STAIRS 2016
D. Pearce and H.S. Pinto (Eds.)

doi:10.3233/978-1-61499-682-8-215

Use of Discourse and Syntactic Features for Gender Identification

Juan Soler-Company [a,1]
[a] *NLP Group, Pompeu Fabra University*

Abstract. Author profiling and Gender Identification have gained relevance in the last few years. The goal of the research in these fields is to extract certain demographic information on the authors of texts by analyzing their writing at several levels. In our work, we address the problem of the identification of the gender (*male* vs. *female*) of the authors of opinion pieces published online. Unlike the overwhelming majority of the proposals, we argue that the use of deeper linguistic features (i.e., syntactic and discourse structure), instead of mainly lexical features leads to a higher accuracy of gender identification. Using such features with supervised machine learning, we achieve very competitive results with accuracies over 84%.

Keywords. Author Profiling, Gender Identification, Text Classification

1. Introduction

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. 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.

Author profiling has gained relevance in the last few years due to the huge amount of data that became available as well as due to its potential applications, for instance, in forensic linguistic applications such as threatening letter analysis, pedophile detection in chat sites or statement analysis in police questionings. It can be also considered as a very powerful marketing tool to adjust the services of a company depending on the demographic characteristics of the main target audience.

In what follows, we address the problem of author gender identification of opinion pieces of online versions of newspapers. Most of the state-of-the-art proposals on author gender identification focus on the use of lexical features. However, lexical features do not capture the stylistic patterns that characterize the writings of an author; rather, they target the content of a writing. Syntactic and discourse features are much more suitable for this purpose. Therefore, in our work, syntactic and discourse features are the main features we draw upon–with a very competitive outcome.

[1] Email: juan.soler@upf.edu

In the next section, we briefly review the related work. In Section 3, we describe our experimental setup and the features we use in the experiments. Section 4 presents and discusses the outcome of the experiments. Section 5, finally draws some conclusions and sketches future lines of our research in this area.

2. Related Work

Author profiling has been addressed in several different works. Two demographic characteristics that attracted the majority of the attention of the field are age and gender.

Some examples can be found in [7,1,11]. In [11], age, gender, geographic origin and occupation identification in Vietnamese blogs are performed. The authors of [2], seek to identify the gender of the authors and the genre of their writing (fiction vs. non-fiction).

In [7], 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.

Identifying the age of the authors of blog posts is the focus in [13] and [6,5,9,10,15, 16,17], focus on the gender of the authors. Both [14] and [12] deal with gender and age identification of blog authors.

Different kinds of genres have been explored in the field. In the case of [20], the texts are informal blog posts; in [6], emails, in [5], tweets, and in [9], chat logs. In [8], the authors also use chat logs, and it is a study on applying author profiling to identify pedophiles in chat forums.

In terms of features, [7] use lexical, character-based and email structure features. In [1], frequent words, function words and parts-of-speech are used. In [11], character and word-based features compose their feature set. The authors of [14] use frequent words, function words and parts-of-speech frequencies as well as specific blog features. In [12], they extract word-based features, punctuation marks and parts-of-speech frequencies. They also analyze the usage of emoticons and polar words. The majority of works of this field focus mainly on content-dependent features.

Below, we present a feature set that mainly captures the dependency syntactic structures and the discourse structure of the texts. It is a fairly novel technique which differentiates this work with the majority of the state of the art.

3. Experimental Setup

In this section, we present the experimental setup that was used for the presented work and introduce the used dataset and the features. We cast the problem of author gender identification as a supervised machine learning problem, where the goal is to distinguish between two classes *male* vs. *female*. For the experiments, we use Weka's implementation of a Bagging classifier using Random Forests as its base classifier. The features are represented in terms of multidimensional vectors, with each feature as a separate dimension and one of the values of a feature as instantiation of its dimension. To obtain more reliable performance figures, we use 10-fold cross validation, such that the outcome of the classification does not depend on which part of the dataset has been used for training and which part for testing.

For the feature extraction part, where raw text is used as input and multidimensional feature vectors that characterize them are outputted, Python and its Natural Language Toolkit (NLTK) were used as well as a dependency parser ([4]) and a discourse parser ([18]).

3.1. Dataset

The data that was used for the experiments is composed of opinion pieces in English obtained from online newspapers. They were crawled, cleaned and manually labeled by the gender of the author of the text[2]. The posts of the dataset are multi thematic, the authors express their opinion on several topics such as: sports, politics, economy or general news.

The sources that were crawled to compile the dataset were the Sun, the Times and the New York Daily. 7148 texts written by 51 different authors compose the dataset. These texts have a mean length of 348.64 words. The corpus is balanced: it contains the same amount of texts written by male and female authors.

3.2. Feature Set

To characterize the authors of the texts, we use a set of features that relies heavily on syntactic and discourse features, which help characterize the style of the authors drawing upon sentence constructions.

In total, six different types of features are used: the used Character-based, Word-based, Sentence-based and Dictionary-based features are described in [17]. In this work, we expand the syntactic features and introduce the discourse features:

Syntactic Features

This group of features accounts for more than 50% of the total number of features. To compute the features that will be presented, the dependency parser described in [4] was used. We can subdivide this group of features in 3 subgroups:

Parts-of-Speech

This subgroup of syntactic features contains the frequency of each parts-of-speech tag. We compute the percentage of words of a text that are classified as each one of the possible parts-of-speech[3].

These features can be very useful to analyze the distribution of word categories per text. A higher usage of adjectives could be seen as an indicator on the expressiveness of a text. The analysis of a text, based on these kind of features can help us find patterns that are gender-specific and that can help distinguishing between texts written by male and female authors.

Dependency Features

This subgroup of features uses the output of the dependency parser. Each sentence is represented by a dependency tree where the arcs are dependencies between words. The dependency tags outputted by the used parser are described in [19].

From the dependency trees we extract the frequency of each one of the individual dependency relations per sentence, the percentage of modifier relations used per tree

[2]This dataset has been made publicly available and can be downloaded at http://bit.ly/1YIMm6S

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

as well as the frequency of adverbial dependencies (they give information on manner, direction, purpose, etc). The ratio of modal verbs with respect to the total number of verbs and the percentage of verbs that are part of a verb chain (such as ”has taken”, ”were thinking”, etc) also are in this feature group.

Tree-Shape Features

The goal of this subgroup of features is to extract information from the shape of the dependency trees. 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. Finally, the ramification factor will be the mean number of children per level. These features will characterize the complexity of the inner structure of the sentences.

We also apply these measures to subordinate and coordinate clauses. The existence of this kind of clauses indicates that a sentence has a certain degree of complexity. If we complement this information with the shape of said clauses, we will measure exactly how complex are these sub-trees.

Discourse Features

To measure the discourse structure of the texts, a discourse parser is used [18]. This parser gets a full text as input, extracts *Elementary Discourse Units* (EDUs) and links them with discourse relations. The output of this parser will be a discourse tree where the leaves will be EDUs, and the relations between them, discourse relations.

We extract a group of features that contains the frequency of each one of the discourse relations per EDU (we divide the number of apparitions of each discourse relation by the number of EDUs per text). The full set of discourse relations contains: Joint, Background, Condition, Evaluation, Summary, Cause, Contrast, Topic-comment, Elaboration, Comparison, Topic-change, Textual-organization, Enablement, Attribution, Explanation, Same-unit and Manner-means. We also measure the shape of these trees by extracting their depth, width and ramification factor. This group of features measures the global discourse structure of texts.

4. Experiment Results and their Discussion

As stated before, we use Weka’s implementation of the ensemble algorithm of Bagging (with random forests as base classifier). We analyze the performance of each group of features individually as well as the whole feature set by computing the accuracy (number of instances where the classifier predicted correctly divided by the total number of instances).

Table 1 shows the performance of the system in different scenarios.

We see that the Character-based group of features proves to be very effective. This group analyzes the usage of punctuation marks such as periods, commas, colons or semicolons. The frequency of commas, for example, can be seen as a highly stylistic choice and the good performance of this group of features tells us that there are clear patterns that differentiate between genders. It is also shown that our syntactic group of features is very effective, analyzing the structure of the phrases gives us valuable information that helps the classifier predict the gender of the authors effectively. The performance of the system with the full set of features achieves very competitive results, predicting correctly in more than 84% of the cases.

Used Features	Accuracy
Character-Based	71.23%
Word-Based	69.51%
Sentence-Based	55.74%
Dictionary-Based	54.86%
Syntactic	76.79%
Discourse-Based	64.32%
Full Set	**84.65%**
Stopword Baseline	66.96%

Table 1. Performance of the gender identification system using different subsets of features.

The performance is compared to a baseline that consists in individual frequencies of stopwords found in the texts (we consider the list of stopwords provided by the NLTK Python Toolkit). It is clear that our approach outperforms the stopword list classification baseline by a large margin (see [3] for an example of gender identification using stopword frequencies).

To analyze what features were the most distinctive in the classification process, we computed the information gain of every feature. The 20 features that were most distinctive were the following:

• colons, • quotations, • syntactic width, • first person plural pronouns, • usage of commas, • usage of pronouns, • subordinate clause frequency, • standard deviation of word length, • syntactic ramification factor, • usage of stop words, • usage of exclamations, • discursive ramification factor, • discursive depth, • usage of nouns, • usage of elaborations, • coordinate clause width, • words per sentence, • vocabulary richness, • usage of hyphens, • usage of appositions

Even though the results of the discourse features by themselves were not impressive, we can see that some of the characteristics of the discourse trees are among the features with more information gain. It is also confirmed that punctuations are very distinctive. It is interesting that the vocabulary richness is relevant in the classification process (looking into the feature values we saw that female authors tend to have richer vocabulary). The high information gain on the frequency of subordinate clause frequency, syntactic width, and syntactic ramification factor tells us that measuring the complexity of the syntactic structures and analyzing the syntactic trees, generates very useful features to differentiate between genders.

5. Conclusions and Future Work

We presented a very effective feature set that is composed mainly of syntactic and discourse features that predicted the gender of the author of texts correctly in more than 84% of the cases. This approach focuses mostly on characterizing the structure of the texts instead on focusing on the content itself. With this approach, we detect stylistic patterns that distinguish between male and female authors effectively. In the future, we want to focus on author profiling, classifying the authors of texts not only by gender but by age, sexual orientation, profession, native language and other demographic characteristics.

References

[1] Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler, 'Automatically profiling the author of an anonymous text', *Communications of the ACM*, **52**(2), (2009).

[2] Shlomo Argamon and Anat Rachel Shimoni, 'Automatically categorizing written texts by author gender', *Literary and Linguistic Computing*, **17**, (2003).

[3] Arun, Saradha, V. Suresh, Murty, and C. E. Veni Madhavan, 'Stopwords and Stylometry : A Latent Dirichlet Allocation Approach', in *NIPS Workshop on Applications for Topic Models: Text and Beyond*, (2009).

[4] Bernd Bohnet, 'Very high accuracy and fast dependency parsing is not a contradiction', in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, (2010).

[5] John D. Burger, John Henderson, George Kim, and Guido Zarrella, 'Discriminating gender on twitter', in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (2011).

[6] Na Cheng Na Cheng, Xiaoling Chen Xiaoling Chen, R Chandramouli, and K P Subbalakshmi, 'Gender identification from E-mails', *2009 IEEE Symposium on Computational Intelligence and Data Mining*, (2009).

[7] Dominique Estival, Tanja Gaustad, Son B. Pham, Will Radford, and Ben Hutchinson, 'Author Profiling for English Emails', in *Proceedings of the Australasian Language Technology Workshop*, (2007).

[8] Aditi Gupta, Ponnurangam Kumaraguru, and Ashish Sureka, 'Characterizing pedophile conversations on the internet using online grooming', *arXiv preprint arXiv:1208.4324*, (2012).

[9] Tayfun Kucukyilmaz, Berkant Barla Cambazoglu, Cevdet Aykanat, and Fazli Can, 'Chat mining for gender prediction.', in *ADVIS*, (2006).

[10] Arjun Mukherjee and Bing Liu, 'Improving gender classification of blog authors.', in *Proceedings of the International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (2012).

[11] Dang Duc Pham, Giang Binh Tran, and Son Bao Pham, 'Author Profiling for Vietnamese Blogs', *2009 International Conference on Asian Language Processing*, (2009).

[12] Francisco Rangel and Paolo 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).

[13] Sara Rosenthal and Kathleen McKeown, 'Age Prediction in Blogs: A Study of Style, Content, and Online Behavior in Pre- and Post-Social Media Generations', in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, (2011).

[14] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker, 'Effects of age and gender on blogging.', in *Proceedings of the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, (2006).

[15] Juan Soler-Company and Leo 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).

[16] Juan Soler-Company and Leo Wanner, 'Multiple language gender identification for blog posts', in *Proceedings of the 37th Annual Cognitive Science Society Meeting (COGSCI'15)*, (2015).

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

[18] Mihai Surdeanu, Thomas Hicks, and Marco 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).

[19] Mihai Surdeanu, Richard Johansson, Adam Meyers, Llu'ıs Màrquez, and Joakim 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*, (2008).

[20] Cathy Zhang and Pengyu Zhang, 'Predicting gender from blog posts', *Technical Report. University of Massachusetts Amherst, USA*, (2010).