Computational Linguistics in the Netherlands Journal 4 (2014) 171-190 Submitted 06/2014; Published 12/2014

Gender Recognition on Dutch Tweets

Hans van Halteren HVH@LET.RU.NL
Nander Speerstra NANDERSPEERSTRA@LIVE.NL

Radboud University Nijmegen, CLS, Linguistics

Abstract

In this paper, we investigate gender recognition on Dutch Twitter material, using a corpus consisting of the full Tweet production (as far as present in the TwiNL data set) of 600 users (known to be human individuals) over 2011 and 2012. We experimented with several authorship profiling techniques and various recognition features, using Tweet text only, in order to determine how well they could distinguish between male and female authors of Tweets. We achieved the best results, 95.5% correct assignment in a 5-fold cross-validation on our corpus, with Support Vector Regression on all token unigrams. Two other machine learning systems, Linguistic Profiling and TiMBL, come close to this result, at least when the input is first preprocessed with PCA.

1. Introduction

In the Netherlands, we have a rather unique resource in the form of the TwiNL data set: a daily updated collection that probably contains at least 30% of the Dutch public tweet production since 2011 (Tjong Kim Sang and van den Bosch 2013). However, as any collection that is harvested automatically, its usability is reduced by a lack of reliable metadata. In this case, the Twitter profiles of the authors are available, but these consist of freeform text rather than fixed information fields. And, obviously, it is unknown to which degree the information that is present is true. The resource would become even more useful if we could deduce complete and correct metadata from the various available information sources, such as the provided metadata, user relations, profile photos, and the text of the tweets. In this paper, we start modestly, by attempting to derive just the gender of the authors[1] automatically, purely on the basis of the content of their tweets, using author profiling techniques.

For our experiment, we selected 600 authors for whom we were able to determine with a high degree of certainty a) that they were human individuals and b) what gender they were. We then experimented with several author profiling techniques, namely Support Vector Regression (as provided by LIBSVM; (Chang and Lin 2011)), Linguistic Profiling (LP; (van Halteren 2004)), and TiMBL (Daelemans et al. 2004), with and without preprocessing the input vectors with Principal Component Analysis (PCA; (Pearson 1901); (Hotelling 1933)). We also varied the recognition features provided to the techniques, using both character and token n-grams. For all techniques and features, we ran the same 5-fold cross-validation experiments in order to determine how well they could be used to distinguish between male and female authors of tweets.

In the following sections, we first present some previous work on gender recognition (Section 2). Then we describe our experimental data and the evaluation method (Section 3), after which we proceed to describe the various author profiling strategies that we investigated (Section 4). Then follow the results (Section 5), and Section 6 concludes the paper.

1. For whom we already know that they are an individual person rather than, say, a husband and wife couple or a board of editors for an official Twitterfeed.

2. Gender Recognition

Gender recognition is a subtask in the general field of authorship recognition and profiling, which has reached maturity in the last decades (for an overview, see e.g. (Juola 2008) and (Koppel et al. 2009)). Currently the field is getting an impulse for further development now that vast data sets of user generated data is becoming available. Narayanan et al. (2012) show that authorship recognition is also possible (to some degree) if the number of candidate authors is as high as 100,000 (as compared to the usually less than ten in traditional studies). Even so, there are circumstances where outright recognition is not an option, but where one must be content with profiling, i.e. the identification of author traits like gender, age and geographical background. In this paper we restrict ourselves to gender recognition, and it is also this aspect we will discuss further in this section.

A group which is very active in studying gender recognition (among other traits) on the basis of text is that around Moshe Koppel. In (Koppel et al. 2002) they report gender recognition on formal written texts taken from the British National Corpus (and also give a good overview of previous work), reaching about 80% correct attributions using function words and parts of speech. Later, in 2004, the group collected a Blog Authorship Corpus (BAC; (Schler et al. 2006)), containing about 700,000 posts to blogger.com (in total about 140 million words) by almost 20,000 bloggers. For each blogger, metadata is present, including the blogger's self-provided gender, age, industry and astrological sign. This corpus has been used extensively since. The creators themselves used it for various classification tasks, including gender recognition (Koppel et al. 2009). They report an overall accuracy of 76.1%. Slightly more information seems to be coming from content (75.1% accuracy) than from style (72.0% accuracy). However, even style appears to mirror content. We see the women focusing on personal matters, leading to important content words like *love* and *boyfriend*, and important style words like *I* and other personal pronouns. The men, on the other hand, seem to be more interested in computers, leading to important content words like *software* and *game*, and correspondingly more determiners and prepositions. One gets the impression that gender recognition is more sociological than linguistic, showing what women and men were blogging about back in 2004. A later study (Goswami et al. 2009) managed to increase the gender recognition quality to 89.2%, using sentence length, 35 non-dictionary words, and 52 slang words. The authors do not report the set of slang words, but the non-dictionary words appear to be more related to style than to content, showing that purely linguistic behaviour can contribute information for gender recognition as well.

Gender recognition has also already been applied to Tweets. Rao et al. (2010) examined various traits of authors from India tweeting in English, combining character N-grams and sociolinguistic features like manner of laughing, honorifics, and smiley use. With lexical N-grams, they reached an accuracy of 67.7%, which the combination with the sociolinguistic features increased to 72.33%. Burger et al. (2011) attempted to recognize gender in tweets from a whole set of languages, using word and character N-grams as features for machine learning with Support Vector Machines (SVM), Naive Bayes and Balanced Winnow2. Their highest score when using just text features was 75.5%, testing on all the tweets by each author (with a train set of 3.3 million tweets and a test set of about 418,000 tweets).[2] Fink et al. (2012) used SVMlight to classify gender on Nigerian twitter accounts, with tweets in English, with a minimum of 50 tweets. Their features were hash tags, token unigrams and psychometric measurements provided by the Linguistic Inquiry of Word Count software (LIWC; (Pennebaker et al. 2007)). Although LIWC appears a very interesting addition, it hardly adds anything to the classification. With only token unigrams, the recognition accuracy was 80.5%, while using all features together increased this only slightly to 80.6%. Bamman et al. (2014) examined about 9 million tweets by 14,000 Twitter users tweeting in American English. They used lexical features, and present a very good breakdown of various word types. When using all user tweets, they reached an accuracy of 88.0%. An interesting observation is that there is a clear class of misclassified users who have a majority of opposite gender users in their social network.

2. When adding more information sources, such as profile fields, they reach an accuracy of 92.0%.

For Tweets in Dutch, we first look at the official user interface for the TwiNL data set, http://www.twiqs.nl. Among other things, it shows gender and age statistics for the users producing the tweets found for user specified searches. These statistics are derived from the users' profile information by way of some heuristics. For gender, the system checks the profile for about 150 common male and 150 common female first names, as well as for gender related words, such as *father*, *mother*, *wife* and *husband*. If no cue is found in a user's profile, no gender is assigned. The general quality of the assignment is unknown, but in the (for this purpose) rather unrepresentative sample of users we considered for our own gender assignment corpus (see below), we find that about 44% of the users are assigned a gender, which is correct in about 87% of the cases. Another system that predicts the gender for Dutch Twitter users is TweetGenie (http://www.tweetgenie.nl), that one can provide with a Twitter user name, after which the gender and age are estimated, based on the user's last 200 tweets. The age component of the system is described in (Nguyen et al. 2013). The authors apply logistic and linear regression on counts of token unigrams occurring at least 10 times in their corpus. The paper does not describe the gender component, but the first author has informed us that the accuracy of the gender recognition on the basis of 200 tweets is about 87% (Nguyen, personal communication).[3]

In later experiments, Nguyen et al. (2014) did a crowdsourcing experiment, in which they asked human participants to guess the gender and age on the basis of 20 to 40 tweets. When using a majority vote to represent the crowd's opinion, the crowd's perception of the gender on the basis of the tweets coincided with the actual gender in about 84% of the cases. The conclusion is not so much, however, that humans are also not perfect at guessing age on the basis of language use, but rather that there is a distinction between the biological and the social identity of authors, and language use is more likely to represent the social one (cf. also (Bamman et al. 2014)). Although we agree with Nguyen et al. on this, we will still take the biological gender as the gold standard in this paper, as our eventual goal is creating metadata for the TwiNL collection.

3. Experimental Data and Evaluation

In this section, we first describe the corpus that we used in our experiments (Section 3.1). Then we outline how we evaluated the various strategies (Section 3.2).

3.1 Corpus Used in the Experiments

We selected our experimental material from the TwiNL data set (Tjong Kim Sang and van den Bosch 2013), which was collected by searching for tweets with any of a number of probably Dutch words, after which a character n-gram language filter was applied. The collection is estimated to contain 30-40% of all public Dutch tweets. From this material, we considered all tweets with a date stamp in 2011 and 2012. In all, there were about 23 million users present. Of these, we only considered the ones who produced 2 to 10 tweets on average per day over 2011 and 2012. The minimum ensured a sufficient amount of text (1500-7300 tweets) for classification; the maximum served to avoid very high volume users, who might be professional (multi-user/edited) feeds or even twitterbots. This restriction brought the number of users down to about 270,000.

We then progressed to the selection of individual users. We aimed for 600 users. We selected 500 of these so that they get a gender assignment in TwiQS, for comparison, but we also wanted to include unmarked users in case these would be different in nature. All users, obviously, should be individuals, and for each the gender should be clear. From the about 120,000 users who are assigned a gender by TwiQS, we took a random selection in such a manner that the volume distribution (i.e. from 2 to 10 tweets per day average) is equally spread throughout the range and approximately equal for men and women. We checked gender manually for all selected users, mostly on the basis

3. As in our own experiment, this measurement is based on Twitter accounts where the user is known to be a human individual.

of the profile texts and profile photo's, and only included those for which we were convinced of the gender.[4] Later even more detailed rechecks, after a few extremely unlikely classification results, served to clean up the (hopefully) last gender assignment errors.[5] The final corpus is not completely balanced for gender, but consists of the production of 320 women and 280 men. However, as research shows a higher number of female users in all as well (Heil and Piskorski 2009), we do not view this as a problem.

From each user's tweets, we removed all retweets, as these did not contain original text by the author. Then, as several of our features were based on tokens, we tokenized all text samples, using our own specialized tokenizer for tweets. Apart from normal tokens like words, numbers and dates, it is also able to recognize a wide variety of emoticons. The tokenizer is able to identify hashtags and Twitter user names to the extent that these conform to the conventions used in Twitter, i.e. the hash (#) resp. at (@) sign are followed by a series of letters, digits and underscores. URLs and email addresses are not completely covered. The tokenizer counts on clear markers for these, e.g. http, www or one of a number of domain names for URLs. Assuming that any sequence including periods is likely to be a URL proves unwise, given that spacing between normal words is often irregular. And actually checking the existence of a proposed URL was computationally infeasible for the amount of text we intended to process. Finally, as the use of capitalization and diacritics is quite haphazard in the tweets, the tokenizer strips all words of diacritics and transforms them to lower case.

3.2 Evaluation

We divided our corpus in five parts, each containing (approximately) the same number of male and female authors.[6] We used this division in all experiments, each time using four parts as training material and one as test material. For those techniques where hyperparameters need to be selected, we used a leave-one-out strategy on the test material. For each test author, we determined the optimal hyperparameter settings with regard to the classification of all other authors in the same part of the corpus, in effect using these as development material.

In this way, we derived a classification score for each author without the system having any direct or indirect access to the actual gender of the author. We then measured for which percentage of the authors in the corpus this score was in agreement with the actual gender. These percentages are presented below in Section 5.

4. Profiling Strategies

In this section, we describe the strategies that we investigated for the gender recognition task. As we approached the task from a machine learning viewpoint, we needed to select text features to be provided as input to the machine learning systems, as well as machine learning systems which are to use this input for classification. We first describe the features we used (Section 4.1). Then we explain how we used the three selected machine learning systems to classify the authors (Section 4.2).

4.1 Machine Learning Features

We restricted ourselves to lexical features for our experiments. The use of syntax or even higher level features is (for now) impossible as the language use on Twitter deviates too much from standard Dutch, and we have no tools to provide reliable analyses. However, even with purely lexical features,

4. On the examined users, the gender assignment of TwiQS proved about 87% correct.
5. Several errors could be traced back to the fact that the account had moved on to another user since 2012.
6. We could have used different dividing strategies, but chose balanced folds in order to give a equal chance to all machine learning techniques, also those that have trouble with unbalanced data. If, in any application, unbalanced collections are expected, the effects of biases, and corrections for them, will have to be investigated.

there are still various options from which to choose. Most of them rely on the tokenization described above. We will illustrate the options we explored with the tweet

> *@ANONYMISED Hahaha...ik geloof dat meneer Bee heeft ingezet op een plan vóór het slapen ;-))ng*
>
> <@> hahaha ... I believe that mister B has opted for a plan before sleeping <smiley> ng

which after preprocessing becomes

> *<@> hahaha ... ik geloof dat meneer bee heeft ingezet op een plan voor het slapen <smiley> ng*

The first option for machine learning features is a traditional one.

Top 100 Function Words The most frequent function words (see (Kestemont 2014) for an overview). We used the 100 most frequent, as measured on our tweet collection, of which the example tweet contains the words *ik*, *dat*, *heeft*, *op*, *een*, *voor*, and *het*.

Then, we used a set of feature types based on token n-grams, with which we already had previous experience (Van Bael and van Halteren 2007). For all feature types, we used only those features which were observed with at least 5 authors in our whole collection (for skip bigrams 10 authors).

Unigrams Single tokens, similar to the top function words, but then using all tokens instead of a subset. About 47K features. In the example tweet, we find e.g. *ik*, *ingezet*, and *<smiley>*.

Bigrams Two adjacent tokens. About 265K features. In the example tweet, e.g. *heeft ingezet*, *slapen <smiley>*, and _ *<@>*, where the double underscore represents the start of the tweet.

Trigrams Three adjacent tokens. About 355K features. In the example tweet, e.g. *op een plan* and _ *<@> hahaha*.

Skip bigrams Two tokens in the tweet, but not adjacent, without any restrictions on the gap size. About 580K features. In the example tweet, e.g. *dat heeft* and *hahaha <smiley>*.

Finally, we included feature types based on character n-grams (following (Kjell et al. 1994)). We used the n-grams with n from 1 to 5, again only when the n-gram was observed with at least 5 authors. However, we used two types of character n-grams. The first set is derived from the tokenizer output, and can be viewed as a kind of normalized character n-grams.

Normalized 1-gram About 350 features. In the example tweet, e.g. *i* and *@*.

Normalized 2-gram About 4K features. In the example tweet, e.g. *ik* and twice *ng*.

Normalized 3-gram About 36K features. In the example tweet, e.g. *gez* and *n_v*, where the underscore represents a space.

Normalized 4-gram About 160K features. In the example tweet, e.g. *slap* and *op_e*.

Normalized 5-gram About 420K features. In the example tweet, e.g. *ingez* and __<@>_, now with a double underscore for the beginning of the tweet and a single underscore for a space.

The second set of character n-grams is derived from the original tweets. This type of character n-gram has the clear advantage of not needing any preprocessing in the form of tokenization.

Original 1-gram About 420 features. In the example tweet, e.g. *e* and *;*.

Original 2-gram About 8K features. In the example tweet, e.g. *Be* and *_@*.

Original 3-gram About 77K features. In the example tweet, e.g. *ing* and *))n*.

Original 4-gram About 260K features. In the example tweet, e.g. *plan* and *;-))*.

Original 5-gram About 580K features. In the example tweet, e.g. *r_Bee* and *a...i*.

4.2 Machine Learning Techniques

Having determined the features we would be working with, we next needed to select a machine learning system. Again, we decided to explore more than one option, but here we preferred more focus and restricted ourselves to three systems. Our primary choice for classification was the use of Support Vector Machines, viz. LIBSVM (Chang and Lin, 2001). We chose Support Vector Regression (ν-SVR to be exact) with an RBF kernel, as it had shown the best results in several research projects (e.g. (van Halteren 2008)). With these main choices, we performed a grid search for well-performing hyperparameters, with the following investigated values: the cost factor C is set to respectively 1/32, 1, 32, 1024, and 32768, γ to 1/4, 1/2, 1, 2 and 4 times LIBSVM's default of one divided by the number of features, and ν to 0.1, 0.3, 0.5 and 0.7.

The second classification system was Linguistic Profiling (LP; (van Halteren 2004)), which was specifically designed for authorship recognition and profiling. Roughly speaking, it classifies on the basis of noticeable over- and underuse of specific features. Before being used in comparisons, all feature counts were normalized to counts per 1000 words, and then transformed to Z-scores with regard to the average and standard deviation within each feature. LP has four hyperparameter settings, three of which weight the relative importance of each feature/dimension in the feature vector when comparing a text's feature vector to the profile vector (in this case the average of the feature vectors for all the training texts for a given gender), and one determining the threshold for feature Z-scores to be taken into account. Here the grid search investigated: the hyperparameter emphasizing the difference between text feature and profile feature to polynomial exponents set to 0.1, 0.4, 0.7, ..., 2.7 and 3; the hyperparameters for emphasizing text feature size to 0 or 1; the hyperparameter for emphasizing profile feature size to -1, 0, 1, and 2; and the threshold hyperparameter also to 0 or 1.

Finally, we added TiMBL (Daelemans et al. 2004), a k-nearest neighbour classification system, which is used extensively in-house for various machine learning tasks, but which we had so far not used for authorship tasks. As the input features are numerical, we used IB1 with k equal to 5 so that we can derive a confidence value. The only hyperparameters we varied in the grid search are the metric (Numerical and Cosine distance) and the weighting (no weighting, information gain, gain ratio, chi-square, shared variance, and standard deviation).

However, the high dimensionality of our vectors presented us with a problem. For such high numbers of features, it is known that k-NN learning is unlikely to yield useful results (Beyer et al. 1999). This meant that, if we still wanted to use k-NN, we would have to reduce the dimensionality of our feature vectors. We chose to use Principal Component Analysis (PCA; (Pearson 1901), (Hotelling 1933)).[7] And, now we had the principal component vectors, we decided also to provide them to SVR and LP. For each system, we provided the first N principal components for various N. In effect, this N is a further hyperparameter, which we varied from 1 to the total number of components (usually 600, as there are 600 authors), using a stepsize of 1 from 1 to 10, and then slowly increasing the stepsize to a maximum of 20 when over 300.

Rather than using fixed hyperparameters, we let the control shell choose them automatically in a grid search procedure, based on development data. When running the underlying systems

7. To be exact, we used the function *prcomp* in R (R Development Core Team 2008), with the instruction $scale = TRUE$ to force normalization of the vectors before the principal components were determined. As scaling is not possible when there are columns with constant values, such columns were removed first.

themselves, we used various hyperparameter settings, as listed above. For each setting and author, the systems report both a selected class and a floating point score, which can be used as a confidence score.[8] For each individual author, the control shell examined the scores for all other authors in the same fold.[9] It then calculated a class separation value, namely the difference between the mean scores for each of the two classes (male and female), divided by the sum of the two standard deviations.[10] The optimal hyperparameter settings are assumed to be those where the two classes are separated most, i.e. where the class separation value is highest. In order to improve the robustness of the hyperparameter selection, the best three settings were chosen and used for classifying the current author in question.

A final detail that we exploited is that SVR and LP are asymmetric in the modeling of the classes. For LP, this is by design. A model, called "profile", is constructed for each individual class, and the system determines for each author to which degree they are similar to the class profile. For SVR, one would expect symmetry, as both classes are modeled simultaneously, and differ merely in the sign of the numeric class identifier. However, we do observe different behaviour when reversing the signs. For this reason, we did all classification with SVR and LP twice, once building a male model and once a female model. For both models the control shell calculated a final score, starting with the three outputs for the best hyperparameter settings. It normalized these by expressing them as the number of non-model class standard deviations over the threshold, which was set at the class separation value. The control shell then weighted each score by multiplying it by the class separation value on the development data for the settings in question, and derived the final score by averaging. It then chose the class for which the final score is highest. In this way, we also get two confidence values, viz. the model score for the chosen class (how male/female the author writes) and the difference between the two scores (how much more female/male the author writes than male/female).

5. Results

In this section, we will present the overall results of the gender recognition. We start with the accuracy of the various features and systems (Section 5.1). Then we will focus on the effect of preprocessing the input vectors with PCA (Section 5.2). After this, we examine the classification of individual authors (Section 5.3 and the distinguishing power of features (Section 5.4).

5.1 Overall Quality

Table 1 shows the accuracy of the recognition, using the desccribed features and systems. For the systems, both SVR and LP are used with the original case vectors as well as with PCA preprocessing, where TiMBL, for reasons mentioned above, is used only with preprocessed vectors. For the measurements with PCA, the number of principal components provided to the classification system is learned from the development data. Below, in Section 5.2, we will examine what the systems are capable of at fixed numbers of principal components.

Starting with the systems, we see that SVR (using original vectors) consistently outperforms the other two. For only one feature type, character trigrams, LP with PCA manages to reach a higher accuracy than SVR, but the difference is not statistically significant. LP and TiMBL are closely matched, although LP appears to be slightly better when combined with PCA, but the next section will shed new light on this comparison. From the measurements here, we can conclude that LP profits from PCA preprocessing, but SVR is better off with the original vectors.

8. For SVR and LP, these are rather varied, but TiMBL's confidence value consists of the proportion of selected class cases among the nearest neighbours, which with k at 5 is practically always 0.6, 0.8, or 1.0.
9. This gives the best chances that the selected optimal hyperparameters generalize to the author in question.
10. The class separation value is a variant of Cohen's d (Cohen 1988). Where Cohen assumes the two distributions have the same standard deviation, we use the sum of the two, practically always different, standard deviations.

Table 1: Accuracy Percentages for various Feature Types and Techniques. For each feature type, the best percentage is bolded, and all percentages are italicized that are not statistically significantly different at the 5% level.

Feature type	Techniques				
	Support Vector Regression		Linguistic Profiling		TiMBL
	original	with PCA	original	with PCA	with PCA
Top 100 Function Words	**84.8**	*83.7*	76.7	84.7	75.8
Token Unigram	**95.5**	*94.3*	91.8	93.0	91.7
Token Bigram	**94.5**	*93.7*	89.8	91.0	91.8
Token Trigram	**92.5**	89.7	89.0	*92.0*	87.2
Token Skip Bigram	**93.5**	*92.7*	88.3	*92.7*	*92.3*
Char 1-gram-n	**81.8**	*79.5*	77.2	*79.7*	76.2
Char 2-gram-n	*88.7*	**89.7**	84.2	*86.2*	78.8
Char 3-gram-n	**94.0**	90.5	89.2	89.3	86.5
Char 4-gram-n	**94.2**	92.2	89.0	91.0	89.2
Char 5-gram-n	**94.3**	*93.0*	90.5	*93.5*	91.2
Char 1-gram-o	*81.7*	**82.2**	76.8	77.8	75.2
Char 2-gram-o	**86.0**	*85.0*	*83.2*	*85.3*	81.2
Char 3-gram-o	*86.6*	*86.0*	*86.7*	**88.0**	*85.5*
Char 4-gram-o	**92.0**	*90.5*	86.7	89.0	87.3
Char 5-gram-o	**92.8**	*92.3*	89.7	*92.2*	88.7

As for features types, the token unigrams are clearly the best choice. In fact, for all the tokens n-grams, it would seem that the further one goes away from the unigrams, the worse the accuracy gets. An explanation for this might be that recognition is mostly on the basis of the content of the tweet, and unigrams represent the content most clearly. Possibly, the other n-grams are just mirroring this quality of the unigrams, with the effectiveness of the mirror depending on how well unigrams are represented in the n-grams. Below (Section 5.4), we will have a closer look at this hypothesis.

For the character n-grams, our first observation is that the normalized versions are always better than the original versions. This means that the content of the n-grams is more important than their form. This is in accordance with the hypothesis just suggested for the token n-grams, as normalization too brings the character n-grams closer to token unigrams. The best performing character n-grams (normalized 5-grams), will be most closely linked to the token unigrams, with some token bigrams thrown in, as well as a smidgen of the use of morphological processes. However, we cannot conclude that what is wiped away by the normalization, use of diacritics, capitals and spacing, holds no information for the gender recognition. To test that, we would have to experiment with a new feature types, modeling exactly the difference between the normalized and the original form.

5.2 Effects of PCA

In the measurements above, the number of principal components provided to the classification systems was learned on the basis of the development sets. This number was treated as just another hyperparameter to be selected. As a result, the systems' accuracy was partly dependent on the quality of the hyperparameter selection mechanism. In this section, we want to investigate how strong this dependency may have been.

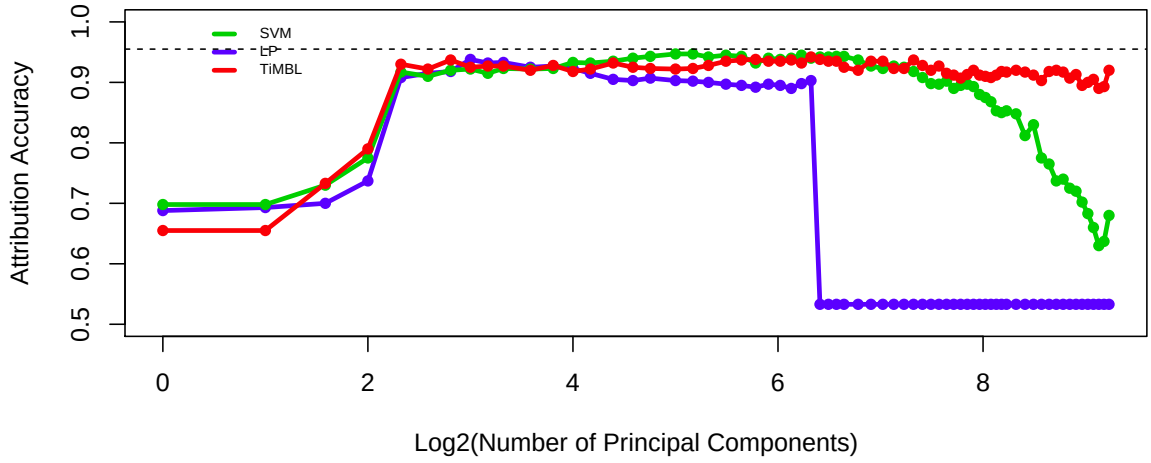

Figure 1: Recognition accuracy as a function of the number of principal components provided to the systems, using token unigrams. The dotted line represents the accuracy of SVR without PCA preprocessing.

Figures 1, 2, and 3 show accuracy measurements for the token unigrams, token bigrams, and normalized character 5-grams, for all three systems at various numbers of principal components. The dotted line is at the accuracy of SVR without PCA. For the unigrams, SVR reaches its peak (94.7%) around 30-40 principal components, with a second peak around 80-90. TiMBL closely follows SVR, but only reaches its best score (94.2%) at the latter peak (80-90). Interestingly, it is SVR that degrades at higher numbers of principal components, while TiMBL, said to need fewer dimensions, manages to hold on to the recognition quality. LP peaks much earlier (93.8%) at only 8-10 principal components. However, it does not manage to achieve good results with the 80-100 principal components that were best for the other two systems. Furthermore, LP appears to suffer some kind of mathematical breakdown for higher numbers of components. If we look at these measurements, it would seem we should prefer TiMBL over LP, which is in contradiction to what we see in Table 1. Although LP performs worse than it could on fixed numbers of principal components, its more detailed confidence score allows a better hyperparameter selection, on average selecting around 9 principal components, where TiMBL chooses a wide range of numbers, and generally far lower than is optimal. We expect that the performance with TiMBL can be improved greatly with the development of a better hyperparameter selection mechanism.

For the bigrams (Figure 2), we see much the same picture, although there are differences in the details. SVR now already reaches its peak (94.3%) at 10 principal components, and stays at almost the same quality until around 200. TiMBL peaks a bit later at 200 with 94.7%, even slightly higher than SVR without PCA. And LP just mirrors its behaviour with unigrams. For the normalized character 5-grams, SVR is clearly better than TiMBL, with peaks (94.2%) from 40 to 100. LP keeps its peak at 10, but now even lower than for the token n-grams (92.8%).

All in all, we can conclude that SVR without PCA is still the best choice. However, all systems are in principle able to reach the same quality (i.e. not significantly lower) with the optimal number of principal components. Even with an automatically selected number, LP already profits clearly

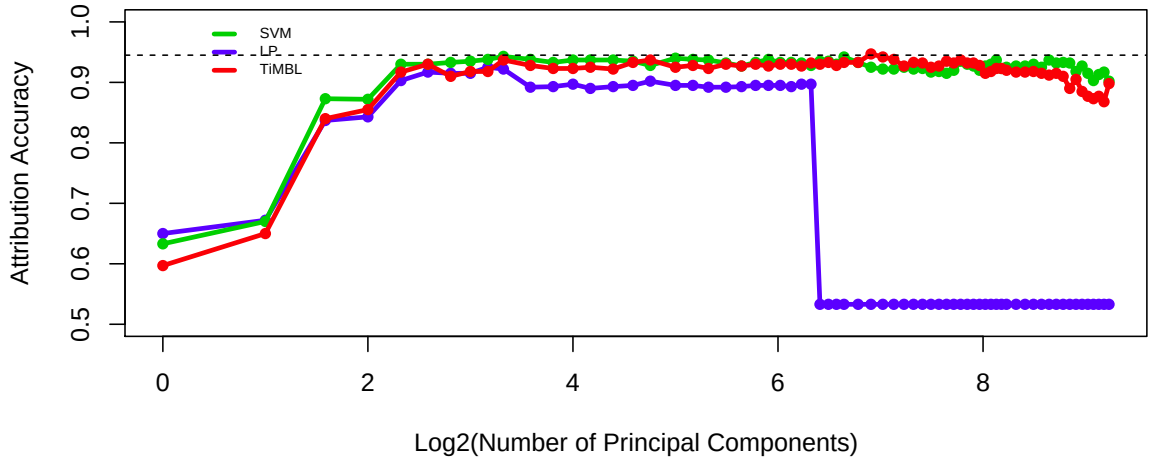

Figure 2: Recognition accuracy as a function of the number of principal components provided to the systems, using token bigrams. The dotted line represents the accuracy of SVR without PCA preprocessing.

from PCA, but (for this task) lags behind. And TiMBL is currently underperforming, but might be a challenger to SVR when provided with a better hyperparameter selection mechanism.

5.3 Analysis of Author Classifications

In this section, we will examine some aspects of author classifications. We will focus on the token n-grams and the normalized character 5-grams. As for systems, we will involve all five systems in the discussion. However, our starting point will always be SVR with token unigrams, this being the best performing combination. We will only look at the final scores for each combination, and forgo the extra detail of any underlying separate male and female model scores (which we have for SVR and LP; see above). As can be seen in Figure 4, the two scores for SVR match almost completely anyway (Pearson Correlation -0.993).[11]

The major exception to the symmetry is author 543, lying clearly in the male area, but quite a bit above the dotted line (at around -2,4 in Figure 4). When we look at his tweets, we see a kind of financial blog, which is an exception in the population we have in our corpus. The exception also leads to more varied classification by the different systems, yielding a wide range of scores. SVR tends to place him clearly in the male area with all the feature types, with unigrams at the extreme with a score of -3.497.[12] SVR with PCA on the other hand, is less convinced, and even classifies him as female for unigrams (1.136) and skipgrams (3.946). LP and TiMBL also show scores all over the range.

Figure 4 shows that the male population contains some more extreme exponents than the female population. The most obvious male is author 430, with a resounding -6.050. Looking at his texts, we indeed see a prototypical (young) male Twitter user: the addressed topics mainly consist of soccer, gaming, school, and music (all of which we will see again below, when examining the most gender

11. This is rather different for LP, but the focus is on SVR here.
12. From this point on in the discussion, we will present female confidence as positive numbers and male as negative.

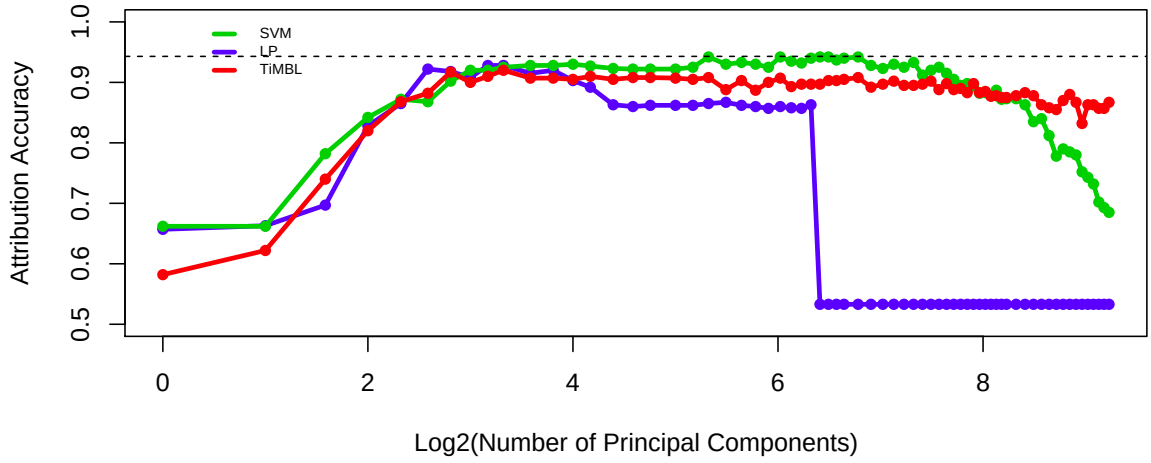

Figure 3: Recognition accuracy as a function of the number of principal components provided to the systems, using normalized character 5-grams. The dotted line represents the accuracy of SVR without PCA preprocessing.

specific unigrams). All systems have no trouble recognizing him as a male, with the lowest scores (around 1) for the top 100 function words. If we look at the rest of the top males (Table 2), we may see more varied topics, but the wide recognizability stays. Unigrams are mostly closely mirrored by the character 5-grams, as could already be suspected from the content of these two feature types. For the other feature types, we see some variation, but most scores are found near the top of the lists.

Table 2: Top ranking males in SVR on token unigrams, with ranks and scores for SVR with various feature types.

Feature type	430	344	564	335	454
Unigram	1: -6.050	2: -5.243	3: -4.886	4: -4.356	5: -4.218
Bigram	1: -5.776	20: -2.875	12: -30.72	8: -3.225	6: -3.375
Trigram	1: -3.916	64: -1.523	49: -1.663	15: -2.338	17: -2.277
Skipgram	1: -3.186	8: -2.808	3: -3.123	21: -2.191	15: -2.385
Char 5-gram	1: -5.228	3: -4.398	4: -4.239	6: -3.754	5: -4.001
Top 100 Function	4: -1.946	144: -0.602	42: -1.114	5: -1.903	31: -1.211

On the female side, everything is less extreme. The best recognizable female, author 264, is not as focused as her male counterpart. There is much more variation in the topics, but most of it is clearly girl talk (of the type described in Section 5.4), again putting the best recognition at a prototypical young Twitter user. In scores, too, we see far more variation. Even the character 5-grams have ranks up to 40 for this top-5.

Another interesting group of authors is formed by the misclassified ones. Taking again SVR on unigrams as our starting point, this group contains 11 males and 16 females. We show the 5 most

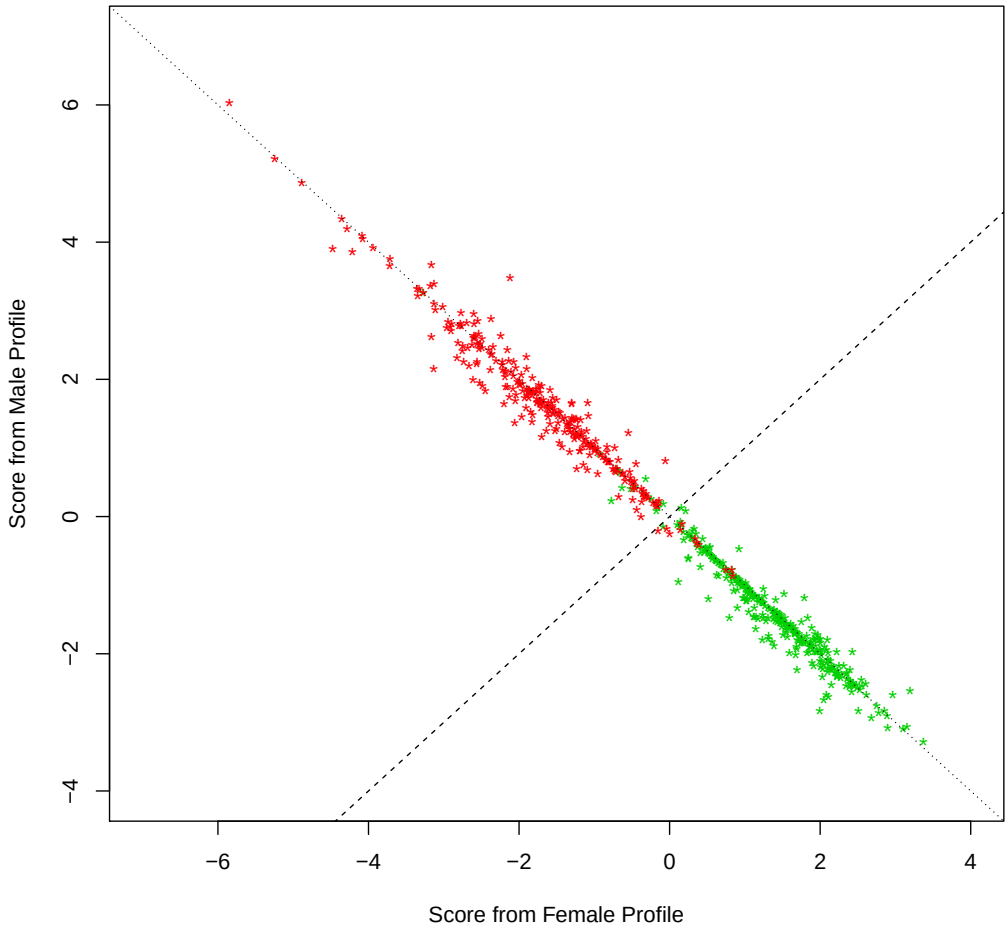

Figure 4: Confidence scores for gender assignment with regard to the female and male profiles built by SVR on the basis of token unigrams. The dashed line represents the separation threshold, i.e. a higher score for one gender than for the other. The dotted line represents exactly opposite scores for the two genders.

Table 3: Top ranking females in SVR on token unigrams, with ranks and scores for SVR with various feature types.

Feature type	264	13	75	43	298
Unigram	1: 3.372	2: 3.196	3: 3.154	4: 3.102	5: 2.964
Bigram	6: 3.011	11: 2.847	78: 2.001	33: 2.472	43: 2.377
Trigram	43: 2.147	49: 2.070	179: 1.136	69: 1.911	157: 1.252
Skipgram	10: 2.546	11: 2.525	13: 2.462	118: 1.405	76: 1.704
Char 5-gram	2: 3.319	9: 2.847	14: 2.687	40: 2.194	29: 2.383
Top 100 Function	9: 1.658	43: 1.215	140: 0.679	88: 0.904	174: 0.553

Table 4: Most strongly misclassified males in SVR on token unigrams, with scores for SVR with various feature types.

Feature type	352	355	386	566	389
Unigram	0.841	0.826	0.753	0.388	0.362
Bigram	0.361	0.047	1.214	0.348	1.472
Trigram	0.300	-0.723	1.016	0.339	1.226
Skipgram	0.104	1.059	0.655	0.750	2.080
Char 5-gram	0.819	0.345	0.870	0.147	1.388
Top 100 Function	1.088	0.163	0.477	0.711	1.508

strongly misclassified ones of each gender in Tables 4 and 5. With one exception (author 355 is recognized as male when using trigrams), all feature types agree on the misclassification. This may support our hypothesis that all feature types are doing more or less the same. But it might also mean that the gender just influences all feature types to a similar degree. In addition, the recognition is of course also influenced by our particular selection of authors, as we will see shortly. Apart from the general agreement on the final decision, the feature types vary widely in the scores assigned, but this also allows for both conclusions.

The male which is attributed the most female score is author 352. On (re)examination, we see a clearly male first name and also profile photo. However, his Twitter network contains mostly female friends. This apparently colours not only the discussion topics, which might be expected, but also the general language use.[13] Another interesting case is author 389. The unigrams do not judge him to write in an extremely female way, but all other feature types do. When looking at his tweets, we

13. This has also been remarked by Bamman et al. (2014).

Table 5: Most strongly misclassified females in SVR on token unigrams, with scores for SVR with various feature types.

Feature type	103	261	49	226	61
Unigram	-3.285	-0.933	-0.700	-0.691	-0.659
Bigram	-1.689	-0.105	-1.149	-0.895	-0.547
Trigram	-0.830	-0.033	-0.747	-0.494	-0.415
Skipgram	-1.243	-0.226	-0.508	-1.027	-0.435
Char 5-gram	-2.914	-0.203	-0.672	-1.009	-0.814
Top 100 Function	-0.667	-0.170	-0.099	-0.683	-0.589

get the impression that Dutch is not his native language, which is supported by his name. There is an extreme number of misspellings (even for Twitter), which may possibly confuse the systems' models.

The most extreme misclassification is reserved for a female, author 103. This turns out to be Judith Sargentini, a member of the European Parliament, who tweets under the name @judithineuropa.[14] Although clearly female, she is judged as rather strongly male (-3.285) when using unigrams and character 5-grams, and male by all system-feature combinations[15] except three. LP with PCA on skipgrams assigns her a female score of 1.321 and SVR with PCA (just as with author 543) arrives at a clearly female attribution with character 5-grams (4.554) and unigrams (5.149). In this case, it would seem that the systems are thrown off by the political texts. If we search for the word *parlement* ("parliament") in our corpus, which is used 40 times by Sargentini, we find two more female authors (each using it once), as compared to 21 male authors (with up to 9 uses). Apparently, in our sample, politics is a male thing.[16] It is intriguing that both here and with the male financial blogger, the erroneous misclassification with unigrams is reversed when using PCA on the unigrams. We did a quick spot check with author 113, a girl who plays soccer and is therefore also misclassified often; here, the PCA version agrees with and misclassified even stronger than the original unigrams (-0.707 versus -0.248). In later research, when we will try to identify the various user types on Twitter, we will certainly have another look at this phenomenon.

5.4 Features

In the analysis so far, we have wondered several times what kind of features are responsible for the rather accurate classification. Are they mostly targeting the content of the tweets, i.e. related to the activities of the authors in real life, or the style, i.e. the way they use the basic building blocks of the Dutch language? In this section, we will attempt to get closer to the answer to this question. Again, we take the token unigrams as a starting point. However, looking at SVR is not an option here. Because of the way in which SVR does its classification, hyperplane separation in a transformed version of the vector space, it is impossible to determine which features do the most work. Instead, we will just look at the distribution of the various features over the female and male texts.

Figure 5 shows all token unigrams. The ones used more by women are plotted in green, those used more by men in red. The position in the plot represents the relative number of men and women who used the token at least once somewhere in their tweets. However, for classification, it is more important how often the token is used by each gender. We represent this quality by the class separation value that we described in Section 4.2, and show it in the form of font size, i.e. the more distinguishing tokens are bigger. As the separation value and the percentages are generally correlated, the bigger tokens are found further away from the diagonal, while the area close to the diagonal contains mostly unimportant (and therefore unreadable) tokens.

On the female side, we see a representation of the world of the prototypical young female Twitter user. It is a very emotional place, with *omg* ("Oh My God!") in a central position, but also containing giggling (*hihi*) and lots of emotionally loaded adjectives, such as *lief* and *lieve* ("sweet"), *schattig* ("cute"), *leuk* and *leuke* ("nice"). And also some more negative emotions, such as *haat* ("hate") and *pijn* ("pain"). Next we see personal care, with *nagels* ("nails"), *nagellak* ("nail polish"), *makeup* ("makeup"), *mascara* ("mascara"), and *krullen* ("curls"). Clearly, *shopping* is also important, as is watching soaps on television (*gtst*). The age is reconfirmed by the endearingly high presence of *mama* and *papa*. As for style, the only real factor is *echt* ("really"). The word *haar* may be the pronoun "her", but just as well the noun "hair", and in both cases it is actually more related to the

14. Identity disclosed with permission.
15. And by TweetGenie as well.
16. An alternative hypothesis was that Sargentini does not write her own tweets, but assigns this task to a male press spokesperson. However, we received confirmation that she writes almost all her tweets herself (Sargentini, personal communication).

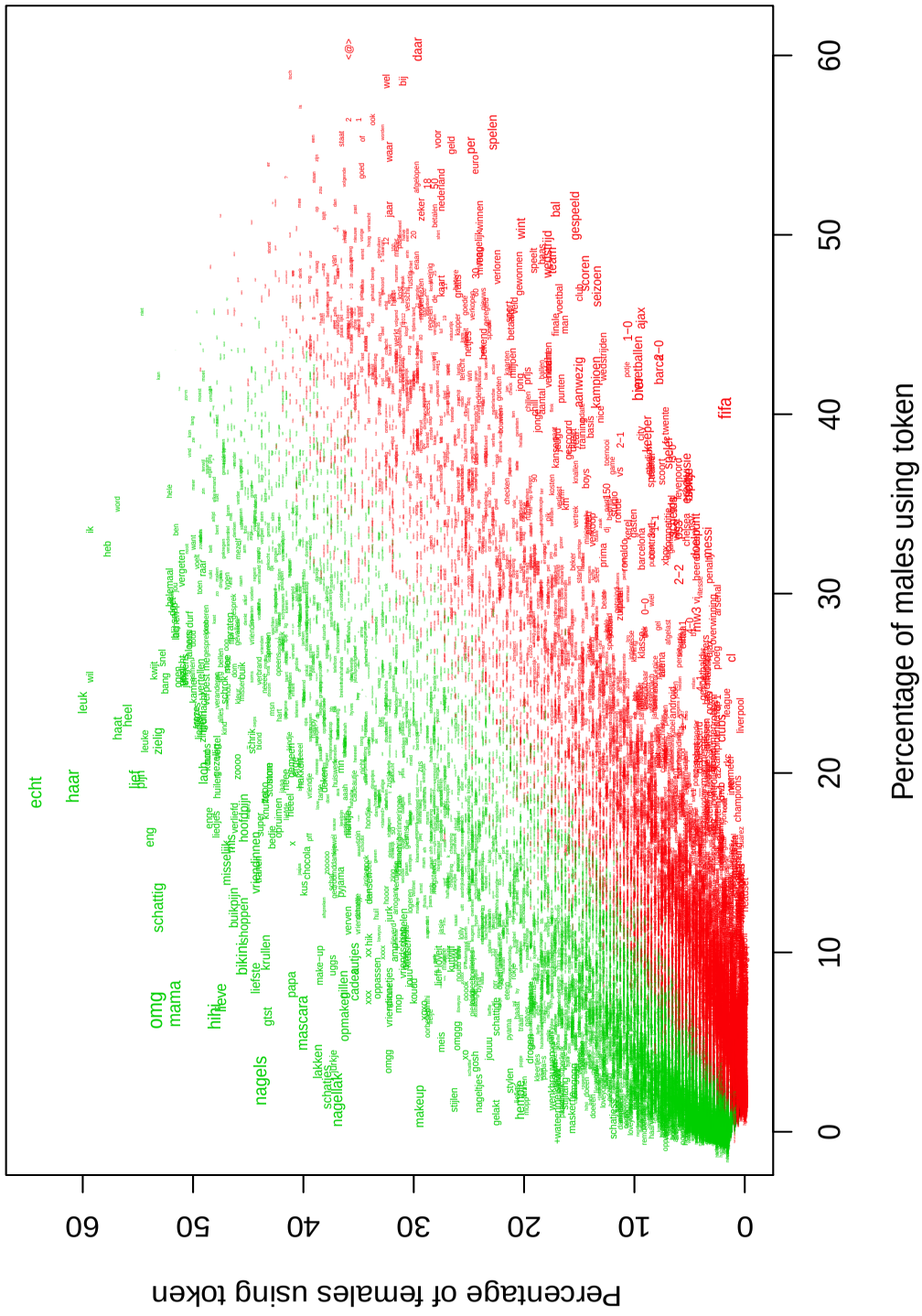

Figure 5: Percentages of use of tokens by female and male authors. The font size of the words indicates to which degree they differentiate between the gender when also taking into account the relative frequencies of occurrence.

already observed activities. We do also see more expressions of self, with *ik* ("I") and first person verbs such as *wil* ("want") and *heb* ("have"), but these are much less distinguishing.

On the male side, we see a rather different world. Apart from *bier* ("beer"), we see an enormous number of soccer-related words, with *fifa* at the most extreme, and then club names, scores, competitions, playing, winning, losing, etc. On the right edge of the plot, though, we do also observe some function words. The location adverbs *daar* ("there") and *waar* ("where") appear to be a more male thing, as well as some prepositions like *per* ("per"), *bij* ("at", "near") and *voor* ("for", "before"). Finally, mentioning other users (<@>) is apparently more often done by men.

For both genders, the tokens are dominated by the young person's world. It is no wonder that classifying different types of authors, such as politicians and financial bloggers, is more problematic.

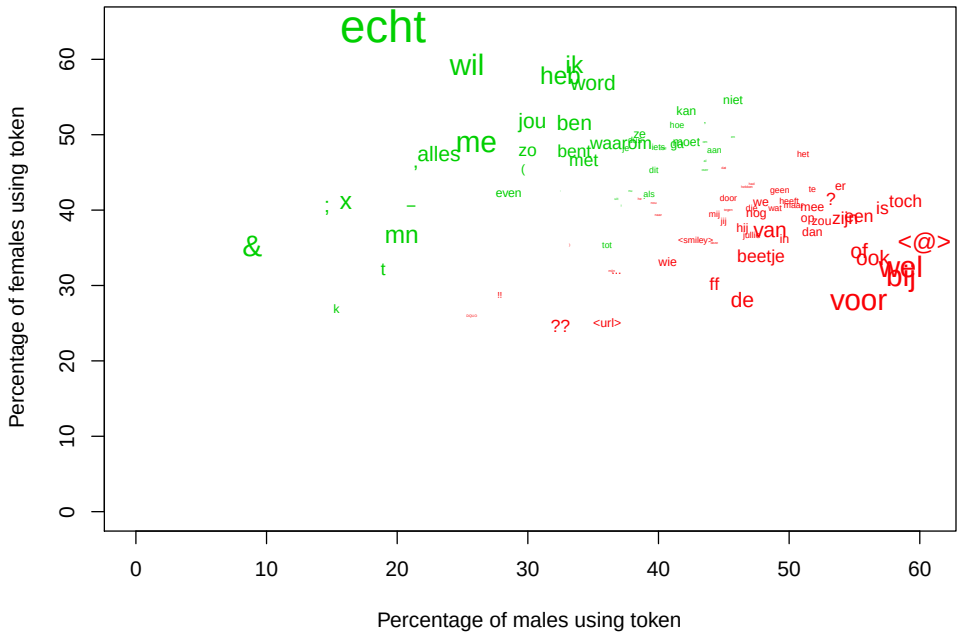

Figure 6: Percentages of use of the 100 most frequent function words and punctuation by female and male authors. The font size of the words indicates to which degree they differentiate between the gender when also taking into account the relative frequencies of occurrence.

Although most distinguishing tokens appear to be related to content, we do observe some style-related tokens. In Figure 6, we show a plot for the top 100 function words (or rather tokens), which was the only feature type focusing on style in our experiments. We can now observe various distinguishing tokens which were so far lost in the dense cloud of words. They correspond to what earlier research (see Section 2) has observed. Females show more personal pronouns, such as the already mentioned *ik* ("I"), but also *me* ("me") and *jou* ("you"), as well as the reduced possessive pronoun *mn* (*mijn*; "my"). Males write more objective structures, with the mentioned prepositions, and also articles like *de* ("the") and *een* ("a"). Here we find also find more third person constructions, with *is* ("is"), *hij* ("he") and *zijn* ("his", or plural "are").

Looking at the bigrams, which we will not plot here, we see a few more style-related constructions appearing. On the female side, we see *niet meer* ("not anymore"), *ik mis* ("I miss") and the

intensifying adverb *zo* combined with various adjectives: *zo moe* ("so tired"), *zo blij* ("so happy"), *zo zielig* ("so pathetic"), and *zo leuk* ("so nice"). On the male side, there are also mostly combinations of already observed unigrams, but also the more pragmatic ending of tweets with the word *man*, in *man !* (",man!"), *nee man* ("no, man"), *niet man* ("not, man"), and *goed man* ("good, man").

All in all, there appear to be quite a few features related to style after all. Furthermore, the top 100 function words are doing quite well, with 84.8%, seeing how few features there are compared to the full set of unigrams. On the other hand, we cannot escape the impression that even these style features are more often related to what is being tweeted about, than to personal writing style.

6. Conclusion and Future Work

We have investigated how well the gender of authors on Twitter can be determined on the basis of token or character n-grams. We find that recognition is possible with a high accuracy, up to 95.5% on our data set (but see dicussion below). Furthermore, some of the errors are probably related to the fact that the authors in question are different from the typical Twitter users dominating our data set. The best feature type for recognition appears to be the token unigrams, with the most distinguishing tokens linked to the typical activities of the dominant Twitter users. As for classification systems, Support Vector Regression clearly performs best with all feature types.

During our investigation into gender recognition, we have also experimented with the use of Principal Component Analysis as a preprocessing step to classification. It was already known that this step was necessary for k-NN learning. We found that SVR is actually hampered rather than helped by the preprocessing. Its accuracy degrades when using PCA, although often not significantly. For Linguistic Profiling, PCA increases accuracy, in some cases enabling it to reach a score which is no longer significantly worse than that of SVR. TiMBL, even with PCA, does not reach the same accuracy level, and only accomplishes scores similar to SVR's scores for token skip bigrams and unnormalized character trigrams. However, TiMBL's lower quality is mostly a matter of hyperparameter selection. The number of principal components provided to the learners was determined automatically on the basis of development data. When we examined the systems' accuracy for fixed numbers of principal components, TiMBL was often at the same accuracy level as SVR, and it was LP that was falling behind.

It has remained unclear to which degree gender can be recognized on the basis of style features. Although the use of all unigrams for classification yields far better results than the use of the 100 most frequent function words, the latter are certainly not doing badly. Furthermore, our closer examination in Section 5.4 may imply that it is not the quality, but the number of features that is the reason for the difference in accuracy. We will revisit this question when we have larger n-gram sets available which can be assumed to be largely domain-independent.[17]

Finally, if we look back at our original goal, the automatic estimation of metadata for the TwiNL data set, we must conclude that we have made a significant step forward, but still only a modest one. Not only did we predict just one user trait, but we also considered just a very select class of users, namely individual users with a significant tweet volume. We will still need to test the minimum number of words on which the classifier can maintain its current high quality. Furthermore, we will need to build classifiers to distinguish between individual user accounts, shared user accounts, accounts controlled by boards of editors, and tweetbots. It may also be useful to distinguish between different uses of Twitter, such as professional communication and social chitchat, and build separate metadata estimators for these different uses. Even more importantly, we will need to look beyond very specific lexical features. If we base metadata on a limited number of such features, we will never be able to use the resulting data for studying language use or social behaviour. If we would try, we would fall victim to circular reasoning, such as observing that only men ever play soccer,

17. We are currently laying the basis for the construction of such sets in other work (van Halteren and Oostdijk Submitted).

since this is the information we put in with our metadata determination. Therefore, if we ever want to automatically add metadata, it will have to be with as many information sources as possible, preferably only using that metadata on which various sources agree.

References

Bamman, David, Jacob Eisenstein, and Tyler Schnoebelen (2014), Gender identity and lexical variation in social media, *Journal of Sociolinguistics*.

Beyer, Kevin, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft (1999), When is "Nearest Neighbor" meaningful?, *In Int. Conf. on Database Theory*, pp. 217–235.

Burger, John D., John Henderson, George Kim, and Guido Zarrella (2011), Discriminating gender on Twitter, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK, July 27–31, 2011*, pp. 1301–1309.

Chang, Chih-Chung and Chih-Jen Lin (2011), LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology* **2** (27), pp. 1–27.

Cohen, Jacob (1988), *Statistical Power Analysis for the Behavioral Sciences (second ed.)*, Delft: Now Publishers.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot, and Antal Van den Bosch (2004), Timbl: Tilburg memory-based learner, *Technical Report ILK-0209*, Tilburg University.

Fink, C., J. Kopecky, and M. Morawski (2012), Inferring gender from the content of tweets: A region specific example, *Proceedings of the International AAAI Conference on Weblogs and Social Media, North America, May 2012*.

Goswami, Sumi, Sudeshna Sarkar, and Mayur Rustagi (2009), Stylometric analysis of bloggers' age and gender, *Proceedings ICWSM 2009*.

Heil, B. and M. J. Piskorski (2009), New Twitter research: Men follow men and nobody tweets, *Harvard Business Review*.

Hotelling, H. (1933), Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology* **24**, pp. 417–441 and 498–520.

Juola, Patrick (2008), *Authorship Attribution*, Lawrence Erlbaum Associates.

Kestemont, M. (2014), Function words in authorship attribution. from black magic to theory?, *Proceedings of the Third Computational Linguistics for Literature Workshop, co-located with EACL 2014 – the 14th Conference of the European Chapter of the Association for Computational Linguistics (27 April 2014, Gothenburg, Sweden)*, pp. 59–66.

Kjell, Bradley, W. Addison Woods, and Ophir Frieder (1994), Discrimination of authorship using visualization, *Inf. Process. Manage.* **30** (1), pp. 141–150.

Koppel, Moshe, Jonathan Schler, and Shlomo Argamon (2009), Computational methods in authorship attribution, *J. Am. Soc. Inf. Sci. Technol.* **60** (1), pp. 9–26.

Koppel, Moshe, Schlomo Argamon, and Anata Rachel Shimony (2002), Automatically categorizing written texts by author gender, *Literary and Linguistic Computing* **17** (4), pp. 401–412.

Narayanan, Arvind, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Eui Chul, Richard Shin, and Dawn Song (2012), On the feasibility of internet-scale author identification, *Proceedings of the 33rd conference on IEEE Symposium on Security and Privacy. IEEE*.

Nguyen, D., D. Trieschnigg, A. S. Doğruöz, R. Gravel, M. Theune, T. Meder, and F.M.G. de Jong (2014), Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment, *Proceedings of COLING 2014*.

Nguyen, Dong, Rilana Gravel, Dolf Trieschnigg, and Theo Meder (2013), "How old do you think I am?": A study of language and age in twitter, *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM 2013)*.

Pearson, K. (1901), On lines and planes of closest fit to systems of points in space, *Philosophical Magazine* **2** (11), pp. 559–572.

Pennebaker, J.W., C.K. Chunk, M. Ireland, A. Gonzales, and R.J. Bootk (2007), *The development and psychometric properties of LIWC2007, Software Manual*.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing. http://www.R-project.org.

Rao, D., D. Yarowsky, A. Shreevats, and M. Gupta (2010), Classifying latent user attributes in Twitter, *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pp. 37–44.

Schler, J., M. Koppel, S. Argamon, and J. Pennebaker (2006), Effects of age and gender on blogging, *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

Tjong Kim Sang, Erik and Antal van den Bosch (2013), Dealing with big data: the case of Twitter, *Computational Linguistics in the Netherlands Journal* **3**, pp. 121–134.

Van Bael, Christophe and Hans van Halteren (2007), Speaker classification by means of orthographic and broad phonetic transcriptions of speech, *Speaker Classification (2)*, pp. 293–307.

van Halteren, Hans (2004), Linguistic Profiling for authorship recognition and verification, *Proceedings ACL 2004*, pp. 199–206.

van Halteren, Hans (2008), Source language markers in Europarl translations, *Proceedings of COLING2008, 22nd International Conference on Computational Linguistics*, pp. 937–944.

van Halteren, Hans and Nelleke Oostdijk (Submitted), Word distribution in Dutch Tweets, *Nederlandse Taal en Letterkunde*.