Text-Based Age and Gender Prediction for Online Safety Monitoring

Janneke van de Loo and Guy De Pauw and Walter Daelemans
CLiPS - Computational Linguistics Group – University of Antwerp
Prinsstraat 13, 2000 Antwerpen, Belgium
firstname.lastname@uantwerpen.be

ABSTRACT

This paper explores the capabilities of text-based age and gender prediction geared towards the application of detecting harmful content and conduct on social media. More specifically, we focus on the use case of detecting sexual predators who try to “groom” children online and possibly provide false age and gender information in their user profiles. We perform age and gender classification experiments on a dataset of nearly 380,000 Dutch chat posts from a social network. We evaluate and compare binary age classifiers trained to separate younger and older authors according to different age boundaries and find that macro-averaged F-scores increase when the age boundary is raised. Furthermore, we show that use-case applicable performance levels can be achieved for the classification of minors versus adults, thereby providing a useful component in a cybersecurity monitoring tool for social network moderators.

KEYWORDS

author profiling, cybersecurity, social media, online safety, grooming

1 INTRODUCTION

The advantages, fun, and opportunities social media bring for children are offset by significant potential dangers. According to a pan-European survey[1], children spend a lot of time on social media interaction without parental supervision (in their bedroom or as mobile users) and are relatively often exposed to dangerous situations. Twelve percent of 9 to 16 year old youngsters report having been bothered or upset during social media use, mainly by exposure to bullying or unwanted sexual content. Although much less frequent, they all too often report attempts at grooming by adults. In this paper we show how author profiling, a text mining area, can be applied to the detection of harmful content in social media, and illustrate this by means of age and gender profiling for the detection of grooming by pedophiles in social media.
The last decade has seen a large improvement in the accuracy and applicability of techniques for knowledge discovery from text (also called *text mining* or *text analytics*). The type of knowledge extracted can be factual, for example for use in medical expert systems (IBM’s Watson for oncology application is a good example [1]), or it can be subjective as in the many sentiment analysis applications where opinions or sentiments of authors are targeted (see [2,3] for applications to political media coverage analysis and economic prediction).

In this paper, we look at a more recent type of knowledge discovery from text, namely author profiling: the extraction of demographic and psychological characteristics of authors from text they have written [4,5]. This is often called *computational stylometry* [6]. By analyzing the linguistic properties of text, “metadata” such as age, gender, region, and personality traits of the author can be estimated on the basis of machine learned models trained on text samples written by authors for which the profile is known. Author profiling has established itself as a text analytics subarea with its own conferences and shared task competitions, for instance the shared tasks at the PAN workshops [7,8,9,10]. Many applications of author profiling have been proposed, ranging from demographic marketing to forensic detection tasks such as those described in this paper.

In the **AMiCA**-project[2], author profiling is used as one of the modules in a system for detecting three

[1] Available from http://www.eukidsonline.net

[2] http://www.amicaproject.be

harmful situations for children in social networks: depression and suicide announcements [10], cyberbullying [12], and sexually transgressive behavior (including grooming by pedophiles [13]). All three of these applications involve content-based text analysis. For example, to detect suicidal children, negative emotions or suicide announcements should be recognized in text, and cyberbullying mostly involves the expression of threats, defamation or insults. But in addition to this factual knowledge extraction, profile information can help as well. Especially age, gender, and personality information can improve the detection of these harmful events when combined with the factual knowledge. For example, there are clear correlations between gender and personality on the one hand, and the probability of being a victim or bully in cyberbullying events, and there are links between personality and risk of depression and suicidal behavior.

This paper is concerned with the application of author profiling information (in this case the detection of age and gender) in the **AMiCA** use case concerned with identifying grooming by pedophiles in social networks. It combines a module detecting sexual content and the specific vocabulary of grooming with a module comparing the profile provided by the user to the profile that is induced from the text produced by this user. The architecture is as follows: when there is a mismatch between the induced and the provided profile (for example a provided profile of a 14-year-old girl does not match with the induced profile of an adult male), the content of the interaction is analyzed by a classifier detecting sexual content and grooming behavior, and if that classifier also returns a positive result, the interaction is reported to the moderator of the social network.

In this monitoring support set-up, it is important that the text analysis classifiers return high recall rather than high precision: it is better to err on the side of false positives than on the side of false negatives, as there will be manual inspection by the moderator anyway before taking action. Sufficiently high recall ensures that no harmful cases are missed, while even modest precision dramatically reduces the number of interactions that need to be manually monitored.

A crucial component in this profiling application is the set-up of the age and gender detection task. The success of age classification partly depends on the age classes that are being distinguished. In our current set-up, we carry out binary age prediction, i.e. determining whether authors are older or younger than a specific age boundary. Working with only two classes (minors versus adults) not only ties in with the intended application, but also serves to maximize classification accuracy. Furthermore, the age boundary itself can be easily adapted to any number that is relevant to the specific use case at hand, based on legal constraints (e.g. the legal age of sexual consent) or age related statistics (e.g. sexual offense rates across age groups). The goal of this paper is to show that the profiling module can be optimized to achieve accuracy levels that are useful for our decision support system for social network moderators.

We will start with a brief overview of related age and gender prediction research (Section 2), followed by a description of the dataset and the methods employed in our current research (Section 3). In Section 4, we present the results of our age and gender classification experiments and discuss the implications of these results for the use case of sexual predator detection. We finish with concluding remarks and outline our plans for future research in Section 5.

2 RELATED RESEARCH

Early work in automatic author profiling was done by [4], who categorized formal written texts from the British National Corpus by author gender. A few years later, age and gender prediction studies became increasingly focused on informal online social media texts, especially on blogs (e.g. [14,15,16,17] and tweets [18,19,20,21], but also on chat posts [13] and YouTube comments [22]. This trend is also reflected in the author profiling tasks that were organized at PAN 2013, 2014 and 2015 [8,9,10].

Various supervised machine learning algorithms have been employed, using a variety of textual features, such as character n-grams, token n-grams, part-of-speech n-grams, specific token subsets (e.g. emoticons, internet acronyms, function

Table 1. The list of classification experiments and the associated classes.

Task	Classes
Age	YOUNGER < age_boundary ≤ OLDER
Gender	♀ - ♂
Age & Gender	YOUNGER♀ - OLDER♂ - YOUNGER♂ - OLDER♀

words, LIWC[3] dictionaries), readability features (e.g. average word and sentence length), character-based stylistic features (e.g. capitalization, character repetitions, punctuation), and extracted topics. In some cases, extratextual profile features were used as well, for instance the number of friends and followers [22,18,24], background colors [25], and posted images [25]. Our current study is limited to features extracted from the written texts.

Binary age prediction, as researched in this paper, was first performed by [22], who predicted whether bloggers were under or over 18. They experimented with shallow textual features based on character counts, language models, and meta-information such as the number of friends. The resulting performance was not far above the majority baseline, however. [27] carried out binary age prediction experiments on transcribed telephone conversations and [18] on tweets. They used the age boundaries 40 and 30, respectively, to separate the two age classes, and in both studies the features used for classification were token n-grams and sociolinguistic features. [13] used token and character n-grams to predict whether authors of Flemish Dutch chat posts from the social network Netlog, were under or over 16. In addition, they studied the effect of increasing the gap between the older and the younger age group. This paper presents experiments on expanded data sets from the same social network. [28] studied the task of predicting whether blog authors were born before or after a specific year. Like in the present study, they experimented with different class boundaries, but they used birth year rather than age to define those boundaries, as their aim was to find a boundary between two generations. This is an important difference, since the blog data per author included posts written at different ages, over periods of up to ten years. In contrast, our research aims at an application that distinctly requires an age-oriented approach, due to the legal context of the application.

3 MATERIAL AND METHODS

In order to detect illegal grooming activity involving minors, we need reliable age and gender assignment to determine the ages of the participants in the conversation, or to detect mismatches between the (possibly false) profile provided by a user and the demographic data as inferred from the text.

3.1 Experimental Setup

We conduct age and gender prediction experiments on a dataset of social network posts, which is described in subsection 3.2. Three types of prediction experiments are carried out: age prediction, gender prediction and combined age and gender prediction. Age prediction is defined as a binary classification task, i.e. predicting whether an author is older or younger than a certain age boundary. We vary this boundary between 16 and 28. The experiment types and the associated classes are listed in Table 1.

For each experiment type, we carry out five-fold cross-validation experiments (i.e. using 80% of the data for training and 20% for testing in each fold), both on the full, unbalanced, dataset and on data subsets balanced for age class and gender. The experiments on the full dataset showcase the performance on real-life data, while the experiments on the balanced data subsets allow us to perform a more detailed analysis of the observed effects, by factoring out effects of class imbalance.

3.2 Dataset

The full dataset consists of 379,769 chat posts from the Belgian social networking website Netlog[4]. The posts are interpersonal chat messages which were posted in the public social networking environment (as opposed to private messages, which could not be made available by Netlog).

[3] http://www.liwc.net

[4] Netlog ceased activitities in 2014 and has since been merged with Twoo.

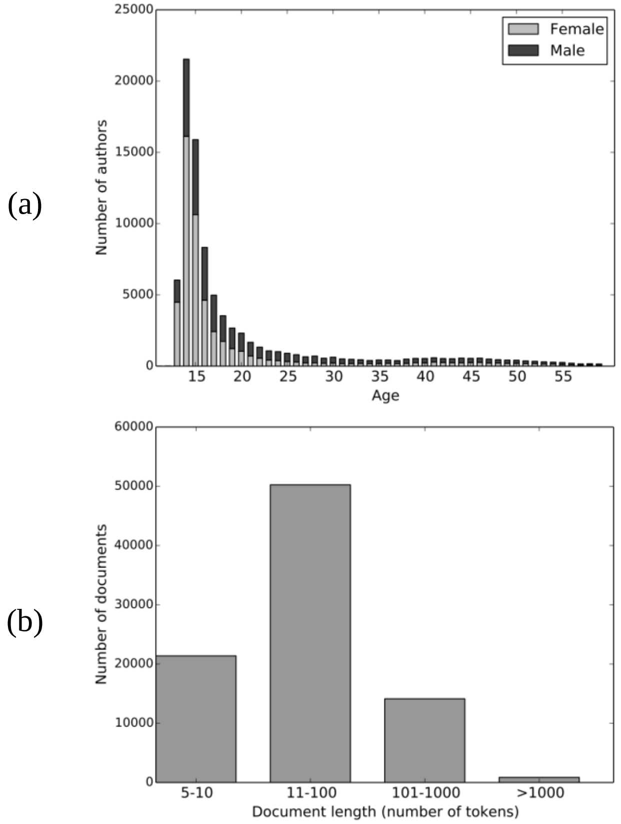

Figure 1. Distribution information for the full dataset.

They were posted between November 2010 and February 2011 by 86,610 different users. For each user, the self-reported age, gender and location was available in the Netlog user profile. Only Dutch posts (classified as such by a language identification system[5]) from Belgian users were included, with a minimum post length of 5 tokens. The dataset contains a large amount of non-standard language, typical of user-generated content. The non-standard forms include spelling errors, unofficial abbreviations (some of which are common in internet language) and various creative spelling variants, which often adopt characteristics from colloquial speech, including regional dialect influences [29].

The ages of the users range from 11 to 59. Figure 1(a) shows the age and gender distribution of the users in the dataset. There is a very high peak at the ages of 14 and 15, with over 20,000 and over 15,000 users respectively, whereas for the 25+ ages, the dataset contains fewer than 1,000 users per age category. For the ages 11 and 12, the number of authors is below 10. In the ages 13 to 15, females are markedly overrepresented, with percentages between 69% and 75%, whereas between the ages 23 and 32, males are slightly overrepresented: between 60% and 67% of the users in those age categories are male.

For our profiling experiments, we concatenated all posts per user into one document, thus yielding 86,610 single-user documents for training and testing. The distribution of document lengths is shown in Figure 1(b). The group of documents with 11 to 100 tokens is largest (about 50,000 documents) and only a small number of documents consist of more than 1,000 tokens.

3.2 Balanced Data Subset

For the prediction of age classes (older or younger with respect to a specific age boundary) and the prediction of combined age and gender classes, we constructed data subsets that are balanced for age class and gender; one for each age boundary. So for instance, for the prediction of the age class with respect to the age boundary of 16 (-16 or 16+), we constructed a data subset with equal amounts of -16 female authors, -16 male authors, 16+ female authors, and 16+ male authors in each of the five partitions. For each age boundary, the number of randomly selected documents per class per partition was the same: 1,165 documents.

This resulted in partitions of 4,660 documents each, so 23,300 documents in total per balanced data subset. The resulting age and gender distributions in four of the balanced subsets (viz. the subsets for age boundaries 16, 18, 22 and 28, respectively) are shown in Figure 2. Due to the random selection of documents per class, the original age distribution within each class is preserved.

For the prediction of gender only, a data subset balanced only for gender was constructed. This data subset consists of 7,006 documents per gender per partition, so 14,012 per partition and 70,060 in total.

3.2 Classifier and Features

The classifier we used is a Support Vector Machine (SVM) classifier with a linear kernel, viz.

[5] Available from http://textgain.com

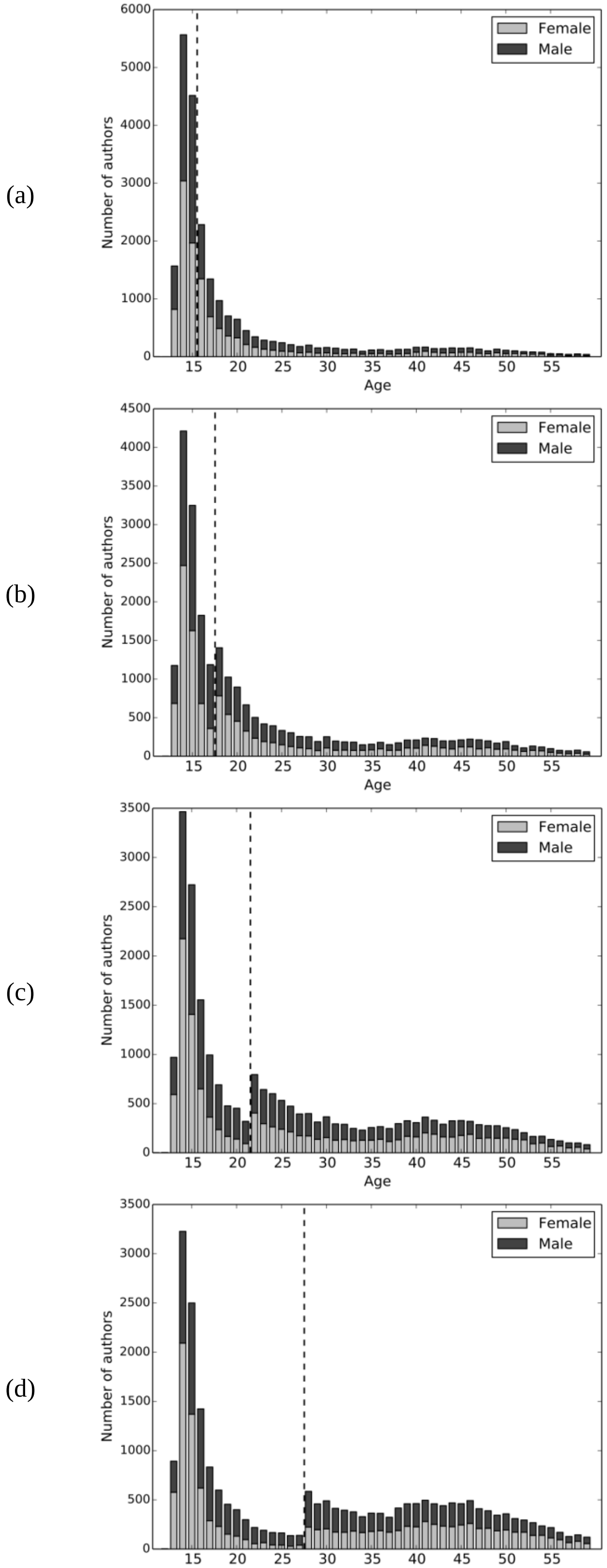

Figure 2. Age and gender distributions in four of the data subsets balanced for age class and gender, viz. the subsets balanced according to age boundaries 16 (a), 18 (b), 22 (c), and 28 (d). In each graph, the age boundary is marked with a vertical dashed line.

scikit-learn's LinearSVC classifier [30]. For each fold, the classifier's parameter *C* was tuned in a 3-fold cross-validation grid search on the training set.

The features used for classification are token and character n-grams. Token n-grams have been widely used for age and gender prediction and have shown good results [14,18,13,20]. Before tokenization, we carried out a number of text preprocessing steps. All uppercase alphabetic characters were converted to lowercase and character repetitions were reduced to a maximum of 3 (e.g. 'hiiiii' → 'hiii'), to obtain a certain level of generalization across different varieties of the same word. For generalization purposes, emoticons, URLs, e-mail addresses, and links to photos and videos were replaced by a single special character.

The character n-grams, on the other hand, were collected from the original, raw text, i.e. without carrying out the preprocessing steps discussed above. The character n-grams can capture many stylistic characteristics, such as (parts of) emoticons, character repetitions, capitalization and morphological features. In addition, the fact that they capture parts of words renders them more robust to spelling variants and errors, which are numerous in these chat data. Furthermore, they capture stylistic tendencies that authors are often less aware of, making it harder for sexual predators to deceive the system. Some other age and gender prediction studies in which character n-grams have been successfully used, are [13], [31], and [32].

Only the n-grams with the highest relative frequencies in the training set were selected, imposing a threshold on the total number of n-grams of each type to be considered by the classifier. An n-gram's relative frequency is the count of the n-gram normalized by the total number of n-grams (of that type) in the document. We selected a relatively high number of character n-grams compared to the number of token n-grams, as the number of high-frequency character n-grams is relatively large.

The list of features is thus as follows:

- the 2,500 most frequent token unigrams;
- the 2,500 most frequent token bigrams;
- the 5,000 most frequent character trigrams;
- the 5,000 most frequent character tetragrams.

3.2 Evaluation Measures

For each fold in each five-fold cross-validation experiment, several evaluation scores were calculated, based on the system's age and gender predictions for the test documents. The calculated scores are precision, recall and F_1-score per class and the overall accuracy and macro-averaged F_1-score. Scores were averaged across the five folds. Accuracy scores were compared to baseline accuracy scores produced by a system that always predicts the majority class.

The age and gender information provided in the users' Netlog profiles was used as gold standard class information. Although this profile information is not fully reliable, we assume that the portion of obfuscated profile information in our data is sufficiently small to appropriately train and evaluate the classifier.

4 RESULTS AND DISCUSSION

In this section, we discuss the results for the three prediction tasks: age prediction, gender prediction and combined age and gender prediction. For each task, we perform five-fold cross-validation experiments, and compare the results produced with the full, unbalanced, dataset to the results produced with balanced subsets. The reported scores are the average scores across five folds. In the graphs that include error bars, these error bars indicate the 95% confidence intervals, based on the standard deviations of the scores across the five folds.

4.2 Age Prediction

Figure 3(a) displays the age prediction scores for the different age boundaries on the full, unbalanced dataset. The accuracy rises from 76.7% with age boundary 16 to 91.7% with age boundary 28. The curve is quite steep in the beginning and starts to level off towards the end. The macro-averaged F-score reaches a maximum of 84.8% at age boundary 23 and slowly decreases after that.

As we see in Figure 3(b), the rise in the accuracy score is mainly due to increased precision and recall scores for the younger class. This is caused partially by the growing class imbalance: as the age boundary rises, the portion of instances in the younger class grows, which has a positive effect on the scores for this class. The growing class imbalance is also reflected in the rise of the majority baseline in Figure 3(a). Still, the accuracy curve in Figure 3(a) remains far above the baseline. The precision and recall scores for the older class remain reasonably stable and show a moderate decrease at the end, which causes the slight decline in the macro-averaged F-score after age boundary 23.

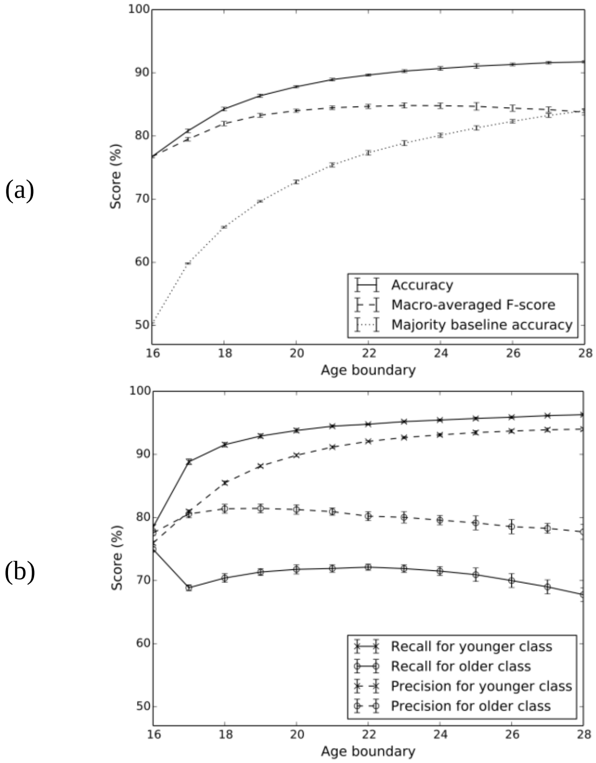

Figure 3. Age prediction scores (overall (a) and per class (b)) with the unbalanced dataset.

Figure 4 shows the results when the effect of increasing class imbalance is eliminated. In Figure 4(b), we can see that with data subsets balanced for age class (and gender), the precision and recall scores for both classes rise. Consequently, the macro-averaged F-score also keeps rising.

In addition to the scores per age class, it is also important to know how the age classifiers perform for authors of specific ages. Figure 5 shows the age prediction accuracies per age for four different age boundaries, produced with the unbalanced dataset. Figure 6 shows the same for four balanced data subsets; they are the same subsets for which the age distributions are depicted in Figure 2. All

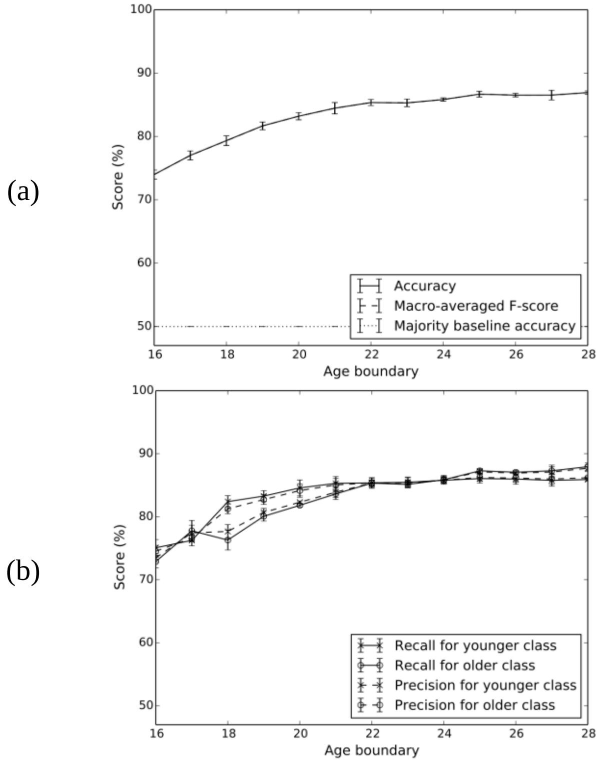

Figure 4. Age prediction scores with the balanced data subsets. In (a), the macro-averaged F-scores are not visible, as they almost fully overlap with the accuracy scores. The majority baseline accuracy in (a) is at a constant level of 50%, regardless of the age boundary.

graphs in Figure 5 and Figure 6 show a clear accuracy drop around the chosen age boundary. This means that texts by authors with ages close to an age boundary are harder to classify, because they are relatively similar to texts by authors close to the other side of the boundary.

With the unbalanced dataset (Figure 5), the minimum of the drop is always at the age just above the age boundary and the drop is much steeper on the left side than on the right side. As the age boundary rises, the drop gets wider, especially on the right side, which means that the scores for the higher ages decrease. The shape of the drop is partly related to the age distribution in the dataset, as can be seen when comparing the graphs for the unbalanced dataset (Figure 5) with those for the balanced data subsets (Figure 6) and by relating them to the age distributions in Figure 1(a) and Figure 2.

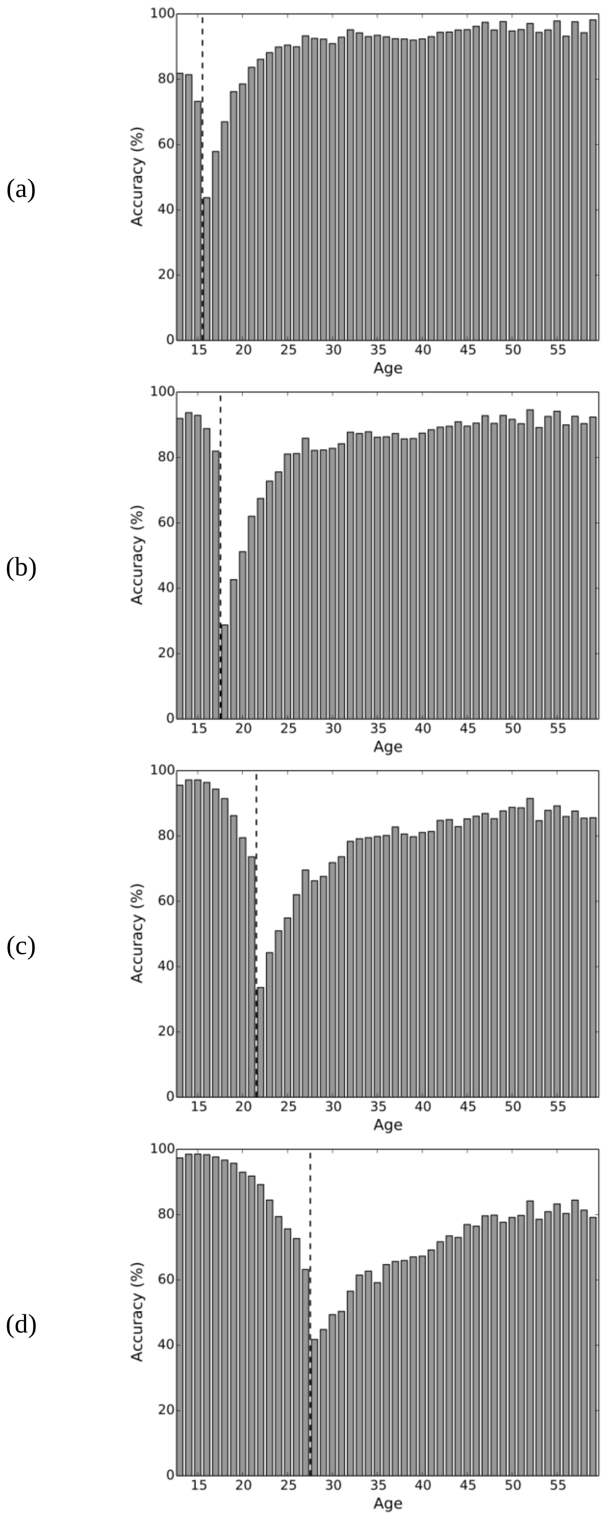

Figure 5. Age prediction: accuracy scores for the unbalanced dataset, when predicting age with respect to four different age boundaries: 16 (a), 18 (b), 22 (c), and 28 (d). In each graph, the ages on the x-axis are the true ages of the authors, the scores on the y-axis are the age prediction accuracies produced for the authors of that specific age, and the vertical dashed line marks the age boundary.

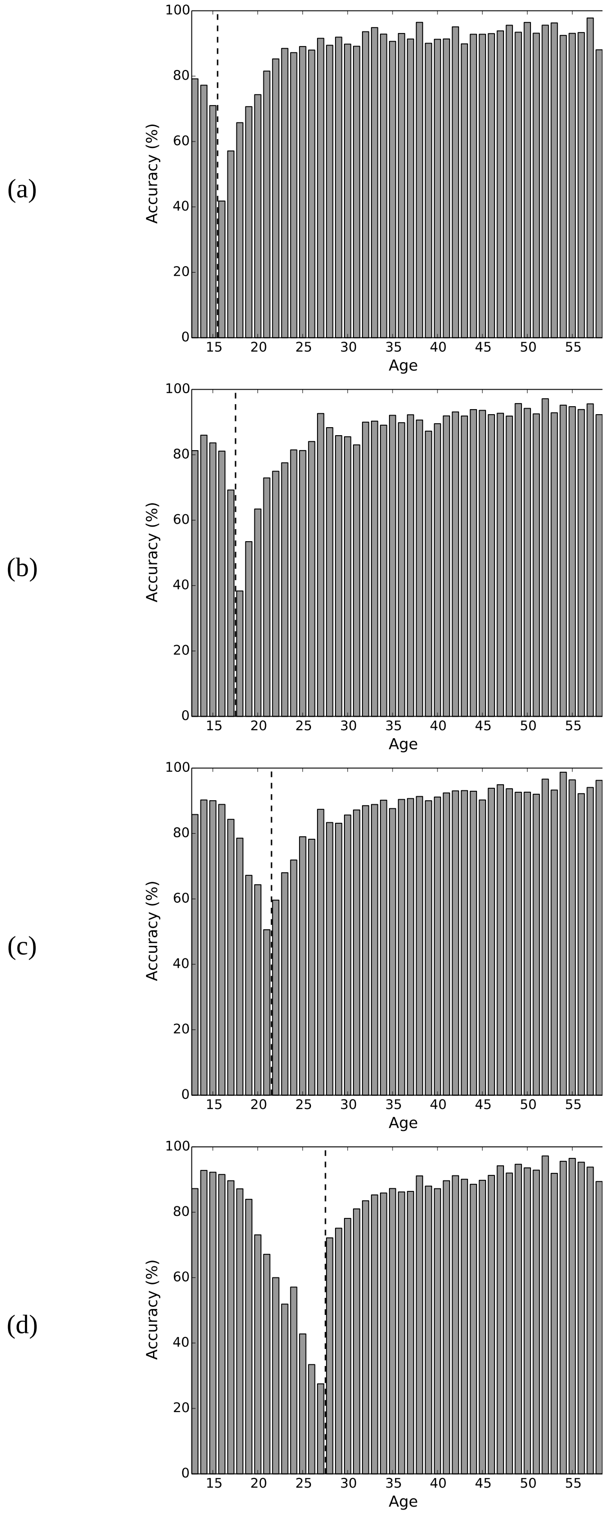

Figure 6. Age prediction: accuracy scores per age with the balanced data subsets, when predicting age with respect to four different age boundaries: 16 (a), 18 (b), 22 (c), and 28 (d). In each graph, the ages on the x-axis are the true ages of the authors, the scores on the y-axis are the age prediction accuracies produced for the authors of that specific age, and the vertical dashed line marks the age boundary.

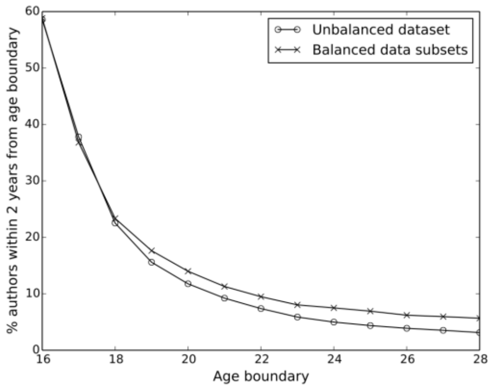

Figure 7. Percentage of authors within 2 years from the age boundary in the unbalanced dataset and in the balanced data subsets.

The effect of closeness to age boundary also plays a role in the increase of the accuracy scores in Figure 3 and Figure 4. As the age boundary increases, the percentage of authors close to it decreases, since the high peak in the age distribution is situated at the lower ages (cf. Figure 1(a)). This effect is shown in Figure 7; it is not only present in the unbalanced dataset, but also in the balanced data subsets. As a result, the average age prediction accuracy rises when the age boundary increases, since the accuracy scores are relatively high for authors further from the age boundary. However, it is unknown to what extent this factor influenced the scores, and excluding both this factor and the factor of class imbalance at the same time is impossible with this dataset.

As expected, another important factor that affects the age classification performance is the length of the document that is classified: on average, longer documents are classified more accurately than shorter documents. Figure 8 shows the macro-averaged F-scores for different document length categories, produced with the full dataset. For the most frequent document length category in the dataset, with documents of 11 to 100 tokens (see Figure 1(b)), the macro-averaged F-scores range between 77.8% and 85.6%, depending on the age boundary. However, for short documents, with only 5 to 10 tokens, the macro-averaged F-scores are still reasonable: they are between 68.4% and 76.9%. When the documents contain more than 1,000 tokens, scores are above 90% for all age boundaries except age boundary 16 (they range between 86.8% and 93.4%).

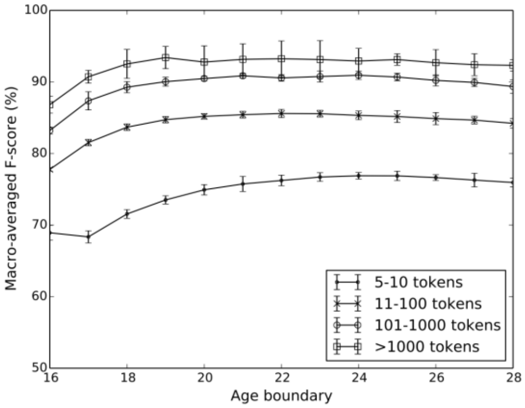

Figure 8. Age prediction scores per document length (i.e. number of tokens in the document) with the unbalanced dataset.

4.2 Gender Prediction

The gender prediction scores are shown in Table 2. The accuracy with the data subset balanced for gender is very similar to the accuracy with the full dataset, even slightly higher, although the dataset is a bit smaller (70,060 vs. 86,610 documents). With the full dataset, the precision and recall scores for the female class are higher than the scores for the male class, especially the recall scores (79.7% vs. 53.0%). This is probably due to the class imbalance (51,269 female authors vs. 35,341 male authors), as with the balanced data subsets, the recall for the female class is *lower* than the recall for the male class (65.8% vs. 72.7%).

4.3 Combined Age and Gender Prediction

Figure 9 displays the overall scores for the combined age and gender prediction task. These scores were produced by training the system on the four combined classes (cf. Table 1) and then predicting the same four classes in the test set. We also carried out experiments in which we predicted age and gender separately (with two systems, trained on the separate binary classes) and then combined the resulting age and gender predictions afterwards. This resulted in very similar scores (not shown here), only with a much larger variance across folds.

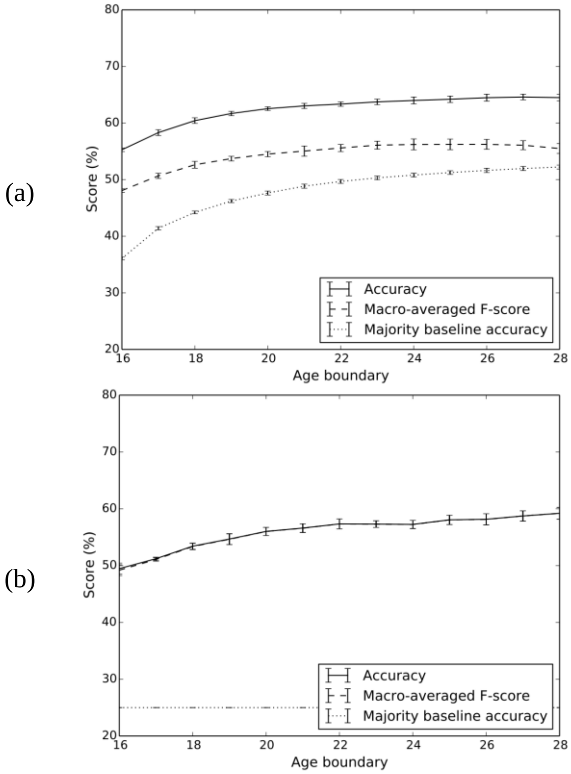

Figure 9. Combined age and gender prediction: overall scores per age boundary, produced with the unbalanced dataset (a) and the balanced data subsets (b). In (b), the macro-averaged F-scores are not visible, as they almost fully overlap with the accuracy scores. The majority baseline accuracy in (b) is at a constant level of 25%, regardless of the age boundary.

In Figure 9(a), which shows the scores with the unbalanced dataset, we see the accuracy score increase again as the age boundary rises, as in Figure 3(a), but the curve starts to level off earlier and the differences are smaller. The accuracy scores range between 55.3% (at age boundary 16) and 64.6% (at age boundary 27) and exceed the majority baseline accuracies by 12.3% to 19.2%. The macro-averaged F-score reaches its maximum of 56.2% at age boundary 26 and then slowly starts

Table 2. Gender prediction scores on the full dataset and on the data subset balanced for gender, averaged across five folds. Macro F = macro-averaged F-score.

	Overall Scores		Scores for ♀			Scores for ♂		
Dataset	**Accuracy**	**Macro-F**	**Precision**	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
Full	68.8	66.6	71.1	79.7	75.1	64.2	53.0	58.0
Balanced	69.3	69.2	70.7	65.8	68.2	68.0	72.7	70.3

degrading again. With the balanced subsets (cf. Figure 9(b)), the accuracy rises from 49.5% to 59.2%, as does the macro-averaged F-score.

4.3 Consequences for the Application

For the application of sexual predator detection, one of the most important aims is to distinguish authors above and below the age of consent for sexual activity, which is currently age 16 in Belgium. We need a -16 classifier to detect the potential victims and a 16+ classifier to detect the potential offenders. For both classifiers, a high recall is most important, but the precision should also be reasonably high to minimize the number of manual interventions by moderators. In addition, accurate classification of -16 and 18+ authors has the highest priority.

With our current unbalanced dataset, the recall of the -16 classifier (with -16 as the positive class) is 78.5% and its precision is 76.0% (cf. the scores for the younger class in Figure 3(b) at age boundary 16). As we can see in Figure 5(a), a large part of the errors pertain to authors that are just above the boundary. Since our focus is mainly on the correct classification of -16 and 18+ authors, we also calculated a more lenient precision score, which excludes the 16-year-olds and 17-year-olds from the false positives. This score, which we call “precision -18”, computes the percentage of -18 authors within the group of authors classified as -16. For our -16 classifier, the “precision -18” score is 91.1%, i.e. much higher than the standard precision score (“precision -16”). This illustrates that a large portion of the false positive authors are 16 or 17 years old.

The 16+ classifier (with 16+ as the positive class) has a recall of 75.0% and a precision of 77.5% on our dataset (cf. the scores for the older class in Figure 3(b) at age boundary 16). Also for this classifier, we calculated a more lenient score, focusing on the 18+ authors. The recall for 18+ authors (“recall 18+”) is 86.6%, which is again much higher than the recall score for 16+ authors. The high recall for 18+ authors is also visible in Figure 5(a); this figure shows the accuracy scores for the specific ages, which correspond to the recall scores per age. As age increases, the recall scores rise, until they start leveling off after age 25 and consistently stay between 90% and 98%.

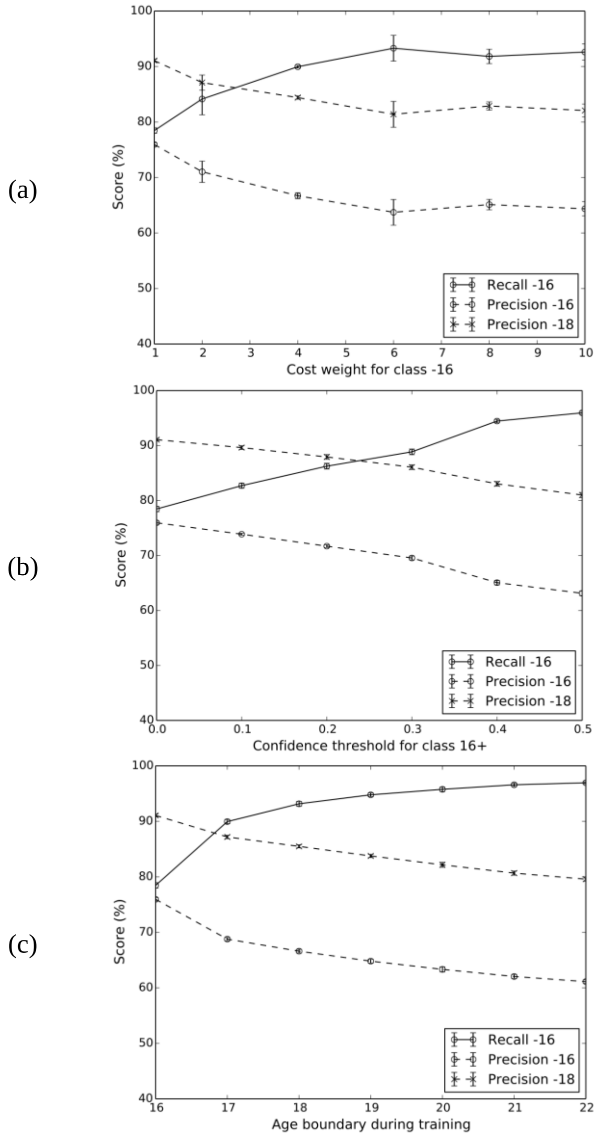

Figure 10. Precision and recall scores for the -16 classifier when using different methods to improve the recall of the class -16: (a) changing cost weight, (b) changing confidence threshold, and (c) changing age boundary during training.

The recall of the -16 classifier, which is important for our application, could be increased in several ways. A standard method for improving recall is to use a higher cost weight for the positive class (-16) during training. Another way is to use the confidence scores that are produced by the SVM classifier (based on an instance’s distance to the hyperplane): if an instance is classified as 16+ with a low confidence score, below a specific threshold, classify it as -16 instead of 16+. A third option is to increase the age boundary that is

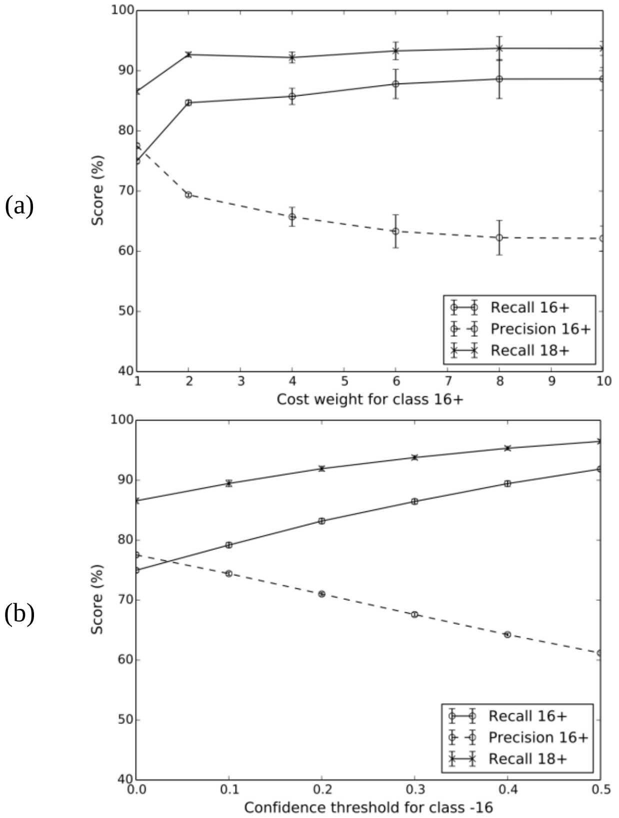

Figure 11. Precision and recall scores for the 16+ classifier when using different methods to improve the recall of the class 16+: (a) changing cost weight, (b) changing confidence threshold.

used for training. As Figure 5 shows, the recall for ages 13 to 15 gets higher when the age boundary rises. So we can train the classifier with instances labeled with the classes "younger" and "older" according to a higher age boundary (e.g. boundary 17), use this classifier to label the test set, and evaluate the resulting labels "younger" and "older" according to the age boundary 16.

We applied the three methods in 5-fold cross-validation experiments on our full, unbalanced dataset, to explore the effects of the different methods on the precision and recall scores. The results are shown in Figure 10. We see that we can achieve recall scores between 90% and 95% with precision scores between 65% and 70%. In addition, the gap between the standard precision score ("precision -16") and the "precision -18" score is very large, especially when we increase the age boundary during training (Figure 10(c)), which means that a large portion of the false positives consists of 16-year-old and 17-year-old authors. With age boundary 17, for instance, recall is 90.0%, with a "precision -16" score of 68.8% and a "precision -18" score of 87.2%.

With cost weight 4 and confidence score 0.3, the recall scores are very similar to the recall with age boundary 17, but the precision scores are less favorable. With cost weight 4, both "precision -16" and "precision -18" are lower (66.7% and 84.4%, respectively). With confidence threshold 0.3, "precision -16" is comparable but the gap between "precision -16" and "precision -18" is a bit smaller (18.4% vs. 16.5%). These tendencies also apply at other comparable recall scores.

Although these exploratory experiments do not show how these results generalize to new data (since we did not use a development set to tune towards high recall in these experiments), the results do show that all three methods are worth considering to improve recall of the -16 class in our final application and that practically usable performance levels can be attained using these methods. Other methods that could be considered are cost-sensitive learning methods such as cost-proportionate rejection sampling [33], in which the negative class is repeatedly downsampled and results are combined in an ensemble set-up.

The methods that use cost weights and confidence thresholds can also be used to improve the recall of the 16+ classifier. The resulting precision and recall scores are shown in Figure 11. The method of moving the age boundary cannot be used here, since we can only increase the age boundary with our current dataset, which decreases the recall for the 16+ class. When we compare the average scores of the two methods for which recall scores for class 16+ are similar, we see that the precision scores and "recall 18+" scores are either very similar or more favorable for the method with the adapted confidence thresholds. Notably, adapting the cost weights yields a much larger score variance across folds and therefore less stable results. With confidence threshold 0.2, recall is 83.2% for 16+ authors and 91.9% for 18+ authors, at a precision of 71.0% for 16+ authors. These are also practically usable scores for detecting potential child groomers.

In addition, gender prediction can be used to detect disagreement between the self-reported gender in a user's profile and the profiling system's gender prediction for that user. Usually, when a user provides false gender information for child grooming purposes, the user is a man who pretends to be female. Unfortunately, the recall for male authors with our unbalanced dataset was only 53.0% (cf. Table 2). Also here, the recall for the male class could be improved by using techniques such as increasing the cost weight for the male class or increasing the confidence threshold for the female class. When combined with content-based information and age prediction, gender discrepancy can be a useful extra cue for sexual predator detection.

4.4 Additional Features

So far, we have only considered character and token n-grams as relevant features towards classification. In a final set of experiments, we investigated the applicability of additional features, provided by the CLiPS profiling software PROFL[6]. These are part-of-speech n-grams, sentiment features (polarity score), LIWC-features and features that quantify general stylistic properties such as average word length, number of emoticons and the like. Furthermore, we also experimented using only character or token n-grams to study their effectiveness in isolation.

Table 3 displays the results of these experiments. Using character and token n-grams in isolation hurts the accuracy of both gender and age prediction. But while the additional features do not aid age classification, they do yield significant advances for gender prediction. Additional experiments are needed to fully explore the spectrum of available features for these classification tasks.

Table 3. Effect of additional features for age and gender prediction. Results in bold indicate statistically significant results, as measured using approximate randomization testing.

	age	**gender**
Baseline: character + token n-grams	*76.7*	*68.8*
+ pos n-grams	**76.4**	**69.3**
+ PROFL-stylistic features	**76.2**	**69.5**
+ LIWC	76.7	**69.5**
+ sentiment	76.7	**69.8**
+ all of the above	**76.1**	**69.2**
only character n-grams	**76.4**	**66.2**
only token n-grams	**74.7**	**61.3**

5 CONCLUSION AND FUTURE WORK

We explored the capabilities of a text-based age and gender profiling system for application in a monitoring environment to secure the online safety of (young) social media users. More specifically, our research focused on the task of detecting sexual predators who try to “groom” children on social networking websites, often providing false age and/or gender information to get closer to their young targets. We presented results of age and gender prediction experiments on a dataset of almost 380,000 Dutch chat posts written by 86,610 users on the social networking platform Netlog.

The age prediction task was set up as a binary classification task, i.e. predicting whether an author is under or over a specific age. The age boundary that separates the two classes can be adapted to the specific use case at hand, based on legal constraints (e.g. the legal age of sexual consent) and age related statistics (e.g. grooming statistics). We carried out age prediction experiments with a range of different age boundaries and found that that macro-averaged F-scores improved as the age boundary increased. This effect persisted when we used data subsets that were balanced for age category and gender.

In addition, we presented a detailed analysis of the system’s performance for authors of different ages, showing that classification errors were mainly concentrated around the age boundary. The consequences of our findings for the application were discussed, zooming in on the case study of detecting sexual offenders and their minor victims according to Belgium’s current legal age of sexual consent, age 16. We found that practically usable recall and precision scores could be achieved for both the -16 and the 16+ classifier, especially when tuning the system towards a high recall. Furthermore, gender prediction, although not yielding high performance in our present experiments, can provide useful information when applied in addition to content analysis and age prediction.

[6] http://amicaproject.be/profl

In future research, we plan to further extend and optimize the feature set used for age and gender classification and extend our experiments to different datasets, including more recently collected chat data and other social media genres such as blog posts and tweets. We will also further test methods for high-recall tuning, including cost-sensitive learning techniques such as cost-proportionate rejection sampling [33]. In addition, we would like to study the applicability of linear regression for age prediction in the AMiCA system. In recent work, [20] produced encouraging results predicting author age on twitter using this technique. Finally, a crucial step is to test our profiling system on real-life sexual predator data, to investigate to what extent the method also works when people pretend to be someone of a different age and/or gender. In these tests, we will also add a text analysis component to detect sexual and grooming-related content in conversations.

ACKNOWLEDGMENTS

The research described in this paper was funded by IWT-SBO grant 120007 (AMiCA). We would like to thank Massive Media for providing the Netlog data.

REFERENCES

1. Peter Bach, Marjorie Glass Zauderer, Ayca Gucalp, Andrew S Epstein, Larry Norton, Andrew David Seidman, Aryeh Caroline, Alexander Grigorenko, Aleksandra Bartashnik, Isaac Wagner, Jeffrey Keesing, Martin Kohn, Franny Hsiao, Mark Megerian, Rick J Stevens, Jennifer Malin, John Whitney, Mark G. Kris. Beyond jeopardy!: Harnessing IBM’s Watson to improve oncology decision making. *Journal of Clinical Oncology*, 31(suppl;abstract 6508), 2013.
2. Enric Junque de Fortuny, Tom De Smedt, David Martens, and Walter Daelemans. Media coverage in times of political crisis: A text mining approach. *Expert Systems with Applications*, 39(14):11616–11622, 2012.
3. Enric Junque de Fortuny, Tom De Smedt, David Martens, and Walter Daelemans. Evaluating and understanding text-based stock price prediction models. *Information Processing & Management*, 50(2):426–441, 2014.
4. Moshe Koppel, Shlomo Argamon, and Anat R. Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17:401–412, 2002.
5. James W Pennebaker and Lori D Stone. Words of wisdom: language use over the life span. *Journal of personality and social psychology*, 85(2):291–301, 2003.
6. Walter Daelemans. Explanation in computational stylometry. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing, volume 7817 of Lecture Notes in Computer Science*, pages 451–462. Berlin, Heidelberg, 2013. Springer Berlin.
7. Giacomo Inches and Fabio Crestani. Overview of the international sexual predator identification competition at PAN-2012. In Pamela Forner and Jussi Karlgren and Christa Womser-Hacker, editors, *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers*, Rome, Italy, 2012. CLEF.
8. Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. Overview of the author profiling task at PAN-2013. In Pamela Forner, Roberto Navigli and Dan Tufis, editors, *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers*, pages 352–365, Valencia, Spain, 2013. CELCT.
9. Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. Overview of the 2nd author profiling task at PAN-2014. In Linda Cappellato and Nicola Ferro and Martin Halvey and Wessel Kraaij, editors, *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers*, pages 898–927, Sheffield, UK, 2014. CEUR-WS.org.
10. Francisco Rangel, Fabio Celli, Paolo Rosso, Martin Pottast, Benno Stein, Walter Daelemans. Overview of the 3rd Author Profiling Task at PAN 2015. In Linda Cappelato and Nicola Ferro and Gareth Jones and Eric San Juan, editors, *CLEF 2015 Labs and Workshops, Notebook Papers*, Toulouse, France, 2015. CEUR-WS.org
11. Bart Desmet and Veronique Hoste. Emotion detection in suicide notes. *Expert Systems With Applications*, 40(16):6351–6358, 2013.
12. Cynthia Van Hee, Ben Verhoeven, Julie Mennes, Els Lefever, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. Detection and fine-grained classification of cyberbullying events. In Galia Angelova and Kalina Bontcheva and Ruslan Mitkov, editors, *Proceedings of the 10th Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria, 2015. Association for Computational Linguistics
13. Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, pages 37–44, New York, USA, 2011. ACM.
14. Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium*, pages 199–205, Menlo Park, USA, 2006. The AAAI Press.

15. Shlomo Argamon, Moshe Koppel, James W Pennbaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123, 2009.
16. Cathy Zhang and Pengyu Zhang. Predicting gender from blog posts. Technical report, University of Massachusetts Amherst, USA, 2010.
17. Arjun Mukherjee and Bing Liu. Improving gender classification of blog authors. In Jun'ichi Tsujii and James Henderson and Marius Pasca, editors, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 207–217, Cambridge, MA, 2010. Association for Computational Linguistics.
18. Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*, pages 37–44, New York, USA, 2010. ACM.
19. Shane Bergsma and Benjamin Van Durme. Using conceptual class attributes to characterize social media users. In Pascale Fung and Massimo Poesio, editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages 710–720, Sofia, Bulgaria, 2013. Association for Computational Linguistics.
20. Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. ”How Old Do You Think I Am?”; A Study of Language and Age in Twitter. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pages 439–448, Cambridge, USA, 2013. AAAI Press.
21. David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, 2014.
22. Katja Filippova. User demographics and language in an implicit social network. In Jun'ichi Tsujii and James Henderson and Marius Pasca, editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1478–1488, Jeju Island, Korea, 2012. Association for Computational Linguistics.
23. John D. Burger and John C. Henderson. An exploration of observable features related to blogger age. In *Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium*, pages 15–20, Menlo Park, USA, 2006. The AAAI Press.
24. Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pages 387–390, Palo Alto, USA, 2012. The AAAI Press.
25. Xiang Yan and Ling Yan. Gender classification of weblog authors. In *Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium*, pages 228–230, Menlo Park, USA, 2006. The AAAI Press.
26. Shigeyuki Sakaki, Yasuhide Miura, Xiaojun Ma, Keigo Hattori, and Tomoko Ohkuma. Twitter user gender inference using combined analysis of text and image processing. In *Proceedings of the Third Workshop on Vision and Language*, pages 54–61, 2014. Dublin City University and the Association for Computational Linguistics
27. Nikesh Garera and David Yarowsky. Modeling latent biographic attributes in conversational genres. In Keh-Yih Su and Jian Su and Janyce Wiebe and Haizhou Li, editors, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 710–718, Suntec, Singapore, 2009. Association for Computational Linguistics.
28. Sara Rosenthal and Kathleen McKeown. Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In Yuji Matsumoto and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 763–772, Portland, USA, 2011. Association for Computational Linguistics.
29. Sarah Schulz, Guy De Pauw, Orphee De Clercq, Véronique Hoste Bart Desmet, Walter Daelemans, and Lieve Macken. Multi-modular text normalization of dutch user-generated content. *ACM Transactions on Intelligent Systems and Technology*, 2016 (in press).
30. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
31. John D. Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on Twitter. In David Yarowsky and Timothy Baldwin and Anna Korhonen and Karen Livescu and Steven Bethard, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP ’11, pages 1301–1309, Stroudsburg, USA, 2011. Association for Computational Linguistics.
32. Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. Gender attribution: Tracing stylometric evidence beyond topic and genre. In Sharon Goldwater and Christopher Manning, editors, *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 78–86, Portland, USA, June 2011. Association for Computational Linguistics.
33. Bianca Zadrozny, John Langford, and Naoki Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Proceedings of the Third IEEE International Conference on Data Mining*, ICDM ’03, pages

435–442, Washington, USA, 2003. IEEE Computer Society.