposed last week was cross linking Twitter profiles with LinkedIn, in order to save time and establish relatively reliably what occupation a user has. Implementing this in an automated manner is hard, because LinkedIn closed off their machine readable search interface last autumn. Doing it by hand will cost a lot of time and looking at other options may therefore be wise.

- Labels and features
  - The labels for the Twitter users will be the average income for their occupation.
  - I will largely use the same features that are used in Flekova et al. (2016), given that the specific feature or metric is available for Dutch. I will adhere to the feature groups that they've used: surface, readability, syntax and style. The exact features need to be determined.

- Additional data
  - I need a data set with occupations and their respective average income, so I can link Twitter profiles to the average income for their occupation, which enables us to establish ground truth.
  - The Dutch Bureau of Statistics (CBS) has data on average yearly and hourly salary per industry following their own *Standaard Bedrijfsindeling* (Standard Classification of Businesses), but it does not seem to provide more granular income statistics on first sight. They do have a Dutch implementation of the International Standard Classification of Occupations, created by the International Labour Organization (ILO). Therefore, I do suspect that there is more data available from CBS. I will contact them to make sure they have more detailed statistics on average income.

- Data pre-processing and usage
  - Preoiuc-Pietro et al. (2015) used the Trendminer pipeline to prepare their corpus. As we will be handling Dutch tweets it may be wise to employ a different pipeline to prepare the tweets. There is a pipeline available that is used with the Dutch-language *twitter2* corpus of the University of Groningen, which may be usable for our purposes.

3 Expected output and evaluation

(i) Indicate how your output data (results of your research) will look like.
(ii) Specify the way you intend to evaluate your results.

- Output data

  Correlations between features and income level

- Evaluation

  Significance: the first step in determining if a feature has predictive power is the significance of that feature in determining the income level of a user.

  Strength of correlation: we can interpret *Pearson's r* by the use of some rules of thumb if the relation is between a feature and income is linear. There are a number rules of thumb like the ones proposed by 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.

Preoiuc-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., and Aletras, N. (2015). Studying user income through language, behaviour and affect in social media. *PLoS ONE*, 10(9):1–17.