Computational Linguistics in the Netherlands Journal 7 (2017) 79-92 Submitted 05/2017; Published 12/2017

Profiling Dutch Authors on Twitter: Discovering Political Preference and Income Level

Reinder Gerard van Dalen R.G.VAN.DALEN@STUDENT.RUG.NL
Léon Redmar Melein L.R.MELEIN@STUDENT.RUG.NL
Barbara Plank B.PLANK@RUG.NL

University of Groningen, The Netherlands

Abstract

Research in author profiling has primarily focused on English-speaking users and attributes like age, gender and occupation. We present first experiments on automatic profiling Dutch Twitter users for two less-studied attributes, namely their political preference and income level (low vs high). We create two novel corpora using distant supervision, evaluate the corpus creation approach, and train predictive models for each attribute. Our empirical evaluation shows that distant supervision is surprisingly reliable and political preference and income level of Dutch users can be predicted relatively accurately from the linguistic input. We also discuss which features are predictive for income and political preference, respectively.

1. Introduction

The widespread use of social media has enabled researchers to study human behavior at a unprecedented scale. Recent research has shown interest in the interplay of language use and user attributes. A diverse set of user attributes (or traits/factors) are shown to be predictable from users' linguistic input. Factors studied so far include gender, age, personality or income, to name but a few (Mairesse and Walker 2006, Luyckx and Daelemans 2008, Rao et al. 2010, Rosenthal and McKeown 2011, Nguyen et al. 2011, Eisenstein et al. 2011, Volkova et al. 2013, Alowibdi et al. 2013, Ciot et al. 2013, Plank and Hovy 2015, Volkova et al. 2015, Verhoeven et al. 2016, Preoţiuc-Pietro et al. 2015a, Flekova et al. 2016).

Predicting traits can play an important role for a wide range of applications, ranging from automatically tailoring customer service communication to personalized machine translation (Mirkin et al. 2015, Rabinovich et al. 2017). We present first results of profiling Dutch Twitter users for two less-studied author attributes, namely, their political preference and income level.

For politicians, social media platforms can be interesting as a channel for their campaign. Not only because of the shifting from offline to online discussions, but also because of the way one can reach specific groups. Political preference is the first attribute of interest in this study. Second, income prediction is a relatively new aspect of author profiling. Very recent research on English (Flekova et al. 2016) has linked Twitter users to occupations and their respective average incomes, and obtained promising results for income prediction. To the best of our knowledge, there is no comparable research for Dutch speakers yet.

It was not clear upfront to what degree users disclose occupation and political preference, and whether information on linking occupation to income is retrievable, so that these in turn can be leveraged as a weak supervision signal for training automatic prediction models for Dutch social media. To summarize, the research questions of this study are the following:

- To what extent is it possible to accurately predict the income level of a Dutch Twitter user? Which stylometric features are predictive?
- To what extent is it possible to automatically classify Dutch Twitter users based on political preference using their tweets?

To answer these two research questions, two novel datasets are presented containing Dutch Twitter data annotated for income level or political affiliation. We employ *distant supervision* as data collection method, following recent work (Preotiuc-Pietro et al. 2015b, Flekova et al. 2016). Distant supervision is a weak annotation method that exploits implicit links to gather annotated data. It was first introduced in affective computing by leveraging hashtags as proxy for emotion labeling (Read 2005, Go et al. 2009, Pak and Paroubek 2010). Two subquestions that we explore are: Is distant supervision an accurate technique to automatically annotate Dutch Twitter users for income level and political preference? What are the most informative features?

Both studies originated as separate works.[1] They follow the same general idea: user profile information is queried to obtain self-disclosed information on a user's occupation or political preference. In this way, possibly large but noisy amounts of data can be collected. The exact steps that were necessary to obtain the corpora and the experiments using the obtained data differ for the two cases, as will become clear later on.

In the following section we describe the data collection and annotation process in further detail. The rest of this article is organized as follows. Section 3 introduces method, features and experimental setup. We then present the results and investigate predictive features as well as discuss limitations of the present study in Section 5. Finally, we discuss related methods in Section 6 and end with the conclusions.

2. Data

We collected two large datasets of tweets, one labeled for user income $\mathcal{D}_I$, the other for political preference $\mathcal{D}_P$. Each dataset consists of the tweets of Twitter users that were identified by distant supervision. In particular, the two steps involved were:

1. Querying a large in-house Twitter corpus for user profiles (biographies). The task is to extract relevant users whose profile matches an attribute of interest (political preference or occupation, see examples in Figure 1). For political affiliation this step resulted directly in the annotated data (preferred political party); in the income case there was an additional step, as the occupation still needed to be linked to income classes (discussed below).

2. Retrieve the most recent tweets for the users. Once the list of relevant users was identified, this second step queries the Twitter API to retrieve the most recent tweets for a given user. For this step, we assumed that the income, political preference and associated biography information are relatively static and therefore remain the same during the collection period. The retrieval of tweets resulted in the final corpora.

An overview of the collected data as well as the sample used in the experiments is given in Table 1. Next we describe the corpus creation in further detail.

Income: $\mathcal{D}_I$ The primary data set for this research is a corpus of Dutch Twitter users with their 500 most recent tweets, categorized by income class, which was created for the purpose of this study. As a starting point, user profiles were obtained from a large in-house corpus of Dutch tweets. This corpus contains a subset of Dutch tweets provided by Twitter's Streaming API, which constitutes approximately one percent of the public messages posted on Twitter. In order to gather user profiles, all tweets from September 1st till September 5th, 2016 were used.[2] For each tweet, username and biography (user profile) were extracted and 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 number of hours worked by the average Dutch worker per year, to compute the average yearly

1. The research presented in this article is the result of two Bachelor thesis projects at the University of Groningen.
2. This was the time span of the first month of this project.

	Income level $\mathcal{D}_I$	**Political preference $\mathcal{D}_P$**
Total corpus:		
Number of users	3,070	3,802
Number of tweets	2.7 million	1.9 million
Users per class:	high: 1571, low: 1499	VVD: 728, PvdA: 647, CDA: 635, SP: 300, GL: 430, D66: 785, CU: 277
Used in experiments:		
Number of users	2,000	2,000
Number of tweets	1 million	200k / 1 million

Table 1: Statistics of the datasets.

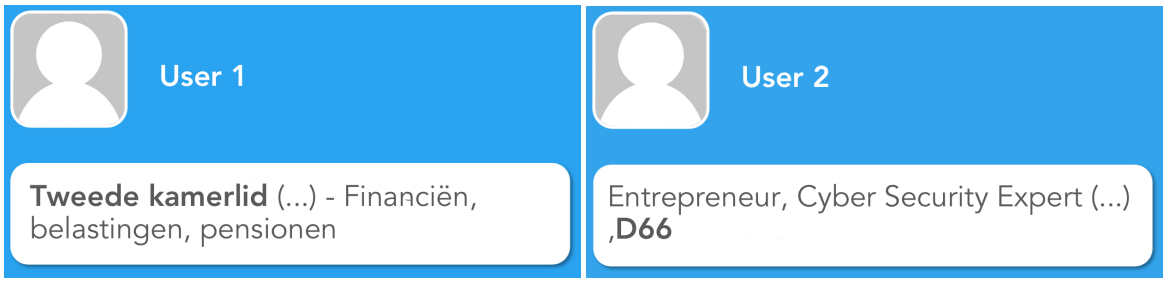

Figure 1: Example of profiles annotated via distant supervision. Left: income level, Right: political preference.

income. All users with a known occupation were labeled with their average yearly income. This resulted in a collection of 36,113 users with suggested occupations and incomes.

Note that in case a user has multiple occupations, we use the first one we find, assuming it is the most relevant. With the found title, we look up the user's occupational class to calculate average hourly income for that class. There are three additional data sources needed in order to get the final income annotations.

In more detail, first, we use a list of occupational titles and their respective classes from Statistics Netherlands (2014a) to look up the occupation of a user. These classes correspond with classes in the International Standard Classification of Occupations (International Labor Office 2013). We had to preprocess the file to make it ready for the matching process (e.g., "assistent-, coach" was converted to "coach" and"assistent-coach"). 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. We use the two-digit classes, as the incomes corresponding to almost all of them are known. For most three- and four-digit classes, incomes are not provided by Statistics Netherlands. Finally, to derive the average yearly income we looked up the average worked hours per year in The Netherlands. According to the European Observatory of Working Life (2015) the average Dutch worker works for 1677 hours a year.

After removing user accounts which meanwhile no longer existed, private accounts and accounts with less than 1000 tweets, 21,862 users were still available. These users were divided into two income classes, high (above €34,500) and low (below €34,500). The split point is the modal income in the Netherlands in 2014. The incomes amongst the two-digit groups vary enough to warrant a viable two-class split of our data, which we aimed at here (we leave a more fine-grained analysis for future work). Afterwards, 1,500 users were randomly selected from each group and their tweets were gathered using the Twitter REST API. Retweets and non-Dutch tweets (as explained in the next

section) were left out of the collection. Users with less than 500 Dutch tweets were discarded. From the remaining users, a thousand were randomly selected per class for further use in this research. They constitute the dataset $\mathcal{D}_I$.

Political preference: $\mathcal{D}_P$ Tweets from September 2015 were used to gather users from whom the political preference can be retrieved. The data from this month is used because the Dutch government traditionally presents the next years' budget and policy in the month of September.

From all users that tweeted in September (1,242,805 users), a search is done based on the users' profile description. If this description contains one of the political parties active in the Netherlands (VVD; PvdA; PVV; CDA; SP; GroenLinks; D66; ChristenUnie; Partij voor de Dieren; 50PLUS; SGP; VNL; DENK) it was added to the list of users. Variations in notation of the different parties were taken into account during the search process (note that negation was not explicitely handled). This user-finding step resulted in a list of 16,977 Dutch Twitter users. This list was filtered so there were no users left that mentioned more than one party in their profile description. Parties represented by less than 500 users were excluded from further research. This filtering resulted in a list of 7,284 users that stated something about one of these seven parties in their profile description: VVD; PvdA; CDA; SP; GroenLinks; D66; ChristenUnie.

Finally, from the list of users we downloaded their most recent 500 tweets that were available at the current time of the project. Users with less than 500 tweets were excluded from the corpus. When extracting the tweets from Twitter, retweets were excluded. Beside that, tweets were checked for being Dutch using the Python module *langid*.[3] Non-Dutch tweets were excluded from the final corpus. This way of collecting the data results in corpus $\mathcal{D}_P$ of 3,802 users with 1,901,000 tweets.

We sampled parties who had at least 400 users, resulting in a balanced dataset contained five classes: VVD, PvdA, CDA, GroenLinks and D66.[4] This dataset of 2,000 users constitutes the dataset $\mathcal{D}_P$. We use two setups for this task, as described in Section 3.1. The distribution of the final corpus is given in Table 1.

Evaluation of distantly supervised data creation To evaluate the distant supervision method of corpus annotation, we manually annotated a random sample for each task and calculated the accuracy of the annotation.

For income prediction, a random sample of 100 users per class was manually annotated by one of the authors of this article. The labels were considered correct if they appeared in the biography of a user, the user was a human and the occupational title was used to indicate a paying occupation, not a hobby or study. The accuracy over the whole group of 200 users was 74.5 percent, with 70 percent in the low class and 79 percent in the high class. In 17 cases, the labels were wrong because the account was simply not used by a person but by a company. As there is no reliable way to distinguish between human and non-human users, we disregard these cases. The overall accuracy without these cases had been 81.4 percent. Finally, in four miscellaneous cases users described an internship or former occupation in a non-trivially detectable way. These results confirm that our distant supervision method yields viable training data for income prediction, even though it is far from perfect. Possible future improvements of the method will be discussed in Section 5.

For political preference, 500 users were randomly selected and manually annotated by another of the authors of this article. The evaluator annotated each profile description by choosing one out of six labels: VVD; PvdA; CDA; GL; D66; Niet Duidelijk (Not Clear). Only 8 out 500 cases were not correctly classified by the distant supervision technique, resulting in indecisive cases. There were no confusions between parties. Thus for political preference the estimated accuracy is 98.4%. Examples of non clear-cut profile descriptions that were assigned the label *Niet Duidelijk* can be found in Figure 2 (right). The first example shows a user that is a journalist. The journalist says he or she writes especially about social affairs, finance and the CDA. So it is not clear what the political preference of this user is. In the second example, a user says that he or she is a liberal

3. `https://www.github.com/saffsd/langid.py`
4. We leave experimentation with the original distribution of parties for future work.

and/or in favour of D66. It is not clear if his or her political preference is D66, it could also be a liberal party like the VVD. The third example shows the abbreviation GL, that stands for the political party GroenLinks. In the profile description, this abbreviation is used to state a zip code, not the political party. The last profile description once again shows the abbreviation GL. In this example, the abbreviation is used to indicate the initials of a user and not the political party.

Figure 2: Twitter user profile descriptions that were assigned the label *Niet Duidelijk* (unclear) for occupation (left) and political preference (right).

3. Method

This section describes the setup for the predictive models. In all experiments we use a logistic regression classifier implemented in `sklearn` (Pedregosa et al. 2011) with default parameters.[5] We use 10-fold cross-validation throughout the experiments.

3.1 Preprocessing

After non-Dutch tweets and retweets were removed, the corpora were preprocessed as follows. URLs, hashtags and usernames are removed (in the income case) and normalized to a common token (for political preference prediction). For political party prediction, very predictive features such as party names, party member names and city names were additionally normalized. Tweets were then tokenized with the NLTK TweetTokenizer (Bird et al. 2009) in both cases. As for income prediction we also needed sentence-based measures for some of the features, therefore $\mathcal{D}_I$ was also sentence split by the Dutch model of the NLTK sentence tokenizer.

For income prediction we use 500 tweets per user as data instance. For political preference prediction we have two setups: classification using 500 tweets per user and using only 100 tweets, to inspect the impact on prediction of the available amount of data per user.

3.2 Features

For political preference prediction we examine the following features: word unigrams, word bigrams, and word trigrams; as well as their combinations. Details can be found in Table 3.2. They are all used as binary indicator features.

For income prediction we use three sets of features inspired by recent work (Flekova et al. 2016): surface, readability and word n-gram features.[6] All individual features can be found in Table 3.2. The

5. This choice was motivated by the fact that we have many features and fewer instances, in which a simple model like logistic regression typically shows robust results.
6. Originally, we planned to implement all four feature groups used in Flekova et al. (2016). Due to time constraints, we could only implement surface and readability features. To compensate for the lack of features, we added a one extra group: word n-grams.

readability metrics were derived using their respective implementations in the readability library (van Cranenburgh 2016). They all have some commonality in the way they are calculated, but differ in measuring scale and intended application. The word N-grams were used as binary indicator features.

	Income ($\mathcal{R}_I$)	**Political preference** ($\mathcal{R}_P$)
Feature group	*Individual features*	
Surface	Average word length	-
	Length of a user's tweets in characters	
	Length of a user's tweets in words	
	Ratio of words longer than 5 characters	
	Type-token ratio	
Readability	Automated Readability Index	-
	Coleman-Liau Index	
	Flesch-Kincaid Grade Level	
	Flesch Reading Ease	
	Gunning-Fog Index	
	LIX Index	
	SMOG Index	
Word N-grams	Unigrams	Unigrams
	Bigrams	Bigrams
	Trigrams	Trigrams
	Unigrams and bigrams	Unigrams and bigrams
	Unigrams and trigrams	Unigrams and trigrams
	Bigrams and trigrams	Bigrams and trigrams
	Unigrams, bigrams and trigrams	Unigrams, bigrams and trigrams

Table 2: Features used for income and political preference.

4. Results

Political preference ($\mathcal{R}_P$) To interpret the results of the created classifier, a random baseline is calculated. The random baseline for political preference prediction results in an accuracy of 20%.

In Table 3 the accuracy of the different setups are summarized. More data per users results in better results. The accuracies of the setup with 500 tweets per user is higher for every feature or combination of features than the setup with fewer tweets per user, as can be seen in Table 3. In the setup with 100 tweets per user, the highest accuracy is 48%. The lowest accuracy with the setup with the fewest tweets is 35%. In the setup with 500 tweets per user, the accuracy is in all cases higher than 50%. The highest accuracy is 66% and comes down to 50%. The classifier performs better than the random baseline of 20%.

The unigram features result in the best representation. In the setup with 100 tweets per user, bigram features add some accuracy as well. Unigrams and bigrams combined ensure the highest accuracy for the setup with the fewest tweets. In the setup with 500 tweets per user, bigram features did not add any value to the unigram features. The unigrams ensure the highest accuracy in the setup with the most data. Bigrams or trigrams separately did not ensure a high accuracy. Likewise, the combination of trigrams with other features didn't ensure a high accuracy. Only unigrams and the combination of unigrams and bigrams, as described in the earlier case, ensured a high accuracy.

The breakdown per party is given in Table 4 for the best feature setup (unigrams). GroenLinks is the party which was identified most accurately.

Features	100 tweets per user	500 tweets per user
Baseline	0.20	0.20
Unigrams	0.46	**0.66**
Bigrams	0.41	0.58
Trigrams	0.35	0.50
Unigrams and bigrams	**0.48**	0.64
Unigrams and trigrams	0.47	0.63
Bigrams and trigrams	0.41	0.57
Unigrams, bigrams and trigrams	0.47	0.62

Table 3: Results of the **political preference** $\mathcal{R}_P$ classifier using *10-Fold Cross Validation*.

	CDA	D66	GL	PvdA	VVD
Precision	0.66	0.58	0.74	0.63	0.68
Recall	0.68	0.55	0.75	0.61	0.72
F1-Score	0.67	0.56	0.75	0.62	0.7
Accuracy	0.66				

Table 4: Precision, Recall, F1-Score and Accuracy of the setup with 500 tweets per user and unigram features using *10-Fold Cross Validation*.

Income level ($\mathcal{R}_I$) In the case of income level prediction, the random baseline is 0.50 due to the two-class setup. A number of different feature combinations were tested with 10-fold cross validation. The most important outcomes are included in Table 4. All setups outperformed the baseline, but the extent differs quite a lot.

Two feature groups resulted in an F1 score of 0.72. These are *unigrams and bigrams* and *unigrams, bigrams and trigrams*. This created the need for a way to distinguish among them. We selected the most robust combination by analysing the standard deviation of the F1 scores across the ten folds of the validation for each group. The analysis highlighted the setup with unigrams, bigrams and trigrams combined as the most robust method.

In contrast to prior work (Flekova et al. 2016), adding readability features did surprisingly not further help in our setup. We leave investigating reasons for this to future work.

Features	F1-score
Baseline	0.50
Surface	0.56
Readability	0.57
Word N-grams (n=1)	0.70
Word N-grams (n=1; 2)	0.72
Word N-grams (n=1; 2; 3)	**0.72**

Table 5: Results of the **income level** $\mathcal{R}_I$ classifier.

5. Discussion

In this section we discuss which features were most predictive for the respective tasks, and discuss limitations of the current work.

Feature Analysis: Political preference For the best setup, we analyzed the 100 most informative features (highest absolute coefficients, either positively or negatively loaded). These informative features are gathered from the setup with 500 tweets per user and unigrams as features.

In this paragraph, the most positive and negative features of the *VVD class* are described. The features of this class are described because they were the most indicative ones. A word cloud[7] with the most positive informative features can be seen in Figure 3. The most negative informative features are presented in Figure 4.

The *VVD* is a liberal (*liberaal*) party that supports a small government with as little as possible rules (*regels*). Interesting is that the users who are classified in the *VVD* class, talk about the left-wing term *linkse*. It could be that they talk about it frequently in an opposing manner. Other typical right wing *VVD* topics that can be identified in the feature list, are *defensie* (defence) and *politie* (police). The *VVD* party propagated supporting entrepreneurs (*ondernemers*). The positive features in the word cloud below, do support this. An example is the feature *OZB*, that stands for *Onroerend Zaak Belasting* (Literaly translated: real estate tax). The *VVD* wants the taxes (*belastingen* and costs (*lasten*) to be low, especially the *OZB*, which is a tax for entrepreneurs. It is interesting to see that the stereotype of *VVD* supporters, people who like spending money and living a cosy and luxurious life, is supported by some features: *hapje* (snack), *euro* (euro), *shoppen* (shopping) and *lunchen* (lunch) and *café* (pub).

In the negative feature list, features that support other political parties can be recognised. *Vluchtelingen* (refugees), *duurzaamheid* (sustainability), *sociaal* (social), *energie* (energy), *geloof* (religion), *scholen* (schools), *armoede* (poverty), *milieu* (environment) and *samenleving* (society) are features that are more indicative for the other classes. These features are supported more by left-wing parties. The feature *gas* (gas) is interesting to see, because in certain areas of The Netherlands, there are frequent earthquakes because of the gas extraction in these areas. The political party *VVD* has always supported the gas extraction because of the economical benefits. That this feature is in the negative list, supports that users classified as *VVD* do not tweet much about this topic. The unigram *jongeren* (youth) could be in the negative feature list, because the *VVD* is not known as a party for the youth. A party like *D66* is more known as a party for this audience.

Figure 3: Most positive informative features of the *VVD class* in the political preference data.

Feature Analysis: Income level As the income level classifier used the same classification algorithm, we can also get an insight in the most predictive features for both high and low income classes. A full overview can be found in Table 6.

7. Word clouds are created using the tool http://www.wordle.net

Figure 4: Most negative informative features of the *VVD class* in the political preference data.

For the low class, words like *training* and the informal, third person singular form of *you* ("je") stand out when looking at the cross validation results. In the high class, words like *heard* ("gehoord"), *happy with* ("blij met") and *books* ("boeken") turned out to have a high predictive value.

Table 6: Most informative features for both classes of the **income level** $\mathcal{R}_I$ classifier.

High class		**Low class**	
Original word	*Gloss*	*Original word*	*Gloss*
blij met	happy with	als je	if you
maar ook	but also	training	training
boeken	books	super	super
gehoord	heard	met je	with you
huh	huh	1 van	one of

Limitations An important limitation of the study on IncomePrediction is the level of detail available in the income data. Statistics Netherlands provides income levels only for classes of professions, not individual incomes of occupations. The so-obtained class average provide a coarse view only. Due to the aforementioned considerations and the fact that data gathering and annotation turned out be quite challenging for this task, we restricted ourself to a two class (high/low income). We originally planned to go for the six class setup given by the Statistics Netherlands scheme, leaving it as of now for future work.

Our distant supervision labeling strategy worked reasonably well, as demonstrated in Section 2. However, it is far from perfect. One problem is that our method cannot distinguish between human and non-human users. Although this is such a complex topic that it would warrant a separate study, adding this capability would help to reduce the noise in our dataset. Another problem is that our labeling method relies on just one heuristic: it will always take the first occupation it can find, even if there are multiple occupations mentioned by a user. It currently has no way to disambiguate between multiple occupations. For PoliticalPreference prediction, it turns out that there is a bias in our dataset, i.e., an overrepresentation of actual politicians. After manual inspection we discovered that of the users that state their political preference many are politicians, which results in

a skewed distribution. Future work on political affiliation prediction could extend the data gathering step by looking at actual tweets[8] rather than only user profiles.

6. Related Work

The studies that come closest to our approach are that of Sylwester and Purver (2015) and Flekova et al. (2016).

Sylwester and Purver (2015) investigated psychological differences between Twitter users of different political orientations. They hypothesized that the language used by liberals emphasizes their perception of uniqueness, contains more swear words, more anxiety-related words and more feeling-related words than conservatives' language. Conversely, they predicted that the language of conservatives hypothesized group membership and contains more references to achievement and religion than liberals' language. With respect to the language use on Twitter and politics, a lot of previous research has been carried out. In 2012, Tjong Kim Sang and Bos (2012) used Twitter to predict the 2011 Dutch senate election results, achieving results that were close to the actual election outcomes. Tumasjan et al. (2010) did a similar study the German federal elections. They also found that Twitter messages plausibly reflect the offline political landscape. Beside predicting elections, Wang et al. (2012) created a system for real-time analysis of public sentiment towards the presidential candidates in the 2012 U.S. election as expressed on Twitter. The system they created offered a new and timely perspective on the dynamics of the electoral process and public opinion to the civilians, the media, politicians and scholars. Also, the use of Twitter for campaigning was shown by Enli and Skogerbø (2013). In this paper, they concluded that Norwegian politicians were using social media platforms like Facebook and Twitter for marketing purposes.

Research into author attributes beyond demographic variables is relatively new, especially in the field of income prediction. All prior work in that direction has focused on English-speaking Twitter users. Related work on non-English contexts mostly focuses on age and gender, and is relatively recent (van Halteren and Speerstra 2014, van Halteren 2015, Verhoeven et al. 2017, Ljubešić et al. 2017), to name but a few.

The most recent and relevant 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 attribute, a separate data set was used. Flekova et al. codified stylistic variation into a number of features, which were grouped into four categories: surface, readability, syntax and style. After performing a ten-fold cross validation with both linear and non-linear regression methods, they discovered that readability metrics like the Flesch Reading Ease metric and the relative use of pronouns correlated stronger with income than age. 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 Preoţiuc-Pietro et al. (2015a). They used the corpus to classify users according to their occupational class. The occupational titles and classes used were gathered from the UK Standard Occupational Classification (SOC) (Office for National Statistics 2010). The SOC is a hierarchical classification of occupations. It consists of four levels, starting with nine very general classes and terminating in hundreds of very specific classes. Each level is indicated with a different number of digits. The coarsest level is indicated with one digit and the finest level with four digits (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 the Twitter REST API was used to find at most 200 users for each occupation. The accumulated users were divided into the three-digit groups that they belong to. Users that were companies, had no description or had a contradicting description, were removed

8. Or hashtags, as done by a very recent study that uses hashtags to identify general political direction (Tatman et al. 2017).

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.

Rangel and Rosso (2013) also studied the relation between age, gender and stylistic features of users. They presented a way to identify age and gender of authors based on their use of language, using an SVM-based approach. Argamon et al. (2009) describes how to know as much as possible about an anonymous author of a text using different features.

7. Conclusions

This study has been a first exploration of the possibilities of profiling Dutch authors on their income and their political preference on the basis of their tweets. Two novel corpora were collected ($\mathcal{D}_P$ and $\mathcal{D}_I$, for PoliticalPreference and IncomePrediction, respectively).

Distant supervision was a surprisingly accurate method for weak data annotation. We estimated the accuracy of the method for both tasks, achieving an accuracy of 98.4% for political preference classification and 74.5% for profession annotation.

Our results show that both user attributes can be predicted relatively well, considerably above baseline. For political preference classification our best setup achieves an accuracy of 66% (in contrast, the random baseline achieves 20%). While this is better than gambling, it is not accurate enough yet to rely on the classifier for downstream applications. For income-class prediction, the best setup reaches an F1-score of 72%. Readability measures turned out to be less informative, in contrast to prior findings (Flekova et al. 2016). The classifier based on solely word unigrams, bigrams and trigrams combined proves the most robust, providing the highest average F1-score with the lowest standard deviation for income prediction.

The datasets created in this study are a starting point in the field of automatically classifying Dutch Twitter users based on political preference or income using their tweets. We hope that this work stimulates further research in this direction, examining additional features (e.g., syntactic features) and gathering more data in combination with alternative labeling strategies (e.g., including labeling tweets).

While author profiling has potential benefits in tailoring customer services or political agendas, it is important to be aware of potential social implications of such methods (Hovy and Spruit 2016), since the step to the dark side is often not far, given the sheer amounts of data available these days.

References

Alowibdi, Jalal S, Ugo A Buy, and Philip Yu (2013), Empirical evaluation of profile characteristics for gender classification on twitter, *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, Vol. 1, IEEE, pp. 365–369.

Argamon, Shlomo, Moshe Koppel, James W Pennebaker, and Jonathan Schler (2009), Automatically profiling the author of an anonymous text, *Communications of the ACM* **52** (2), pp. 119–123, ACM.

Bird, Steven, Ewan Klein, and Edward Loper (2009), *Natural Language Processing with Python*, O'Reilly Media.

Ciot, Morgane, Morgan Sonderegger, and Derek Ruths (2013), Gender inference of twitter users in non-english contexts, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 1136–1145. http://aclweb.org/anthology/D13-1114.

Eisenstein, Jacob, A. Noah Smith, and P. Eric Xing (2011), Discovering sociolinguistic associations with structured sparsity, *Proceedings of the 49th Annual Meeting of the Association for Compu-*

tational Linguistics: Human Language Technologies, Association for Computational Linguistics, pp. 1365–1374. http://aclweb.org/anthology/P11-1137.

Enli, Gunn Sara and Eli Skogerbø (2013), Personalized campaigns in party-centred politics: Twitter and facebook as arenas for political communication, *Information, Communication & Society* **16** (5), pp. 757–774, Taylor & Francis.

European Observatory of Working Life (2015), *Developments in collectively agreed working time 2014*.

Flekova, Lucie, Lyle Ungar, and Daniel Preotiuc-Pietro (2016), Exploring stylistic variation with age and income on twitter, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL*, pp. 313–319.

Go, Alec, Richa Bhayani, and Lei Huang (2009), Twitter sentiment classification using distant supervision, *CS224N Project Report, Stanford* **1**, pp. 12.

Hovy, Dirk and L. Shannon Spruit (2016), The social impact of natural language processing, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, pp. 591–598. http://aclweb.org/anthology/P16-2096.

International Labor Office (2013), *International Standard Classification of Occupations 2008 (ISCO-08)*, International Labor Office. https://www.amazon.com/International-Standard-Classification-Occupations-ISCO-08/dp/9221259528

Ljubešić, Nikola, Darja Fišer, and Tomaž Erjavec (2017), Language-independent gender prediction on twitter, *Proceedings of the Second Workshop on NLP and Computational Social Science*, Association for Computational Linguistics, Vancouver, Canada, pp. 1–6. http://www.aclweb.org/anthology/W17-2901.

Luyckx, Kim and Walter Daelemans (2008), Personae: a corpus for author and personality prediction from text, *LREC 2008*. http://aclweb.org/anthology/L08-1030.

Mairesse, François and Marilyn Walker (2006), Automatic recognition of personality in conversation, *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. http://aclweb.org/anthology/N06-2022.

Mirkin, Shachar, Scott Nowson, Caroline Brun, and Julien Perez (2015), Motivating personality-aware machine translation, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Lisbon, Portugal, pp. 1102–1108. http://aclweb.org/anthology/D15-1130.

Nguyen, Dong, A. Noah Smith, and P. Carolyn Rosè (2011), *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Association for Computational Linguistics, chapter Author Age Prediction from Text using Linear Regression, pp. 115–123. http://aclweb.org/anthology/W11-1515.

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

Pak, Alexander and Patrick Paroubek (2010), Twitter as a corpus for sentiment analysis and opinion mining., *LREc*, Vol. 10, pp. 1320–1326.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011), Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12**, pp. 2825–2830.

Plank, Barbara and Dirk Hovy (2015), Personality Traits on Twitter—Or—How to Get 1,500 Personality Tests in a Week, *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Association for Computational Linguistics, Lisboa, Portugal, pp. 92–98. http://aclweb.org/anthology/W15-2913.

Preoţiuc-Pietro, Daniel, Vasileios Lampos, and Nikolaos Aletras (2015a), An analysis of the user occupational class through twitter content, *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)*, Association for Computational Linguistics, pp. 1754–1764. http://aclweb.org/anthology/P15-1169.

Preotiuc-Pietro, Daniel, Vasileios Lampos, and Nikolaos Aletras (2015b), An analysis of the user occupational class through twitter content, *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)*, Association for Computational Linguistics, Beijing, China, pp. 1754–1764. http://www.aclweb.org/anthology/P15-1169.

Rabinovich, Ella, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner (2017), Personalized machine translation: Preserving original author traits, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Association for Computational Linguistics, Valencia, Spain, pp. 1074–1084.

Rangel, Francisco and Paolo Rosso (2013), Use of language and author profiling: Identification of gender and age, *Natural Language Processing and Cognitive Science* p. 177.

Rao, Delip, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta (2010), Classifying latent user attributes in twitter, *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, ACM, pp. 37–44.

Read, Jonathon (2005), Using emoticons to reduce dependency in machine learning techniques for sentiment classification, *Proceedings of the ACL Student Research Workshop*, ACLstudent '05, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 43–48. http://dl.acm.org/citation.cfm?id=1628960.1628969.

Rosenthal, Sara and Kathleen McKeown (2011), Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, pp. 763–772. http://aclweb.org/anthology/P11-1077.

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.

Sylwester, Karolina and Matthew Purver (2015), Twitter language use reflects psychological differences between democrats and republicans, *PloS one* **10** (9), pp. e0137422, Public Library of Science.

Tatman, Rachael, Leo Stewart, Amandalynne Paullada, and Emma Spiro (2017), Non-lexical features encode political affiliation on twitter, *Proceedings of the Second Workshop on NLP and Computational Social Science*, pp. 63–67.

Tjong Kim Sang, Erik and Johan Bos (2012), Predicting the 2011 dutch senate election results with twitter, *Proceedings of the Workshop on Semantic Analysis in Social Media*, Association for Computational Linguistics, pp. 53–60.

Tumasjan, Andranik, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe (2010), Predicting elections with twitter: What 140 characters reveal about political sentiment., *ICWSM* **10**, pp. 178–185.

van Cranenburgh, Andreas (2016), Readability. Retrieved from https://github.com/andreasvc/readability. https://github.com/andreasvc/readability.

van Halteren, Hans (2015), Metadata induction on a dutch twitter corpus: Initial phases, *Computational Linguistics in the Netherlands Journal* **5**, pp. 37–48.

van Halteren, Hans and Nander Speerstra (2014), Gender recognition on dutch tweets, *Computational Linguistics in the Netherlands Journal* **4**, pp. 171–190.

Verhoeven, Ben, Iza Škrjanec, and Senja Pollak (2017), Gender profiling for slovene twitter communication: The influence of gender marking, content and style, *BSNLP 2017* p. 119.

Verhoeven, Ben, Walter Daelemans, and Barbara Plank (2016), Twisty: A multilingual twitter stylometry corpus for gender and personality profiling., *LREC*.

Volkova, Svitlana, Theresa Wilson, and David Yarowsky (2013), Exploring demographic language variations to improve multilingual sentiment analysis in social media, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 1815–1827. http://aclweb.org/anthology/D13-1187.

Volkova, Svitlana, Yoram Bachrach, Michael Armstrong, and Vijay Sharma (2015), Inferring latent user properties from texts published in social media (demo), *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI)*, Austin, TX.

Wang, Hao, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan (2012), A system for real-time twitter sentiment analysis of 2012 us presidential election cycle, *Proceedings of the ACL 2012 System Demonstrations*, Association for Computational Linguistics, pp. 115–120.