# How to Use Less Features and Reach Better Performance in Author Gender Identification

**Juan Soler Company[1], Leo Wanner[2,1]**
[1]NLP Group, Department of Information and Communication Technologies, Universitat Pompeu Fabra,
[2]Catalan Institute for Research and Advanced Studies (ICREA)
[1]C/ Roc Boronat, 138, 08018 Barcelona, Spain
juan.soler@upf.edu, leo.wanner@upf.edu

**Abstract**

Over the last years, author profiling in general and author gender identification in particular have become a popular research area due to their potential attractive applications that range from forensic investigations to online marketing studies. However, nearly all state-of-the-art works in the area still very much depend on the datasets they were trained and tested on, since they heavily draw on content features, mostly a large number of recurrent words or combinations of words extracted from the training sets. We show that using a small number of features that mainly depend on the structure of the texts we can outperform other approaches that depend mainly on the content of the texts and that use a huge number of features in the process of identifying if the author of a text is a man or a woman. Our system has been tested against a dataset constructed for our work as well as against two datasets that were previously used in other papers.

**Keywords:** Author Profiling, Gender Identification, Machine Learning

## 1. Introduction

Author profiling in general and author gender identification in particular are an increasingly popular research area in corpus-oriented computational linguistics. Their range of potential applications spans from forensic investigations to online commercial client complaint analysis. The basic assumption underlying the research on author profiling is that authors with specific common characteristics express themselves similarly, i.e., have a similar writing style, such that analyzing texts written by various authors one will be able to classify the texts with respect to these characteristics (and thus assign them to author profiles). The characteristics can be age, educational level, geographical or societal origin, native tongue, or gender. In the extreme case, the sought class of authors is a single individual, as in some forensic applications. Most works on author profiling are defined as supervised machine learning (ML) problems, using surface-oriented features: function words, most frequent words, triples and/or pairs of frequently co-occurring words, part of speech (POS) n-grams, punctuation marks, etc.; see, e.g., (Argamon et al., 2009; Koppel et al., 2000; Burger et al., 2011; Schler et al., 2006) . Only a few use also syntactic features; see, e.g., (Cheng et al., 2009). What all of them have in common is that the number of features is rather large. It may easily be higher than 1000 and is seldom lower than 500. As a consequence, very often dimension reduction is applied to minimize the complexity—which, in its turn, has negative consequences for the accuracy and transparency of the outcome.

It is thus desirable to come up with an approach that uses less features and can still compete with state-of-the-art proposals in terms of performance. This goal can be achieved only if more distinctive and more generic features than those commonly used in the literature are exploited. Features of this kind are likely to be rather of a structural than content-oriented nature. In what follows, we present our work on author gender recognition (woman vs. man) that uses a small collection of 83 features in total (with syntactic dependency features constituting the biggest share (namely 67) of them) and shows a very competitive performance due to the structural nature of these features.

For classification, we use WEKA's Bagging variant. The corpus on which we carried out our experiments is a collection of postings of a New York (NY) Times opinion blog. This blog is extremely multi-thematic, with the authors commenting on science, philosophy, ongoing political and economic affairs in the US and worldwide, etc. We can assume that all texts are well-written and in US English. The corpus is balanced; it contains 836 texts written by more than 100 male and 836 texts written by more than 100 female authors.

Our experiments show that combining all of the 83 features we achieve an accuracy of 82.72%, but also that some additional feature engineering pays off: using a subset of features, we achieve an even higher accuracy of 82.83%. The good news is also that already with 14 features we achieve an accuracy of 70.28%. Compared to, e.g., Argamon et al. (2009), who use more than 1.000 features and achieve an accuracy of 76.1%, this is very encouraging.

To validate that the performance of our technique is not biased by the dataset that has been compiled specifically for our experiments, we furthermore ran experiments on an external blog post dataset presented in (Mukherjee and Liu, 2010). The outcome has been equally positive: our technique outperformed Mukherjee and Liu (2010) by 9%.

The remainder of the paper is structured as follows. In the next section, we present the features we use for our experiments. Section 3 describes the experiments and their results. In Section 4, we then explore how our approach performs when not only gender but also age of the author is to be predicted. Section 5, finally, outlines the directions of the future work we plan in this area.

## 2. Features

For our supervised machine learning experiments, we used five different types of features. These features have been derived from an empirical study of a development corpus sample on the assumption that these features are most distinctive for the writing styles of women and men. The combination of features we used attempts to reflect the writing of men and women from the most basic level (usage of characters) to a more global level (sentence structure). The five groups of features in question are:

1. Character-based features
2. Word-based features
3. Sentence-based features
4. Dictionary-based features
5. Syntactic features

**Character-based features** capture the frequency of punctuation marks (comma, full stop, interrogation and exclamation marks), the frequency of upper case characters and the total number of characters per text. The use of these features (e.g., the comma setting or the use of upper case to express salience of a word) is to a major extent motivated by individual stylistic preferences. Therefore, we wanted to assess whether general patterns of character features can be identified and used to differentiate the writing styles of men and women.

**Word-based features** capture the word distribution: the total number of words per text, the number of different words per text, the number of acronyms, the number of stop words, etc. These features are motivated by the assumption that wordiness, richness of vocabulary, or the tendency to use abbreviations may be a factor for gender differentiation.

**Sentence-based features** simply capture the number of sentences in a text and the number of words per sentence. This group tries to capture in a very simple way the structure of the text at a higher level.

**Dictionary-based features** are the percentages of the words in a text from one of the three dictionaries we used. The first two of these dictionaries are polarity dictionaries, i.e., lists of words that are classified as emotionally "positive" or "negative". These dictionaries were used for the first time in (Hu and Liu, 2004) and contain approximately 6800 words classified by their polarity. Our assumption was that women are more expressive and that their emotional involvement during the process of writing is considerably higher than that of men. As a consequence, women's writings should contain more positive and negative words. It turned out to be true in our particular case (i.e., in the corpora we used). Furthermore, we assumed that men tend to tell stories focusing on what happened, while women focus more on how they felt when these stories happened, instead of focusing on the story itself.

The third dictionary contains "patriotic" words such as "US", "Americans", etc. We introduced this dictionary based on the observation that our development corpus contained a considerable number of such words (not necessarily only in blog postings that talked about politics). Our hypothesis here was that men's writings would have higher rates of this kind of words—which also proved right.

**Syntactic features** capture the grammatical dependency functions (subject, direct object, modifier, determiner, etc.) and the average length of the dependencies (the distance between the head and the dependent in words in the linearized sentence) . Using the syntactic information as features helped us determine who builds more complex sentences, men or women, and what kind of dependencies are used more often by each gender. Syntactic features constitute the largest group of features of our approach.

Table 1 displays the number of features of each type we used in our experiments.

| Feature Category | #Features |
|---|---|
| Character-based | 6 |
| Word-based | 5 |
| Sentence-based | 2 |
| Dictionary-based | 3 |
| Syntactic | 67 |

Table 1: Distribution of features across categories

While the first three types of features can be considered as standard features that are used in many state-of-the-art proposals on author profiling, the last two types are rather novel. Thus, although a number of approaches claim to use syntactic features, usually these features are Parts of Speech (PoS) of the words and PoS combinations, i.e., morpho-syntactic categories, rather than syntactic dependencies between words. Also, many works in author profiling and author gender identification use dictionaries to analyze the content of the texts. The novelty here is to use polarity dictionaries to measure the expressiveness of the authors and use this information to distinguish between genders—something which is more often used in Sentiment Analysis. These two groups of features, syntactic features and dictionary-based features, which capture the expressiveness/emotionality of a text and the syntactic stylistic idiosyncrasies, proved to be very effective for gender identification.

Let us furthermore note that from all the features that are used in our work, only 3 (the dictionary-based features) depend on the actual content of the text.

## 3. Experiments

In this section, we outline our experiments, i.e., the experiment setup and the results we obtained when running the experiments.

### 3.1. Experiment Setup

As already mentioned above, we used for gender identification Weka's Bagging classifier, with REPTree (a fast decision tree learning algorithm) as base classifier. For feature extraction, Python and its Natural Language Toolkit (NLTK) were used. The output of the feature extraction is represented as an ARFF file, in which all the texts are represented in terms of multi-dimensional vectors, with each feature as a separate dimension and one of the values of

a feature as instantiation of its dimension. The ARFF file was fed to WEKA for classification. To obtain more reliable performance figures, we used 10-fold cross validation, such that the outcome of the classification does not depend on which specific part of the dataset was used for training and which part for testing.

To explore the relevance of the different types of features both in combination with other features and in isolation, we ran a number of experiments, each of them with a specific feature set; see the first column of Table 2 for the different feature sets that we used in our experiments. These experiments were first run on the test dataset of the NY Times Opinion Blog corpus. Table 2 lists the accuracy figures obtained on this dataset when using the different feature sets.

| **Feature combination** | **#Features** | **Accuracy (%)** |
|---|---|---|
| Sentence-based (S) | 2 | 56.81 |
| Dictionary-based (D) | 3 | 59.75 |
| S + D | 5 | 60.59 |
| Word-based (W) | 5 | 63.63 |
| Character-based (C) | 6 | 64.53 |
| C + S | 8 | 64.71 |
| W + C | 11 | 66.45 |
| C + S + D | 11 | 66.63 |
| C + D | 9 | 66.99 |
| W + D | 8 | 67.46 |
| W + S + D | 10 | 68.18 |
| W + C + S + D | 16 | 69.86 |
| W + C + D | 14 | 70.28 |
| Syntactic (Y) | 67 | 77.03 |
| Y + D | 70 | 77.39 |
| Y + W | 72 | 77.87 |
| Y + S | 69 | 78.35 |
| Y + S + W | 74 | 80.32 |
| Y + C | 73 | 81.16 |
| Y + C + W | 78 | 82.12 |
| Y + C + S | 75 | 82.35 |
| Y + C + S + W + D | 83 | 82.72 |
| Y + C + S + W | 80 | 82.83 |

Table 2: Performance of our approach on the NY Times blog dataset when using different feature sets

Since the NY Times blog dataset was compiled in the scope of our work, we wanted to ensure that our classification is not biased towards this dataset, i.e., that it shows a similar performance on a different, independent dataset. For this purpose, we ran some experiments on the dataset described and used by Mukherjee and Liu (2010), which is an informal blog post dataset. With this new dataset, we obtained an accuracy of 97.08%, in the best case—compared to 88.56% reported in (Mukherjee and Liu, 2010).

Table 3 shows the accuracies of our approach on the Mukherjee and Liu (2010) dataset, with the same differentiation of combinations of features as for the NY Times Blog dataset.

In order to assess the importance of structural features for gender identification, we carried out an additional experiment with a totally different approach to gender identification, namely using features as a "bag of words". The individual posts were thus considered as vectors where each dimension stands for the percentage of the occurrence of a specific common word in them. To obtain the set of common words, we discarded stop words and calculated the *tf*idf* measure for all the remaining words. The 1000, 2000 and 3000 words with higher *tf*idf* values were used for classification. The experiment was run on the NY Times blog dataset; the classifier was again the Bagging implementation of WEKA, with a REPTree classifier as base classifier. The results of this experiment are summarized in Table 4.

| **Feature comb.** | **#Features** | **Accuracy (%)** |
|---|---|---|
| Sentence-based (S) | 2 | 63.74 |
| Dictionary-based (D) | 3 | 74.70 |
| Word-based (W) | 5 | 77.32 |
| S + D | 5 | 84.39 |
| Y + S | 69 | 95.07 |
| Syntactic (Y) | 67 | 95.08 |
| Y + W | 72 | 95.08 |
| Y + S + W | 74 | 95.08 |
| Y + C | 73 | 95.16 |
| Y + C + W | 78 | 95.16 |
| Y + C + S | 75 | 95.16 |
| Y + C + S + W | 80 | 95.16 |
| Y + D | 70 | 95.46 |
| Y + C + S + W + D | 83 | 95.50 |
| W + S + D | 10 | 95.66 |
| W + D | 8 | 95.77 |
| W + C + S + D | 16 | 96.5 |
| W + C + D | 14 | 96.5 |
| C + S | 8 | 96.54 |
| C + S + D | 11 | 96.62 |
| Character-based (C) | 6 | 96.66 |
| C + D | 9 | 96.66 |
| W + C | 11 | 97.08 |

Table 3: Performance of our approach on the Mukherjee and Liu (2010) dataset

| **#Features** | **Accuracy (%)** |
|---|---|
| 1000 | 66.09 |
| 2000 | 72.49 |
| 3000 | 73.80 |

Table 4: Performance of the Bag-of-words approach on the NY Times blog dataset

### 3.2. Discussion of the results

Table 2 shows that using the whole set of features for the NY Times blog dataset, we obtain an accuracy of 82.72%. This is an accuracy that is definitely within the range of the accuracies achieved by the state-of-the-art approaches in this area. However, it is remarkable that this accuracy is achieved using a much smaller number of features than in most of the state-of-the-art approaches. It is also impor-

tant to highlight that using only 14 features, an accuracy of 70.28% is achieved. Table 2 furthermore shows that the use of syntactic dependency features pays off. Using only this group of features, we achieve an accuracy of 77.03%. This gives us a hint that there are important differences in how men and women syntactically structure their sentences.

Table 3 further proves this hypothesis on a different dataset. The usage of only syntactic features gives us an accuracy of 95.08%. It is also remarkable that nearly every combination of features we use achieves a higher accuracy than Mukherjee and Liu (2010)—a fact that tells us that the chosen features are definitely relevant features for gender identification. In this case, the use of only 11 features gives us the best result. Word-based and character-based features were extremely effective in this experiment, giving us an accuracy of 97.08%. In contrast, when only content features are used in a bag-of-words approach, the performance decreases significantly (see Table 4), despite the enormous increase of the number of features. In other words, the use of mere lists of words (as, e.g., Zhang and Zhang (2010) do) implies a significantly more complex feature management and leads to worse performance than the use of context-independent structural sentence features.

## 4. One Step further towards Author Profiling

After proving that our feature set leads to a good performance in two different scenarios of gender identification, we explored whether it can be equally used for other parameters of author profiling such as age prediction. For this purpose, we used another collection of blog posts as dataset. This dataset was compiled and described by Schler et al. (2006). It was also used by Argamon et al. (2009). The dataset is composed of informal blog posts extracted from *blogger.com*. The blogs are tagged by the gender and the age of the author and are thus ideal for our experiments. The ages are grouped into three classes: 1. "teens", which represents the authors whose age ranges from 13 to 17, 2. "twenties", which goes from age 23 to 27, and 3. "thirties", which captures the authors who are older than 30.

In contrast to the NY Times blog posts, these blogs are not well structured and written. They contain many orthographic errors, slang expressions, abbreviations, emoticons, etc.

We performed two different classification runs: one in which the classifier predicted whether the author is a man or a woman, and the other that determined in which of the three age classes the author is situated. The classification runs were performed using a balanced subset of the dataset that was composed of 5955 posts. The classifier that was used in these experiments was SMO, a variant of support vector machines with a radial kernel.

Using the same set of features as used for our initial experiments on gender identification, we obtain the figures outlined in Table 5 (as baseline, we use the accuracy of a random classifier; this is adequate since we use the same number of blogs for each category in both classifications).

Although the performance of gender identification has been in this run considerably lower than the performance we achieved on the NY Times blog, age classification still improves the baseline by 22,96%.

| | **Gender** | **Age** |
|---|---|---|
| **Accuracy** | 68.09% | 55.96% |
| **Baseline** | 50% | 33% |

Table 5: Performance of our approach with the same set of features when classifying by Gender and Age

The analysis of the results further reveals that due to the numerous orthographic and syntactic errors encountered in the dataset, the performance of Bohnet (2010) dependency parser, which we used in our experiments decreased significantly. Since the syntactic features constitute the majority of our feature set, our hypothesis was that the decrease in the performance of the parser was (at least partially) responsible for the lower performance of our gender/age identification.

To tackle this problem, we used a shallow parser that is a simplification of our original dependency parser and that was expected to be more tolerant to faulty texts.

Furthermore, several other features were added to boost the performance in the case of age classification. These features were:

1. number of orthographic errors per word,
2. percentage of discourse markers,
3. frequency of curse words and abbreviations,
4. usage of passive voice,
5. further dictionary based features that measured the usage of words related to school, college, duties and leisure time.

The total number of features that were used in this extended dataset thus grew to 100.

After the new features were introduced, the performance of the classification was as shown in Table 6:

| | **Gender** | **Age** |
|---|---|---|
| **Accuracy** | 66.97% | 62.92% |
| **Baseline** | 50% | 33% |

Table 6: Performance of our approach when classifying by Gender and Age with an extended feature set

As can be observed, the introduction of these new features leads to a slight decrease of accuracy in gender identification, but, at the same time, to a considerable increase in age classification.

It is obvious that the quality of texts influences the performance of author profiling and that in order to capture idiosyncratic features of a specific genre (such as recurrent orthographic and syntactic mistakes), specific features must be drawn upon.

## 5. Conclusions and Future Work

The most obvious conclusion from our work is that a small set of distinctive features that characterize blog postings—including traditional word-oriented features, but also structural sentential features and dictionary features that capture"positive" and "negative" words (as used in sentiment analysis) and "patriotic" words help distinguish between writings of men and women. The differences between these writings can be observed at many different levels. Thus, we can see a difference in how specific punctuation marks, words, and grammatical means are used, as well as in the expressiveness of the writings.

We compared our approach with a bag-of-words approach in which only content-features were used for classification. This comparison revealed that the use of thousands of features that capture only the content of the writings does not lead to a higher accuracy. The quality (i.e., the distinctiveness) of the features is more important than their quantity.

Our approach performed well in three different datasets; the use of the same feature set for age identification improved the baseline by more than 20%. With a small number of additional features that were more age-oriented, the accuracy improved the baseline even by more than 29%. In other words, even this small experiment shows that author profiling using a small set of features is a feasible goal, but that these features must also capture the idiosyncrasies of the authors.

In our future work, we plan to expand our research to author profiling in general. We will explore the identification of the age, native tongue and education level of the authors, using genre-independent techniques that deliver a good outcome with a small amount of distinctive features.

## 6. References


Argamon, S., Koppel, M., Pennebaker, J. W., and Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119.

Bohnet, B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 89–97, Stroudsburg, PA, USA. Association for Computational Linguistics.

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating Gender on Twitter. *Test*, 146:1301–1309.

Cheng, N. C. N., Chen, X. C. X., Chandramouli, R., and Subbalakshmi, K. P. (2009). Gender identification from E-mails. *2009 IEEE Symposium on Computational Intelligence and Data Mining*.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Koppel, M., Argamon, S., and Gan, R. (2000). Automatically Categorizing Written Texts by Author Gender.

Mukherjee, A. and Liu, B. (2010). Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 207–217, Stroudsburg, PA, USA. Association for Computational Linguistics.

Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. W. (2006). Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 199–205. AAAI.

Zhang, C. and Zhang, P. (2010). Predicting gender from blog posts. pages 1–10.