Pattern Recognition Letters 105 (2018) 87–95

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

On the role of syntactic dependencies and discourse relations for author and gender identification

Juan Soler-Company[a,*], Leo Wanner[b]

[a] Pompeu Fabra University, Carrer de Roc Boronat 138, Barcelona 08018, Spain
[b] Pompeu Fabra University and ICREA, Carrer de Roc Boronat 138, Barcelona 08018, Spain

ARTICLE INFO

Article history:
Available online 6 December 2017

MSC:
41A05
41A10
65D05
65D17

Keywords:
Author profiling
Author identification
Gender identification
Text classification

ABSTRACT

Author and author gender identification are two major tasks in the context of profiling of authors of written material. Author identification (or, more precisely, "authorship attribution") copes with the assignment of the author, who is to be chosen from a given list of author names, to a piece of written material. Gender identification deals with the prediction of the gender of the author (male vs. female). Both tasks are very relevant to a number of applications, including, e.g., plagiarism and deception detection, document authenticity verification, and blackmailing. State of the art in both fields tends to rely mainly upon lexical and token (sequence) distribution features. But this means to neglect numerous linguistic studies that clearly indicate the high relevance of "deep linguistic", i.e., syntactic and discourse, features to the characterization of the style of an author or a group of authors. Our work on author and gender identification confirms this relevance. We show with two different genres, namely blog posts and literary writings, that the use of deep linguistic features is very effective. It leads to $>78\%$ (in the case of blog posts) and $>91\%$ (in the case of literary writings) of accuracy in author identification and $>89\%$ (blog posts) and $>90\%$ (literary writings) of accuracy in gender identification.

1. Introduction

It is an attractive idea to be able to profile the author of a text, i.e., to identify or characterize them, since there is a large number of potential practical applications, ranging from forensic investigations to marketing studies. In the case of forensics, author profiling can be used, for instance, to detect potential pedophiles in chat rooms, to validate crime confessions, to characterize authors of calls for criminal actions, etc. In the case of marketing studies, it can be used to analyze customer content and feedback data. Even in literary studies, the question of authorship is an issue; see, e.g., [16] for an analysis of Shakespeare's and Fletcher's works from this perspective.

Author profiling deals with the extraction of demographic information on the author of a written text. The research in this field is based on the assumption that authors with similar demographic characteristics express themselves in terms of common or similar patterns because they have been exposed to similar influences. Author profiling can be generic (such as, e.g., identification of the gender of the author) or very targeted (such as, e.g., identification of the author themselves). But no matter what its scope is, one of the crucial tasks that is to be tackled is feature engineering: The process of selecting the features that describe best the style of an individual author or a group of authors and thus help distinguish this author or these authors from others. For instance, gender-specific patterns can be identified when writings of male and female authors are analyzed, profession- or social background-specific patterns can be determined when opinion pieces are examined, etc. In the case of author and author gender identification, some of the commonly used features in the state of the art have been so far word *n*-gram frequency, function word usage, word length distribution, digit frequency, etc. Nearly all of these features draw upon specific tokens and their distribution within a writing. However, these features are not sufficient to capture a writing style in its entireness. Thus, it is well-known from linguistics and philology that syntax and discourse are important aspects of the writing style of an author; cf. [8]. But only a small number of works in the field incorporate, for instance, syntactic features in author identification experiments, and when they do, they usually draw upon information in constituency parse trees (e.g., part-of-speech tag distribution, frequency of passive constructions, frequency of NPs containing a noun in singular, or distribution of punctuation marks); see, e.g., [20,36].

[*] Corresponding author.
E-mail addresses: juan.soler@upf.edu (J. Soler-Company), leo.wanner@upf.edu (L. Wanner).

https://doi.org/10.1016/j.patrec.2017.12.006
0167-8655/© 2017 Elsevier B.V. All rights reserved.

Our goal is to prove that syntactic dependency and discourse features are of high relevance to the task of author profiling. In other words, our aim is not to propose a novel machine learning model or to argue that we discovered features unknown so far (obviously, syntactic and discourse features are looked at, for instance, in parsing respectively discourse analysis applications). Rather, the novelty of our proposal lies in innovative feature engineering: we aim to show that the predominant use of deep linguistic features that have been largely neglected so far for the tasks of author and gender identification significantly outperforms state-of-the-art proposals, which tend to focus on lexical features.

To achieve our goal, we explore the use of primarily deep linguistic (syntactic dependency and discourse) features, which we combine with a small number of more traditional features, for author and author gender identification in the context of two different genres (opinion pieces in online versions of newspapers and literary oeuvres) and show that indeed even rather straightforward off-the-shelf machine learning models achieve very competitive performance and beat, e.g., the best performing system in a recent shared task competition on the topic. Furthermore we show that the use of deep linguistic features allows for a drastic reduction of the number of features (from >1000 to about 200) – which is crucial for applications for which training material is scarce.

The remainder of the paper is structured as follows. Section 2 reviews briefly the related work in the field. Section 3 introduces the set of features we use in our experiments, before in Section 4 the experimental setup is presented. Section 5 outlines the results of the experiments, which are discussed in Section 6. Section 7, finally, draws some conclusions and outlines our future work in the area of author profiling and identification.

2. Related work

An annually organized author profiling shared task drives the research in the area of author profiling. The summary of the outcome of the competition reflects well the state of the art in the area; see, e.g., [34]. An outline of the state of the art can be also found in survey articles such as [33]. In what follows, we focus on the review of the work that is related to ours.

One of the first proposals that implemented data-driven author identification was Mosteller and Wallace [18], who attempted to clarify the authorship of the Federalist Papers, drawing upon function words and using Naïve Bayes classification. Years after, this problem was retaken by Holmes and Forsynth [13], who used high-frequency words and genetic algorithms.

Two demographic characteristics that attracted most of the attention in the field have been so far age and gender; cf., e.g., [1,10,22] for illustration. In [22], age, gender, geographic origin and occupation identification in Vietnamese blogs are addressed. The identification of the age of the authors of blog posts has also been tackled, e.g., in [25]. Schler et al. [26] and Rangel and Rosso [24] deal with age and gender identification of blog authors. Cheng et al. [7], Burger et al. [5], Kucukyilmaz et al. [15], Mukherjee and Liu [19], Soler-Company and Wanner [29–31] focus on the gender of the authors. Argamon and Shimoni [2] seek to identify the gender of the authors and the genre of their writing (fiction vs. nonfiction). In [10], the authors predict gender, age, native language, country of origin and psychometric traits of email authors. This approach is similar to [1], where gender, age, native language and personality identification are performed.

Most approaches draw upon character-based and word-based features such as word or character n-grams; see, e.g., [21,22]. Estival et al. [10] use lexical, character-based and email structure features; in [1], frequent words, function words and part-of-speech tags are used–as also in [26], who furthermore use specific blog features. Rangel and Rosso [24] extract word-based features, punctuation marks and part-of-speech frequencies, and, in addition, analyze the usage of emoticons and polar words. Instead of open class words, function words are also used to distinguish between authors, see [41].

Some works attempt to avoid purely character- or word-based features. For instance, [23] use context-free grammars, Stamatatos et al. [35] focus on style markers and [3] use the frequencies of syntactic rewriting rules to distinguish between authors. A combination of shallow linguistic (trigrams, function word frequencies) and deep linguistic features (context-free grammar production frequencies and features derived from semantic graphs) are used in [11]. Syntactic n-grams[1] are used to distinguish between three literary authors using 39 documents retrieved from the Gutenberg project[2] in [28].

Different kinds of genres have been explored in the field. In the case of [40], the texts are informal blog posts; in [7], emails; in [5]; tweets, and in [15], chat logs. In [12], the authors also work with chat logs in a study on applying author profiling for identification of pedophiles in chat forums.

In our work, we also deal with blogs, as many others do. In our case, these are opinion pieces. As second genre, we use literary material, as, e.g., [16]. As far as features are concerned, we already pointed that we mainly draw upon syntactic dependency structures and discourse structures, which is in contrast to most of the other works in the field. The syntactic and discourse features are complemented by comparatively few lexical features. In total, we use somewhat less than 200 features, while n-gram or bag-of-words models tend to use more than 1,000. This is certainly a further advantage in the light of practical applications.

3. Capturing the style of the writing

As already mentioned, there is a tendency in the areas of gender and author identification to use lexical (or content-dependent) features such as frequent words or character/word n-grams. However, models that exploit these features are not scalable and significantly depend on the domain of the used corpora. On the other hand, deeper syntactic features such as sentence dependency structure, the frequency of specific phrasal or dependency patterns are relevant characteristics of the writing style of an author. For instance, in [9] the authors argue for, e.g., (adjectival) premodification, (PP) postmodification, subordination, coordination, etc. as style-influencing elements. Burstein et al. [6] does so for discourse structure features. We take this into account in that we assign to syntactic and discourse features an important role. More generally, we rely mainly upon linguistic structure rather than on lexis. Apart from capturing the style significantly better than lexis in general, linguistic structure also reflects to a higher degree the unconscious style elements (consider, e.g., the distribution of specific punctuation marks or the syntactic complexity of the statements) and is thus less prone to manipulation.

The set of features that we propose can be divided in six groups: character-based, word-based, sentence-based, dictionary-based, syntactic and discourse features. The first four have been extracted using Python and its Natural Language Toolkit (NLTK). To obtain dependency trees from which the syntactic features are then extracted, Bohnet and Nivre's [4] statistical dependency parser is used. The discourse features are extracted from discourse structure trees obtained using Surdeanu et al.'s [37] discourse parser.

All features have been extracted automatically. The feature extraction software code is written in Python; it is available

[1] Syntactic n-grams are defined as linearized paths of length n in syntactic trees.

[2] https://www.gutenberg.org.

Table 1
Feature summary.

Feature group name	Number of features
Character-based	12
Word-based	11
Sentence-based	3
Dictionary-based	22
Syntactic	127
Discourse	20
Total number of features	195

at https://github.com/joanSolCom/author_profiling_tools. The code provides data structures to manage the input data, to extract every type of features we use and to generate outputs for machine learning toolkits such as Weka and scikit-learn that have been used to perform the classification.

Table 1 gives an overview of the number of features in each group. As can be observed, the total number of features is rather low compared to the related work (which in some cases reaches considerably more than 1000). Also note the high number of deep linguistic features (147 from 195).

In what follows, we introduce the different groups of features.

Character-based features capture the usage of punctuation marks: Commas, periods, parenthesis, exclamation and question marks, hyphens, colons, semi-colons, quotations and other symbols such as the percentage sign, the ampersand, the plus sign, and the dollar sign. They are calculated as the ratio between the frequency of a character in question and the total number of characters. The percentage of upper case characters and the frequency of the use of numbers is also computed.

Word-based features are composed by features that are again calculated as ratios. This time between the mean number of characters per word, the standard deviation in word length, the difference between the longest and shortest word, the vocabulary richness (i.e., the number of different words), the number of different acronyms, the number of stop words and first person pronouns (both singular and plural) and the total number of words in the writing.

Sentence-based features are: The mean number of words per sentence, the standard deviation in sentence length, and the difference between the longest and shortest sentence.

Dictionary-based features are based on lexical items distributed across a number of different dictionaries (each of which contains a specific kind of lexical item): interjections, discourse markers, positive/negative words, abbreviations, curse words and emotion words. As positive and negative word dictionaries, we use the publicly available sentiment analysis lexicons provided by Hu and Liu [14]. The emotion word lexicons belong to a publicly available resource called “Depeche Mood”, which provides dictionaries that contain words that evoke the following emotions: fear, amusement, anger, annoyance, indifference, happiness, inspiration and sadness; for more information, see [32]. The other mentioned dictionaries were specifically compiled for this work.

The dictionary-based features are the ratios between the frequency of each of the lexical items from the above dictionaries in a writing and the total number of words in this writing. For each of the considered emotions, apart from the relative emotion word frequency ratio, two further features are computed: the mean number of words per writing that correspond to this specific emotion and the percentage of the emotion words that belong to this emotion.

Syntactic features account for more than 65% of the total number of features. This group of features can be divided into three subgroups:

Part of speech: This subgroup of syntactic features covers the frequency of each part-of-speech (PoS) tag as the ratio between the count of a given tag in a writing and the total number of tags.[3]

PoS-based features can be very useful for the analysis of the distribution of word categories per text. For instance, a higher usage of adjectives could be seen as an indicator of the expressiveness of a text. The analysis of a text based on this kind of features can help us find patterns that are gender-specific and thus be instrumental for the distinction between writings of male and female authors.

Dependency features: For the compilation of this subgroup of features, we use the syntactic trees provided by the dependency parser.[4]

From the dependency trees, we extract the frequency of each one of the individual dependency relations per sentence, the percentage of modifier relations per tree as well as the frequency of adverbial dependencies (they reveal whether an author tends to elaborate on manner, direction, purpose, etc. of an action). The ratio between the frequency of modal verbs and the total number of verbs and the percentage of verbs that are part of a complex verbal construction (such as *has taken, were thinking*, etc.) are also in this feature group.

Tree-shape features: The goal of this subgroup of features is to capture information about the shape of the dependency trees and thus the complexity of the inner structure of the sentences. We measure their width, depth and ramification factor. The depth is the maximum number of nodes between the root and a leaf node. We consider the width as the maximum number of siblings at a level of the tree. The ramification factor is the mean number of children per level.

We equally apply these measures to subordinate and coordinate clauses, whose existence indicates that a sentence has a certain degree of complexity. When we complement the complexity figures of the clauses with the figures concerning their shape, we measure exactly how complex these subtrees are.

Discourse features are not as numerous as syntactic features, but equally relevant to our proposal.

Surdeanu et al.’s[37] discourse parser extracts from a given writing *Elementary Discourse Units* (EDUs) and links them via discourse relations such that the final output is a discourse tree where the leaves are EDUs and the relations between them are discourse relations. The discourse features capture, on the one hand, the frequency of each one of the discourse relations per EDU (we divide the number of occurrences of each discourse relation by the number of EDUs per text). The full set of discourse relations consists of:

> Joint, Background, Condition, Evaluation, Summary, Cause, Contrast, Topic-comment, Elaboration, Comparison, Topic-change, Textual-organization, Enablement, Attribution, Explanation, Same-unit, and Manner-means.

On the other hand, the discourse features reflect the shape of the discourse trees – as already in the case of syntactic trees, in terms of their depth, width and ramification factor.

4. Experimental setup

We consider author identification as a multi (23 and 16 respectively)-class classification problem and gender identification as a binary classification problem. For the classification experi-

[3] The tag set that was used can be found in http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html.

[4] The dependency relation tags used by Bohnet and Nivre’s [4] parser in our setup are described in [38].

ments, we use Weka's implementation of LibSVM (with a linear kernel), using 10-fold cross-validation.

The first dataset (henceforth referred to as "BlogDataset") that is used for our experiments is composed of 4284 journalistic posts of 23 authors from the blogs of the British newspapers *The Guardian'*, *The Independent*, and *The Daily Mail*. For each author, between 80 and 250 posts are available. Each post of the dataset is tagged with the name of its author and his/her gender.

To prove the effectiveness of our feature set in another scenario, a second dataset (henceforth referred to as "LiteraryDataset") is used. This dataset is a corpus of 1570 chapters of 48 different novels, written by 16 different authors (three books per author). All the authors are American from roughly the same time period. The chosen authors are the following:

> Herman Melville, Louisa May Alcott, Susan Warner, Henry James, Susan Glaspell, Willa Cather, Edith Wharton, Frances Harper, Kate Chopin, William Dean Howells, Mark Twain, John Pendleton Kennedy, Frank Norris, James Fenimore Cooper, Nathaniel Hawthorne and Charlotte Perkins Gilman.

To contrast the performance achieved with our features against the performance achieved with some of the features discussed and used in the state-of-the-art literature, we implemented six different baselines and applied them to both classification problems. These baselines are existing methods, previously applied to author profiling and identification approaches. The first baseline uses normalized frequencies of function words, used by Zhao and Zobel [41]. Since the list of function words that was used in the original is not available, we merged the lists available at http://myweb.tiscali.co.uk/wordscape/museum/funcword.html and http://www.sequencepublishing.com/academic.html. The next two baselines use normalized frequencies of parts of speech and normalized frequencies of stop words[5] respectively. The last baselines are a combination of the above baselines and the majority class classifier. For all baselines, we use the same classifier as used with our model, as well as the same dataset.

5. Results

Two experiments are performed on each dataset. In the first experiment, the goal is to predict which one of the 23 respectively 16 possible authors wrote each one of the texts of the datasets. The second experiment uses the same feature set and corpora using another labeling of the data. In that case, the goal is to predict the gender of the author of the texts (male vs. female). Let us, in what follows, summarize the results of the experiments.

The results of the first experiment on the BlogDataset are shown in Table 2. The performance of our model on each individual feature group, the full set, selected feature combinations as well as the performance of the baselines are presented. The outcome of the first experiment on the LiteraryDataset is reflected in Table 3.

The second experiment on the BlogDataset resulted in figures displayed in Table 4. Table 5 shows the outcome of the second experiment on the LiteraryDataset.

Table 2
Results of the author identification experiment on the BlogDataset.

Features used	Accuracy
Full set	**78.16%**
Character-based (CB)	62.18%
Word-based (WB)	51.72%
Sentence-based (SB)	19.84%
Dictionary-based (DB)	27.84%
Syntactic (Syn)	68.44%
Discourse (Disc)	40.11%
Syn + Disc	70.14%
CB + WB + SB + DB	76.42%
CB + WB + SB + DB + Disc	76.17%
Majority class baseline	5.95%
Function word baseline (FW)	66.99%
Stopword (SW) baseline	65.81%
Parts of speech (PoS) baseline	63.30%
FW + PoS	72.45%
SW + PoS	72.78%

Table 3
Results of the author identification experiment on the LiteraryDataset.

Features used	Accuracy
Full set	91,08%
Character-based (CB)	76.05%
Word-based (WB)	65.79%
Sentence-based (SB)	29.36%
Dictionary-based (DB)	68.53%
Syntactic (Syn)	91.01%
Discourse (Disc)	64.39%
Syn + Disc	92.99%
CB + WB + SB + DB	93.69%
CB + WB + SB + DB + Disc	**95,03%**
Majority class baseline	12.42%
Function word baseline (FW)	86.36%
Stopword (SW) baseline	85.85%
Parts of speech (PoS) baseline	81.01%
FW + PoS	86.96%
SW + PoS	87.29%

Table 4
Results of the gender identification experiment on the BlogDataset.

Features used	Accuracy
Full Set	**89.97%**
Character-based (CB)	87.91%
Word-based (WB)	81.18%
Sentence-based (SB)	65.01%
Dictionary-based (DB)	71.45%
Syntactic (Syn)	85.17%
Discourse (Disc)	75.34%
Syn + Disc	85.92%
CB + WB + SB + DB	88.09%
CB + WB + SB + DB + Disc	89.22%
Majority class baseline	64.11%
Function word baseline (FW)	81.72%
Stopword (SW) baseline	81.46%
Parts of speech (PoS) baseline	81.53%
FW + PoS	82.67%
SW + PoS	82.88%

6. Discussion

In what follows, we discuss the results displayed in the tables above in both the author identification and gender identification experiments.

In the author identification experiment, Table 2 shows that the classifier trained on our features outperforms the baselines when applied to the BlogDataset. The obvious explanation for this is that the accuracy of an approach that uses function words (as some baselines do) depends heavily on the choice of the words in the precompiled list, although targeted selection of function and stop words is a very powerful means to characterize some of the stylistic choices of the authors and a challenging baseline. Adding part-of-speech information improves the baselines' accuracy, leading to a better accuracy than achieved by some of our feature combi-

[5] The list of stop words that is used is available in Python's Natural Language Toolkit (NLTK).

Table 5
Results of the gender identification experiment on the LiteraryDataset.

Features used	Accuracy
Full set	90,71%
Character-based (CB)	81.02%
Word-based (WB)	78.79%
Sentence-based (SB)	73.88%
Dictionary-based (DB)	84.39%
Syntactic (Syn)	90.76%
Discourse (Disc)	75.22%
Syn + Disc	91.46%
CB + WB + SB + DB	90.95%
CB + WB + SB + DB + Disc	**91.78%**
Majority class baseline	52.22%
Function word baseline (FW)	52.73%
Stopword (SW) baseline	52.73%
Parts of speech (PoS) baseline	74.84%
FW + PoS	76.81%
SW + PoS	76.87%

nations. But, in any case, our full feature set outperforms them, which shows that our feature choice is effective.

In the case of author identification applied to the LiteraryDataset (cf. Table 3), the model on the full feature set achieves a very high accuracy of 91,08%, which outperforms the baselines (which in this case are also very challenging). Syntactic features perform similarly to the full set of features; when discourse features are added to them, the model reaches 92.99%, i.e., higher than the full set. The synergy between discourse features and the other types of features is clear: even though with 64.39%the discourse features do not perform well on their own, they improve the accuracy of each feature combination they are part of. The best performing feature combination omits the syntactic features and achieves a very high value of 95.03% accuracy. This means that even the high discrimination potential of syntactic features can be compensated by an appropriate combination features.

To further compare the performance of our features with the performance of the function word baseline in the author identification experiment using the BlogDataset, cf. the confusion matrices in Fig. 1. The confusion matrices are very illustrative in that they show where the classifier erred and what the cause for it was.

The function word list approach works reasonably well in cases in which 200 or more texts from one author are available for training. With authors that have less than 100 texts, the results are much worse. This behavior can be observed in several cases. For example, in the case of the class “r”, our model predicts correctly 53 instances of the class, while the baseline approach predicts only 7 of them correctly. The situation is similar with the classes “w”, “o”, “s” and “e”. That is, features that incorporate syntactic phenomena lead to a more accurate author identification.

The use of function words is partly also stylistically motivated. But partly their use is purely grammatical (as, e.g., in the case of governed prepositions). Therefore, a larger training dataset is necessary to adequately cover the stylistic use of function words.

The confusion matrices in Fig. 2 reflect cases where the classifier predicts wrongly using our model as well as using the function word baseline on the LiteraryDataset. According to these confusion matrices, the authors that have smaller amounts of instances are again harder to predict for the function word baseline, compared to our model. This phenomenon can be observed clearly in the case of the writings of Charlotte Perkins Gilman, where 7 instances were correctly predicted by the baseline, compared to the 32 of our system, and in the case of the writings of Frances Harper (24 vs. 64).

One of the first confusions that was also studied in literary circles is the reciprocal confusion between some of the writings by Willa Cather and Kate Chopin. Both authors were contemporary and had interesting exchanges: Cather wrote an essay to criticize publicly Chopin’s novel “The Awakening” and even published “O Pioneers” as a response. Both books shared many similarities and, in general, both authors wrote about sensitive intelligent women who want to be independent (and in many cases fail to do so). Both authors have female characters that try to push female social boundaries and in some cases try to make a living as a man would. It is rather obvious that both authors had an influence on each other, which could explain why they are confused in our confusion matrix.[6]

Nathaniel Hawthorne is confused with Herman Melville by our model. These two authors were directly in contact; several references to each other are found in their writings (Moby Dick was directly dedicated to Hawthorne) and shared homosexual undertones in their writings. Moreover, both authors exchanged many letters in which their affection is clearly shown (some could be interpreted directly as love letters). So the influence between these two authors is also clear.[7]

John Pendleton Kennedy is confused with James Fenimore Cooper. In this case, a fact that could influence this confusion is the friendship between both and the time spent in the U.S. army by the two authors. Kennedy is known for his contributions to a genre made popular by Cooper, namely historical romances in the early American days. These facts could have influenced the writing style of Kennedy and make it similar to Cooper’s in some cases, which would explain this confusion.[8]

As far as gender identification is concerned, we can see in Table 4 that our model predicts correctly the gender of the authors in 89,97% of the cases. This outperforms every presented baseline. It is very interesting to observe that in this case, the syntactic features by themselves are also able to outperform every baseline, showing that the syntactic structure is very stylistic, and that clear patterns exist that are gender-specific. Character-based features achieve 87.91% of accuracy, which is very close to the performance of the full set. Comma, semi-colon and period usage has been proven in the past to be very stylistic; the performance of this feature group indicates that it is also true in this work.

Table 5 shows that the system’s performance on the prediction of the gender of the authors of literary texts is also very high. In this case, it outperforms every baseline by a very large margin. The use of function and stop words is ineffective for distinguishing between genders in this case. When part-of-speech information is used by the baselines, their performance improves drastically, showing that in a gender identification experiment, the word category usage is a distinctive characteristic of the writings of male and female authors. The table also shows that syntactic features are very distinctive, outperforming the full set of features and improving their performance when combined with discourse features.

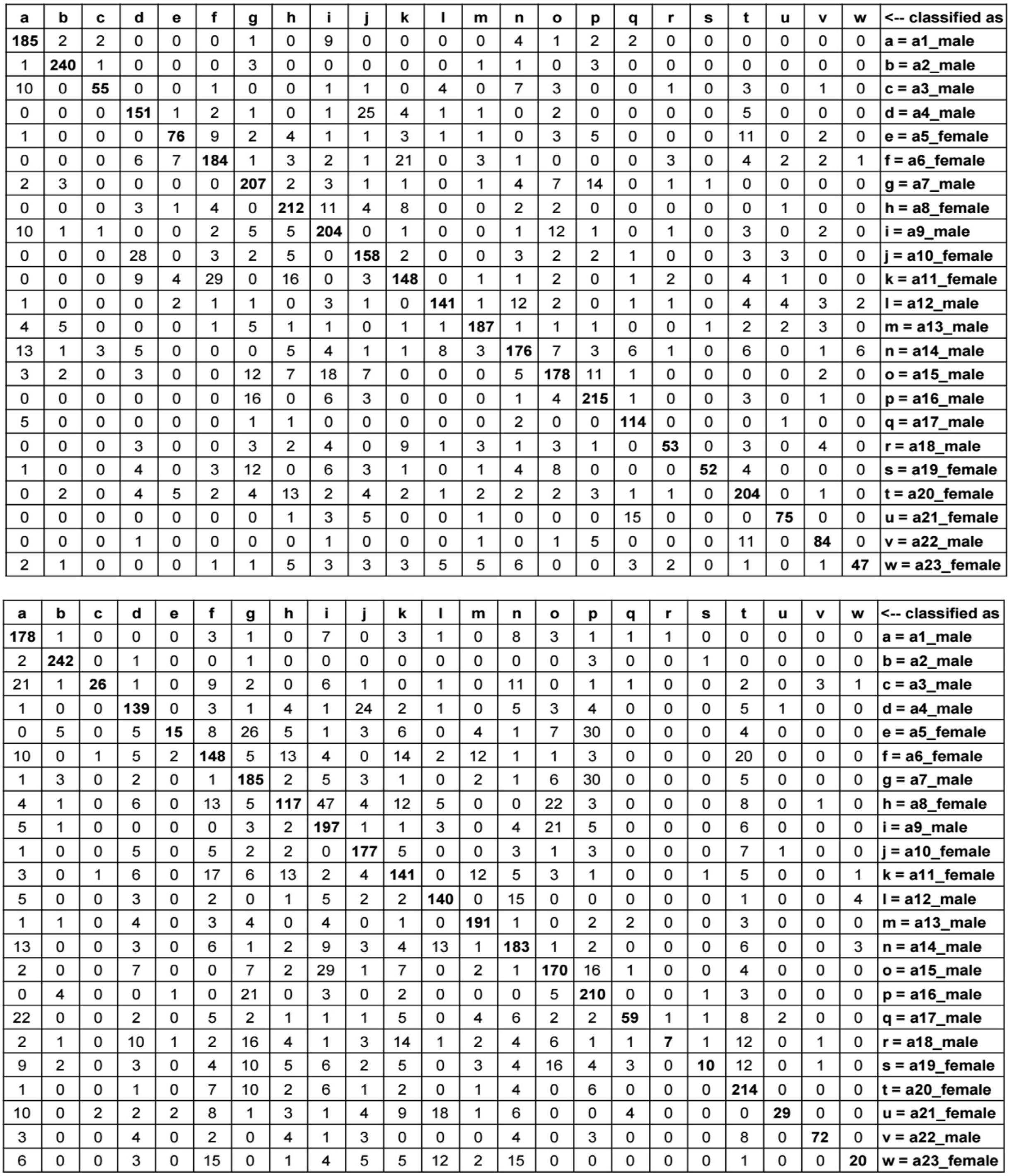
After successfully analyzing the results of both gender and author identification in literary and blog texts, a résumé on the relevance of specific features in each experiment and dataset needs to be drawn. To do so, we computed the information gain of each feature in each one of the presented scenarios. Table 6 shows the most relevant features per experiment.[9]

[6] More information about this confusion can be found in http://realismandnaturalism.blogspot.com.es/2011/10/kate-chopin.html and [27].

[7] For more information on the topic, see http://www.hawthorneinsalem.org/ScholarsForum/MMD2461.html, https://www.theguardian.com/books/2011/jan/30/herman-melville-mark-twain-parini and http://rictornorton.co.uk/melvill2.htm.

[8] More information on the topic can be found in http://docsouth.unc.edu/southlit/kennedy/bio.html and http://www.knowsouthernhistory.net/Culture/Literature/john_pendleton_kennedy.htm.

[9] To facilitate the understanding of the features that are presented, some clarifications must be made. The upper-cased feature names are either part-of-speech tags or syntactic dependencies and their specific meaning is the following: ‘NNP’



a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	<-- classified as
185	2	2	0	0	0	1	0	9	0	0	0	0	4	1	2	2	0	0	0	0	0	0	a = a1_male
1	**240**	1	0	0	0	3	0	0	0	0	0	1	1	0	3	0	0	0	0	0	0	0	b = a2_male
10	0	**55**	0	0	1	0	0	1	1	0	4	0	7	3	0	0	1	0	3	0	1	0	c = a3_male
0	0	0	**151**	1	2	1	0	1	25	4	1	1	0	2	0	0	0	0	5	0	0	0	d = a4_male
1	0	0	0	**76**	9	2	4	1	1	3	1	1	0	3	5	0	0	0	11	0	2	0	e = a5_female
0	0	0	6	7	**184**	1	3	2	1	21	0	3	1	0	0	0	3	0	4	2	2	1	f = a6_female
2	3	0	0	0	0	**207**	2	3	1	1	0	1	4	7	14	0	1	1	0	0	0	0	g = a7_male
0	0	0	3	1	4	0	**212**	11	4	8	0	0	2	2	0	0	0	0	0	1	0	0	h = a8_female
10	1	1	0	0	2	5	5	**204**	0	1	0	0	1	12	1	0	1	0	3	0	2	0	i = a9_male
0	0	0	28	0	3	2	5	0	**158**	2	0	0	3	2	2	1	0	0	3	3	0	0	j = a10_female
0	0	0	9	4	29	0	16	0	3	**148**	0	1	1	2	0	1	2	0	4	1	0	0	k = a11_female
1	0	0	0	2	1	1	0	3	1	0	**141**	1	12	2	0	1	1	0	4	4	3	2	l = a12_male
4	5	0	0	0	1	5	1	1	0	1	1	**187**	1	1	1	0	0	1	2	2	3	0	m = a13_male
13	1	3	5	0	0	0	5	4	1	1	8	3	**176**	7	3	6	1	0	6	0	1	6	n = a14_male
3	2	0	3	0	0	12	7	18	7	0	0	0	5	**178**	11	1	0	0	0	0	2	0	o = a15_male
0	0	0	0	0	0	16	0	6	3	0	0	0	1	4	**215**	1	0	0	3	0	1	0	p = a16_male
5	0	0	0	0	0	1	1	0	0	0	0	0	2	0	0	**114**	0	0	0	1	0	0	q = a17_male
0	0	0	3	0	0	3	2	4	0	9	1	3	1	3	1	0	**53**	0	3	0	4	0	r = a18_male
1	0	0	4	0	3	12	0	6	3	1	0	1	4	8	0	0	0	**52**	4	0	0	0	s = a19_female
0	2	0	4	5	2	4	13	2	4	2	1	2	2	2	3	1	1	0	**204**	0	1	0	t = a20_female
0	0	0	0	0	0	0	1	3	5	0	0	1	0	0	0	15	0	0	0	**75**	0	0	u = a21_female
0	0	0	1	0	0	0	0	1	0	0	0	1	0	1	5	0	0	0	11	0	**84**	0	v = a22_male
2	1	0	0	0	1	1	5	3	3	3	5	5	6	0	0	3	2	0	1	0	1	**47**	w = a23_female

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	<-- classified as
178	1	0	0	0	3	1	0	7	0	3	1	0	8	3	1	1	1	0	0	0	0	0	a = a1_male
2	**242**	0	1	0	0	1	0	0	0	0	0	0	0	0	3	0	0	1	0	0	0	0	b = a2_male
21	1	**26**	1	0	9	2	0	6	1	0	1	0	11	0	1	1	0	0	2	0	3	1	c = a3_male
1	0	0	**139**	0	3	1	4	1	24	2	1	0	5	3	4	0	0	0	5	1	0	0	d = a4_male
0	5	0	5	**15**	8	26	5	1	3	6	0	4	1	7	30	0	0	0	4	0	0	0	e = a5_female
10	0	1	5	2	**148**	5	13	4	0	14	2	12	1	1	3	0	0	0	20	0	0	0	f = a6_female
1	3	0	2	0	1	**185**	2	5	3	1	0	2	1	6	30	0	0	0	5	0	0	0	g = a7_male
4	1	0	6	0	13	5	**117**	47	4	12	5	0	0	22	3	0	0	0	8	0	1	0	h = a8_female
5	1	0	0	0	0	3	2	**197**	1	1	3	0	4	21	5	0	0	0	6	0	0	0	i = a9_male
1	0	0	5	0	5	2	2	0	**177**	5	0	0	3	1	3	0	0	0	7	1	0	0	j = a10_female
3	0	1	6	0	17	6	13	2	4	**141**	0	12	5	3	1	0	0	1	5	0	0	1	k = a11_female
5	0	0	3	0	2	0	1	5	2	2	**140**	0	15	0	0	0	0	0	1	0	0	4	l = a12_male
1	1	0	4	0	3	4	0	4	0	1	0	**191**	1	0	2	2	0	0	3	0	0	0	m = a13_male
13	0	0	3	0	6	1	2	9	3	4	13	1	**183**	1	2	0	0	0	6	0	0	3	n = a14_male
2	0	0	7	0	0	7	2	29	1	7	0	2	1	**170**	16	1	0	0	4	0	0	0	o = a15_male
0	4	0	0	1	0	21	0	3	0	2	0	0	0	5	**210**	0	0	1	3	0	0	0	p = a16_male
22	0	0	2	0	5	2	1	1	1	5	0	4	6	2	2	**59**	1	1	8	2	0	0	q = a17_male
2	1	0	10	1	2	16	4	1	3	14	1	2	4	6	1	1	**7**	1	12	0	1	0	r = a18_male
9	2	0	3	0	4	10	5	6	2	5	0	3	4	16	4	3	0	**10**	12	0	1	0	s = a19_female
1	0	0	1	0	7	10	2	6	1	2	0	1	4	0	6	0	0	0	**214**	0	0	0	t = a20_female
10	0	2	2	2	8	1	3	1	4	9	18	1	6	0	0	4	0	0	0	**29**	0	0	u = a21_female
3	0	0	4	0	2	0	4	1	3	0	0	0	4	0	3	0	0	0	8	0	**72**	0	v = a22_male
6	0	0	3	0	15	0	1	4	5	5	12	2	15	0	0	0	0	0	1	0	0	**20**	w = a23_female

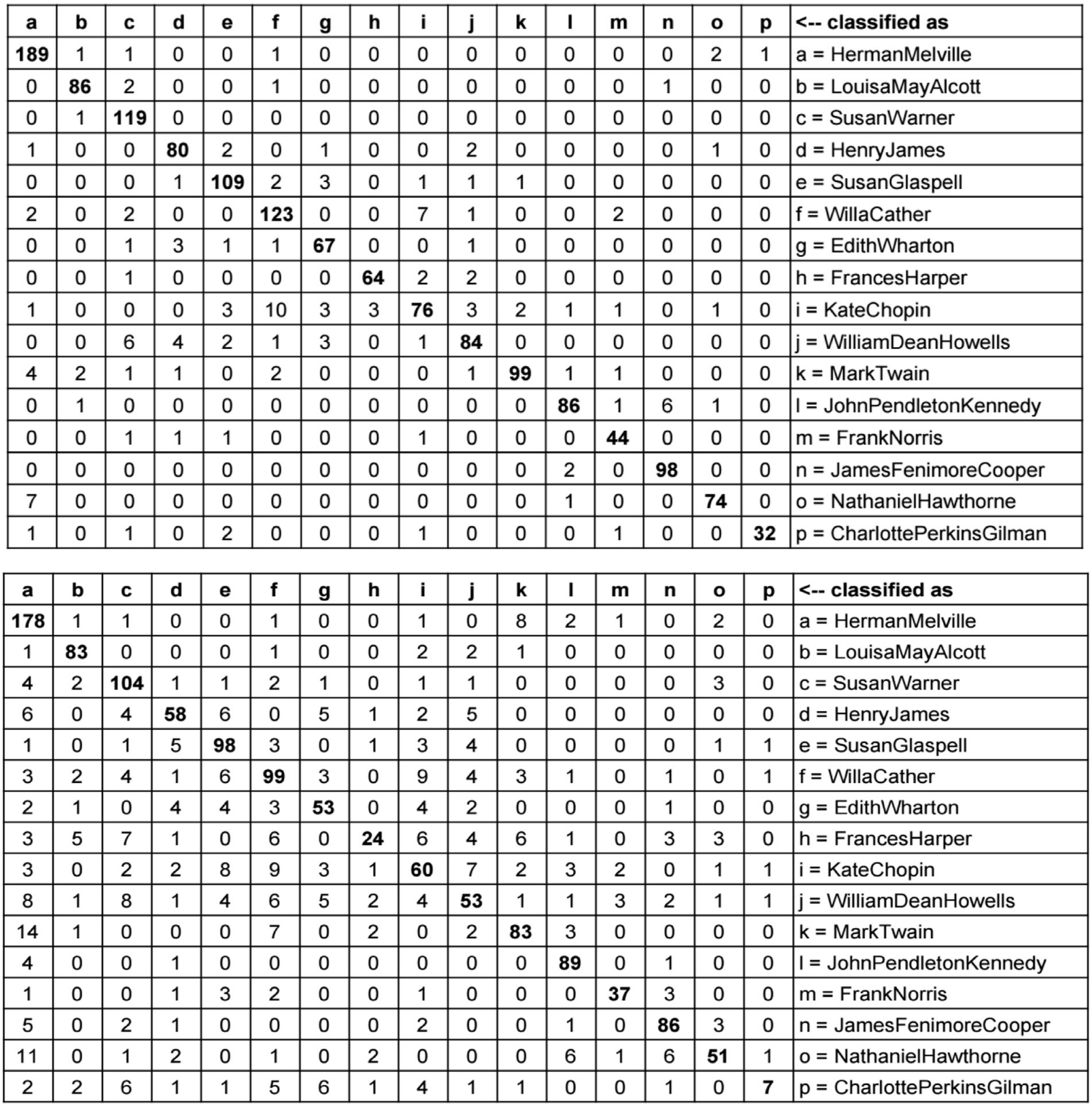
Fig. 1. Confusion matrices of our model (top matrix) with the FWords baseline (bottom matrix) on the BlogDataset.

It can be observed that some of the features are relevant in most experiments (as, e.g., comma and quotation usage, the mean number of characters per word, usage of quotations or the vocabulary richness), while others are specific to some experiments (as, e.g., the usage of negative and emotion words). Negative and emotion words are relevant only in the gender identification experiments, which could be explained by the different perception of emotions and the way they are described by each gender.

Note that there are also patterns which depend on the data that are being used. Thus, first person singular pronoun usage is relevant and distinctive for the BlogDataset (while it is not distinctive for the LiteraryDataset), which indicates that some of the authors express personal opinions in their blog posts, while others are more neutral in their discourse.

The tendency for the use of passive voice is a distinctive trait of authors in the LiteraryDataset, as is the syntactic width, discourse

stands for "singular proper nouns"; 'CD' for "cardinal numbers"; 'POS' for words with a possessive ending; 'NN' for "singular nouns"; 'IN' for "prepositions"; 'WP$' for "possesive wh-pronouns"; 'PRP$' for "possessive pronouns"; 'NMOD' for "nominal modifiers"; 'PMOD' for "preposition modifiers"; 'LOC' for "locative adverbials"; 'APPO' for "appositions"; 'LGS' for "logical subjects of a passive verbs"; 'PRP' for "adverbials of purpose"; and 'PRN' for "parenthetical constructions".



a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	<-- classified as
189	1	1	0	0	1	0	0	0	0	0	0	0	0	2	1	a = HermanMelville
0	**86**	2	0	0	1	0	0	0	0	0	0	0	1	0	0	b = LouisaMayAlcott
0	1	**119**	0	0	0	0	0	0	0	0	0	0	0	0	0	c = SusanWarner
1	0	0	**80**	2	0	1	0	0	2	0	0	0	0	1	0	d = HenryJames
0	0	0	1	**109**	2	3	0	1	1	1	0	0	0	0	0	e = SusanGlaspell
2	0	2	0	0	**123**	0	0	7	1	0	0	2	0	0	0	f = WillaCather
0	0	1	3	1	1	**67**	0	0	1	0	0	0	0	0	0	g = EdithWharton
0	0	1	0	0	0	0	**64**	2	2	0	0	0	0	0	0	h = FrancesHarper
1	0	0	0	3	10	3	3	**76**	3	2	1	1	0	1	0	i = KateChopin
0	0	6	4	2	1	3	0	1	**84**	0	0	0	0	0	0	j = WilliamDeanHowells
4	2	1	1	0	2	0	0	0	1	**99**	1	1	0	0	0	k = MarkTwain
0	1	0	0	0	0	0	0	0	0	0	**86**	1	6	1	0	l = JohnPendletonKennedy
0	0	1	1	1	0	0	0	1	0	0	0	**44**	0	0	0	m = FrankNorris
0	0	0	0	0	0	0	0	0	0	0	2	0	**98**	0	0	n = JamesFenimoreCooper
7	0	0	0	0	0	0	0	0	0	0	1	0	0	**74**	0	o = NathanielHawthorne
1	0	1	0	2	0	0	0	1	0	0	0	1	0	0	**32**	p = CharlottePerkinsGilman

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	<-- classified as
178	1	1	0	0	1	0	0	1	0	8	2	1	0	2	0	a = HermanMelville
1	**83**	0	0	0	1	0	0	2	2	1	0	0	0	0	0	b = LouisaMayAlcott
4	2	**104**	1	1	2	1	0	1	1	0	0	0	0	3	0	c = SusanWarner
6	0	4	**58**	6	0	5	1	2	5	0	0	0	0	0	0	d = HenryJames
1	0	1	5	**98**	3	0	1	3	4	0	0	0	0	1	1	e = SusanGlaspell
3	2	4	1	6	**99**	3	0	9	4	3	1	0	1	0	1	f = WillaCather
2	1	0	4	4	3	**53**	0	4	2	0	0	0	1	0	0	g = EdithWharton
3	5	7	1	0	6	0	**24**	6	4	6	1	0	3	3	0	h = FrancesHarper
3	0	2	2	8	9	3	1	**60**	7	2	3	2	0	1	1	i = KateChopin
8	1	8	1	4	6	5	2	4	**53**	1	1	3	2	1	1	j = WilliamDeanHowells
14	1	0	0	0	7	0	2	0	2	**83**	3	0	0	0	0	k = MarkTwain
4	0	0	1	0	0	0	0	0	0	0	**89**	0	1	0	0	l = JohnPendletonKennedy
1	0	0	1	3	2	0	0	1	0	0	0	**37**	3	0	0	m = FrankNorris
5	0	2	1	0	0	0	0	2	0	0	1	0	**86**	3	0	n = JamesFenimoreCooper
11	0	1	2	0	1	0	2	0	0	0	6	1	6	**51**	1	o = NathanielHawthorne
2	2	6	1	1	5	6	1	4	1	1	0	0	1	0	**7**	p = CharlottePerkinsGilman

Fig. 2. Confusion matrices of our model (top matrix) with the FWords baseline (bottom matrix) on the LiteraryDataset.

ramification factor, width and depth and the usage of modifier relations. All of the mentioned features characterize the writing complexity and inner structure of the texts. Their relevance indicates that deep linguistic features are a very powerful profiling tool to distinguish between writing styles.

The main conclusion that can be drawn from the most relevant features per experiment is that our feature set is versatile enough to achieve high accuracies in different scenarios, although in each of these scenarios different features are most distinctive. The feature set has not been adapted to each specific classification problem and dataset and still facilitates high values of accuracy in every shown case, beating competitive baselines.

To further prove that our feature engineering exercise is effective, we compare our approach with the results of the 2014 PAN shared task competition on author verification (see [34] for more details). The competition provided a dataset that contains pairs of text instances where one instance is known to be written by a specific author and the goal is to determine whether the other instance is also written by the same author. Note that the task of author verification is different from the task of author identification, so some adjustment is needed. To apply our model in this context, we compute the feature values for each pair of known–anonymous instances and subtract the feature values of the known instance from the features of the anonymous one; the feature values are normalized. As a result, a feature difference vector for each pair is computed. The vector is labeled so as to indicate whether both instances were written by the same author or not.

The task performance measure is computed by multiplying the area under the ROC curve (AUC) and the "c@1" score, which is a metric that takes into account unpredicted instances. In our case, the classifier outputs a prediction for each test instance, such that the c@1 score is equivalent to accuracy. In Table 7, the performance of our model, compared to the winner and second ranked of the English literary text section of the shared task (cf. [17] and [39] for details), is shown.

Table 6
20 features with more information gain per scenario.

Author Id. LiteraryDataset	Author Id. BlogDataset	Gender Id. LiteraryDataset	Gender Id. BlogDataset
Semicolons	Disc. RamFactor	Periods	Quotations
Commas	Quotations	Words Per Sent STD	NNP
Chars Per Word STD	Disc. Width	Commas	Upper Cases
Periods	Disc. Depth	Chars Per Word STD	Vocab. Richness
Hyphens	Vocab. Richness	Curse Words	FPers Sing Prons
Words Per Sent STD	Uppers	Words Per Sent	Disc. RamFactor
Vocab. Richness	CD	Words Per Sent Range	Disc. Width
Determiners	Past Verbs	Indifference Words	POS
Disc. Depth	Modifier Relations	Indifference Ratio	Disc. Depth
NN	FPers Sing Prons	NMOD	Commas
NMOD	Words Per Sent	IN	CD
Syn. Width	Colon Usage	PMOD	LOC
Disc. RamFactor	Conjunctions	NN	Percentage
Disc. Width	Chars Per Word	Afraid Ratio	Word Range
Present Verbs	NNP	Inspired Words	Negative Words
Modifier Relations	PMOD	APPO	Chars Per Word
LGS	Commas	Elaboration	WP$
PRP$	Word Range	Upper Cases	Condition
Chars Per Word	NNS	Angry Ratio	Past Verbs
Quotations	PRP	PRN	Superlatives

Table 7
Performance of our model compared to other participants on the 2014 PAN Literary dataset.

Approach	Final score	AUC	c@1
Our model	**0.671**	**0.866**	**0.775**
Modaressi & Gross	0.508	0.711	0.715
Zamani et al.	0.476	0.733	0.650
META-CLASSIFIER	0.472	0.732	0.645
BASELINE	0.202	0.453	0.445

Our model outperforms the task baseline as well as the best performing approach of the shared task, the META-CLASSIFIER (MC), by a large margin. The task baseline is the best-performing language-independent approach of the PAN-2013 shared task. MC is an ensemble of all systems that participated in the task in that it uses for its decision the averaged probability scores of all of them.

7. Conclusions and future work

We have shown that a relatively small set of features, composed mainly of deep linguistic features such as syntactic and discourse features, is very competitive in the author and gender identification tasks. The feature set was applied using standard machine learning techniques in two different scenarios: using blog posts and literary texts. In both cases, the accuracy was very competitive, outperforming all the implemented baselines by a large margin and achieving impressive accuracy values. In addition, we have shown that the use of our features in the context of the 2014 PAN shared task beats the winner. This is promising and could have great impact in the applications oriented towards plagiarism detection, forensic linguistic investigation, literary studies or even marketing studies.

In the future, we plan to introduce new demographic traits, including age, sexual orientation, geographic origin and academic background. The analysis of the specific confusions between authors is likely to lead to very interesting literary insights, such that we also plan to expand our work on the prediction of authors of literary writings by further authors of different origin, time period and genre. From the technical angle, we research unsupervised methods and their potential application to real-world scenarios in forensics.

Acknowledgment

The first author is supported by a PhD grant of the Department of Information and Communication Technologies, Pompeu Fabra University.

References

[1] S. Argamon, M. Koppel, J.W. Pennebaker, J. Schler, Automatically profiling the author of an anonymous text, Commun. ACM 52 (2) (2009) 119.
[2] S. Argamon, A.R. Shimoni, Automatically categorizing written texts by author gender, Lit. Ling. Comput. 17 (2003) 401–412.
[3] H. Baayen, H. van Halteren, F. Tweedie, Outside the cave of shadows: using syntactic annotation to enhance authorship attribution, Lit. Ling. Comput. 11 (3) (1996) 121–132, doi:10.1093/llc/11.3.121.
[4] B. Bohnet, J. Nivre, A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics, 2012, pp. 1455–1465.
[5] J.D. Burger, J. Henderson, G. Kim, G. Zarrella, Discriminating gender on twitter, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2011.
[6] J. Burstein, D. Marcu, K. Knight, Finding the WRITE stuff: automatic identification of discourse structure in student essays, IEEE Intell. Syst. 18 (1) (2003) 32–39.
[7] N.C.N. Cheng, X.C.X. Chen, R. Chandramouli, K.P. Subbalakshmi, Gender identification from E-mails, 2009 IEEE Symposium on Computational Intelligence and Data Mining, 2009.
[8] D. Crystal, D. Davy, Investigating English Style, Longman Group Ltd., London, 1969.
[9] C. DiMarco, G. Hirst, A computational theory of goal-directed style in syntax, Comput. Ling. 19 (3) (1993) 451–499.
[10] D. Estival, T. Gaustad, S.B. Pham, W. Radford, B. Hutchinson, Author profiling for english emails, in: Proceedings of the Australasian Language Technology Workshop, 2007.
[11] M. Gamon, Linguistic correlates of style: authorship classification with deep linguistic analysis features, in: International Conference on Computational Linguistics, 2004.
[12] A. Gupta, P. Kumaraguru, A. Sureka, Characterizing pedophile conversations on the internet using online grooming, arXiv preprint arXiv:1208.4324 (2012).
[13] D.I. Holmes, R.S. Forsynth, The federalist revisited: new directions in authorship attribution, Lit. Ling. Comput. 10 (2) (1995) 111–127.
[14] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD '04, ACM, New York, NY, USA, 2004, pp. 168–177, doi:10.1145/1014052.1014073.
[15] T. Kucukyilmaz, B.B. Cambazoglu, C. Aykanat, F. Can, Chat mining for gender prediction., ADVIS, 2006.
[16] R.A. Matthews, T.V. Merriam, Neural computation in stylometry i: an application to the works of shakespeare and fletcher, Lit. Ling. Comput. 8 (4) (1993) 203–209.
[17] P. Modaresi, P. Gross, A language independent author verifier using fuzzy c-means clustering., in: CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, CEUR-WS.org, Sheffield, UK, 2014, pp. 877–897.

[18] F. Mosteller, D.L. Wallace, Inference in an authorship problem, J. Am. Stat. Assoc. 58 (302) (1963) 275–309.

[19] A. Mukherjee, B. Liu, Improving gender classification of blog authors, in: Proceedings of the International Conference on Empirical Methods in Natural Language Processing (EMNLP), 2012.

[20] A. Narayanan, H. Paskov, N.Z. Gong, J. Bethencourt, E. Stefanov, E.C.R. Shin, D. Song, On the feasibility of internet-scale author identification, in: Proceedings of the 2012 IEEE Symposium on Security and Privacy, in: SP '12, IEEE Computer Society, Washington, DC, USA, 2012, pp. 300–314, doi:10.1109/SP.2012.46.

[21] F. Peng, D. Schuurmans, S. Wang, V. Keselj, Language independent authorship attribution using character level language models, in: Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics, in: EACL '03, 1, Association for Computational Linguistics, Stroudsburg, PA, USA, 2003, pp. 267–274, doi:10.3115/1067807.1067843.

[22] D.D. Pham, G.B. Tran, S.B. Pham, Author profiling for vietnamese blogs, in: 2009 International Conference on Asian Language Processing, 2009.

[23] S. Raghavan, A. Kovashka, R. Mooney, Authorship attribution using probabilistic context-free grammars, in: Proceedings of the ACL 2010 Conference Short Papers, in: ACLShort '10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 38–42.

[24] F. Rangel, P. Rosso, Use of language and author profiling: Identification of gender and age, in: Proceedings of the 10th International Workshop on Natural Language Processing and Cognitive Science, 2013.

[25] S. Rosenthal, K. McKeown, Age prediction in Blogs: a study of style, content, and online behavior in pre- and post-social media generations, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2011.

[26] J. Schler, M. Koppel, S. Argamon, J.W. Pennebaker, Effects of age and gender on blogging., in: Proceedings of the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, AAAI, 2006, pp. 199–205.

[27] C.A. Schneider, Willa cather's" o pioneers!" as a response to kate chopin's" the awakening" (2005).

[28] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, L. Chanona-Hernández, Syntactic n-grams as machine learning features for natural language processing, Expert Syst. Appl. 41 (3) (2014) 853–860. Methods and Applications of Artificial and Computational Intelligence. https://doi.org/10.1016/j.eswa.2013.08.015.

[29] J. Soler-Company, L. Wanner, How to use less features and reach better performance in author gender identification, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 2014.

[30] J. Soler-Company, L. Wanner, Multiple language gender identification for blog posts, in: Proceedings of the 37th Annual Cognitive Science Society Meeting (COGSCI'15), 2015.

[31] J. Soler-Company, L. Wanner, A semi-supervised approach for gender identification, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016.

[32] J. Staiano, M. Guerini, Depechemood: a lexicon for emotion analysis from crowd-annotated news, CoRR, abs/1405.1605, 2014.

[33] E. Stamatatos, A survey of modern authorship attribution methods, J. Am. Soc. for Inf. Sci. Technol. 60 (3) (2009) 538–556, doi:10.1002/asi.21001.

[34] E. Stamatatos, W. Daelemans, B. Verhoeven, M. Potthast, B. Stein, P. Juola, M.A. Sanchez-Perez, A. Barrón-Cedeño, Overview of the author identification task at pan 2014, Analysis 13 (2014) 31.

[35] E. Stamatatos, N. Fakotakis, G. Kokkinakis, Computer-based authorship attribution without lexical measures, Comput. Hum. 35 (2) (2001) 193–214.

[36] E. Stamatatos, G. Kokkinakis, N. Fakotakis, Automatic text categorization in terms of genre and author, Comput. Ling. 26 (4) (2000) 471–495, doi:10.1162/089120100750105920.

[37] M. Surdeanu, T. Hicks, M.A. Valenzuela-Escárcega, Two practical rhetorical structure theory parsers, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT): Software Demonstrations, 2015.

[38] M. Surdeanu, R. Johansson, A. Meyers, L. Màrquez, J. Nivre, The conll-2008 shared task on joing parsing of syntactic and semantic dependencies, in: Proceedings of the Twelfth Conference on Computational Natural Language Learning, in: CoNLL '08, Association for Computational Linguistics, Stroudsburg, PA, USA, 2008, pp. 159–177.

[39] H. Zamani, H.N. Esfahani, P. Babaie, S. Abnar, M. Dehghani, A. Shakery, Authorship identification using dynamic selection of features from probabilistic feature set, in: CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, CEUR-WS.org, Sheffield, UK, 2014, pp. 877–897.

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

[41] Y. Zhao, J. Zobel, Effective and Scalable Authorship Attribution Using Function Words, in: G. Lee, A. Yamada, H. Meng, S. Myaeng (Eds.), Information Retrieval Technology, Lecture Notes in Computer Science, 3689, Springer Berlin Heidelberg, 2005, pp. 174–189, doi:10.1007/11562382_14.