On the Potential of Grammar Features for Automated Author Profiling

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 and serve as input features for common machine learning algorithms. An extensive evaluation using a state-of-the-art test set containing several thousand English web blogs investigates on the optimal parameter and classifier configuration. Promising results indicate that the proposed feature can be used as a standalone, significant characteristic to automatically predict the gender and age of authors. Finally, further evaluations incorporating other commonly used word-based features like the number of stop words, the type-token-ratio or different readability indices strengthen the high potential of grammar analysis for automated author profiling.**

***Keywords*—Author Profiling; Text Classification; Grammar Trees; Textual Features; Machine Learning.**

PREFACE

The following article extends previous work on profiling the gender and age of authors [1].

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 [2] 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 hidden intentionally.

In contrast to traditional authorship attribution approaches [3] 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 [4][5][6], but also other demographic information like cultural background or psychological analyses are examined in recent approaches [7][8]. 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*) [9], where, e.g., the number of possible perpetrators can be reduced. Moreover, especially nowadays in the area of cybercrime [10], recent approaches investigate the content of e-mails [11], suicide letters or try to automatically expose sexual predators from chat logs [12].

In this paper, a novel style feature to automatically extract the gender and age of authors of text documents is presented and compared to other common text features. Using results of previous work in the field of intrinsic plagiarism detection [13] and authorship attribution [14], 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 then examined by utilizing modern machine learning approaches like support vector machines, decision trees or Naive Bayes classifiers. Finally, the results gained from pure grammar analysis are complemented with and compared to other commonly used lexical features like the type-token-ratio or different readability indices.

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. In order to put the results into perspective, Section V integrates and evaluates commonly used word-based features. Finally, related work is summarized in Section VI and conclusions including future work are outlined in Section VII.

2015, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

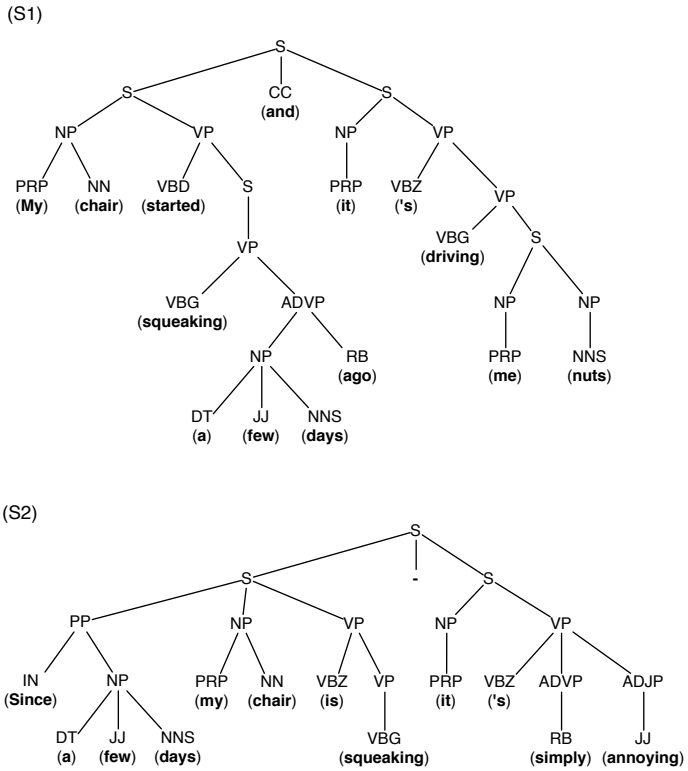

Fig. 1. Grammar Trees of the Semantically Equivalent Sentences (S1) and (S2).

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 [15][16]. 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.

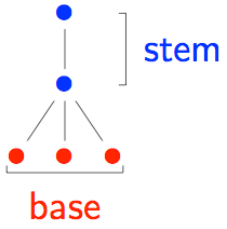

Fig. 2. Structure of a pq-gram Consisting of Stem $p = 2$ and Base $q = 3$.

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. Previous work (e.g., [13]) and 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* [17]. Each sentence is then anaylzed by its grammar, i.e., a full syntax tree is calculated using the *Stanford Parser* [18]. 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 [19], 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 five 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 which is used in the evaluation (see Section IV).
3) Finally, the pq-gram profiles including occurrences and ranks are used as features that are applied to common

2015, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE I
EXAMPLE OF THE FIVE 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
NP-NNP-*-*-*	1.44	4
S-VP-*-*-VBD	1.08	5

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.

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 other efficient approaches in information retrieval (e.g., [20]).

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.

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

C. Test Set

The approach has been evaluated extensively using a state-of-the-art test set created by Schler et. al [6], containing thousands of freely accessible English web blogs. For this evaluation, a subset of approximately 12,000 randomly selected blogs have been used[1], whereby for each blog entry the gender as well as the age of the composer is given.

Regarding the latter, the ages are clustered into three distinct groups, as defined by the original test set [6]: 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, 47% for age and 25% for gender+age profiling, respectively.

[1] in the base study of this work [1] only 8,000 blogs have been used, which led to similar, but slightly different evaluation results

2015, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE IV
TEST DATA DISTRIBUTION.

	female	male	sum
10s	18%	19%	37%
20s	22%	25%	47%
30s	7%	9%	16%
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

Aside from 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 [21], the following classifiers have been utilized:

- Naive Bayes classifier [22]
- Bayes Network using the K2 classifier [23]
- Large Linear Classification using LibLinear [24]
- Support vector machines using LIBSVM with nu-SVC classification [25]
- A k-nearest-neighbours classifier (kNN) using $k = 1$ [26]
- A pruned C4.5 decision tree [27]

E. Results

All possible settings, i.e., combinations of assignments of p and q with classifiers, have been evaluated on the test set 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, for all classifiers the best accuracies could be achieved by using small values for p and q, i.e., $p = q = 2$ in most cases. Finally, the best scores except for gender+age profiling result from the using the combined occurrence-rate and rank feature set.

1) Gender Results: The best result using $p = 2$ and $q = 2$ could be achieved with *LibSVM*, leading to an accuracy of about 68%. It utilizes the combined feature set, whereby males could be identified with 69%. Although the prediction rate is a little worse than those of other approaches (e.g., [6] achieves 80% over the full test set using several style and content features), the result is promising as here only the proposed feature is evaluated and the baseline of 53% could be surpassed clearly.

2) Age Results: Using an identical setting, the maximum accuracy of about 65% results again from using *LibSVM* and the combined feature set. In general, the accuracy for the prediction of the age groups `10s` and `20s` are very solid, but all classifiers except the Bayes approaches have problems predicting the `30s` group. For example, the best configuration achieved a rate of nearly 72% for `10s` and 69% for `20s`, respectively, but could only predict 18% correctly in the eldest group.

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 ([6]), 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%. By incorporating other commonly used features (Section V exemplary adds some lexical features) it can be assumed that a higher accuracy can be achieved (e.g., [4] 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. By reusing the previous assignments $p = 2$ and $q = 2$, an accuracy of 40% could be achieved using the occurrence-rate feature set. In contrast to the isolated gender and age prediction, the combined feature set never led to the highest accuracies for any classification algorithm.

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 visualization of the confusion matrices is presented in Figure 3, while a detailed analysis of the best working classifications is shown in Table VIII in the Appendix. When predicting the gender, the number of false-positives for `females` is slightly higher than for `males`. On the other side, the classification of age groups had massive problems concerning the `30s` group, where only 11.5% have been labeled correctly. The vast majority of this group has been

2015, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

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

			Feature Set									
			Occurrence-Rate			Rank			Combined			
Classifier	**p**	**q**	**male**	**female**	**w. avg**	**male**	**female**	**w. avg**	**male**	**female**	**w. avg**	**Max**
Naive Bayes	2	2	64.7	65.3	65.0	65.3	65.4	**65.3**	65.0	65.3	65.1	65.3
BayesNet	2	2	64.7	65.3	65.0	65.3	65.4	**65.4**	65.0	65.3	65.1	65.4
LibLinear	2	3	69.4	65.6	**67.6**	68.5	64.5	66.7	68.6	64.9	66.9	67.6
LibSVM	2	2	68.2	64.1	66.3	67.2	63.2	65.3	69.4	65.7	**67.7**	**67.7**
kNN	2	2	61.6	56.9	**59.4**	59.9	55.6	57.8	60.6	56.0	58.5	59.4
C4.5	2	3	61.5	57.0	59.4	62.1	57.4	59.9	62.2	58.0	**60.2**	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	2	2	67.4	50.2	40.3	54.3	67.7	50.4	40.9	**54.7**	67.9	49.3	41.0	54.2	54.7
BayesNet	2	2	67.4	50.0	40.2	54.1	67.4	50.0	40.2	54.1	67.8	49.1	41.0	**54.2**	54.2
LibLinear	2	2	68.6	65.8	25.8	**61.6**	67.2	64.0	24.9	60.0	68.2	64.1	29.8	60.6	61.6
LibSVM	2	2	71.2	69.0	16.8	64.4	69.5	67.4	18.0	62.8	71.8	69.1	18.0	**64.7**	**64.7**
kNN	2	3	58.0	57.2	27.5	**52.9**	55.5	56.4	27.0	51.5	56.5	57.3	26.7	52.3	52.9
C4.5	2	2	60.4	57.3	27.9	**53.8**	58.0	54.7	23.9	51.0	60.1	56.1	28.5	53.0	53.8

(b) Results for Age Prediction.

			Feature Set			
Classifier	**p**	**q**	**Occurrence-Rate**	**Rank**	**Combined**	**Max**
Naive Bayes	2	2	35.9	**36.8**	36.0	36.8
BayesNet	2	2	35.9	**36.6**	35.9	36.6
LibLinear	2	2	**40.1**	38.7	39.5	**40.1**
LibSVM	2	4	**39.2**	38.6	39.0	39.2
kNN	2	3	**31.9**	31.3	31.5	31.9
C4.5	3	3	**31.5**	29.7	31.1	31.5

(c) Results for Gender+Age Prediction.

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.

It can be seen clearly that while the classification works reasonably for the gender and age classes `10s` and `20s`, respectively, the approach faces problems attributing the `30s` class. Accordingly, this can be seen in subfigures (b) and (c), where all columns and rows containing the latter class have not been classified correctly.

Summarizing, Figure 4 illustrates the evaluation results for all three classification problems using the different feature sets. As can be seen, all baselines could be exceeded.

V. COMPARISON OF GRAMMAR FEATURES AND COMMON WORD-BASED FEATURES

In order to put the previous results into perspective, they have been compared to the outcomes of commonly used word-based features. The features incorporated are explained in Section V-C, and subsequently Section V-B presents the accuracy gained by using only those features on the same data set. Finally, an evaluation incorporating both the grammar features and the lexcial features is presented in Section V-C.

A. Incorporated Word-Based Features

Because the previously introduced grammar feature operates only on parse trees, the focus for selecting additional features has been laid on any metrics incorporating information on the usage of words, i.e., the vocabulary. In concrete, the following 18 features have been used:

- number of stop words[2], e.g., used in [28], [29], [30]
- number of function and specific words, e.g., used in [31], [32]
  - auxiliary verbs
  - conjunctions
  - determiners
  - prepositions
  - pronouns
  - quantifiers
  - General Service List (GSL)[3] [33]
- vocabulary richness, e.g., used in [34], [35], [36]

[2]gained from http://xpo6.com/list-of-english-stop-words/, visited August 2015

[3]gained from http://www.sequencepublishing.com, visited August 2015

2015, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

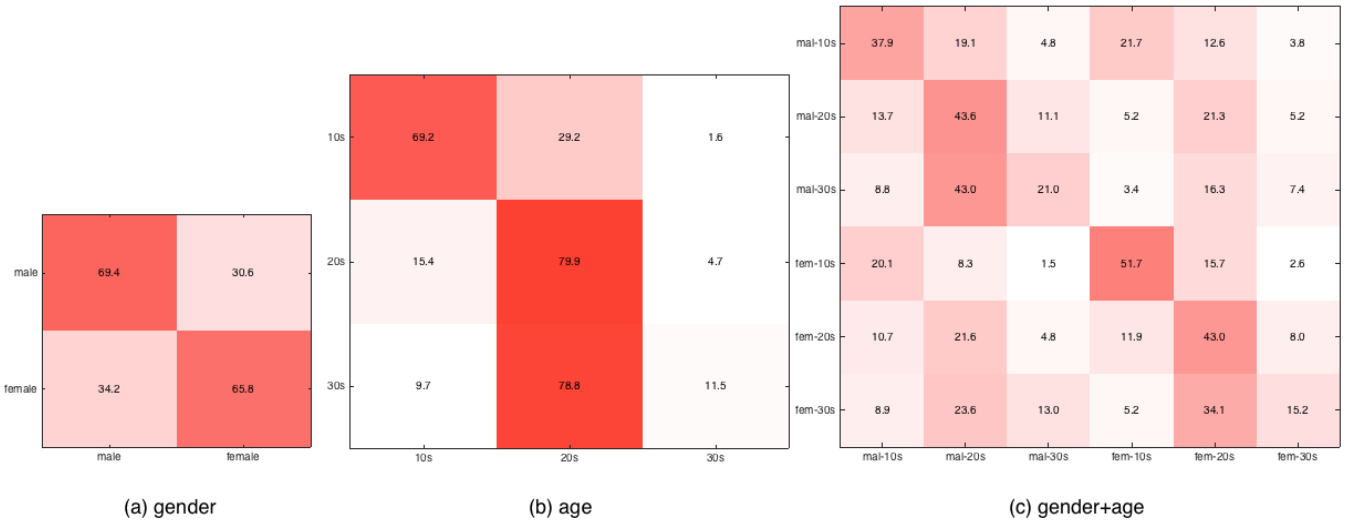

Fig. 3. Confusion Matrices for Profiling Gender, Age and Gender+Age with Grammar.

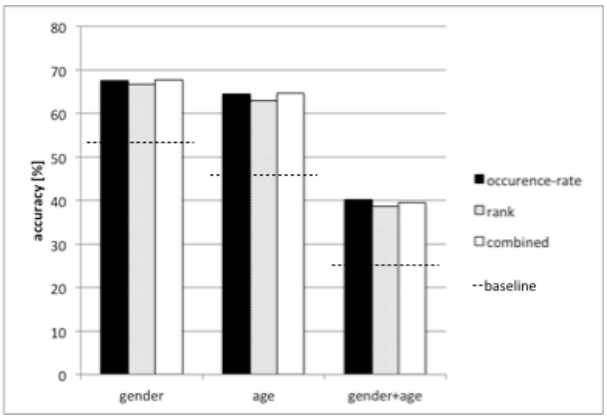

Fig. 4. Summarizing Evaluation Results Using Different Feature Sets.

  - type-token-ratio
  - Honore's H measure [37]
  - hapax legomena
  - hapax dislegomena
- readability metrics, e.g., used in [38], [39]
  - Flesch Reading Ease [40]
  - Flesch-Kincaid Reading Grade [41]
  - SMOG index [42]
  - Automated Readability Index (ARI) [43]
  - Gunning Fog index [44]
  - Coleman-Liau index [45]

B. Feature Evaluation

In order to measure the impact of the selected word-based features when added to the grammar features, the textual features have been evaluated in isolation at first. The result using the same classifiers can be seen in Table VI. For each predicted class, the best accuracy could be gained by the *LibLinear* classifier. Obviously, it can handle the small number of features (18 compared to several thousands) much better than all other algorithms and thus performed significantly better. In case of gender and the combined gender+age classes, nearly identical results as with pure grammar analysis could be achieved. The result for the age class is also comparable, although slightly inferior to the grammar approach. Interestingly and similar to the previously shown grammar results, the correctness of the `30s` group is very low also with the word-based features (and nearly zero in the *LibLinear* case).

C. Combining Grammar and Word-Based Features

In a final evaluation, the grammar features have been enriched with the word-based features. Thereby the latter have been combined with the different grammar feature sets, i.e., the occurrence-rate (`WB + Occurrence-Rate`), rank (`WB + RANK`) and the combined set (`WB + Combined`). The detailed results are presented in Table VII, and a visualized summary of the best results for all examined evaluations using word-based-only, grammar-only and all combined features is depicted in Figures 5-7 in the Appendix. It can be seen that for all three classes the grammar results could be improved as expected - nevertheless, the performance gain is relatively low as discussed later.

Identical to the pure grammar evaluation, the best results for the gender and age classes are produced by the LibSVM framework, and the combined gender+age profiling worked best with LibLinear, respectively. Also, the same pq-gram values, i.e., $p = q = 2$ have been used in all cases, leading to best results when all available word-based features as well as all grammar features have been utilized. While the identification of males and females is relatively balanced, the `30s`-age group has again been detected at a significant lower accuracy.

Although the grammar results could be enhanced by incorporating word-based features, the best accuracies are only slightly improved. At first, this indicates that the proposed

2015, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE VI
EVALUATION RESULTS USING ONLY WORD-BASED FEATURES.

Classifier	male	female	w. avg
NaiveBayes	62.2	61.9	62.0
BayesNet	62.2	61.9	62.0
LibLinear	70.1	64.6	**67.6**
LibSVM	61.4	53.9	58.0
kNN	61.5	57.0	59.4
C4.5	64.8	61.6	63.3

(a) Gender

Classifier	10s	20s	30s	w. avg
NaiveBayes	61.6	54.2	32.2	53.2
BayesNet	61.6	54.2	32.3	53.2
LibLinear	69.3	67.9	0.3	**62.8**
LibSVM	65.4	59.0	30.7	57.0
kNN	61.2	57.6	29.1	54.2
J48	62.3	59.8	19.9	55.1

(b) Age

Classifier	w. avg
NaiveBayes	32.4
BayesNet	32.3
LibLinear	**41.9**
LibSVM	40.2
kNN	31.8
C4.5	32.5

(c) Gender+Age

grammar features are very informative on its own. On the other hand, the reason for the relatively low improvement could be a result of the imbalance of the number of features, i.e., the 18 word-based features compared to the several thousand features resulting from calculating and synchronizing snippets of grammar trees. In order to improve performance, future work should therefore investigate on balancing both types of features, e.g., by applying attribute selection prior to classification. It can be assumed that a lot of the produced grammar features are dispensable, and that a reduction of features leads to better performances of the classification algorithms.

VI. RELATED WORK

The profiling of authors falls under the problem class usually referred to as text categorization [46], whereby an often applied concept is the utilization of different machine learning algorithms based on a previously selected set of features. The main problem types are differentiated between single-label and multi-label classification problems, respectively. Within the single-label text categorization problem the gender and age of the author of a text document has been analyzed frequently. Thereby the first attempts to distinguish between women and men were motivated by sociological studies (e.g., [47]). With the progress in text categorization and authorship attribution, many approaches also tried to automatically detect meta-information like the gender and age of authors, most often by reusing or adapting stylometric features that have been used in other fields. Beside, this core information also many other characteristics have been profiled, including the level of education, the geographical origin or psychological types like extrovertism or neuroticism. In the following, some examples of current profiling approaches are given.

Gender and Age: Probably the best approach that can be directly compared to the results presented in this paper is described in [6]. It leads to slightly better results, but incorporates also many other features that have not been considered in this approach. The fact that grammar-only analyses can lead to nearly similar results is thus promising. The approach is based on the work of [48] that analyzes the gender of the author and also automatically distinguishes between fiction and non-fiction documents, the web blog corpus created by Schler et al. - which is also used in this work - has been created to classify gender and age based on many style and content features [6]. Beside basic features like the frequencies of function words, pronouns, determiners or the average number of words per post, also blogwords (neologisms) like *'lol'*, *'haha'* or *'ur'* as well as the frequency of hyperlinks have been analyzed. With a proposed so-called *Multi-Class Real Winnow* learning algorithm, the gender of the authors of the web blogs could be profiled with an accuracy of 80%, and the age with an an accuracy of 76%, respectively. Similarly to the results described in this work, the authors also report significant problems discriminating mid-twenty year olds from mid-thirty year olds.

An extension to the previous work that additionally attempts to classify the language and personality of a writer has been proposed in [4] by utilizing taxonomies of POS tags combined with other style and content-specific features. By using a Bayesian Multinomial Regression learning algorithm [49] on the same web blog corpus, 76% accuracy on gender and 77% accuracy on age could be gained.

Two new feature sets using POS tag patterns are proposed in [50] to enhance current state-of-the art profiling approaches. In simplified terms, the frequencies of POS-n-grams (where n is not fixed) are collected, rated in terms of significance and used as features if some conditions hold. An evaluation performed also on a (different) blog corpus, the effectiveness of the two new features has been tested. The best result using a support vector machine and incorporating the large number of nearly 24,000 features could enhance the previously described result of Schler et al. by 8%, i.e., reaching 88% on their data set.

In [51], the authors try to automatically expose the gender of writers of Twitter messages by incorporating the huge amount of over 15 million features. The origins of the features are thereby quite simple and can be categorized into character {1-5}-grams and word {1-2}-grams of the actual tweets, complemented by the corresponding n-grams of the user's profile information. As expected, the best result could be gained by using the full name n-grams, reaching an accuracy of 89%.

An interesting approach that also analyzes the gender of web blog authors is presented in [7]. Besides commonly used features in the field of text categorization the focus has been laid on blog-specific features. The approach thereby considers the usage of background colors, emoticons like `;-)` or `:-D`, punctuation marks or fonts. It is shown that the prediction of gender can be enhanced by using these features. Moreover, as a result from the experiment, a list of words which occur in

2015, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

TABLE VII
EVALUATION RESULTS IN PERCENT BY COMBINING GRAMMAR AND WORD-BASED FEATURES

			Feature Set									
			WB + Occurrence-Rate			WB + Rank			WB + Combined			
Classifier	**p**	**q**	**male**	**female**	**w. avg**	**male**	**female**	**w. avg**	**male**	**female**	**w. avg**	**Max**
Naive Bayes	2	2	64.6	65.4	65.0	65.1	65.3	**65.2**	65.0	65.3	65.2	65.2
BayesNet	2	2	64.6	65.4	65.0	65.1	65.4	**65.3**	65.0	65.3	65.2	65.3
LibLinear	2	3	70.3	66.5	**68.5**	69.7	66.0	68.0	69.1	65.6	67.4	68.5
LibSVM	2	2	69.6	65.5	67.7	68.8	64.9	66.9	70.3	66.9	**68.7**	**68.7**
kNN	2	2	61.6	57.3	**59.6**	59.8	55.7	57.9	60.7	56.4	58.6	59.6
C4.5	3	3	61.4	56.8	59.2	61.2	57.1	59.3	61.7	58.0	**60.0**	60.0

(a) Results for Gender Prediction.

			Feature Set											
			WB + Occurrence-Rate				WB + Rank				WB + Combined				
Classifier	**p**	**q**	**10s**	**20s**	**30s**	**w. avg**	**10s**	**20s**	**30s**	**w. avg**	**10s**	**20s**	**30s**	**w. avg**	**Max**
Naive Bayes	2	2	68.0	50.1	40.3	**54.5**	67.9	50.0	40.3	54.5	68.0	49.2	40.7	54.2	54.5
BayesNet	2	2	68.0	50.0	40.2	**54.4**	67.9	49.9	40.4	54.5	68.0	49.0	40.6	54.1	54.5
LibLinear	2	2	69.7	66.0	27.5	**62.2**	69.0	64.8	26.3	61.2	69.3	64.4	30.3	61.2	62.2
LibSVM	2	2	72.1	69.2	16.4	64.8	70.3	68.0	18.1	63.5	72.2	69.3	17.8	**64.9**	**64.9**
kNN	2	3	57.9	57.3	27.7	**52.9**	55.8	56.4	26.5	51.5	56.6	57.5	27.2	52.5	52.9
C4.5	2	2	60.2	57.7	26.9	**53.7**	59.1	53.4	26.7	51.1	60.2	56.4	27.6	53.3	53.7

(b) Results for Age Prediction.

			Feature Set			
Classifier	**p**	**q**	**WB + Occurrence-Rate**	**WB + Rank**	**WB + Combined**	**Max**
Naive Bayes	2	2	35.7	**36.3**	36.0	36.3
BayesNet	2	2	35.7	**36.2**	35.9	36.2
LibLinear	2	2	**41.3**	40.3	40.3	**41.3**
LibSVM	2	3	**39.1**	38.2	37.5	39.1
kNN	2	3	**32.0**	31.3	31.5	32.0
C4.5	2	2	**31.3**	30.8	30.4	31.3

(c) Results for Gender+Age Prediction.

male but rarely/not in female blogs (e.g., *"psst"*, *"income"* or *"wasup"*) and vice versa (e.g., *"muah"*, *"jewelry"* or *"kissme"*) is presented. On the other side, examples of the most gender-discriminant words of the study are: *"peace"*, *"shit"*, *"yo"*, *"man"*, *"fuck"*, *"damn"*.

Other Information: Many studies (e.g., [52]) have analyzed the five psychological traits: neuroticism, extraversion, openness, agreeableness and conscientiousness. Thereby one key problem for verifying such approaches is the lack of test data, i.e., the ground truth is always manually created and thus subjective to some extent. For example, for the evaluation in [53] psychology students have been asked to write a random essay within 20 minutes, whereby the categorization of personality has been made by filling out an additional questionnaire. In another paper [54] web blogs have been psychologically and gender-wise analyzed. Here, 71 bloggers have been asked to submit previously written text, and to additionally fill out a sociobiographic questionnaire as well as an online implementation of a psychological categorization test. By inspecting only eight different POS frequencies like the number of nouns, adjectives or articles, every personality trait of the authors could be predicted with an accuracy of 50-60% in this study.

English emails have been profiled into ten classes including gender, age, geographic origin or level of education as well as into the five psychological traits in [55]. The authors use several character-level, lexical and structural features and report a similar accuracy for gender classification as the outcome presented in this work, but show a worse result for age classification. But it has to be stressed that emails are typically significantly shorter than blogs, and thus the result should not be compared directly.

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 [5]) that a well-defined set of style and content features can be used to expose meta information of chat logs. In a recent workshop [56], participants also gained good results for profiling gender, age and the personality of Twitter users by applying several types of features sets. Nevertheless, the authors in [57] show that the application of common text categorization techniques using natural language processing is challenging - but possible - when facing highly limited data sets.

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 [13], attribute authors to unlabeled text documents [14] and to automatically decompose a multi-author document [58].

2015, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

VII. 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 68% and 65%, respectively.

An extensive evaluation compared the outcome of the proposed grammar features with the performance of a selected set of commonly used word-based metrics like the type-token-ratio or frequencies of stop words. Results show that - in isolation - the pq-gram features perform better than the word-based statistics, and that the best performance can be achieved by combining all features. In order to reduce the large number of features resulting from grammar trees, future work should investigate whether a prior attribute selection algorithm can further improve accuracies.

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, besides the exemplarily selected word-based features used in this work. In addition to the utilization of other common lexical, syntactic or complexity features, detailed metrics 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. The approach could additionally also benefit from applying sentiment analysis.

Finally, research should 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.

Appendix

In this section alternative result views are presented.

TABLE VIII
Confusion Matrices of the Best Results For Gender and Age Profiling.

	classified as [%]	
	male	**female**
male	**69.4**	30.6
female	34.2	**65.8**

(a) Gender

	classified as [%]		
	10s	**20s**	**30s**
10s	**69.2**	29.2	1.6
20s	15.4	**79.9**	4.7
30s	9.7	**78.8**	11.5

(b) Age

	classified as [%]					
	mal-10s	**mal-20s**	**mal-30s**	**fem-10s**	**fem-20s**	**fem-30s**
mal-10s	**37.9**	19.1	4.8	21.7	12.6	3.8
mal-20s	13.7	**43.6**	11.1	5.2	21.3	5.2
mal-30s	8.8	**43.0**	21.0	3.4	16.3	7.4
fem-10s	20.1	8.3	1.5	**51.7**	15.7	2.6
fem-20s	10.7	21.6	4.8	11.9	**43.0**	8.0
fem-30s	8.9	23.6	13.0	5.2	**34.1**	15.2

(c) Gender And Age

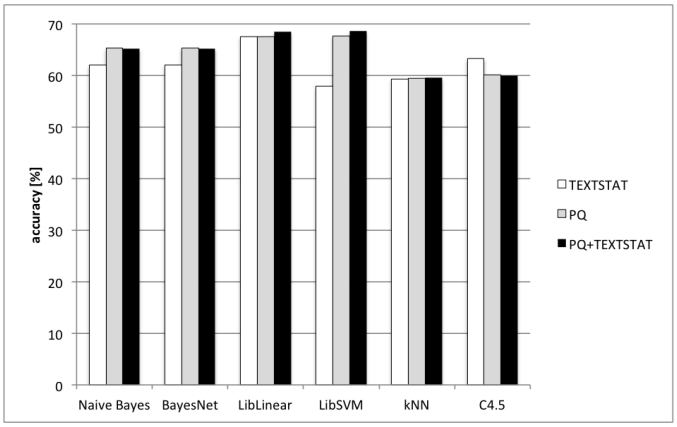

Fig. 5. Best Evaluation Results For Gender Using All Features.

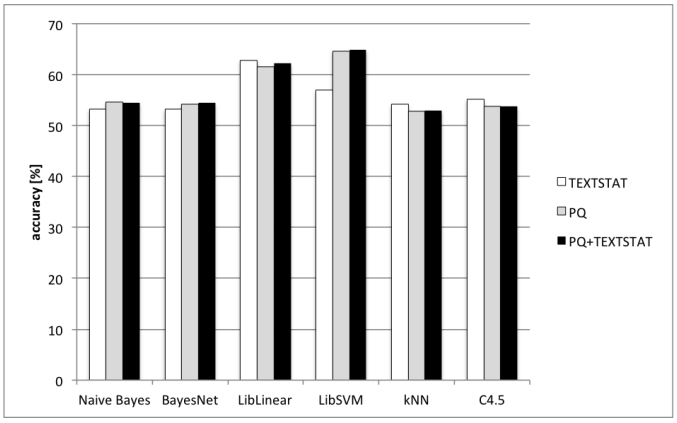

Fig. 6. Best Evaluation Results For Age Using All Features..

2015, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

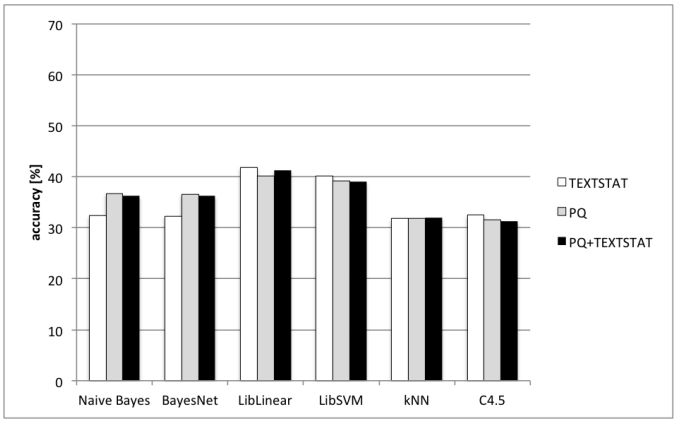

Fig. 7. Best Evaluation Results For Gender+Age Using All Features.

REFERENCES

[1] M. Tschuggnall and G. Specht, "What grammar tells about gender and age of authors," in *Proceedings of the 4th International Conference on Advances in Information Mining and Management (IMMM)*, Paris, France, July 2014, pp. 30–35.

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

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

[4] 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.

[5] 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.

[6] 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.

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

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

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

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

[11] 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.

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

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

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

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

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

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

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

[28] F. Mosteller and D. Wallace, *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, 1964.

[29] D. I. Holmes, "The evolution of stylometry in humanities scholarship," *Literary and Linguistic Computing*, vol. 13, no. 3, pp. 111–117, 1998.

[30] S. Argamon, M. Šarić, and S. S. Stein, "Style mining of electronic messages for multiple authorship discrimination: First results," in *Proceedings of the 9th International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. Washington, DC, USA: ACM, August 2003, pp. 475–480.

[31] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 378–393, 2006.

[32] M. Koppel and J. Schler, "Exploiting stylistic idiosyncrasies for authorship attribution," in *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, vol. 69, Acapulco, Mexico, August 2003, pp. 72–80.

[33] W. Michael, "A general service list of english words," 1953.

[34] C. U. Yule, *The Statistical Study of Literary Vocabulary*. Cambridge University Press, 1943.

[35] C. E. Chaski, "Empirical evaluations of language-based author identification techniques," *Forensic Linguistics*, vol. 8, pp. 1–65, 2001.

[36] F. J. Tweedie and R. H. Baayen, "How variable may a constant be? measures of lexical richness in perspective," *Computers and the Humanities*, vol. 32, no. 5, pp. 323–352, 1998.

[37] A. Honoré, "Some simple measures of richness of vocabulary," *Association for literary and linguistic computing bulletin*, vol. 7, no. 2, pp. 172–177, 1979.

[38] W. Daelemans and V. Hoste, "Stylene: an environment for stylometry and readability research for dutch," 2013.

[39] G. Lynch, "A supervised learning approach towards profiling the preservation of authorial style in literary translations," *Proceedings 25th COLING*, pp. 376–386, 2014.

[40] R. Flesch, "A new readability yardstick." *Journal of applied psychology*, vol. 32, no. 3, p. 221, 1948.

[41] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," DTIC Document, Tech. Rep., 1975.

[42] G. H. McLaughlin, "Smog grading: A new readability formula," *Journal of reading*, vol. 12, no. 8, pp. 639–646, 1969.

[43] R. Senter and E. Smith, "Automated readability index," DTIC Document, Tech. Rep., 1967.

[44] R. Gunning, "{The Technique of Clear Writing}," 1952.

[45] M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring." *Journal of Applied Psychology*, vol. 60, no. 2, p. 283, 1975.

2015, © Copyright by authors, Published under agreement with IARIA - www.iaria.org

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

[47] N. Besnier, "Language and affect," *Annual Review of Anthropology*, vol. 19, no. 1, pp. 419–451, 1990.

[48] 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.

[49] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale bayesian logistic regression for text categorization," *Technometrics*, vol. 49, no. 3, pp. 291–304, 2007.

[50] 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.

[51] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on twitter," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Edinburgh, UK: Association for Computational Linguistics, July 2011, pp. 1301–1309.

[52] S. Argamon, S. Dhawle, M. Koppel, and J. W. Pennebaker, "Lexical predictors of personality type," in *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*, St. Louis, Missouri, USA, June 2005.

[53] J. W. Pennebaker and L. A. King, "Linguistic styles: Language use as an individual difference." *Journal of Personality and Social Psychology*, vol. 77, no. 6, p. 1296, 1999.

[54] S. Nowson, J. Oberlander, and A. J. Gill, "Weblogs, genres and individual differences," in *Proceedings of the 27th Annual Conference of the Cognitive Science Society*. Stresa, Italy: Citeseer, July 2005.

[55] 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.

[56] F. Rangel, P. Rosso, M. Potthast, B. Stein, and W. Daelemans, "Overview of the 3rd author profiling task at pan 2015," in *Proceedings of CLEF*, 2015.

[57] 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.

[58] M. Tschuggnall and G. Specht, "Automatic decomposition of multi-author documents using grammar analysis," in *Proceedings of the 26th GI-Workshop on Grundlagen von Datenbanken*. Bozen, Italy: CEUR-WS, October 2014.

2015, © Copyright by authors, Published under agreement with IARIA - www.iaria.org