Predicting Age and Gender in Online Social Networks

Claudia Peersman
Antwerp University & Artesis
CLiPS
Lange Winkelstraat 40
BE-2000 Antwerp (Belgium)
(+32) (0)3 265 5225
claudia.peersman@ua.ac.be

Walter Daelemans
Antwerp University
CLiPS
Lange Winkelstraat 40
BE-2000 Antwerp (Belgium)
(+32) (0)3 265 5222
walter.daelemans@ua.ac.be

Leona Van Vaerenbergh
Artesis University College Antwerp
Translating and Interpreting
Schildersstraat 41
BE-2000 Antwerp (Belgium)
(+32) (0)3 240 1901
leona.vanvaerenbergh@artesis.be

ABSTRACT
A common characteristic of communication on online social networks is that it happens via short messages, often using non-standard language variations. These characteristics make this type of text a challenging text genre for natural language processing. Moreover, in these digital communities it is easy to provide a false name, age, gender and location in order to hide one's true identity, providing criminals such as pedophiles with new possibilities to groom their victims. It would therefore be useful if user profiles can be checked on the basis of text analysis, and false profiles flagged for monitoring. This paper presents an exploratory study in which we apply a text categorization approach for the prediction of age and gender on a corpus of chat texts, which we collected from the Belgian social networking site Netlog. We examine which types of features are most informative for a reliable prediction of age and gender on this difficult text type and perform experiments with different data set sizes in order to acquire more insight into the minimum data size requirements for this task.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *text analysis*.

General Terms
Experimentation.

Keywords
Text categorization, stylometry, age and gender prediction, cyber-pedophilia detection, social media.

1. INTRODUCTION
In recent years, online social networks like Facebook, MySpace, Bebo, Hyves and Netlog have expanded impressively and have enabled millions of users of all ages to develop and support personal and professional relations. However, a common characteristic of these digital communities is that it is easy to provide a false name, age, gender and location in order to hide one's true identity, providing criminals such as pedophiles with new possibilities to groom their victims. When attempting to detect these Internet predators, both law enforcement agencies and social network moderators are confronted with two main problems: (i) the vast number of profiles and communications on social networks make manual analyses virtually impossible and (ii) Internet predators often create a false identity, posing as adolescents, in order to make contact with their victims. Therefore, efficient automated methods for identity detection and checking are becoming necessary.

Recent advances in natural language processing technology have enabled computational linguists to predict an author's age (group) and gender in several text genres by automatically analysing the variation of linguistic characteristics. However, in social networks, computational linguists are confronted with several issues. First of all, little information about the users' gender, age, social class, race, geographical location, etc., is available to researchers. Most online social networks do not provide open access to the users' profile data, so it is difficult to collect training data for this task. Secondly, communication in online social networks typically occurs via posts on guestbooks, blogs, walls, etc. These are typically very short messages, often containing non-standard language usage, which makes this type of text a challenging text genre for natural language processing. Finally, given the speed at which chat language has originated globally and continues to develop, especially among adolescents, a third challenge in automatically detecting false profiles on social networks will be the constant retraining of the machine learning algorithms in order to pick up new variations of chat language usage that are linked to age and/or gender.

This paper presents a study on a corpus of 1,537,283 Flemish Dutch posts from the Belgian social networking site Netlog, which we were able to obtain together with the users' profile data. We investigate the feasibility of automatically predicting age and gender on short chat messages and examine which types of features are most informative for this application. Since it is our main objective to develop a useful component in a pedophile detection system, contrary to previous research on age prediction (see Section 2), which mainly focused on predicting age groups (e.g. 10s, 20s, 30s), we focus on classifying adults versus adolescents. This way, the system should be able to detect adults posing as adolescents and flag their profiles for monitoring. We include gender detection in our study because the vast majority of pedophiles are male, and mismatches between profile gender and predicted gender can therefore also be useful. Furthermore, (predicted) gender can be a helpful information source in

SMUC'11, October 28, 2011, Glasgow, Scotland, UK.

constructing more accurate classifiers for age. Finally, we also perform experiments with different data set sizes in order to acquire more insight into the minimum data set size requirements for future retraining experiments.

Our approach to this computational stylometry task is based on text categorization, and involves the creation of document representations based on a selected set of (patterns) of features, feature selection using statistical techniques, and classification using machine learning algorithms.

Section 2 of this paper surveys the literature on age and gender prediction, together with a summary of prior research on text categorization with short texts and the effect of data set size. In Section 3 we present our data and give an overview of its characteristics. In Section 4 and 5 we describe our computational approach and we present our results. We conclude this paper in Section 6 by summarizing and discussing our research contributions and by anticipating on our future research.

2. RELATED WORK

We address three main problems in this paper: (i) the feasibility of detecting age and gender using a text categorization approach on the text genre of chat, (ii) the usefulness of the approach when confronted with very short texts, often containing non-standard language usage, and (iii) the minimum training data size that is required for a reliable performance. We describe related research in these areas.

2.1 Age and Gender Related Research

Recent advances in natural language processing technology have enabled computational linguists to perform automatic linguistic analyses of lexical, morphological, and syntactic properties of texts. Consequently the study of the variation of linguistic characteristics in texts according to the authors' age or gender has already become feasible on literary corpora (e.g. Pennebaker et al., 2001; Pennebaker and Stone, 2003; Holmes and Meyerhoff, 2003; and Burger and Henderson, 2006). Argamon et al. (2002) were able to predict authors' gender with approximately 80% accuracy, analyzing a large corpus of formal written texts (fiction and non-fiction) from the British National Corpus. Using 42,000 words on average as training data, they found that best performance was achieved when combining both function words distributions and part-of-speech n-grams as features.

With regard to Computer-mediated Communication (CMC), there have been several studies on gender prediction of blogs (e.g. Sarawgi et al., 2011; Mukherjee and Liu, 2010; Argamon et al., 2007; Nowson et al., 2007; Yan and Yan, 2006). Argamon, et al. (2002) found that women tend to use more personal pronouns and negation, along with some function words like "for", "with" and "and", whereas male authors use determiners and numbers, along with "he" and "of" more frequently. Herring and Paolillo (2006) then selected a set of male and female preferential specific word forms, following the model of Argamon et al. (2002), to investigate whether gender or genre is a stronger predictor of linguistic variation in weblogs, but only found a correlation between language variation and genre. Only one study (so far) by Zhang and Zhang (2010) provides experiments with short segments of blog posts (15 tokens per segment, 10,000 segments), which produced a best accuracy of 72.10% for gender prediction.

With regard to age prediction on CMC, Tam and Martel's (2009) Support Vector Machine model was able to yield a 0.996 f-score when distinguishing teens from adults using word trigram features, but they used an unbalanced data set with 1263 documents in the adult class versus only 465 in the teens class. Rosenthal and McKeown (2011) were able to predict if a blog author was part of a pre- or post social media generation with an accuracy of 81.57%.

Few studies have investigated both age and gender related language variation. Schler et al. (2006) gathered a corpus of over 71,000 blogs and extracted style-based features (non-dictionary words, parts-of-speech, function words and hyperlinks) and content-based features (content-based single words with the greatest information gain) for both age and gender prediction of the blogs' authors. Their research showed that, regardless of gender, language usage in blogs correlates with age: pronouns and assent/negation become scarcer with age, while prepositions and determiners become more frequent. Argamon et al. (2007) and Goswami et al. (2009) expanded the research of Schler et al. (2006) by adding non-dictionary words and the average sentence length as features. They found that teenage bloggers tend to use more non-dictionary words than adult bloggers do. Furthermore, the stylistic difference in usage of non-dictionary words combined with content words allowed to predict the age group (10s, 20s, 30s or higher) with an accuracy of 80.32% and gender with an accuracy of 89.18%. The average sentence length did not correlate with age or gender. Using a linear regression model based on shallow text features and adding gender as feature, Nguyen et al. (2011) obtained age-based correlations up to 0.74 and mean absolute errors between 4.1 and 6.8 years on a joint corpus of blogs, telephone conversations, and online forum posts. However, all of these studies (except Zhang and Zhang, 2010) have worked with text fragments that contained a minimum of 250 words.

To our knowledge, there is only one previous study that analyses language variation in online social networks. In their demographic study on MySpace, Caverlee and Webb (2008) provide lists of distinguishing words for both age and gender, which they selected using the Mutual Information metric, but they did not apply any machine learning algorithms.

2.2 Working with Small Data Sets

In a wider stylometry context, most traditional studies use large sizes of training data with a limited set of authors, which usually leads to a better performance of the machine learning algorithms. As was mentioned in Section 2.1 the large majority of the previously mentioned studies have worked with text fragments that contained a minimum of 250 words. In stylometry research, the effect of data size has not been researched in much detail yet, since most of those studies tend to focus on long texts or several short texts per author (Luyckx and Daelemans, 2010). Burrows (2007) regards 10,000 words per author to be a reliable minimum for an authorial set.

In the context of authorship attribution, a few studies focus explicitly on data set size. In a study on short texts by the Brontë sisters Hirst and Feiguina (2007) found that using multiple short texts can reduce the problem of dealing with short texts, even when 'short' means only 200 words per author. Sanderson and Guenter (2006) stated that 5000 words in training could be considered a minimum requirement. When reducing the number of words per text fragment to 100 words, Luyckx and Daelemans (2010) reported a dramatic decrease of the performance of the text categorization approach for authorship attribution.

With this paper we show that a text categorization approach to the identification of age and gender can also be used with sufficient reliability in social network communication despite its challenging characteristics (short texts often containing non-standard language). In addition, we introduce a new corpus of Dutch chat language. Dutch chat shows even more non-standard

variation than English chat language in that it more strongly reflects dialectical influence. It is to the description of this corpus and of Dutch chat language that we turn next.

3. THE NETLOG CORPUS

3.1 Structure

Netlog is a Belgian online social networking platform, which focuses mainly on European adolescents and has over 67 million members, utilizing over 37 different languages. Members can create a profile page containing blogs, pictures, videos, events, playlists, etc. that can be shared with other members. For this study, we obtained a collection of 1,537,283 Flemish Dutch Netlog posts containing 18,713,627 tokens (i.e. words, emoticons and punctuation marks) in total, for which we also received information about the age, gender and location of the authors. Compared to most prior studies on CMC, which contained a minimum of 250 words and previous research in stylometry on slices of prose text of no less than 100 words (see Section 2), in this study our posts are much shorter with an average length of 12.2 tokens per post. Table 1 provides an overview of the distribution of the posts for age, gender and the author set size per category. Since it was impossible to perform manual verification of the linked profiles, it is possible that some of the profiles contain false information about gender or age. We assume the number of these data points to be limited and to be considered as noise for our machine learning algorithm.

Table 1. Distribution of Flemish Dutch Netlog posts according to age group, gender and number of authors.

Age group	Female posts	Female authors	Male posts	Male authors
10s	782,431	61,912	477,438	38,051
20s	29,306	6509	67,013	10,181
30s	12,007	2570	25,930	3739
Plus40s	61,490	4866	52,241	5621
Total	**883,993**	**75,857**	**550,400**	**57,592**

3.2 Dutch Chat Language

The language usage in Flemish chat as it occurs in online social networks such as Netlog and in chat conversations, is mainly determined by two maxims, which are applied in order to approximate spoken discourse (Vandekerckhove and Nobels, 2010): (i) write as you speak to ensure the informal character of the conversation and (ii) write as fast as you can in order to ensure a fluent interaction.

Contrary to English CMC, in Dutch chat messages, the first maxim gave rise to the use of regional varieties and dialect forms typical of colloquial speech in general, which normally cannot be found in written language. In Table 2 we provide some examples of the main Flemish Dutch dialects that were present in our Netlog corpus and their equivalent in Standard Dutch and English.

The second maxim leads to the use of different kinds of abbreviations. Chatters often omit letters or even entire words in order to maximize their typing speed. In addition, spelling errors are seldom corrected, punctuation marks are often left out and uppercase is only used to emphasise the content. As Crystal (2001) argued, the speed of turn taking is essential, because otherwise the conversation does not run smoothly and delayed reactions may overlap with other reactions that arrived earlier on the screen. Given that CMC on online social networks usually consist of posting messages on users' blogs, walls or guestbooks, the speed of turn taking is far less important than in real-time chat conversations. Nonetheless, it is this maxim that accounts for the presence of numerous abbreviations and acronyms in our Netlog corpus and, as is illustrated in Table 2, it is even applied on the regional varieties and dialect forms we mentioned before.

Similar to English chat language usage, there is also a marked presence of emoticons or smileys and different kinds of orthographic and typographic conventions (cf. Crystal, 2001).

Table 2. Examples of typical chat elements in the Netlog Corpus.

Variation Type	Netlog example	Standard Dutch	English
West-Flemish dialect	zitn kik omeki zo verre?	Zit ik ineens zo ver?	Am I that far suddenly?
East-Flemish dialect	Est gedon	Is het gedaan?	Is it over/done?
Antwerp dialect	wa hedde gij die schoene gekocht	Waar heb jij die schoenen gekocht?	Where did you buy those shoes?
Flemish-Brabant dialect	we hebbe toch gelache zemen	We hebben toch gelachen, hoor.	We had a good laugh, hadn't we?
Limburg dialect	hou gans veel van u	Ik hou heel veel van jou.	I love you very much.
Letter/word omissions	kwni	Ik weet het niet	I don't know
Abbreviations	wrm,w8	waarom,wacht	why,wait
Acronyms	hjg	hou je goed	take care

The properties of the Flemish language variations in the Netlog corpus we described in this section make this a challenging text genre for which standard language analysis tools do not work. Especially the presence of various dialect forms, proved to be problematic for automatic linguistic analysis in our preliminary experiments. For example, when applying a part-of-speech tagger, "dak" (*roof*) would be tagged as a noun, while in chat language it is the abbreviated/dialect form of "dat ik" (*that I*), so it should be tagged as a combination of a relative and a personal pronoun.

4. EXPERIMENTAL SETUP

4.1 Pre-processing the Data

Posts on the Netlog social network can contain multiple quotes from previous posts (of which we do not have the correct author ID). Therefore, the first step in pre-processing the data consisted of extracting only the last post of each interaction, of which we did have the required metadata, and saving these as separate documents.

In the second step of pre-processing we tokenized our dataset, in which we interpreted general emoticons, Netlog-specific emoticons and punctuation marks as tokens. We also normalized each token to lowercase and reduced all four or more consecutive identical characters to three, so that e.g. "niiice" and "niiiiice" were considered to be the same type.

The third step consisted of grouping our data using the profile data. In Belgium, the legal minimum age for sexual interactions is

set at 16 and the legal age of majority at 18. Therefore, we categorized our corpus into the following subclasses: *min16* (from 11 to 15 years old), *plus16* (16 and older) and *plus18* (18 and older). Because the illegality of e.g. relationships between an 18 year old and a 15 year old is often very difficult to determine without a thorough police investigation, we also decided to create a class for an age group for which there could be no doubt about the illegal character of a sexual interaction with adolescents under 16, i.e. *plus25* (25 and older). As research in psychiatry and behavioural sciences has shown, most individuals that engage in pedophilia are male (e.g. Ryan et al., 2007; Snyder, 2000). Nevertheless, in a study on sexual crimes Snyder (2000) reported that females were the abuser in 6% of all juvenile cases. Moreover, predators can either pose as an adolescent of the same sex in order to gain their victims' trust, or pretend to be an attractive adolescent of the opposite sex to seduce their victims. Therefore, we incorporated the metadata we had for gender and also created the following complex classes: *min16_male*, *min16_female*, *plus25_male* and *plus25_female*.

In order to investigate the minimum amount of training data that is required for a reliable prediction of age and gender on short posts containing non-standard language variations for future retraining, we first set up our experiments with 10,000 posts per class and then subsequently decreased our dataset to 5000 and 1000 posts per class.

4.2 Feature Selection

In text categorization, a document representation (in our case a Netlog post) is composed of different types of features, the selection of which can significantly affect the performance of the machine learning algorithm. Unlike some previous studies on age and/or gender prediction (e.g. Argamon et al., 2007; Goswami et al., 2009), due to the complexity of the Dutch language usage in the Netlog posts we choose not to set up a dictionary of hand-picked features that are considered to be most likely to distinguish between age or gender categories. Moreover, using various feature sets is more interesting as they provide features with varied granularities and diversities. For this study we applied the Chi-square ($\chi 2$) feature selection metric (e.g. Manning and Schütze, 2001).

Features can be based on (n-grams of) bags of tokens or characters, but also on morphological, lexical, syntactic or semantic features. As the non-standard language usage in our corpus provides severe challenges for automatic linguistic analysis (e.g., stemming, lemmatization, part-of-speech tagging), for the current study, we limited our feature set to token and character features only: word unigrams, bigrams and trigrams and character bigrams, trigrams and tetragrams.

For our experiments we built our feature sets by selecting the 1000, 5000, 10,000 and 50,000 features with the highest $\chi 2$-values from the training data.

5. EXPERIMENTS AND RESULTS

In this section we present our results for the classification of Netlog posts according to their authors' age group and gender using the Support Vector Machine learning package Liblinear (Fan et al., 2008). We also describe which types of features are most informative for this application and compare the robustness of these features with different feature set sizes. Subsequently, we provide an overview of the performance as we decrease our data set size from 10,000 instances per class to 5000 and finally 1000 instances per class.

5.1 Adults versus Adolescents

For this study we set up a classification experiment for:

- *min16 vs. plus16*
- *min16 vs. plus18*
- *min16 vs. plus25*
- *min16_male vs. min16_female vs. plus25_male vs. plus25_female*

After selecting the features, we represented each document (in our case each Netlog post) as a sparse binary vector for the SVM classifier. We did not use frequencies in our vectors, as it did not produce better results in preliminary experiments. We evaluated the performance using 10-fold cross validation as experimental regime. This technique selects 10 equally sized subsets of randomly selected documents. Subsequently each subset is used nine times in training and once in test, in order to provide a more reliable estimation of the performance of the system. First we give an overview of our binary classification experiments of *min16 vs. plus16*, *min16 vs. plus18* and *min16 vs. plus25*, followed by a discussion of our 4-way classification experiment of *min16_male vs. min16_female vs. plus25_male vs. plus25_female.* We also discuss three different approaches to including the gender metadata into a binary age-based classification experiment.

5.1.1 Age Classification

For our main data set of 10,000 posts per class, the SVM classifier yielded an accuracy of 71.3% for the age-based classification task of *min16 vs. plus16*. This is to be compared to the random baseline of 50.0 %. As the distance between the age groups increased, the accuracy rose to 80.8 % for *min16 vs. plus18* and even to 88.2 % for *min16 vs. plus25*. An overview of the accuracy results for *min16 vs. plus16*, *min16 vs. plus18* and *min16 vs. plus25* can be found respectively in Figure 1, 2 and 3.

As can be observed in these figures, token features outperformed character features in each age classification task. Furthermore, word unigram features (i.e. words, (Netlog) emoticons and punctuation marks) showed to be the most robust feature type for age prediction. Interesting to see was that for all token feature types the best accuracy results were achieved by using the 50,000 most informative features ($\chi 2$).

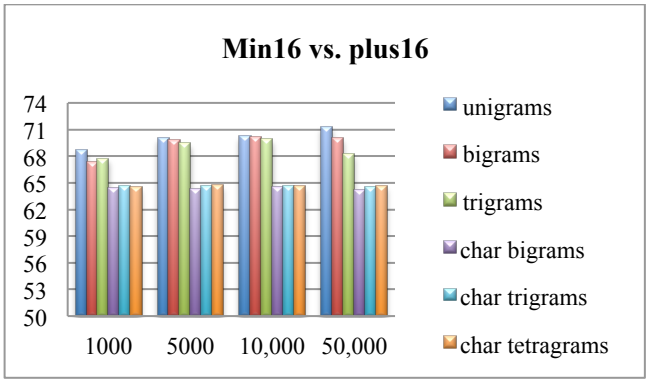


Figure 1. Accuracy results (%) for *min16 vs. plus16* using 1000, 5000, 10,000 and 50,000 most informative features ($\chi 2$).

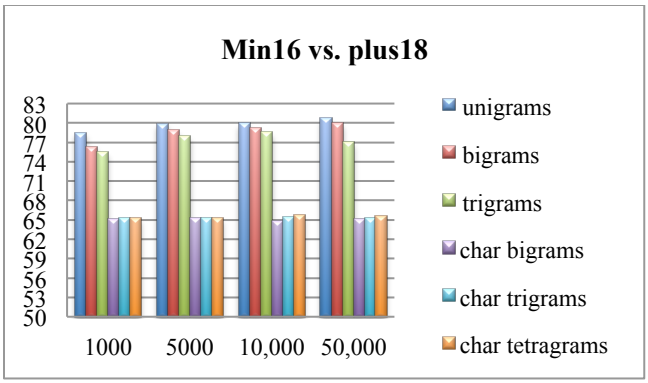


Figure 2. Accuracy results (%) for *min16 vs. plus18* using 1000, 5000, 10,000 and 50,000 most informative features (χ2).

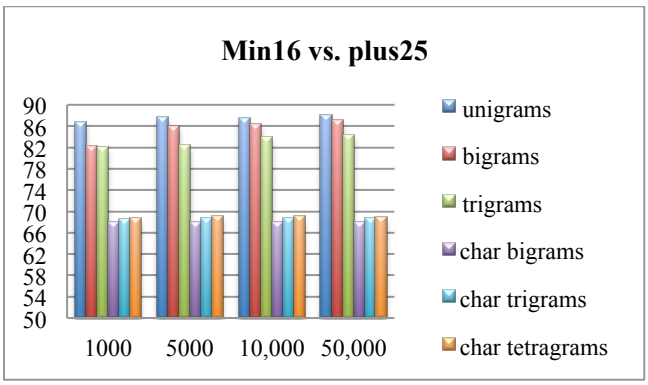


Figure 3. Accuracy results (%) for *min16 vs. plus25* using 1000, 5000, 10,000 and 50,000 most informative features (χ2).

5.1.2 Age Classification Including Gender

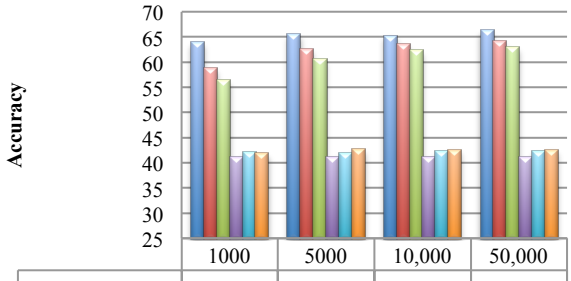
In this section we first present our results for the 4-way classification experiment, in which we included 5000 instances per class. Moreover, we discuss three approaches of including the metadata for gender in order to investigate their effect on age prediction.

For the 4-way age-based classification experiment of *min16_male vs. min16_female vs. plus25_male vs. plus25_female*, the SVM classifier yielded an accuracy of 66.3%. This is to be compared to the random baseline of 25.0 %. Similar to the results for the binary age classification experiments we described in the previous section, token features outperformed character features and word unigrams proved to be the most robust feature type for age prediction including gender metadata. Again the best accuracy results were achieved by using the 50,000 most informative features (χ2). We provide an extensive overview of these results in Figure 4.

As is shown in Table 3, the confusion matrix for our experiment based on the 50,000 most informative word unigrams, indicated very promising results for our main objective of detecting adults posing as adolescents. Of all the adult male posts 5.3% and 6.6% were wrongly classified as an adolescent boy and girl, respectively. The confusion matrix also indicated that the varieties in (chat) language usage were more related to age than to gender, because there was a greater confusion between the gender classes of the same age groups as there was between the age classes of the same gender groups.

Table 3. Confusion matrix for *min16_male vs. min16_female vs. plus25_male vs. plus25_female*, using 50,000 most informative word unigram features (χ2).

Confusion Matrix(%)	Min16 female	Min16 male	Plus25 female	Plus25 male
Min16 female	61.1	20.8	6.1	5.3
Min16 male	26.5	69.4	5.4	6.6
Plus25 female	7.8	4.8	67.8	19.6
Plus25 male	4.6	5.0	20.7	68.5



	1000	5000	10,000	50,000
unigrams	64.0	65.5	65.3	66.3
bigrams	58.7	62.6	63.6	64.2
trigrams	56.3	60.6	62.3	63.0
char bigrams	41.1	41.3	41.2	41.2
char trigrams	42.1	42.0	42.4	42.2
char tetragrams	41.8	42.7	42.6	42.6

Figure 4. Accuracy results (%) for *min16_male vs. min16_female vs. plus25_male vs. plus25_female* using 1000, 5000, 10,000 and 50,000 most informative features (χ2).

In view of a comparative analysis to the results of the binary classification experiment of *min16 vs. plus25* we described in Section 5.1.1 (Data set 1), which was balanced according to age only, we then utilised the data set with the complex classes, which was balanced according to both age and gender (Data set 2), to examine three different approaches of including the metadata for gender in order to investigate their effect on age prediction. We also report the precision, recall and f-scores for these experiments. Given our main research objective of detecting adults posing as adolescents, during these experiments we focused on the scores for the adult class of plus25.

In the first experiment (Exp. 1) we reduced the number of classes in both train and test sets (10-fold cross validation) from the four complex classes to two, so that we could compare the results to those from Data set 1 and examine whether balancing our data set on both age and gender had an effect on performance. We reduced our classes from min16_male and min16_female to min16 and from plus25_male and plus25_female to plus25 and retrained the SVM classifier, using the 50,000 most informative unigram features (χ2). Compared to the results of Data set 1, which was balanced according to age only, the accuracy improved slightly from 88.2 % to 88.8%. More importantly, the precision, recall and f-score for the plus25 class rose from respectively 87.8%, 88.6% and 88.2% to 90.5%, 92.9% and 91.7%.

In the second additional experiment (Exp. 2) we first trained our classifier on our four complex classes and then reduced their number in the classifier's output and our test sets to two age classes in order to examine whether the extra gender information the classifier had acquired during training would lead to a better age prediction on the binary test sets. The results showed an improvement upon Data set 1 to 88.5% accuracy. We found similar results for the precision, recall and f-score of the adult class, i.e. a slight improvement to those of Data set 1 to resp. 88.3%, 88.8% and 88.5%, but they do not exceed the results from Exp. 1.

The third experiment (Exp. 3) consisted of reducing the number of classes in both train and test sets to two age classes, as we did in Exp. 1, and including gender as an extra feature in every instance. Again the results improved upon those of Data set 1, with an accuracy of 88.7% and 91.5% precision, 85.7% recall and 88.5% f-score for the adult class. In Table 4 we provide an overview of the results for our three additional experiments compared to those of Data set1.

After examining these three different approaches of including the metadata for gender in order to investigate their effect on age prediction, for accuracy (88.8%) and recall (92.9%) and f-score (91.7%) of the adult class, the best results were achieved by balancing our data set according to both age and gender: 10,000 instances for both min16 and plus25, including 5000 instances for male and female within each age group. Adding the metadata for gender as an additional feature in each instance produced the best precision score for plus25 (91.5%). All the additional experiments showed improvement to the results of Data set 1, which was only balanced according to age.

Table 4. Results for data set 1, exp. 1 (2 classes in train and test), exp. 2 (4 classes in train, 2 in test) and exp. 3 (gender as feature) using 50,000 word unigram features(χ2).

Scores (%)	Age Group	Data set 1	Data set 2		
			Exp. 1	Exp. 2	Exp. 3
Precision	Min16	**88.5**	85.1	61.1	86.5
	Plus25	87.8	90.5	88.3	**91.5**
Recall	Min16	87.7	80.5	71.5	**92.0**
	Plus25	88.6	**92.9**	88.8	85.7
F-score	Min16	88.1	82.7	65.9	**89.2**
	Plus25	88.2	**91.7**	88.5	88.5
Accuracy		88.2	**88.8**	88.5	88.7

5.2 Data Set Size

As we mentioned in Section 1, another challenge in detecting false profiles on social networks will be the constant retraining of the machine learning algorithms in order to pick up new varieties of chat language usage that are linked to age and/or gender. Therefore, we performed experiments with different data set sizes in order to acquire more insight into the minimum data size that will be required in future retraining experiments. In this section we describe the results that were achieved when decreasing our initial data set of 10,000 instances per class to 5000 and finally 1000 instances per class. We included all of the previously mentioned feature set sizes and types in order to examine whether the best performing set of 50,000 most informative word unigrams (see Section 5.1) is robust to decreasing the data set size.

After randomly extracting 5000 instances per class from our initial dataset of 10,000 instances per class, we retrained the SVM classifier, using the same feature extraction metric (χ2) to extract the 1000, 5000, 10,000 and 50,000 most informative features for each feature type. With only half the size of the first dataset, the SVM classifier still yielded an accuracy of 71.1% for the age-based classification task of *min16 vs. plus16*, indicating a very small decrease of 0.2% compared to the performance of the initial data set. A similar observation was made for *min16 vs. plus18* and *min16 vs. plus25*, for which the accuracy scores decreased only slightly with 0.9% and 0.5% respectively to 79.9 % (*min16 vs. plus18*) and 87.7 % (*min16 vs. plus25*). Again the best results for all feature types were achieved using the 50,000 most informative features. Moreover, the dominance of the word unigrams over the other feature types continues for this data set.

From our reduced data set of 5000 instances per class, we then randomly extracted 1000 posts per class and retrained the SVM classifier with a setup identical to the first two experiments. With only 10% of the original dataset the classifier was still able to improve considerably upon random baseline performance for all binary age-based classification experiments: 65.6% for *min16 vs. plus16*; 77.4% for *min16 vs. plus18* and 85.6% for *min16 vs. plus25*. For this small data set the best results for all feature types were achieved using the 10,000 most informative features and the most robust feature type for all experiments again proved to be the word unigrams. In Figure 5 we provide the learning curves over the three data sets using the best accuracy scores for each age-based classification experiment.

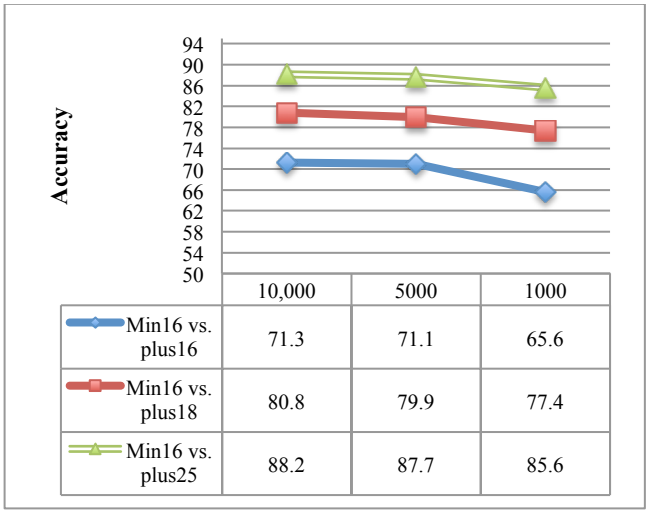


Figure 5. Accuracy results (%) for the binary age-based classification experiments using 10,000, 5000 and 1000 instances per class.

Subsequently, we also calculated the F-scores for the adult classes to investigate to which extend reducing the data set size would affect the performance of detecting them. Comparable to the overall accuracy scores we mentioned before, the F-scores for the adult classes showed a similar pattern of decrease when halving the data set size to 5000 instances per class: from 71.9% to 71.1% (*plus16*), from 80.9% to 80.1% (*plus18*) and from 88.2% to 87.2% (*plus25*). After further reducing our data set size to 1000 instances per class the SVM classifier was again still able to improve upon random baseline and yielded 66.7% for *plus16*, 77.6% for *plus18* and 85.5% for *plus25*. The learning curves over the three data sets using the best F-scores for each age-based classification experiment are provided in Figure 6.

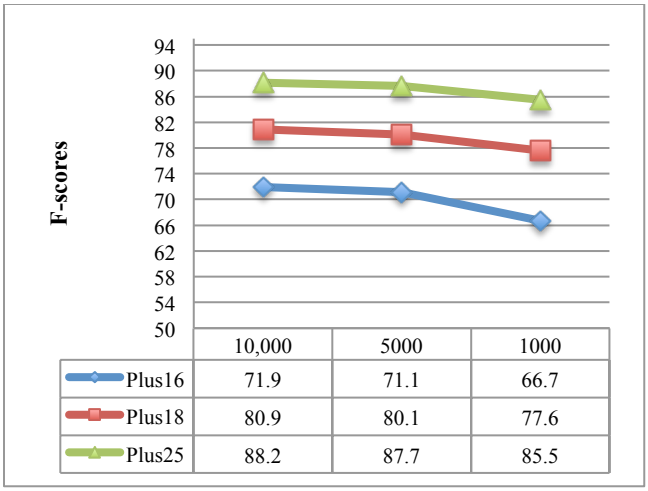


Figure 6. F-scores (%) of the adult class for the binary age-based classification experiments using 10,000, 5000 and 1000 instances per class.

Given that with only 10% of the original dataset the classifier was still able to improve considerably upon random baseline performance for all binary age-based classification experiments, systematic retraining of the machine learning algorithms will definitely be feasible with the collaboration of popular social networks like Netlog, which receive at least 10,000 posts per day.

6. DISCUSSION AND FUTURE RESEARCH

Despite the challenging characteristics of this text genre for natural language processing, in this study we showed that it is feasible to improve upon random baseline performance for age classification using highly limited data sets of on average 12.2 tokens (i.e. words, (Netlog) emoticons and punctuation marks) per instance. Moreover, the choice of (chat) words (i.e. word unigram features like "bro" (*brother*), "grts" (*greetings*), "fotokes" (*pictures*)) seems to be more important for age prediction than the way those words are combined in bigrams or trigrams. When training on the 50,000 most informative word unigram features ($\chi 2$), the SVM classifier showed promising results for classifying adults versus adolescents and yielded an accuracy of 71.3% for the age-based classification task of *min16 vs. plus16*. As the distance between the age groups increased, the accuracy rose to 80.8 % for *min16 vs. plus18* and even to 88.2 % for *min16 vs. plus25*. Since gender could be a helpful information source in constructing a more accurate classifier for age, we subsequently examined three different approaches to including the metadata for gender. First we balanced our data set according to both age and gender. Secondly we trained our classifier on four complex classes, which included both age and gender information, but reduced the number of classes to two, which only related to age, in the test sets and finally we performed a binary classification experiment with gender as additional feature in each instance. The best results were achieved by balancing our data set according to both age and gender with a best accuracy score of 88.8%, best precision score of 91.5%, best recall score of 92.9% and best F-score of 91.7% for the adult class. Adding the metadata for gender as an additional feature in each instance produced the best precision score for plus25 (91.5%). However, all three approaches showed improvement to the results of the similar data set that was only balanced according to age. Finally, we also performed experiments with different data set sizes in order to acquire more insight into the minimum data set size that would be required in future retraining experiments. With only half the size of the first dataset, the SVM classifier still yielded an accuracy of 71.1% for the age-based classification task of *min16 vs. plus16*, indicating a very small decrease of 0.2% compared to the performance of the initial data set. A similar observation was made for *min16 vs. plus18* and *min16 vs. plus25*, for which the accuracy scores decreased only slightly with 0.9% and 0.5% respectively to 79.9 % (*min16 vs. plus18*) and 87.7 % (*min16 vs. plus25*). Given that with only 10% of the original dataset (1000 instances per class) the classifier was still able to improve considerably upon random baseline performance for all binary age-based classification experiments, systematic retraining of the machine learning algorithms will definitely be feasible with the collaboration of popular social networks like Netlog, which receive at least 10,000 posts per day.

As to our future research, given that the non-standard language usage in the Netlog Corpus provided great challenges for stemming, lemmatization and part-of-speech tagging, we will certainly address these issues within our project. We will also explore other machine learning algorithms that have proven to be reliable in e.g. authorship attribution. The data also allows further experimentation on the mutual influences between age and gender in categorization.

Another question that remains is how these trained models will perform when the adult is a sexual predator, possibly posing as an adolescent. In order to answer that, we are collaborating with local and federal law enforcement agencies to collect training data from closed paedophile cases.

7. ACKNOWLEDGMENTS

This study has been carried out in the framework of the DAPHNE project (Defending Against Pedophiles in Heterogeneous Network Environments), funded by the Industrial Research Fund of the Antwerp University in Belgium. We also thank Netlog and Mollom for supplying the data needed for constructing this corpus and the reviewers for their comments that helped improve the manuscript.

8. REFERENCES

[1] Argamon, S., Koppel, M., Fine, J., and Shimoni, A. 2002. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing.* 17, 4 (November 2002), 401-412. DOI=10.1093/llc/17.4.401.

[2] Argamon, S., Koppel, M., Pennebaker, W., and Schler, J. 2007. Mining the Blogosphere: Age, gender and the varieties of self–expression. *First Monday*.12, 9 (September 2007). DOI=http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2003.

[3] Burger, J.D., and Henderson, J.C. 2006. An exploration of observable features related to blogger age. In *Proceedings of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs*. (California, USA, March 27 - 29, 2006).

[4] Burrows, J. 2007. All the way through: testing for authorship in different frequency strata. *Literary and Linguistic Computing*. 22, 1 (2007), 27–47. DOI= http://dx.doi.org/10.1093/llc/fqi067.

[5] Caverlee, J., and Webb, S. 2008. A large-scale study of MySpace: observations and implications for online social networks. In *Proceedings of the 2nd International Conference on Weblogs and Social Media* (Seattle, USA, March 30 - April 2, 2008). ISWCM'08. International AAAI Conference on Weblogs and Social Media. DOI=http://www.aaai.org/Library/ICWSM/2008/icwsm08-012.php.

[6] Crystal, D. 2001. Language and the Internet. Cambridge University Press, Cambridge, NY, USA.

[7] Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., and Lin, C.J. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*. 9 (August, 2008), 1871-1874. DOI=http://doi.acm.org/10.1145/1390681.1442794.

[8] Goswami, S., Sarkar, S., and Rustagi, M. 2009. Stylometric analysis of bloggers' age and gender. In *Proceedings of the Third International ICWSM Conference* (San Jose, USA, May 17 - 20, 2009). ISWCM'09. International AAAI Conference on Weblogs and Social Media. DOI= http://aaai.org/ocs/index.php/ICWSM/09/paper/view/208.

[9] Herring, S.C., and Paolillo, J.C. 2006. Gender and genre variation in weblogs. Journal of Sociolinguistics. 10, 4 (August, 2006), 439 - 459. DOI=10.1111/j.1467-9841.2006.00287.x

[10] Hirst, G., and Feiguina, O. 2007. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*. 22, 4 (October, 2007), 405 - 417. DOI= 10.1093/llc/fqm023.

[11] Holmes, J., and Meyerhoff, M. 2003. The Handbook of Language and Gender. Blackwell, Oxford, UK. DOI= 10.1111/b.9780631225034.2004.x.

[12] Luyckx, K., and Daelemans, W. 2010. The Effect of Author Set Size and Data Size in Authorship Attribution. *Literary and Linguistic Computing*. 26, 1 (August, 2010). DOI= 10.1093/llc/fqq013.

[13] Manning, C.D., and Schütze, H. 2001. Foundations of statistical natural language processing. MIT Press, Cambridge, Massachusetts, USA. DOI=10.1145/601858.601867.

[14] Mukherjee, A., and Liu, B. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (Cambridge, USA, October 9 - 11, 2010). EMNLP '10. Association for Computational Linguistics, Stroudsburg, PA, USA, 207-217. DOI= http://www.aclweb.org/anthology/D10-1021.

[15] Nguyen, D., Smith, N., and Rosé C. 2011. Author Age Prediction from Text using Linear Regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (Portland, USA, 19 – 24 June, 2011). Association for Computational Linguistics, Stroudsburg, PA, USA, 115-123.

[16] Nowson, S., and Oberlander, J. 2007. Identifying more bloggers. Towards large scale personality classification of personal weblogs. In *Proceedings of the 1st International Conference on Weblogs and Social Media* (Boulder, USA, March 26 - 28, 2007). ISWCM'07. International AAAI Conference on Weblogs and Social Media.

[17] Pennebaker, J.W., and Graybeal, A. 2001. Patterns of natural language use: disclosure, personality, and social integration. *Current Directions in Psychological Science.* 10, 3 (2001), 90-93. DOI= 10.1111/1467-8721.00123.

[18] Pennebaker, J.W., and Stone, L.D. 2003. Words of wisdom: Language use over the lifespan. *Journal of Personality and Social Psychology.* 85, 2 (Aug 2003, 2003), 291-301. DOI=10.1037/0022-3514.85.2.

[19] Rosenthal, S., and McKeown, K. 2011. Age Prediction in Blogs: A Study of Style, Content, and Online Behavior in Pre- and Post-Social Media Generations. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning* (Portland, USA, 19 – 24 June, 2011). Association for Computational Linguistics, Stroudsburg, PA, USA, 763-772

[20] Ryan, C., Hall, W., and Hall, R. 2007. A profile of pedophilia: definition, characteristics of offenders, recidivism, treatment outcomes, and forensic issues. In *Mayo Clinic Proceedings.* 82, 4 (April, 2007), 457 - 471. DOI= 10.4065/82.4.457.

[21] Sanderson, C., and Guenter, S. 2006. Short text authorship attribution via sequence kernels, Markov chains and author unmasking: an investigation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (Sydney, Australia, 22 - 23 July , 2006). EMNLP'06. Association for Computational Linguistics, Stroudsburg, PA, USA, 482–491. DOI= http://www.aclweb.org/anthology/W06-1657.

[22] Sarawgi, R., Gajulapalli, K., and Choi, Y. 2011. Gender Attribution: Tracing Stylometric Evidence Beyond Topic and Genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning* (Portland, USA, 19 – 24 June, 2011). Association for Computational Linguistics, Stroudsburg, PA, USA, 78 – 86.

[23] Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. 2006. Effects of age and gender on blogging. In *Proceedings of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs*. (California, USA, March 27 - 29, 2006). DOI=http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.71.216.

[24] Snyder, H.N. 2000. Sexual assault of young children as reported to law enforcement: victim, incident, and offender characteristics. US Departement of Justice, Bureau of Justice Statistics. Washington, DC, USA. Publication NCJ 182990.

[25] Tam, J., and Martell, C. 2009. Age Detection in Chat. In *Proceedings of the 3rd IEEE International Conference on Semantic Computing.* (Berkeley, USA, September 14-16, 2009). DOI=10.1109/ICSC.2009.37.

[26] Vandekerckhove, R., and Nobels, J. 2010. Code eclecticism: Linguistic variation and code alternation in the chat language of Flemish teenagers. *Journal of Sociolinguistics.* 14, 5 (November, 2010), 657 - 677. DOI=10.1111/j.1467-9841.2010.00458.x.

[27] Yan, X., and Yan, L. 2006. Gender classification of weblog authors. In *Proceedings of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs*. (California, USA, March 27 - 29, 2006).

[28] Zhang, C., and Zhang, P. 2010. Predicting gender from blog posts. Technical Report. University of Massachusetts Amherst, USA.