Computational Linguistics in the Netherlands Journal 5 (2015) 37-48 Submitted 06/2015; Published 11/2015

Metadata Induction on a Dutch Twitter Corpus: Initial phases

Hans van Halteren HVH@LET.RU.NL

Radboud University Nijmegen, CLS, Linguistics

Abstract

In this paper, I pose that metadata induction for the TwiNL collection of Dutch tweets should start at a fundamental level, for which I address three types of classification. Identification of users tweeting predominantly in Dutch is shown to be possible with an F-value over 98%. For the identification of individual humans (as opposed to e.g. groups or bots), no classifier is tested, but a number of useful text-based measurements are presented. Finally, the identification of children of school-going age (by far the largest user group in the collection) is shown to be possible on the basis of just unigram counts with high accuracy (93.5%), increasing to very high accuracy (97%) when at least 1,000 tweets are available.

1. Introduction

As was noted at last year's CLIN (van Halteren and Speerstra 2014), the TwiNL collection of Dutch tweets (Tjong Kim Sang and van den Bosch 2013) is a unique resource, that is already very useful for many kinds of research, but the potential of which would be greatly increased if only it came with reliable metadata. Van Halteren and Speerstra (2014) investigated whether it is possible to induce the gender of Twitter users in 2011 and 2012, and were quite succesful (95% correct assignment). However, it turned out that the investigated sample of users was dominated by (high) school kids, and that the more severe gender misclassifications were found for users outside that group. Furthermore, the sample was already filtered (manually) to contain only real people tweeting in Dutch. It was concluded that further metadata induction, such as age and regional background, should be postponed until some more fundamental classifications are in place. First, the TwiNL set would need to be filtered so as to exclude users who are tweeting in languages other than Dutch. Then it would be necessary to distinguish between individual human users on the one hand and accounts where the tweets are produced by groups of people, editorial boards and/or bots on the other hand. Finally, within the individual humans, it would be advantageous to identify any major groupings of users showing comparable language use (or *tribes*, as suggested by Oostdijk and van Halteren (2015)). Once these groupings were available, it was claimed, there would be a better chance of reliably determining more extensive metadata. In this paper, I aim to follow up on these claims.

The best results for metadata induction can be expected when deriving information from all the available sources, such as the provided Twitter metadata (including the Twitter profile text), user relations, profile photos, and the text of the tweets. At the time of writing this paper, I am still collecting the necessary data, and will come back to this in future work. For now, as with the gender recognition experiment last year, I focus on the content of the users' tweets. I address all three of the listed classification tasks, describing the current level of achievement on the portion of TwiNL that covers the period January 1, 2011 to June 30, 2013.

First, in Section 2, I describe the formation of a profile for each user. In the subsequent sections, each of the three classification tasks are presented: language filtering (Section 3), individual recognition (Section 4), and tribal assignment (Section 5). Finally, I summarize the current state and suggest what I think should be the next steps (Section 6).

2. Creating user profiles

From the TwiNL dataset, I extracted all tweets with a date stamp from Januari 1, 2011 to June 30, 2013. The result was a collection of 2,351,637,973 tweets by 41,271,439 users. This included 299,309,557 retweets (13%). As I wanted to base the induction of metadata on the way in which users produce original text, and not on how they select tweets to retweet, I removed users posting only retweets (about 14%), leaving 35,334,045 users.

I then tokenized all tweets with a specialized tokenizer for Dutch tweets. First, as the use of capitalization and diacritics is found to be quite haphazard in tweets, the tokenizer strips all words of diacritics and converts them to lower case, thus producing a kind of normalized form. Then, just like other tokenizers, this tokenizer identifies words, numbers and punctuation. If words do not consist of Latin script or numbers of Arabic digits, the word/number is marked as foreign. Similarly, symbols other than those used in (for Dutch) standard punctuation are marked as symbols.[1] Furthermore, the tokenizer is able to recognize a wide variety of types abundant in social media texts. Thus it is able to identify hashtags and Twitter user mentions 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. Many (but certainly not all) URLs and email addresses are recognized as well. Finally, a substantial fraction of the emoticons used are recognized as such. For user mentions, URLs, and emoticons, the normalized forms <usr>, <url> and <emo> are produced.

Each (normalized) token was also checked against two token lists. The first list was created in previous research (van Halteren and Oostdijk 2015) and consists of 5,520 very high frequency tokens.[2] These tokens are associated with a so-called U-score (van Halteren and Oostdijk 2015), which indicates how generally usable the token is, as opposed to only usable for specific topics/domains. The U-score ranges from 1 (ubiquitous) to 0 (fully dependent on a single topic/domain). The second list is the OpenTaal word list. OpenTaal is a project directed by the Dutch Language Union which aims to make available for free (written) Dutch language resources for use in open source projects (e.g. OpenOffice.org). One of the resources is the OpenTaal word list that was compiled for use with, for example, spelling checkers and grammar checkers. The word list used for the research described in this paper is version 2.10g. It includes some 350,000 word forms, including many frequently used abbreviations and common Dutch proper names.[3] On the basis of the two lists, each token was assigned to one of three groups. The first are the frequent tokens, i.e. the tokens that are in the frequent token list. These are also associated with their U-score. Next are the known tokens, i.e. those listed in the OpenTaal list but not included in the list of frequent tokens. Finally, there are the OOV tokens, i.e. all other tokens. In addition, there is a subgroup of the frequent tokens, namely the ones with a U-score of at least 0.75, below referred to as ubiquitous tokens.[4]

In the experiments, I wanted to rule out those users for which insufficient text is available for measurements. I set the threshold for inclusion to at least 10 tweets and at least 100 tokens, in both cases excluding retweets. For experiments involving author profiling, this is rather little, but for other measurements it should often be enough.

For all 5,883,805 qualifying users, I created a profile, which was to be used for the majority of measurements instead of the full tweet collection. Whereas the full collection is rather unwieldy, the

[1] Sometimes, as the recognition of alphanumeric characters is done by the operators \p{Letter} and \p{Number} in Perl regular expressions, characters of non-alphabetic scripts are marked as symbol rather than taken as part of a foreign word. This will be addressed in the near future by checking on which code page a Unicode character is placed. For the work described here, the problem has not yet been solved.

[2] A rough approximation would be to say that they occur at least once in every 100,000 tokens.

[3] For more information see http://www.opentaal.org/opentaal.

[4] Although this division into groups is useful, it is still unsatisfactory. With the current data and tools, the grouping is restricted to the form of each token. It is impossible to recognize an OOV token as a spelling variant or spelling error for a known word, or a token that seems to be known, e.g. *dak* (”roof”), but in context turns out to be a spelling variant, e.g. *dak* for *dat ik* ('that I'). Once operations such as part-of-speech tagging, word sense disambiguation and spelling normalization for Dutch tweets become available, even more useful user profiles will become possible.

profile contains on average less than 3,000 bytes per user, which can be processed without excessive computer use. For 483,149 highly productive users, with more than 1,000 tweets in the collection, I took a random sample of about 1,000 tweets, again to reduce computing time.[5] For the remaining users, the full tweet collection was used. All further processing on the data was done separately for the highly productive and the less productive users.

The user profile includes:

Username Used as an identifier only.

Volume The user's number of tweets and retweets in the collection and, where appropriate, the number of sampled tweets and retweets (for calculation purposes).

Token use fractions The fractions of all tokens that were marked by the tokenizer as word, foreign word, foreign number, punctuation, symbol, hashtag, user mention, URL or emoticon.

Token use distribution The mean and standard deviation of the number of specific types of token present in each tweet, for the following types: all tokens, frequent tokens, OOV tokens, hashtags, user mentions, and URLs. In addition, the mean and standard deviation of the mean U-score of the frequent tokens per tweet. Furthermore, I complemented each mean-standard deviation pair with a kind of normalized variation indicator. I sorted all users by the mean in question and divided them into buckets of 10,000. Within each bucket, I calculated the coefficient of variation (standard deviation divided by mean) for each user, and then took mean and standard deviation of these coefficients. Each user's variation indicator was then set to the Z-score of the corresponding coefficient of variation with regard to the bucket mean and standard deviation.[6]

Vocabulary distribution The total number of different types (determined on the basis of the normalized version of the token as output by the tokenizer) and, derived from this, the type-token ratio (TTR). Another measure representing vocabulary richness is the percentage of hapaxes. Finally, in this group, there is the so-called non-zipfiness (van Halteren and Oostdijk 2015), which indicates how close the token distribution curve is to the idealized curve predicted by Zipf's Law (Zipf 1935). All three measures are correlated with the sample size, here the total number of tokens sampled for the user, and have therefore been normalized by sorting all users by their sampled number of tokens, dividing the list into buckets of 10,000 users, and replacing each user's measurement by its Z-score with relation to the scores of the users in the same bucket.

TwiNL language indicators Tweets in the TwiNL dataset are marked with the language that is deemed most likely for that tweet (Tjong Kim Sang and van den Bosch 2013). For the profile, these markers are used to determine the fractions of each user's tweets marked as **dutch**, **notdutch** or any other language (**other**), as well as the language suggested most often.[7] The same is done for retweets.

Frequency lists For each of three groups of tokens, namely ubiquitous, known and OOV, a list of the 100 most frequent types.[8]

[5] Similarly, every user's retweets were sampled down to 100.

[6] At low values of the mean, the coefficient of variation shows erratic behaviour. Therefore, for all means under a threshold, namely 5 for the number of all tokens per tweet and 1 for the other measures, I used the standard deviation itself rather than the coefficient of variation.

[7] More on these language indicators can be found in Section 3.

[8] If the user did not produce enough text, these lists might be shorter than 100 items. If the item at rank 100 was part of a series of items with the same frequency, and one might expect that the list included all these items and would simply stretch beyond 100, the maximum of 100 was strictly enforced. In such cases, a random selection from the series of items was included in the list.

Frequency list statistics The fraction of Dutch words (according to OpenTaal) in the 20 most frequent word types for the user.[9] Also, the fraction of hashtags in the list of most frequent OOV types, both in terms of type count and token count.

3. Filtering out the non-Dutch

It is known that the TwiNL collection contains a substantial portion of tweets in other languages than Dutch. As such noise may hinder statistical modelling, it seems wise to try to filter out the non-Dutch content as much as possible.

3.1 TwiNL sampling

For TwiNL, Dutch tweets are being harvested by two strategies, keyword search and user search on the most prolific Dutch users (Tjong Kim Sang and van den Bosch 2013). The keyword search uses a list of 229 Dutch words and hashtags. The harvested tweets are subsequently subjected to two language filters, viz. libTextCat and Twitter's own language identifier, and tweets are accepted as being Dutch if either filter accepted them as such.[10] This strategy is estimated to lead to a minimal amount of noise in the form of non-Dutch tweets (about 2.5%), at the cost of removing almost 9% of genuine Dutch tweets.

Given the automatic collection procedure, the level of noise is inevitable. The common Dutch words used as anchors sometimes also occur in other languages. Especially German tweets are often captured. This is due to shared common frequent words like *als* and *hier*, such as in

#job #psychologe/in Psychologe als Teamleiter: Munchen, Bayern - hier finden Sie freie Stellen, die Arbeitgeber... <url>.

It is the language filters' task to stop such tweets from getting into the final collection, but finding the right balance between excluding as many non-Dutch tweets as possible and including as many Dutch tweets as possible is not easy. The makers of the TwiNL collection are aware of this, and have changed tactics a few times in order to address the problem (Tjong Kim Sang, personal communication). For older sections of the collection, up to September 2012, only tweets clearly recognized as Dutch have been included. Later parts include more tweets, together with an indication of the language proposed by the filtering process, such as **dutch** or **english**, but also including **UNKNOWN** and **notdutch**. The latter occurs with both foreign language tweets, e.g. English, and Dutch language tweets containing multiple non-Dutch tokens. Furthermore, even tweets marked **dutch** are not always really in Dutch. In effect, an additional language filter should check all tweets in the collection.

3.2 Language indicators

Language identification for tweets is generally done at the tweet level. Lui and Baldwin (2014) compare a number of systems and conclude that identification with a combination system is possible at higher than 99% accuracy.[11] Still, although language filters appear to do very well on test sets, they still have trouble with many tweets in practice, as we can also see from the presence of many non-Dutch tweets in the TwiNL collection. As I did not expect to be able to construct a significantly better language filter at the individual tweet level, I decided to address the problem at

[9] Words being sequences of letters, apostrophes and hyphens, including at least one letter.

[10] As from May 2014, the TwiNL collection process no longer uses the two mentioned language filters, but instead the language indication provided by Twitter (http://ifarm.nl/erikt/twinl/2014/05/09/twitter-language-field). However, this is irrelevant here, as the dataset for this paper only contains tweets with a time stamp from January 1, 2011 to June 30, 2013.

[11] This is most probably an overestimation, as the test set appears to be composed of tweets by users whose language is very clearly recognizable, since they were chosen on the basis of a substantial number of their tweets being all recognized as being in the same language.

the user level. For each user, I took all availabe tweets from the sample period and calculated five language indicators.[12]

TwiNL percentage Dutch. If more than 90% of the tweets by the user is marked as **dutch**, the indicator suggests *dutch*.[13] If less than 30% of the tweets by the user is marked as either **dutch** or **notdutch**, the indicator suggests *other*. In all other cases, the indicator is *undecided*.

TwiNL most suggested language. The indicator counts which language is most often listed by TwiNL. If this is **dutch**, the indicator suggests *dutch*. Similarly, **notdutch** leads to *notdutch*. Any other language will yield *other*.

Top-20 in OpenTaal. If the percentage of Dutch words (according to the OpenTaal list) in the user's top-20 words is over 90%, the indicator suggests *dutch*. If the percentage is under 60%, the indicator suggests *other*. In all other cases, the indicator is *undecided*.

Percentage foreign and symbol. The indicator checks the values for the fractions of foreign words, foreign numbers, and symbols. If the individual percentages of foreign (either word or number) or symbol are over 40% of all tokens, or the joint percentage over 60%, the indicator suggests *other*. In all other cases, the indicator is *undecided*.

OOV profile. The indicator considers whether the most frequent OOV words produced by the user indicate a specific user language, leading to a suggestion of *dutch*, *other*, or *undecided*. This indicator will be explained in more detail in the next subsection.

Once all individual indicators have been determined, a voting-type combination process decides on the overall judgement. First the indicators' positions as to whether the user's tweet language should be marked as *dutch* or *other* are combined. For each of the two positions, each indicator can vote either in favour or against. Exceptions are the percentage foreign/and symbol, which can only favour *other*, and the judgement *notdutch* by the TwiNL most suggested language, which can only vote against *dutch*. The two TwiNL-based indicators can contribute one vote each, the other three indicators three votes each.

In order to decide on a final judgement of *dutch*, two votes in favour suffice, but only if there are no votes against. At three votes in favour, one vote against is permitted, and at five votes in favour, two votes against are permitted. Fewer than two votes in favour, or too many votes against, block a judgement of *dutch*.[14]

3.3 OOV profile

The language indicator that examines the OOV profile is based on the idea that tweets which have been inadvertently included in the TwiNL collection, as they are mostly in another language, should contain a high number of frequent words from that other language, most of which will not be listed by OpenTaal and will therefore register as OOV words. For Dutch tweets on the other hand, the frequent OOV words should be frequent Dutch words (or Dutch Twitter words) that have not yet been accepted as standard Dutch words. This would also include a good percentage of

[12] Only the username, tweet text, and TwiNL language indication were available. Unfortunately, the Twitter generated language indicator was not present in this version of the data.

[13] There are two marking systems involved here. The language marking by TwiNL is indicated in bold. The language marking by my system is indicated in italic.

[14] It is clear that the current procedure has been tailored to the dataset at hand. It may well be that changes over time, such as a shifting population of Twitter users and different sampling methods, may necessitate changes in the language identification process. As for different languages than Dutch, a direct port will be difficult as the information sources that could be used here are very likely not present in the same form. In both cases, the choice between manual engineering and machine learning the best way to identify the language on the basis of the available features will depend on both the number and transparency of the features and the availability of a sufficient amount of annotated development material.

spelling variants. This means that the list of most frequently observed OOV words for a user will be markedly different between Dutch and non-Dutch tweeting users, and even between the non-Dutch tweeting users using different languages.[15]

In order to train the recognition of the various languages, I collected users for whom the main tweet language can be determined with high probability. I assumed a user could be classified as Dutch if more than 95% of his/her tweets were marked as **dutch** by TwiNL. On the other hand, if there was a language that was suggested more often than **dutch** by TwiNL for the tweets of a user, I assigned the user the language that was most often suggested. In all, 56 languages were suggested, ranging from 654,640 users for English to 2 users for Tamil. Setting the threshold at 100 users, but also taking out a few languages using mostly non-alphabetic scripts which would be blocked by the indicator considering high percentages of symbols (e.g. Chinese), I continued with 48 languages to compare to Dutch.

The next step was to make an inventory of observed OOV words for each language, i.e. those present in the profiles of the users for which the language was deemed known. As expected, the top words for Dutch consist of very common Dutch Twitter words:

ff xd mn hahaha ok jaa twitter m'n ofzo xx

And other languages show high frequency "normal" words for that language, e.g. the top for French is:

que le et il un qui c'est j'ai elle tu

There are of course overlaps between the various lists, especially for social-media-related words. For example *xd*, which is in second place for Dutch, covering almost 2% of the OOV words, is also present in 44 of the other OOV lists, with the highest presence in Polish, at position 13 covering 1% of the OOV words of the Polish users. Obviously, some selection is in order if the OOV words are to be used fruitfully for language detection.

On the basis of the inventories, I determined for each word whether it could be used as a marker for one or more specific languages. I first calculated the relative value of the word within all the OOV words in an inventory for each language: given a randomly selected word (weighted by frequency) from all OOV words for the language, what is the probability that it is the word in question? Then I compared these probabilities and divided all probabilities for a word by the highest observed probability. In this way, the highest value for each word becomes 1.0. As an example, the word *lelik* is seen most often (relatively; 13 times) as an OOV word for Turkish users, giving it a probability for Turkish of 0.000002, as the 343,887 types in the Turkish OOV list are cumulatively seen 6,4 million times. This frequency is recalculated to 1.0. For Dutch, *lelik* is seen 212 times on about 708 million total OOV observations, leading to a probability of 0.0000003. The division by 0.000002 for Turkish leads to a final score of 0.15 for Dutch. All word-language combinations with a score less than 0.1 were excluded from further processing. Finally, all OOV words seen with fewer than 10 users or observed with 12 or more languages were also excluded.

With the resulting set of language markers, it becomes possible to classify each user.[16] The system takes all OOV words in the user's profile and multiplies each word's relative frequency within the user's OOV words with the various language probabilities.[17] It then sums over all words to arrive at scores for each possible language. If the best score is at least three times higher than the next best, the language associated with the best score is selected, leading to either *dutch* or *other*. If there are languages with positive scores, and Dutch is not among the best three, *other* is selected. The same happens if Dutch was in third place, but its score is three times lower than the best score. In all other cases, the result is *undecided*.

[15] A problem can be expected, though, if we have code-switching users, i.e. users who use multiple languages in their tweets.

[16] The classification is only attempted if there are at least five OOV words for the user.

[17] Given the sometimes low numbers, the count for the word by the user in question is substracted from the language counts.

Table 1: Quality of the recognition of predominantly Dutch tweeting users for various language indicators.

Indicator	Precision	Recall	F-measure
Combination by voting	97.4	99.5	98.5
TwiNL percentage Dutch	93.9	72.8	82.0
TwiNL most suggested	90.6	97.3	94.1
Top-20 in OpenTaal	97.6	78.3	86.9
OOV profile	96.7	97.4	97.0

3.4 Results

As there are 5 language indicators, each yielding three possible values, there are in principle 729 different combinations that can occur. 102 of these occur in practice in our user collection. The number of users varies enormously per combination, as can be expected, with most users (3.2 million; 54%) showing combinations of only *dutch* and *undecided*. For each combination, I took a random sample of 10 users.[18] For all the sampled users, I manually inspected first their profile and, if that was insufficient for classification, then their full set of tweets. This led to a manual assignment of Dutch or Other. In all, I inspected 1142 users. 236 of these were Dutch.[19] 842 were users predominantly using a single other language.[20] 54 users used two or more languages in their tweets, 6 were clearly recognizable as bots, and for 4 accounts it would seem that the later tweets were produced by another author than the earlier tweets.[21]

On the benchmark set, the automatic classification of users as Dutch has a precision of 96.4%. However, an investigation of the false positives shows that they are all code switchers, reused accounts and bots, each of which does produce a significant number of Dutch tweets, which with less strict criteria would imply a precision of 100%.[22] The recall is more disappointing, at a mere 80.1%, leading to an F-measure of 88.3%.

These numbers, though, are misleading. As I already mentioned, the sample is biased towards non-Dutch, whereas the collection as a whole is biased towards Dutch. Assuming that the benchmark classification quality per combination is an acceptable approximation of the quality over the whole collection for that combination, I extrapolated the results to the full collection. The results for the overall classifier, as well as for each individual language indicator, are shown in Table 1.[23] The individual indicators perform acceptable to good. The two that aim for verifying the use of Dutch (TwiNL percentage, Top-20 in OpenTaal) have a good precision, but their recall is far too low.[24] The two selecting Dutch in comparison with other languages (TwiNL most suggested, OOV profile) have a far better recall, but the TwiNL most suggested indicator pays for this in precision. Clearly, the voting combination is the best choice overal, showing a very high F-measure.

[18] If there were fewer than 10 users with the combination, I inspected all of them. Also note that a halfway redesignation of indicators led to more than 10 users for some combinations; I did not resample down to 10, judging that more information is more useful than a complete balance here.

[19] As already indicated, most Dutch users are concentrated in a small number of combinations, so that most of the combinations are linked to users with other languages.

[20] 3 of these used Afrikaans, which is derived from Dutch and therefore has many overlapping words.

[21] Potentially, this can be confirmed by the Twitter metadata, but at the time of writing this paper, this metadata was not yet available to me.

[22] However, only a sample was tested. Still, if we would apply less strict criteria, a precision close to 100% should be feasible.

[23] The percentage of foreign words/numbers and symbols is not listed separately, as it can only vote against Dutch, and never suggest it.

[24] This confirms that the classifier used by TwiNL for language recognition on individual tweets is indeed performing much worse than one would expect based on the literature. It would be interesting to compare to other state-of-the-art language identifiers, but this will have to be relegated to future work.

I therefore chose to keep all users that were identified as being predominantly Dutch by the combination system. Of the 5,883,805 original users, 2,638,224 (45%) were thus filtered out. However, as only a fraction of these users' tweets were included in the TwiNL collection, this filtering only reduced the original 1,985,417,436 tweets by 114,563,665 (5.8%). After filtering, then, there were 3,245,581 users left, who tweeted predominantly in Dutch, and together produced 1,870,853,771 tweets, from which 997,447,015 were used because of sampling only around 1,000 tweets for the highly productive users (478,509 users).

4. Distinguishing individual humans

The induction of metadata like gender and age only makes sense for individual humans, making it necessary to single out such users first.[25] These need to be distinguished from such users as bots,[26] editorial boards, and groups sharing an account.[27]

Previous work has focused on the recognition of bots, as these may be felt as an annoyance and potentially sometimes even a threat. Examples of bot threats are campaigns using bots to influence human opinion (e.g. with political aims) or to create a false impression for systems measuring the state of the world by monitoring Twitter (e.g. to manipulate the stock market).[28] In the early stages, bot recognition focused on features such as posting time patterns, use of hashtags and URLs, and, as the annoyance was mostly connected to spam, spam recognition (Chu et al. 2012). Later, as more subtle bots and bot use started to appear, bot recognition research switched to a more adversarial scenario: recognition should take into account that bots mimic some human behaviour patterns in order not to be recognized.[29] The result is a much larger range of features, including new ones like information derived from the Twitter profile, the structure of the relation network, the use of specific parts of speech in the tweet text, and the use of sentiment words (Ferrara et al. 2015).[30] Most of these features were either unavailable to me at the time of the experiments (lacking Twitter metadata and a reliable POS tagger for Twitter Dutch), or needed a substantial amount of gold truth data for proper training (also unavailable). I therefore postponed the full range of possible features to future work, and for the time being focused on overall text characteristics, both for recognizing bots and groups of humans.

Now, as is well-accepted, all humans have a specific writing style, by which they can be often be distinguished from other humans. Various measurements on their texts, which are taken to characterize this style, can be expected to be distributed in a specific range, and the distributions can be compared to differentiate between users. This being the case, it should also be possible to distinguish between the distributions of single and multiple authors, the latter being more diffuse. However, to test this hypothesis, again a substantial amount of ground truth data would be necessary. Without ground truth data, distinguishing between bots and humans is a more realistic goal. First

[25] Other types of feeds should not be filtered out, but set apart, so that they can be handled through other processes. Unlike in the previous section on language, the aim here is not to rule out specific users, but merely to add metadata, as all types of users can be useful in some types of research.

[26] A special case is formed by accounts where human writing is mixed with automated production, often referred to as "cyborgs" in the literature (e.g. Chu et al. 2012). An example are human users who let most of their tweets be sent by apps, e.g. reporting on today's running session. Ideally, each user should be characterized by the percentage of human-authored tweets, but mostly we see a separate cyborg class in the classification.

[27] Sharing can be simultaneous, a number of people all allowed to post under the same username, but there are also cases where a username gets reused some time after the previous owner stops using it, as we see in the language filtering tests. In principle, these cases should be traceable on the basis of Twitter metadata, but as already mentioned, the current experiments were conducted without access to metadata other than time and username.

[28] See (Ferrara et al. 2015) for a more extensive discussion.

[29] As an example, we see a Dutch user with a perfectly human looking username and profile, containing a common first name and even mentioning being the father of two children. However, this user posted on average about one tweet every three minutes during Februari 2015, practically all news items and commercials, something unimaginable for an (unaided) normal human.

[30] For a fairly complete overview of features applied to user characterization in general, and examples of studies where they were used, see (Cossu et al. 2015).

of all, there should be a much stronger difference between bot and human than between a single human and a group. Bots may be able to produce pieces of text that looks like text written by a human, but they will for sure not produce series of texts that show the same variation as series of texts produced by humans. Furthermore, it is much easier to verify suggested bots. When given a list of suggestions for users that might be bots, it is possible (to a large extent) to manually check that list, which would be much harder for groups of humans writing about a shared topic. I therefore decided to set up an experiment in just this way: I compiled various text characteristics by which bots should be recognizable, let a system provide a list of suggested bots, and checked samples from this list.[31]

4.1 Potential characteristics to recognize bots

As just mentioned, humans produce texts with characteristics with specific distributions. These distributions emerge from the different topics the human tweets about, which most likely stays reasonably constant, the writing styles that are appropriate for those topics, and of course the personal writing style of the author. Now it can be expected that the distributions for bots deviate markedly from those of humans. They may be more consistent, when generating text themselves, or less consistent, when scraping text from random other accounts. They may write about fewer topics than humans, e.g. marketing a specific product, or about more topics, e.g. forwarding all news items from various newsfeeds. But in all cases, I expect bots to be more extreme than humans on several measurements. If I am correct, then bots should be found at the extreme ends of the measurement distributions. In order to attempt to recognize them, I took a number of fields from the user profiles described in Section 2, namely

vol The volume of tweets produced over the 2.5 years covered by the dataset. (1 field)

len Mean tweet length in tokens. (1 field)

pun Mean punctuation-to-word ratio. (1 virtual field, comprising the ratio of 2 actual profile fields)

wor Predominant use of standard constructions as witnessed by a high mean fractions of frequent tokens, mean U-score of frequent tokens, and fraction of Dutch words in the 20 most frequent words, and a low fraction of OOV words. (4 fields)

nzp Mean non-zipfiness. (1 field)

vcr Vocabulary richness, as represented by the type-token ratio and the percentage of hapaxes. (2 fields)

url Mean number of URLs per tweet and fraction of URLs in all tokens. (2 fields)

usr Mean number of user mentions per tweet and fraction of user mentions in all tokens. (2 fields)

hsh Mean number of hashtags per tweet, fraction of hashtags in all tokens, fraction of hashtags in OOV tokens, and fraction of hashtags in OOV types. (4 fields)

con Consistency, as represented by the variation indicator for the numbers per tweet of tokens, URLs, user mentions, hashtags, OOV words, and frequent tokens, and for the mean U-score of frequent tokens. (7 fields)

The determination of whether a value "extreme" is based on what I call the Core Z-score. As it is unknown which fraction of the users is covered by special feeds, it is also unknown to which degree the standard Z-score would be affected by these feeds. However, if we assume that most of the users

[31] Obviously, in this way it is only possible to test precision, and not recall. For recall, again ground truth data would be needed.

Table 2: Users with the most extreme characteristics from the highly productive group(at least 6 characteristics at highest level)

User	vol	len	pun	wor	nzp	vcr	url	usr	hsh	con
prijs_klapper			++	- -	++	- -	++		++	++
_goedkoop			++	-	++	- -	++		++	++
kerstavond			++	-	++	- -	++		++	++
_Kerst_man			++	-	++	- -	++		++	++
Kerstshopping			++	-	++	- -	++		++	++
prijskijk			++	-	++	- -	++		++	++
Noord_Holland_	++			- -	++	++	++		++	+
mijnsexyfoto	++		++	- -	++	+	++		++	
LifeStyle_Heads	++			- -	++	++	++		++	
OnderwerpenInfo	++			- -	++	++	++		++	
watkopenwij	++			- -	++	++	++		++	

are "normal" users, we can also assume that the center of the distribution is dominated by these "normal" users. Therefore, for each measure, I selected those users for whom the measurement did not deviate more than one standard deviation from the mean, i.e. for whom the standard Z-score was between -1 and 1. I then recalculated the mean and standard deviation on the basis of only these "core" users, and calculated the Core Z-score for all users with regard to these recalculated mean and standard deviation.

After an inspection of the resulting numbers, I intuitively set thresholds at Core Z-scores of 4.5 (class I), 9 (class II) and 13.5 (class III). Each of the ten characteristics is marked as being at the highest level of extremity (marked as ++ or – in Tables 2 and 3) if it has at least one field in class III or at least two fields in class II. The exception is hashtag use, as its four fields are highly correlated; here I demand at least two fields in class III or all four in class II. A high level (marked as + or - in Tables 2 and 3) is assigned when at least one field is in class II or at least three fields are in class I. Again, hashtags are different, requiring at least two fields in class II or 4 fields in class I. One might also expect higher requirements for consistency, as this comprises 7 fields, but these are much less correlated and any one field suffices as a strong indicator.

4.2 Results

That extreme values of these characteristics indeed indicate bot authors (or other special feeds) can be seen in Tables 2 and 3. Table 2 shows the users with the highest number of extreme characteristics out of all highly productive users, i.e. the 478,509 users with at least 1,000 tweets in the period covered by the dataset. Table 3 does the same for the other 2,767,072 users. It is unclear whether the user OnlineKakhiel is a bot. Its tweets all consist of a forwarded funny photo with hardly any text (and usually the same), and the forwarder might well be human. Still, the nature of the feed clearly places it under the specialized feeds.

We see that many extreme users have especially high URL and hashtag use, as expected given the primary function of such feeds: to "sell" something and/or to get the reader to click through to a website. The third type of metatoken, user mentions, is only used very frequently by DevC0n, that sends basically the same promotional message directly to many different users. Clearly, this is a marker, but a much rarer one. All characteristics are present for many extreme users. Tweet length is present as well, although it is not visible in the tables, as it occurs for the first time at rank 41 in the highly productive group, observed for user Weerstationblij, which also shows a high punctuation-word ratio, low use of standard constructions, high URL use, and high consistency. The

Table 3: Users with the most extreme characteristics from the less productive group(at least 4 characteristics at highest level and 5 at next level)

User	vol	len	pun	wor	nzp	vcr	url	usr	hsh	con
Bikini_show				- -	++	- -	++		++	
shalinilindveld			++	- -		- -	++		++	
CHEERS12345			++	-	++	- -			++	+
CHEERS130613			++	-	++	- -			++	+
CHEERS112			++	-	++	- -			++	+
gewoonbestellen					++	- -	++		+	++
OnlineKakhiel					++	- -	++		+	++
parkeerkorting					++	- -	++		++	+
DevC0n				-		- -	++	++	++	
ArtsVacatures				- -		-	++		++	- -

Table 4: Scores given in June 2015 by Bot or Not? to the suggested bots. For an indication of the features covered by the various scores, see Ferrara et al. (2015). Note that scores under Content and Sentiment are tentative as Bot or Not? is aimed primarily at English.

User	Network	User	Friends	Timing	Content	Sentiment	Overall
prijs_klapper	65%	1%	79%	35%	36%	57%	24%
_goedkoop	75%	6%	79%	54%	38%	55%	24%
kerstavond	70%	3%	79%	48%	36%	50%	22%
_Kerst_man	74%	2%	79%	55%	39%	56%	26%
Kerstshopping	68%	1%	79%	41%	38%	53%	23%
prijskijk	67%	2%	79%	69%	37%	52%	26%
Noord_Holland_	84%	58%	79%	80%	59%	71%	67%
mijnsexyfoto	51%	35%	79%	64%	62%	17%	54%
LifeStyle_Heads	35%	71%	66%	44%	48%	71%	35%
OnderwerpenInfo	84%	3%	79%	67%	74%	64%	34%
watkopenwij	46%	10%	55%	35%	49%	25%	38%
Bikini_show	58%	24%	79%	81%	60%	49%	35%
shalinilindveld	96%	53%	79%	84%	95%	80%	65%
CHEERS12345	64%	22%	79%	86%	48%	25%	42%
CHEERS130613	60%	28%	79%	79%	48%	30%	41%
CHEERS112	70%	9%	79%	87%	51%	29%	43%
gewoonbestellen	account suspended on check in June 2015						
OnlineKakhiel	68%	82%	70%	37%	56%	67%	62%
parkeerkorting	28%	10%	79%	62%	72%	54%	20%
DevC0n	33%	39%	32%	47%	58%	43%	45%
ArtsVacatures	46%	47%	85%	74%	61%	36%	49%

tables also show that a single characteristic will not suffice as every marker is missing (or weaker) for some of even these most extreme users.

All in all, it would seem that all proposed characteristics can be applied in the recognition of bots and other specialized feeds. Furthermore, seeing the scores that Bot or Not? (Ferrara et al. 2015; http://truthy.indiana.edu/botornot) assigns to the suggested accounts (Table 4), the new characteristics may complement the existing ones nicely.[32] However, the application of the new characteristics is not straightforward. Although we see extreme values overall in specialized feeds, not all such feeds have extreme values on all characteristics, and individual human users may well also have extreme values on the characteristics. The final choice is not just a matter of setting a few thresholds and, as we have seen, a single characteristic will not be sufficient. Which brings us back to the necessity of gold truth metadata, in order to investigate whether machine learning can provide a usable relation between vectors of characteristic scores and the nature of the feed. Obviously, when taking this next step, it is also time to involve a substantial selection of previously suggested features in the recognition process.

5. Identifying Tribes

As shown by Oostdijk and van Halteren (2015), there are clear differences in language use within the Dutch tweet population, which are related to the users producing the tweets and the situation in which they produce them. For processing, both in building a data collection for research and in the research itself, it is advantageous to identify the various "tribes" and process each with the most appropriate techniques and models. Implementing a full tribe recognition system is out of scope for this paper, but I will try to provide a proof of concept by attempting to identify the (very large) tribe of (high) school children.[33] To my knowledge, this exact classification task has not been worked on previously, although there is of course overlap with various age recognition tasks (for example, see Rangel et al. (2014), Nguyen et al. (2013), Peersman et al. (2011)).

5.1 Identification strategy

Given the success of distinguishing between men and women by van Halteren and Speerstra (2014), I decided to use exactly the same approach here, namely applying an SVR-based classifier to vectors representing the unigram counts in the users' tweets.

In the absence of a manually selected Gold Standard dataset for training and testing, I used an automatic selection procedure for experimental data that ought to come very close to the Gold Standard. As all Dutch children are obliged by law to attend school at least up to the age of 16,[34] and all schools give their pupils some kind of homework, it is safe to assume that such children will be tweeting about homework and/or other school-related topics. I therefore selected positive examples, i.e. users who are probably school children (henceforth group S for "school"), by looking for the use of *hw* and *#hw*, the standard abbreviation for *huiswerk* ("homework"). To be exact, *hw* or *#hw* needed to present during the last year in the test period (i.e. from July 2012 to June 2013). Of the highly productive users, 138,899 (29%) showed one of these words in the last year.[35]. Among the less productive users, we find only 145,375 (5%) with *hw* and/or *#hw*.[36] Clearly, the school children are not only a major user group on the Dutch part of Twitter, but they are generally also very prolific writers. For our experiment, there is no shortage of positive examples.

For negative examples (henceforth group O for "other"), I looked for users that did not have any of a number of school-related words in their whole tweet collection, such as *school* ("school"), *leraar* ("teacher"), *hw* ("homework"), *huiswerk* ("homework"), *wiskunde* ("mathematics"), *duits* ("German"), *gs* (short for *geschiedenis*, "history"), and *ak* (short for *aardrijkskunde*, "geography"). As most of these words can also be used in other situations than children talking about their work,

[32] Although it must be taken into account that Bot or Not? is aimed at English.

[33] This first differentation also has practical use, as this tribe rather dominates the population of Dutch Twitter users, and actually needs to be set apart in order to be able to model the other users. At least, this domination was visible in 2011 and 2012. There are signs that more recently the user community is shifting ((Turpijn et al. 2015)). This will be one of the topics of future research.

[34] And possibly even up to the age of 18. They are allowed to stop once they are older than 16 and have a diploma.

[35] Looking over the whole collection period, 270,273 (56%) of the users use *hw* and/or *#hw*

[36] And looking over the whole period, 323,465 (12%).

it was to be expected that this ruled out more than just these school children, but it is safer to err on the side of caution. However, even fewer users than expected were found, especially among the highly productive users, with only 862 (0.2%) without the selected words. In order to be able to select a sufficiently large sample of users for the classification experiment, I decided to apply the criterion loosely for this group (see below). For the less productive users, the situation was better, with 859,779 (31%), and the original criterion of no words from the list could be maintained.

Within each productivity class (highly productive, less productive) I sampled 3,000 users from group S and 3,000 users from group O. The positive examples were sampled randomly from the users for whom *hw* and *#hw* together were (on average) present during the last year at least once in a thousand tokens. For the negative samples, for the less productive users, I also sampled randomly from those using no school-related words; for the highly productive ones, I took the 3,000 users whose use of school-related words was the lowest and who never used *hw* or *#hw*. For each user I built a unigram frequency vector.[37] I used all unigrams found for at least 1 in 1,000 users in the full user group in question (highly and less productive), but excluding all unigrams used in selecting classes S and O, such as *hw* and *gs*. This led to vectors of about 12,000 counts for the highly productive users and about 16,000 counts for the less productive users. In building the vectors, I used the counts from the three token lists included in the user profiles rather than recounting in the full dataset. As the classification method considers both overuse and underuse, there might be a problem for the tokens that were present in a user's tweets but unlisted as they were at a rank higher than 100. I therefore assigned all unlisted tokens the relative frequency of the least frequent token in the corresponding list.

I also chose the best performing machine learning approach in the gender recognition experiment, namely Support Vector Regression (ν-SVR from LIBSVM; Chang and Lin 2001) with an RBF kernel. Rather than using fixed hyperparameters, I let a control shell choose them automatically in a grid search procedure, based on development data. When running the underlying systems themselves, I used various hyperparameter settings: the cost factor C was 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. For each setting and author, the systems reported both a selected class and a floating point score, which could be used as a confidence score. For each individual author, the control shell examined the scores for all other authors in the same fold in the applied 3-fold cross-validation.[38] It then calculated a class separation value, namely the difference between the mean scores for each of the two classes (S and O), divided by the sum of the two standard deviations.[39] The optimal hyperparameter settings were assumed to be those where the two classes were separated most, i.e. where the class separation value was 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.

5.2 Results

When testing the classifier in 3-fold cross-validation (separately for the two productivity classes), the results are very good for the highly productive users, with a precision in recognizing Group S of 96.6% and a recall of 97.1%, hence an F-value of 96.9%. For the less productive users, the scores are somewhat lower, with a precision of 94.3%, a recall of 92.5%, and an F-value of 93.4%. However, these measurements are against the estimated gold standard. I manually checked the 68 users that were misclassified in fold 3 for the highly productive class. Of the 38 "false" accepts, 1 can be shown to be a correct accept after all on the basis of the content, and 13 certainly sound like school children but no irrefutable evidence can be found in the tweet content. Then there are 11 users

[37] I chose unigrams as they proved the best feature type in last year's gender recognition experiment on this type of data (van Halteren and Speerstra 2014).

[38] This gave the best chances that the selected optimal hyperparameters generalize to the author in question.

[39] The class separation value is a variant of Cohen's d (Cohen 1988). Where Cohen assumes the two distributions have the same standard deviation, I used the sum of the two, practically always different, standard deviations.

Table 5: Most typical unigrams for Group S

Unigram	Gloss	Contribution towards S	S users with word	O users with word
echt	really	0.69	99%	56%
ik	I	0.63	100%	71%
<usr>	user mention	0.52	100%	79%
niet	not	0.47	100%	84%
x	kiss	0.44	94%	40%
me	me	0.39	100%	58%
<emo>	emoticon	0.38	98%	59%
slapen	to sleep	0.37	98%	40%
zo	so, soon	0.36	100%	68%
wakker	awake	0.36	98%	41%
moet	must	0.35	99%	57%
”	”	0.34	77%	44%
was	was	0.33	99%	57%
morgen	tomorrow, good morning	0.32	99%	64%
eten	to eat, food	0.32	97%	44%
die	that, those	0.31	100%	77%
maar	but, only	0.31	100%	79%
ben	am	0.30	100%	61%
haha	haha	0.29	95%	42%

that show the language use of school children, but probably are not themselves school children (any more). Finally, there are 13 bots and/or special feeds, which should be identified at another point in the classification process. Depending on how exactly we want to delimit Group S, by physical age or by language use, then, there are either 24 actual false accepts, 11, or none at all. In the 30 “false” rejects, we do see 23 real false rejects, as the tweet content of these 23 shows them to be indeed school children. 2 of them can be shown to be school children at the start of the test period, but to leave school later on. Finally, in this group we find 5 bots and/or special feeds, for some of which the inclusion in Group S was erroneous as they used *hw* in another sense, often somebody's initials.[40] All in all, it would seem that the accuracy scores are actually an underestimation of the true classification accuracy.

If I apply the resulting classifiers, each in their own productivity class, to all the users in the collection (i.e. everybody who produced at least 10 tweets and at least 100 tokens), they estimate that 364,724 (76%) of the highly productive users are school children and 860,849 (31%) of the less productive ones. Over the whole collection, this sums to 1,225,573, i.e. 38%, and these 38% of the users account for 73% of the tweets in the collection.

5.3 Discerning unigrams

The unigrams that contribute most to the classification are listed in Tables 5 and 6. Where the most discerning words for gender recognition consist almost exclusively of content words (van Halteren and Speerstra 2014), this is much less the case here. Apart from typical daily activities (*slapen, wakker, eten*), there are more function-word-like tokens for Group S, connected to personal communication about activities and emotions. For Group O, there is more non-personal content, expressed in a more standard written-text format, as indicated by the presence of articles, prepositions and the remarkably coherent group of punctuation marks. Then, there a some words (*heerlijk, goedemorgen,*

[40] In later research, it turned out that *hw* also represents the weekly newspaper *Harener Weekblad.*

Table 6: Most typical unigramss for Group O

Unigram	Gloss	Contribution towards O	O users with word	S users with word
<url>	URL	0.65	85%	73%
de	the	0.65	97%	100%
een	a, an	0.48	94%	100%
in	in	0.33	93%	100%
voor	for, before	0.32	93%	100%
op	on	0.31	96%	100%
van	of, from	0.29	96%	100%
het	the, it	0.26	94%	99%
onze	our	0.24	43%	28%
heerlijk	delicious	0.24	40%	12%
)	)	0.23	57%	44%
(	(	0.22	54%	50%
!	!	0.22	80%	91%
.	.	0.22	89%	96%
goedemorgen	good morning	0.21	25%	4%
nieuwe	new	0.21	78%	88%
via	by way of	0.21	41%	17%
te	to, too	0.19	88%	97%
uw	your (polite)	0.18	15%	0%

uw) that appear to be out of fashion with the younger crowd, and probably replaced by competing words (e.g. *lekker, morgen, je*). From the contribution scores and the percentages it becomes apparent that Group O is much more diffuse, although the top position of the URL indicates that it contains a large contingent of professional feeds.[41]

6. Conclusion and Future Work

In this paper, I describe some initial experiments concerning the induction of three fundamental metadata fields which are necessary for proper induction of more detailed metadata, namely a) whether the user is tweeting predominantly in Dutch, b) whether the user is an individual human, and c) whether the user is a boy or girl at school-going age.[42] All users with at least 10 tweets and at least 100 tokens present in the TwiNL dataset were considered. Pending the collection of extra-linguistic information from Twitter metadata, all experiments considered only the text of the users' tweets present in the TwiNL dataset.

Identifying whether a user's predominant language on Twitter is Dutch proves to be possible with high accuracy (Section 3). The estimated F-value over the whole tested user population is over 98%. As this is on the basis of text alone, and some potential improvements to the classification method are already suggested, this score can probably be improved upon. Still, even the current classification seems good enough to work with. However, this counts on having a minimum amount of text for a user, which in this paper is set to 10 tweets and 100 tokens. When combining the data-based classification with the metadata-based one, special attention should be given to users with less

[41] But remember that Group O is only a part of the non-school-children users, as it was selected by having no mention of school matters at all, also ruling out e.g. parents and teachers. A proper investigation of interests and language use will have to wait until the tribal classification is more advanced.

[42] The latter fact may seem less fundamental, but certainly is, as this group dominates the Dutch part of Twitter in the period studied.

text than this. Furthermore, there is quite some code switching visible in the tweets, so that other possible next steps would be to classify users not in a binary fashion (Dutch versus non-Dutch) but rather along a scale (percentage of Dutch), and to annotate the individual tokens in the tweets for language.

For recognizing whether a user is an individual human, as opposed to a group of humans, an editorial board, or a bot, a number of text characteristics are shown to be useful in classification (Section 4). A proper determination of the potential classification quality has to wait for the aforementioned collection of Twitter metadata, as well as for a sufficient amount of gold truth development material. However, it is already clear that some thought will have to go into the definition of the classes here. Feeds related to news, the job market, and marketing, tend to show a similar language use whether they are produced automatically by a bot or selected and transmitted by one or more moderators. It could therefore be argued that it is not the humanness of the user that should be marked, but the type of feed, which means that most of the identification of individual humans is merely a subtask of the determination of the tribe hierarchy, as addressed in Section 5. It is as yet unclear whether this will also be able to cover the recognition of feeds authored by more than one human, as this seems to be a completely different dimension of the nature of the text.

With relation to the identification of tribes inside the tweet collection, i.e. tweets by specific users and about specific domains, this paper only investigates the largest of tribes, boys and girls of school age, tweeting predominantly about their daily life and emotions (Section 5). Identification of this tribe is the most urgent, as it dominates the tweet collection and, as a result, hinders the induction of a number of metadata types. For this tribe, the recognition quality on text alone is promising. For highly productive users, recognition is about 97%. That a large amount of text is necessary can be seen from the quality for less productive users, only about 93.5%. However, for all users we can expect an increase in recognition quality when also taking into account extra-linguistic features such as username, relation networks and time patterns in posting.

All in all, the text-based classification appears to be a good basis for metadata induction. In combination with other information sources, such as the already mentioned extra-linguistic features, it should be possible to induce the three addressed metadata fields reliably enough for most research purposes, after which it becomes possible to advance to the induction of further metadata.

References

Chu, Z., S. Gianvecchio, H. Wang, and S. Jajodia (2012), Detecting automation of twitter accounts: Are you a human, bot, or cyborg?, *IEEE Transactions on Dependable and Secure Computing* **9** (6), pp. 811–824.

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

Cossu, Jean-Valère, Vincent Labatut, and Nicolas Dugué (2015), A review of features for the discrimination of twitter users: Application to the prediction of offline influence, *arXiv:1509.06585v1.*

Ferrara, Emilio, Onur Varol, Clayton Davis, Filippo Menczer, and Allessandro Flammii (2015), The rise of social bots, *arXiv:1407.5225v2.*

Lui, Marco and Timothy Baldwin (2014), Accurate language identification of twitter messages, *Proceedings of the EACL 2014 Workshop on Language Analysis in Social Media (LASM 2014), Gothenburg, Sweden*, pp. 17–25.

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).*

Oostdijk, Nelleke and Hans van Halteren (2015), Twitter tribal languages, *Handbook of Twitter for Research.*

Peersman, Claudia, Walter Daelemans, and Leona Van Vaerenbergh (2011), Predicting age and gender in online social networks, *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, SMUC '11, ACM, New York, NY, USA, pp. 37–44.

Rangel, Francisco, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans (2014), Overview of the 2nd author profiling task at pan 2014, *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers.*

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.

Turpijn, Loes, Samantha Kneefel, and Neil van der Veer (2015), Nationale social media onderzoek 2015, *Technical report*, Newcom Research & Consultancy B.V.

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

van Halteren, Hans and Nelleke Oostdijk (2015), Word distributions in Dutch tweets: a quantitative appraisal of the distinction between function and content words, *Tijdschrift voor Nederlandse Taal- en Letterkunde* pp. 189–226.

Zipf, G. (1935), *The psycho-biology of language*, Houghton Mifflin.