What Grammar Tells About Gender and Age of Authors

Michael Tschuggnall and Günther Specht
Databases and Information Systems
Institute of Computer Science, University of Innsbruck, Austria
{michael.tschuggnall, guenther.specht}@uibk.ac.at

***Abstract*—The automatic classification of data has become a major research topic in the last years, and especially the analysis of text has gained interest due to the availability of huge amounts of online documents. In this paper, a novel style feature based on grammar syntax analysis is presented that can be used to automatically profile authors, i.e., to predict gender and age of the originator. Using full grammar trees of the sentences of a document, substructures of the trees are extracted by utilizing pq-grams. The mostly used patterns are then stored in a profile, which serve as input features for common machine learning algorithms. An extensive evaluation using a state-of-the-art test set containing thousands of English web blogs investigates on the optimal parameter and classifier configuration. Finally, promising results indicate that the proposed feature can be used as a significant characteristic to automatically predict the gender and age of authors.**

***Index Terms*—Author Profiling; Text Classification; Grammar Trees; Machine Learning.**

I. Introduction

With the advent of the internet in general and recently especially with social media, users frequently use the numerous possibilities to compose and post text in various ways. Considering current statistics [1] estimating 70 billion pieces of content shared via Facebook or 190 million short messages posted on Twitter every day, the amount of shared textual information is huge. Although the authors of the latter examples are generally known, the information is most often restricted to a user name. Moreover, there also exist cases like anonymized blogs where every information concerning the originator is intentionally hidden.

In contrast to traditional authorship attribution approaches [2] that try to assign one of several known candidate authors to an unlabeled document, the author profiling problem deals with the extraction of useful meta information about the author. Often this information includes gender and age of the originator [3][4][5], but also other demographic information like cultural background or psychological analyses are examined in recent approaches [6][7]. Where the mining of such information can be applied very well to commercial applications by knowing the percentages of gender and age commenting on a new product release for example, it is also of growing importance in juridical applications (*Forensic Linguistics*) [8], where, e.g., the number of possible perpetrators can be reduced. Moreover especially nowadays in the area of cybercrime [9], recent approaches investigate the content of e-mails [10], suicide letters or try to automatically expose sexual predators from chat logs [11].

In this paper, a novel style feature to automatically extract the gender and age of authors of text documents is presented.

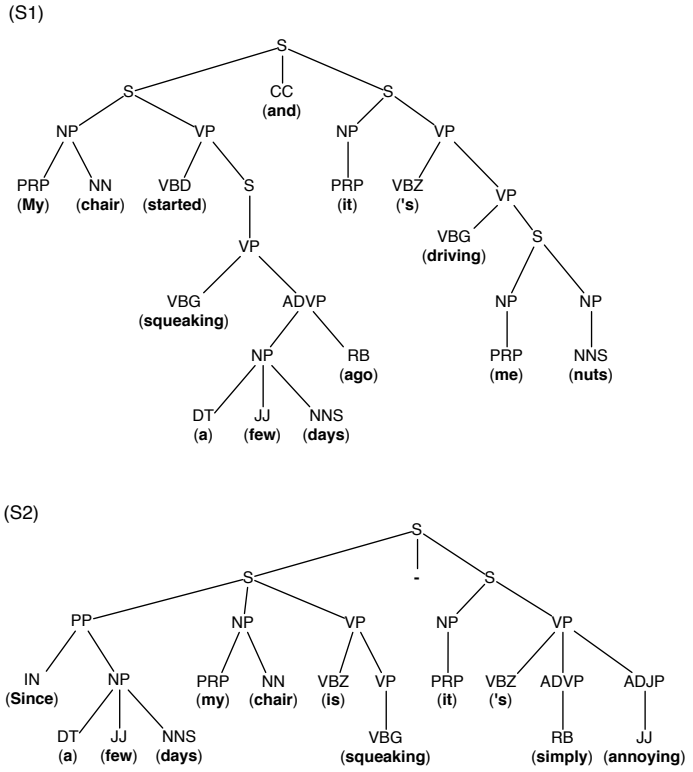

Fig. 1. Grammar Trees of the Semantically Equivalent Sentences (S1) and (S2).

Using results of previous work in the field of intrinsic plagiarism detection [12] and authorship attribution [13], the assumption that individual authors have significantly different writing styles in terms of the syntax that is used to construct sentences has been reused. For example, the following sentence extracted from a web blog: *"My chair started squeaking a few days ago and it's driving me nuts."* (S1) could also be formulated as *"Since a few days my chair is squeaking - it's simply annoying."* (S2) which is semantically equivalent but differs significantly according to the syntax as can be seen in Figure 1. The main idea of this work is to quantify those differences by calculating grammar profiles using pq-grams of full grammar trees, and to evaluate how reliable a prediction of an authors meta information is when solely this grammar feature is used. Given the grammar profiles, the prediction of gender and age, respectively, is finally examined by utilizing modern machine learning approaches like support vector machines, decision trees or Naive Bayes classifiers.

 ISBN: 978-1-61208-364-3

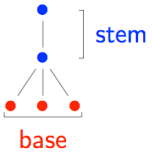

Fig. 2. Structure of a pq-gram Consisting of Stem $p = 2$ and Base $q = 3$.

The rest of this paper is organized as follows: Section II recaps the concept of pq-grams as a fundamental basis of this work, while Section III explains the profiling process in detail. An extensive and promising evaluation using a comprehensive test set of web blogs is presented in Section IV. Finally, related work is summarized in Section V and conclusions including future work are discussed in Section VI.

II. Preliminaries: pq-grams

Similar to n-grams that represent subparts of given length n of a string, pq-grams extract substructures of an ordered, labeled tree [14][15]. The size of a pq-gram is determined by a stem (p) and a base (q) like it is shown in Figure 2. Thereby p defines how much nodes are included vertically, and q defines the number of nodes to be considered horizontally. For example, a valid pq-gram with $p = 2$ and $q = 3$ starting from `PP` at the left side of tree (S2) shown in Figure 1 would be `[PP-NP-DT-JJ-NNS]` (the concrete words are omitted).

The pq-gram index then consists of all possible pq-grams of a tree. In order to obtain all pq-grams, the base is shifted left and right additionally: If then less than p nodes exist horizontally, the corresponding place in the pq-gram is filled with $\star$, indicating a missing node. Applying this idea to the previous example, also the pq-gram `[PP-IN-*-*-*]` (no nodes in the base) is valid, as well as `[PP-NP-*-*-DT]` (base shifted left by two), `[PP-NP-*-DT-JJ]` (base shifted left by one), `[PP-NP-JJ-NNS-*]` (base shifted right by one) and `[PP-NP-NNS-*-*]` (base shifted right by two) have to be considered. As a last example, all leaves have the pq-gram pattern `[`*leaf_label*`-*-*-*-*]`.

Finally, the pq-gram index is the set of all valid pq-grams of a tree, whereby multiple occurrences of the same pq-grams are also present multiple times in the index.

III. Profiling Authors Using pq-gram Indices

The number of choices an author has to formulate a sentence in terms of grammar structure is rather high, and the assumption in this approach is that the concrete choice is made mostly intuitively and unconsciously. Evaluations shown in Section IV reenforce that solely grammar syntax represents a significant feature that can be used to categorize authors.

Basically, the profiling of a given text using pq-grams works as follows:

1) At first the text is parsed and split into single sentences using common sentence boundary detection algorithms, which is currently implemented with *OpenNLP* [16]. Each sentence is then anaylzed by its grammar, i.e., a full syntax tree is calculated using the *Stanford Parser* [17]. For example, Figure 1 depicts the grammar trees resulting from analyzing sentences (S1) and (S2), respectively. The labels of each tree correspond to a part-of-speech (POS) tag of the Penn Treebank set [18], where e.g *NP* corresponds to a noun phrase, *DT* to a determiner or *JJS* to a superlative adjective. In order to examine the building structure of sentences only like it is intended by this work, the concrete words, i.e., the leafs of the tree, are omitted. In case of ambiguity of grammar trees, i.e., if there exist more than one valid parse tree for a sentence, the tree with the highest probability estimated by the parser is chosen.
2) Using the grammar trees of all sentences of the document, the pq-gram index is calculated. As shown in Section II all valid pq-grams of a sentence are extracted and stored into a pq-gram index. By combining all pq-gram indices of all sentences, a pq-gram profile is computed which contains a list of all pq-grams and their corresponding frequency of appearance in the text. Thereby the frequency is normalized by the total number of all appearing pq-grams. As an example, the three mostly used pq-grams using $p = 2$ and $q = 3$ of a sample document are shown in Table I. The profile is sorted descending by the normalized occurrence, and an additional rank value is introduced that simply defines a natural order and is used in the evaluation (see Section IV).

TABLE I
Example of the Three Mostly Used pq-grams of a Sample Document.

pq-gram	Occurrence [%]	Rank
`NP-NN-*-*-*`	2.68	1
`PP-IN-*-*-*`	2.25	2
`NP-DT-*-*-*`	1.99	3

3) Finally, the pq-gram profiles including occurrences and ranks are used as features that are applied to common machine learning algorithms. This step is explained in detail in Section IV.

IV. Evaluation

Basically, the prediction of gender and age of the author of a text document is made by machine learning algorithms. Independent of the classifier used (see Section IV-D), the input consists of a large list of features with appropriate values and a corresponding classification class. The class is used to train the algorithms if the document is part of the training set, as well as for evaluating if the document is part of the test set. Details on the usage of training and test sets, respectively, and on the test corpus in general are explained in Section IV-C.

A. Features

The features that have been used as input for the classifiers consist of the pq-gram profiles described previously. Thereby, each pq-gram represents a feature. To examine the significance of the concrete percentage of occurrence compared to the plain rank, a pq-gram-rank feature has been added additionally.

 ISBN: 978-1-61208-364-3

A small example of a feature list including the correct gender and age classification is depicted in Table II. If a document does not contain a specific feature, i.e., a pq-gram, the feature value for the pq-gram as well as for the corresponding rank is set to -1. For example, the author of document C didn't use the structure `[PP-IN-*-*-*]` to build his/her sentences, so therefore the according feature values are set to -1.

TABLE II
EXAMPLE OF A FEATURE LIST SERVING AS INPUT FOR CLASSIFICATION ALGORITHMS.

Feature	Doc. A	Doc. B	Doc. C
`NP-NN-*-*-*`	2.68	1.89	2.84
`NP-NN-*-*-*--RANK`	1	6	2
`PP-IN-*-*-*`	2.25	0.24	-1
`PP-IN-*-*-*-RANK`	2	153	-1
`NP-DT-*-*-*`	1.99	2.11	1.23
`NP-DT-*-*-*-RANK`	3	2	11
...	...	...	...
correct gender	male	female	male
correct age	20s	10s	30s

Depending on the evaluation setup shown subsequently the number of attributes to be handled by the classification algorithms range between 7,000 and 20,000.

B. Evaluation Setup

The computation of the feature list is an essential part of the approach. Basically, it depends on the assignment of p and q, respectively, that is used for the extraction of pq-grams from sentences. For example, by using $p = 1$ and $q = 0$ the pq-grams would be reduced to single POS tags. Nevertheless, based on results in previous work such configurations have been excluded as they led to insufficient results. The range of both stem and base of pq-grams has been evaluated in the range between 2 and 4, conforming to the size of n-grams that are used in efficient approaches in information retrieval (e.g. [19]).

Considering the huge amount of possible features, especially if $p + q > 6$, the maximum number of sentences per text sample (s_{max}) as well as the maximum number of pq-grams in a profile (pq_{max}) have been limited. Accordingly, only the first 200 sentences of each document have been processed. The final pq-gram profile has then been sorted descending by the rank and limited to the 500 mostly used patterns.

Finally, three different feature sets have been used as input for the machine learning algorithms: the percentage of occurrence of each pq-gram, the rank of each pq-gram, and a combination of both occurrence-rate and rank.

An overview of all settings that have been evaluated can be seen in Table III.

C. Test Set

The approach has been evaluated extensively using a state-of-the-art test set created by Schler et. al [5], containing thousands of freely accessible English web blogs. For this evaluation, a subset of approximately 8,000 randomly selected blogs have been used, whereby for each blog entry the gender as well as the age of the composer is given.

TABLE III
PARAMETER SETUP USED FOR THE EVALUATION.

Parameter	Assignment
p, q	$2 - 4$
s_{max}	200
pq_{max}	500
input feature set	occurrence-rate, rank, combined

Regarding the latter, the ages are clustered into three distinct groups, as defined by the original test set [5]: 13-17 (=10s), 23-27 (=20s) and 33-42 (=30s). The five-year gap between each group is thereby added to gain higher distinguishability. The corpus is fairly balanced with respect to gender, but has a majority in the 20s group and a minority in the 30s group. A detailed information about the class distribution is shown in Table IV. Because of the fact that simply predicting the majority class in all cases would lead to an accuracy of, e.g., 53% for `male`, the baseline which should be exceeded is set accordingly to 53% for gender, 46% for age and 25% for gender+age profiling, respectively.

TABLE IV
TEST DATA DISTRIBUTION.

	female	male	sum
10s	18%	19%	37%
20s	21%	25%	46%
30s	8%	9%	17%
Sum	47%	53%	

Each blog consists of at least 200 English words and has been textually cleaned in the original test data, i.e all unnecessary whitespace characters and HTML tags etc. have already been removed. Hyperlinks have been replaced by the word '`urlLink`'. Nonetheless, because this approach depends on the calculation of grammar trees, the latter tags have been manually removed for the evaluation, as the computation of grammar trees would be falsified.

D. Classifiers

Besides the parameter settings the accuracy of the profiling process depends on the classification algorithm that is used in combination with the set of features that are applied. Therefore, to determine the best working algorithm for this approach, several commonly used methods have been tested. Using the WEKA toolkit as a general framework [20], the following classifiers have been utilized: Naive Bayes classifier [21], Bayes Network using the K2 classifier [22], Large Linear Classification using LibLinear [23], support vector machines using LibSVM with nu-SVC classification [24], k-nearest-neighbours classifier (kNN) [25] using $k = 1$ and a pruned C4.5 decision tree [26].

E. Results

All possible settings, i.e., combinations of assignments of p and q with classifiers, have been evaluated on the test set

 ISBN: 978-1-61208-364-3

using a 10-fold cross validation. For each classifier the best results for predicting the gender, age and both gender and age combined are shown in Table V. The detailed results for each feature set is depicted, as well as the concrete sub results for the individual classes. Note that the average value is weighted, i.e., adjusted to the test data distribution.

In general, the results could significantly exceed the corresponding baselines, which manifests that solely the grammar of authors - analyzed with syntax trees and pq-grams - serves as a distinct feature for author profiling.

Despite of the class to predict, the support vector machine framework *LibSVM* and the large linear classification *LibLinear* worked best, whereas the kNN classifier and the C4.5 decision tree produced worse results. Also, the combined feature set using the occurrence-rate and the rank is always inferior to the isolated subsets, which may possibly be correlated to the large amount of features employed with this set (double the size of the other sets).

1) Gender Results: The best result using $p = 2$ and $q = 3$ could be achieved with *LibSVM*, leading to an accuracy of 69%. It utilizes the occurrence-rate feature set, whereby males could be identified with 71%. Although the prediction rate is a little worse than those of other approaches (e.g. [5] achieves 80% over the full test set using several style and content features), the result is promising as it uses and evaluates only the proposed feature and the baseline of 53% could be surpassed clearly.

2) Age Results: Using an almost identical setting, the maximum accuracy of 63% results again from using *LibSVM* and the occurrence-rate feature set (but with $p = 3$ instead of $p = 2$). In general the accuracy for the prediction of the age groups `10s` and `20s` are very solid, but all classifiers have problems predicting the `30s` group. For example, the best configuration achieved a rate of 70% for `10s` and 68% for `20s`, respectively, but could only predict 5% correctly in the eldest group. While the other algorithms could profile the latter class at a higher accuracy, interestingly the Naive Bayes classifier even missed it totally.

A reason for this may be the unbalanced distribution of the test data, which contains only a small amount of `30s` text samples compared to the other groups. It might be the case that the classifiers would have needed more samples to construct a proper prediction model. Even though the unbalanced test set is an immediate consequence of the original test data distribution ([5]), future work should try to create a smaller, but equally distributed test set in order to examine the source of the problems occurring in the `30s` classification.

As with gender, the age results also significantly exceed the baseline of 43%. Like it can be assumed, by taking also other features into account, a higher accuracy can be achieved (e.g. [3] could reach 77% for age profiling).

3) Gender+Age Results: For this problem, the combinations of gender and age, i.e., six classes, had to be predicted. The baseline coming from the majority class `male-20s` is 25% and could also be surpassed using the *LibLinear* classifier. With relatively large structure fragments resulting from the assignments $p = 4$ and $q = 3$, an accuracy of 39% could be achieved using the rank feature set.

TABLE VI
CONFUSION MATRICES OF THE BEST RESULTS FOR GENDER AND AGE PROFILING.

	classified as [%]	
	female	**male**
female	**30.8**	16.1
male	15.0	**38.1**

(a) Gender

	classified as [%]		
	10s	**20s**	**30s**
10s	**25.3**	11.6	0.4
20s	7.8	**37.5**	0.9
30s	1.7	14.3	**0.5**

(b) Age

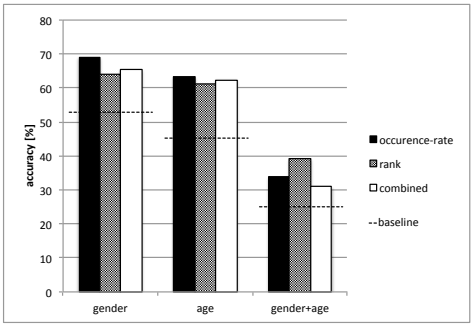

Fig. 3. Summarizing Evaluation Results Using Different Feature Sets.

Due to visibility reasons the details for the individual sub results have been omitted in the table. Nonetheless the experimental data shows that the combined gender and age classification also suffers from predicting the male/female classes of the `30s` age group correctly.

4) Confusion Matrices: A detailed analysis of the best working classifications are shown in the confusion matrices in Table VI. When predicting the gender, the number of false-positives for `male` as well as for `female` are approximately the same. On the other side, the classification of age groups had massive problems concerning the `30s` group, where only 0.5% have been labeled correctly. The majority of this group has been predicted as `20s`, which represents also the majority group of the test set.

As already mentioned, a possible explanation might be the unbalanced test set. This is reinforced by the fact that mostly all false-positives of the `10s` group have also been labeled as `20s`. But what also seems plausible is the hypothesis that the grammar of 13-17 (`10s`) year olds differs significantly from that of 23-27 (`20s`) year olds, where on the other hand the grammatical style of the latter is similar to 33-42 (`30s`) year olds. Intuitively this seems reasonable when looking at sample documents, but future work should investigate further to verify or falsify this assumption.

Summarizing Figure 3 illustrates the evaluation results for all three classification problems using the different feature sets. As can be seen, all baselines could be exceeded.

V. RELATED WORK

Since the advent of the world wide web, offering a huge amount of publicly available documents, the automatic clas-

TABLE V
EVALUATION RESULTS IN PERCENT FOR PROFILING GENDER, AGE AND GENDER+AGE.

			Feature Set									
			Occurrence-Rate			Rank			Combined			
Classifier	**p**	**q**	**female**	**male**	**w. avg**	**female**	**male**	**w. avg**	**female**	**male**	**w. avg**	**Max**
Naive Bayes	4	4	65.9	66.4	66.2	66.4	67.2	**66.8**	65.8	66.1	66.0	66.8
BayesNet	2	4	66.5	67.2	66.8	67.5	68.2	**67.8**	67.4	67.7	67.6	67.8
LibLinear	3	2	61.4	65.7	**63.7**	59.2	64.2	61.8	60.0	64.4	62.4	63.7
LibSVM	2	2	66.5	71.1	**69.0**	61.5	66.3	64.0	62.5	67.8	65.3	**69.0**
kNN	2	2	54.6	62.9	**59.2**	48.6	53.6	51.2	47.7	57.9	53.2	59.2
C4.5 tree	4	2	55.3	61.1	58.4	58.1	62.0	**60.2**	56.4	61.7	59.2	60.2

(a) Results for Gender Prediction.

			Feature Set											
			Occurrence-Rate				Rank				Combined				
Classifier	**p**	**q**	**10s**	**20s**	**30s**	**w. avg**	**10s**	**20s**	**30s**	**w. avg**	**10s**	**20s**	**30s**	**w. avg**	**Max**
Naive Bayes	3	3	39.4	64.6	0.0	52.7	38.2	63.9	0.0	51.9	40.0	65.1	0.0	**53.2**	53.2
BayesNet	2	4	67.6	48.1	39.7	**53.4**	66.7	48.0	39.7	53.1	67.5	47.5	40.3	**53.4**	53.4
LibLinear	2	2	62.3	58.7	24.7	54.8	61.9	55.6	26.1	53.0	63.7	59.1	25.6	**56.6**	56.6
LibSVM	3	2	70.1	68.4	5.0	**63.2**	67.0	66.1	19.9	61.1	68.2	67.5	18.0	62.4	**63.2**
kNN	3	3	54.4	53.3	25.1	48.9	51.2	56.8	27.2	49.8	53.5	56.8	26.5	**50.5**	50.5
C4.5 tree	2	4	52.9	52.3	24.8	48.2	56.1	53.5	26.4	**50.2**	56.9	51.6	24.8	49.3	50.2

(b) Results for Age Prediction.

			Feature Set			
Classifier	**p**	**q**	**Occurrence-Rate**	**Rank**	**Combined**	**Max**
Naive Bayes	4	2	34.8	**35.7**	35.1	35.7
BayesNet	2	4	36.1	**36.4**	36.0	36.4
LibLinear	4	3	33.9	**39.1**	30.9	**39.1**
LibSVM	4	2	**37.2**	34.5	25.8	37.2
kNN	2	2	**32.6**	25.6	25.5	32.6
C4.5 tree	3	3	31.1	**28.2**	27.1	31.1

(c) Results for Combined Gender+Age Prediction.

sification of text has gained more and more interest in the information retrieval field. An often applied concept in order to categorize documents into predefined classes is the utilization of different machine learning algorithms [27], like it is used in this paper. Thereby the problem types are differentiated between single-label and multi-label classification problems, respectively, where the first type assigns only one label for a document (e.g. the gender or age of the author) and the latter type is allowed to assign more labels (e.g. the content type of an article: sports, religion, science, etc.) [28].

Within the single-label text categorization problem the gender and age of the author of a text document has been analyzed frequently. Based on the work of [29] that analyzes the gender of the author and also automatically distinguishes between fiction and non-fiction documents, the web blog corpus used in this approach has been created to classify gender and age based on many style and content features [5]. Here, also blogwords (neologisms) like *'lol'*, *'haha'* or *'ur'* as well as the frequency of hyperlinks have been analyzed. An extension that additionally attempts to classify the language and personality (e.g. neuroticism or extraversion) of a writer has been proposed in [3] by utilizing taxonomies of POS tags combined with other style and content-specific features. Two new feature sets using POS tag patterns are proposed in [30] to enhance current state-of-the art classification approaches.

An interesting approach that also analyzes web blogs is presented in [6]. Besides commonly used features in the field of text categorization the focus has been laid on blog-specific features such as the usage of background colors, emoticons, punctuation marks or fonts. It is shown that the prediction of gender can be enhanced by using these features.

English emails have been profiled into ten classes including gender, age, geographic origin or level of education as well as into five psychological traits in [31]. The authors use several character-level, lexical and structural features and report a similar accuracy for gender classification as the outcome presented in this paper, but show a worse result for age classification (note that emails are typically significantly shorter than blogs).

With the recent rise of social media networks, also content such as chat lines, Facebook postings or tweets have been analyzed and automatically profiled. It is shown (e.g. in [4] or [32]) that a well-defined set of style and content features can be used to expose meta information of chat logs, also in other languages such as Spanish. Nevertheless, the authors in [33] show that the application of common text categorization techniques using natural language processing is challenging - but possible - when facing highly limited data sets. It is demonstrated that even for text samples containing only approximately 12 tokens, the classification of gender and age is feasible.

A related problem in the field of forensic linguistics has recently been investigated in a scientific workshop [11]: Given

 ISBN: 978-1-61208-364-3

the task to automatically expose sexual predators from chat logs, several approaches showed promising results.

The analysis of grammar trees with pq-grams has also been used in previous work, where it has been shown that the grammar of authors is also a feasible criteria to intrinsically expose plagiarism [12] and to correctly attribute authors to unlabeled text documents [13].

VI. Conclusion and Future Work

In this paper, a novel feature that can be used to automatically profile the author of a text document is presented. Based on full grammar trees, it utilizes substructures of these trees by using pq-grams. State-of-the-art machine learning algorithms are finally applied on pq-gram profiles to learn and predict the gender and age of the originator. An extensive evaluation using a state-of-the art test set shows that pq-grams can be used as significant features in text classification, whereby gender and age can be predicted with an accuracy of 69% and 63%, respectively. With respect to the fact that the experiment in this paper solely uses the presented feature, the results are promising.

Evaluation results showed that the approach has problems predicting the `30s` age group. Although hypothesis explaining the problem have been stated, they should be verified or falsified in detail by utilizing a different test set.

In order to build a reliable text classification approach, the grammar feature should be combined with other commonly used style and content feature sets in future work. Besides the utilization of common lexical, syntactic or complexity features, the usage of vocabulary or neologisms should be considered, especially when analyzing online content. Moreover it should be investigated whether the proposed feature is also applicable to shorter text samples such as chat logs or even single-line Twitter postings.

Research should finally also examine whether pq-gram profiles are also exploitable to other languages, especially as syntactically more complex languages like German or French may lead to even better results due to the higher amount of grammar rules available.

References

[1] "Statistic Brain Research Institute," http://www.statisticbrain.com/social-networking-statistics, visited February 2014.

[2] E. Stamatatos, "A Survey of Modern Authorship Attribution Methods," *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 3, pp. 538–556, Mar. 2009.

[3] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, "Automatically Profiling the Author of an Anonymous Text," *Commun. ACM*, vol. 52, no. 2, pp. 119–123, Feb. 2009.

[4] L. Flekova and I. Gurevych, "Can We Hide in the Web? Large Scale Simultaneous Age and Gender Author Profiling in Social Media," *Notebook Papers of CLEF 13 Labs and Workshops*, 2006.

[5] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker, "Effects of Age and Gender on Blogging," in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006, pp. 199–205.

[6] X. Yan and L. Yan, "Gender Classification of Weblog Authors," in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006, pp. 228–230.

[7] J. Noecker, M. Ryan, and P. Juola, "Psychological Profiling Through Textual Analysis," *Literary and Linguistic Computing*, 2013.

[8] J. Gibbons, *Forensic Linguistics: An Introduction to Language in the Justice System*. Blackwell Pub., 2003.

[9] S. Nirkhi and R. Dharaskar, "Comparative Study of Authorship Identification Techniques for Cyber Forensics Analysis," *International Journal*, 2013.

[10] E. E. Abdallah, A. E. Abdallah, M. Bsoul, A. F. Otoom, and E. Al-Daoud, "Simplified Features for Email Authorship Identification," *International Journal of Security and Networks*, vol. 8, no. 2, pp. 72–81, 2013.

[11] G. Inches and F. Crestani, "Overview of the International Sexual Predator Identification Competition at PAN-2012," in *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[12] M. Tschuggnall and G. Specht, "Using Grammar-Profiles to Intrinsically Expose Plagiarism in Text Documents," in *NLDB*, 2013, pp. 297–302.

[13] ——, "Countering Plagiarism by Exposing Irregularities in Authors Grammars," in *EISIC, European Intelligence and Security Informatics Conference, Uppsala, Sweden*, 2013, pp. 15–22.

[14] N. Augsten, M. Böhlen, and J. Gamper, "The pq-Gram Distance between Ordered Labeled Trees," *ACM Transactions On Database Systems (TODS)*, vol. 35, no. 1, p. 4, 2010.

[15] S. Helmer, N. Augsten, and M. Böhlen, "Measuring Structural Similarity of Semistructured Data Based on Information-theoretic Approaches," *The VLDB Journal—The International Journal on Very Large Data Bases*, vol. 21, no. 5, pp. 677–702, 2012.

[16] The Apache Software Foundation, "Apache OpenNLP," http://incubator.apache.org/opennlp, visited February 2014.

[17] D. Klein and C. D. Manning, "Accurate Unlexicalized Parsing," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ser. ACL '03, Stroudsburg, PA, USA, 2003, pp. 423–430.

[18] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, "Building a Large Annotated Corpus of English: The Penn Treebank," *Computational Linguistics*, vol. 19, pp. 313–330, Jun. 1993.

[19] E. Stamatatos, "Intrinsic Plagiarism Detection Using Character n-gram Profiles," in *CLEF (Notebook Papers/Labs/Workshop)*, 2009.

[20] M. Hall et al., "The WEKA Data Mining Software: an Update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[21] G. H. John and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers," in *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1995, pp. 338–345.

[22] G. F. Cooper and E. Herskovits, "A Bayesian Method for the Induction of Probabilistic Networks From Data," *Machine learning*, vol. 9, no. 4, pp. 309–347, 1992.

[23] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A Library For Large Linear Classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[24] C.-C. Chang and C.-J. Lin, "LIBSVM: a Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[25] D. Aha and D. Kibler, "Instance-Based Learning Algorithms," *Machine Learning*, vol. 6, pp. 37–66, 1991.

[26] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Series in Machine Learning, 1993.

[27] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[28] G. Tsoumakas and I. Katakis, "Multi-Label Classification: An Overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.

[29] M. Koppel, S. Argamon, and A. R. Shimoni, "Automatically Categorizing Written Texts by Author Gender," *Literary and Linguistic Computing*, vol. 17, no. 4, pp. 401–412, 2002.

[30] A. Mukherjee and B. Liu, "Improving Gender Classification of Blog Authors," in *Proceedings of the 2010 Conference on Empirical Methods in NLP*. Association for Computational Linguistics, 2010, pp. 207–217.

[31] D. Estival, T. Gaustad, S. B. Pham, W. Radford, and B. Hutchinson, "Author Profiling for English Emails," in *Proceedings of the 10th Conference of the Pacific Assocation for Computational Linguistics*, 2007, pp. 263–272.

[32] M. Meina et al., "Ensemble-Based Classification for Author Profiling Using Various Features," *Notebook Papers of CLEF*, 2013.

[33] C. Peersman, W. Daelemans, and L. Van Vaerenbergh, "Predicting Age and Gender in Online Social Networks," in *Proceedings of the 3rd international workshop on Search and mining user-generated contents*. ACM, 2011, pp. 37–44.

 ISBN: 978-1-61208-364-3