Multiple Language Gender Identification for Blog Posts

Juan Soler-Company (juan.soler@upf.edu)
Department of Information and Communication Technologies
Pompeu Fabra University, Barcelona

Leo Wanner (leo.wanner@upf.edu)
Catalan Institute for Research and Advanced Studies (ICREA) and
Department of Information and Communication Technologies
Pompeu Fabra University, Barcelona

Abstract

In data-driven gender identification, it has been so far largely assumed that the same types of (mostly content-oriented) data features can be used to differentiate between male and female authors. In most cases, this distinction is done in a monolingual scenario. In this work, we discuss a set of features that distinguish between genders in six different datasets of blog data in English, Spanish, French, German, Italian and Catalan with accuracies that range from 77% to 88%. Using a reduced set of language-independent structural features in a multilingual scenario we first identify the gender and then the gender and language of the author, achieving accuracies higher than 74%.

Keywords: Natural Language Processing; Text Categorization; Author Profiling; Gender Identification

Introduction

Identification of the gender of the author of a written (or spoken) discourse has become a popular research topic in empirical computational linguistics. This research presupposes that men and women think, talk and write differently. But how differently? Does the language background also influence the difference of how men and women write? It is known that an average English sentence has a less complex syntactic structure than a German sentence. Does the assumed difference in the complexity of the syntactic structures in English and German lead to idiosyncrasies in gender identification in English and German?

The vast majority of approaches to data-driven gender identification have been so far on English; rather few are on other languages; see, e.g., (Estival, Gaustad, Pham, Radford, & Hutchinson, 2007) on Arabic, (Rangel & Rosso, 2013) on Spanish, (Kucukyilmaz, Cambazoglu, Aykanat, & Can, 2006) on Turkish and (Pham, Tran, & Pham, 2009) on Vietnamese, and there are practically no systematic language-contrastive experiments. In author profiling research, some works attempt to recognize the native language of English learners. This is usually done by error analysis of the writings of learners with the goal to find parallelisms between the errors and the characteristics of another language. If such a parallelism is identified, the language in question is hypothesized to be the native language of the writer (Koppel, Schler, & Zigdon, 2005; Wong & Dras, 2009). However, the restriction to learner texts and the use of idiosyncratic mistakes as features limit the value of these works for general language background studies in the context of author profiling.

In order to shed some light on the above questions, we carried out three experiments on blog post corpora in Catalan, English, French, German, Italian and Spanish, interpreting the problem of gender and language identification as a supervised classification problem: (i) classification of blog posts in each of these languages with respect to the gender of their authors (man vs. woman); (ii) classification of all posts joined into one multilingual data set with respect to the gender of the writers; and (iii) classification of all posts with respect to gender and language of the author at the same time (as, e.g., 'male English', 'female Spanish', etc.).[1]

For the first experiment, we use a series of structural features (including grammatical function features of the kind 'subject', 'direct.object', etc., which reflect language-specific grammatical tags). For experiments (ii) and (iii), we use strictly language-independent, universal features, such that the classification procedure does not have any explicit language clues. In none of the experiments, content-oriented features (as, e.g., the most common words or *n*-grams) are used, since content-oriented features let gender identification heavily depend on the training dataset and make it hardly comparable across languages. This makes our proposal different from the vast majority of the state-of-the-art approaches to gender identification, which all heavily draw on content-oriented features.

In the next section, the features that are used in the experiments are presented. Then, we describe the experiments and discuss their outcome. A brief summarization of the related work in the area of gender identification and author profiling precedes some conclusions from the presented work and the outline of the future work we plan in this area.

Feature set

The overwhelming majority of the approaches to data-driven gender recognition and author profiling usually use large quantities of content-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. Some approaches additionally use syntactic fea-

[1] We are aware that gender is not necessarily binary; see, e.g., (Lorber, 2011) and other numerous studies in sociolinguistics. Still, in the experiments presented in this paper, we will work with the *gender binary* assumption. In the future, we plan to explore gender as spectrum in the context of author profiling.

tures. For cross-language language background studies, as in our case, these features are not appropriate. What we need, are features that are entirely or at least to a certain extent (as, e.g., grammatical functions) language and content-independent. These are structural features. For our work, we use four different types of mostly content-independent features: (i) character-based features, (ii) word-based features, (iii) sentence-based features, and (iv) syntactic features. Table 1 displays a summary of the number of features of each type that were used.

Table 1: Feature number overview

Type	# Features
Character-based Features	15
Word-based Features	14
Sentence-based Features	2
Syntactic Features	22–65

Character-based features consist of the ratios of

• comma, • dot, • colon, • semicolon, • exclamation mark, • question mark, • opening/ closing parenthesis, • opening/closing bracket, • quotation mark, • plus sign, • minus sign, • hyphen, • percentage sign, • dollar sign, and • numerals

per post (i.e., the frequency of their occurrence in a post divided by the total number of characters in this post).

Word-based features consist of the ratios of

• interjections, • affirmation and negation words, • first person singular and first person plural pronouns, • stop words, • proper nouns, • acronyms, • words with less than five characters, • five or more characters per post, and • different words (vocabulary richness)

per post as well as • the average number of characters per word, and • total number of words in a post.

Sentence-based features are composed by two features only: • the total number of sentences in a post and • the average number of words per sentence in a post.

Syntactic features constitute the largest group of our features. They consist of the frequencies of individual dependency relations in the dependency trees of the sentences in the post as well as the mean width and depth of the dependency trees. The depth of the trees is defined as the longest path between the root and one of the leaves. The width is the maximum number of siblings at any of the depths of the tree. The depth and width of dependency trees can be interpreted as a measure of the complexity of the structure of the corresponding sentences.

To obtain the dependency trees, (Bohnet, 2010)' statistical dependency parser is used. The dependency tag sets differ from language to language and are also of different granularity (from 22 for French to 65 for English). As a result, the number of syntactic features differs from language to language.

Experimental Setup

For the supervised classification experiments, we use Weka's Bagging classifier with Random Forests as base classifier.[2]

The features are captured in a file in which all blog posts 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. 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.

Data sets

As already mentioned, we experiment with Catalan, English, French, German, Italian and Spanish texts. For the compilation of the data sets, the same methodology was used for all six languages. We searched for blogs in which the authors were known, such that their gender could be deduced for validation of the performance of our algorithm. For this purpose, we looked for blog sections of online newspapers and magazines listed in Table 2.

The blog posts were crawled, cleaned from html-tags, and tagged manually with a gender (man vs. woman) and language tag. To avoid distortion, in all six data sets, the distribution between male and female authors has been balanced (50%). The topics of the blogs are quite diverse, ranging from politics to sports, even about television, theater, fashion and many other topics. All posts are well structured and well written and most of the times an opinion is expressed. Table 3 summarizes the number of texts per dataset that were crawled.

Experiments and their results

As outlined in the Introduction, we carried out three different experiments, taking as baseline in all three random classification.

In the first experiment, we carried out gender identification for each language dataset separately. Table 4 displays the performance of our classifier in this experiment.

For the second and third experiments, the six datasets were merged, such that the resulting dataset is composed of 29117 texts by male and female authors in Catalan, English, French, German, Italian and Spanish. Furthermore, the set of features has been reduced to 27 language-independent features: all punctuation features, the frequency of the usage of acronyms, the frequency of the usage of first person singular/plural pronouns, the frequency of the usage of stop words, the mean number of words per sentence, characters per word,

[2]Weka is University of Waikato's a public machine learning platform that offers a great variety of different classification algorithms for data mining (Hall et al., 2009).

Table 2: Data set sources

Catalan	El Punt, Avui, Ara, Mes, Directe
English	Sun, Times, New York Daily
French	L'Express, Le Monde
German	Die Welt, Süddeutsche Zeitung, Frankfurter Allgemeine Zeitung, Compact, taz
Italian	Corriere della Sera, Il Messaggero, Il Post
Spanish	Publico, El Mundo, La Vanguardia, 20minutos, ABC, El Periódico

Table 3: Overview of the data sets

	English	**Spanish**	**German**	**French**	**Catalan**	**Italian**
Number of Posts	7148	5794	3564	4310	4078	4265
Number of Authors	51	101	127	18	33	43
Mean length (words)	348.64	612.02	500.77	364.11	404.31	263.45

Table 4: Performance of the monolingual gender identification classifier ('Acc' stands for "accuracy of our algorithm", 'BS' for "baseline accuracy" and '#Feat' for "number of features")

	Eng	**Sp**	**Ger**	**Fr**	**Cat**	**It**
Acc	0.80	0.88	0.77	0.83	0.88	0.86
BS	0.50	0.50	0.50	0.50	0.50	0.50
#Feat	96	83	73	52	79	52

Table 5: Reference Corpora

Language	**Corpus**
Catalan	Cess_cat
English	Brown
French	Baf
German	Tiger
Italian	Turin university treebank
Spanish	Cess_esp

the percentage of words that are more (and less) than 5 characters and the percentage of words that start/end with vowel/consonant.[3]. They are language-independent in the sense that they appear in all of the languages we consider—although they are, obviously, instantiated differently. But since we count only their appearance, not their concrete instantiations, they can indeed be considered universal.

In order to avoid the influence of idiosyncratic characteristics of a language[4] on these features, the feature values are normalized: each value is divided by the value of the corresponding reference feature obtained from a reference corpus of the language in question. As a consequence, we obtain for each text a feature profile that reflects the author's personal writing style rather than a language-inherent bias. Table 5 lists the used reference corpora.

In order to be able to normalize features during the experiments, i.e., when we classify a test dataset (and thus do not know the language of a text), we implemented a language prediction procedure. The procedure is based on the similarity of the feature values to each of the corresponding reference feature values: the more similar the values, the more likely the language of the reference features is to be used for normalization.

In the second experiment, the texts in the merged dataset have been classified with respect to the gender of the authors of the texts. The difference between this experiment and the first one is that in this case the classification is carried out with language-independent features only, on a multilingual dataset using feature normalization as described above. The results of this experiment can be seen in Table 6.

Table 6: Results of multilingual gender identification

	Merged Dataset
Accuracy	77.01%
Baseline	50.19%

In the third experiment, the texts in the merged dataset were classified with respect to twelve different classes: 'catalan_male', 'catalan_female', 'english_male', 'english_female', ... The purpose of this experiment has been to assess to what extent we can identify the gender and language of an author in one single dataset analyzing only the writing style of the authors. If this is feasible (again, without any dictionaries or language-dependent features), it can be feasible to identify the native language of an author not only in language learner texts, but also in well-written texts. The results of this experiment are displayed in Table 7. The

[3]Syntactic features cannot be used here because the dependency relation tag sets are language-specific.

[4]For instance, in German punctuation is much more grammaticalized than in English, where it is highly style-driven. This leads to a higher relative frequency of, e.g., commas and semicolons in German. The same occurs with capitalization: in German, nouns are capitalized.

baseline is low because the number classes that are used in this classification process is rather large (recall that we use random classification as baseline).

Table 7: Performance of the joint gender and language identification experiment

	Merged Dataset
Accuracy	74.67%
Baseline	12.26%

Discussion

The results of the first experiment show that a set of features that captures mainly the syntactic structure and writing style of an author (rather than the vocabulary and thus content, as does the majority of the state-of-the-art proposals) achieves state-of-the-art accuracy not only, e.g., for English, where such features are more freely used, but also for French, German, etc., where punctuation is much more regularized (such that gender identification is a priori more difficult). The fact that the same features worked very well for all languages can be seen as clear evidence that there are common patterns that distinguish the writing style of both genders for all six languages considered.

The performance figures of the second and third experiments show that a small number of structural features can be used for gender identification with a competitive outcome, and that the writings of the authors of different genders show idiosyncratic patterns of language-independent features that allow for the identification of the language in which they are written. Due to the fact that the use of these patterns by an author is, as a rule, subconscious, it can be hypothesized that it is realistic to assume that it is feasible to identify the gender and native tongue of the author when he or she writes in a foreign language. The hypothesis would be that the writers carry over their writing style from their native language to their writings in a foreign tongue.

Figures 1 and 2 show the contribution of the individual features to the writing style of both genders in our six languages. Each axis represents the normalized mean value of a feature for men and women. Figure 1 shows the contribution of the punctuation features, while Figure 2 captures the word-oriented features. Remember that the normalized features are calculated as the ratio between actual feature values and the reference feature values. Both graphs have the mean values of the features represented in a logarithmic scale.

Both figures reveal there are several differences between languages at a punctuation and word level, and these differences are what makes both gender and gender and language identification possible. In Figure 1, the main differences are observed in the use of quotation marks of German writers relatively to the other languages. There are also some deviations in the writings of Italian men and women with respect to the use of exclamation marks.

In Figure 1, it is revealing to compare Spanish and Catalan. Even though these two languages are quite similar, we see that the way men and women deviate from the reference features in both languages is different. The deviation in the usage of quotation marks, semicolons, question marks and dots is quite different if we compare the writings of the opposed genders. It can be also observed that French women deviate more than men in all punctuation features.

The style of German authors deviates most from the style of the other authors: the values of the features of German authors are smaller than in the other languages in both cases. This means that the deviation from the reference features in German authors is smaller than in the other languages. We can hypothesize that this could be due to the cultural influences. The lack of space prevents us from entering into more details here.

We also observe that the difference between genders is larger in the first figure than in the second one. Punctuation features can be considered highly stylistic features that are used in a subconscious way and as a result, the difference between the values of these features and the reference features is larger than in the case of word-oriented features.

Some interesting language-contrastive observations of the distribution of features can also be extracted. Thus, the distribution of word-oriented features in all Romance languages that we considered in our experiments is rather similar. Since we eliminated the linguistic bias by normalization, we can hypothesize that this similarity is again due to cultural influences.

Related Work

The problem of author profiling has been addressed in several works. See, for instance, (Estival et al., 2007; Koppel et al., 2005; Argamon, Koppel, Pennebaker, & Schler, 2009; Pham et al., 2009). (Estival et al., 2007) deal with gender, age, native language, country of origin and psychometric traits identification of email authors—similar to (Argamon et al., 2009), who do gender, age, native language and personality identification. In (Pham et al., 2009), the age, gender, geographic origin, and occupation of the authors of blogs in Vietnamese is worked on, while (Argamon & Shimoni, 2003) seek to identify the gender of the authors and the genre of their writing (fiction vs. non-fiction).

In particular two parameters in author profiles attracted so far the attention of the field: age and gender. Cf., e.g., (Rosenthal & McKeown, 2011), who focus on the age of the authors of blog posts and (Zhang & Zhang, 2010; Cheng, Chen, Chandramouli, & Subbalakshmi, 2009; Burger, Henderson, Kim, & Zarrella, 2011; Kucukyilmaz et al., 2006; Mukherjee & Liu, 2012), who focus on gender of the authors. (Schler, Koppel, Argamon, & Pennebaker, 2006) and (Rangel & Rosso, 2013) deal with both gender and age identification of blog authors. In the case of (Zhang & Zhang, 2010), the texts are informal blog posts; (Cheng et al., 2009) work on emails, (Burger et al., 2011) on tweets, and (Kucukyilmaz et al., 2006) on chat logs.

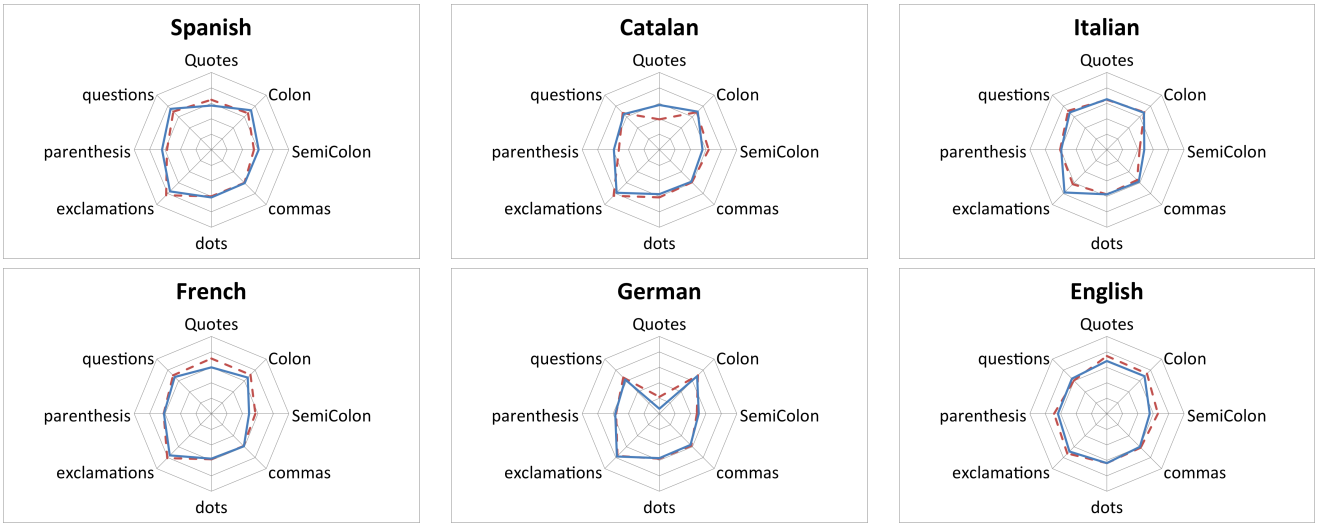

Figure 1: Distribution of punctuation features in the posts of men and women across languages; solid line (male), dotted line (female)

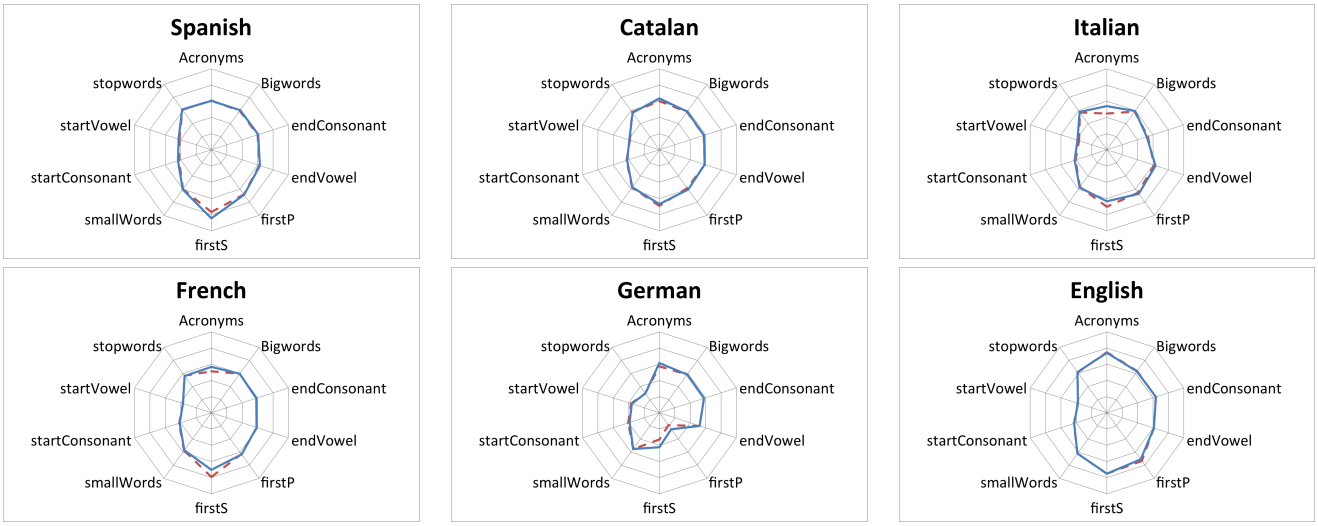

Figure 2: Distribution of word-oriented features in the posts of men and women across languages; solid line (male), dotted line (female); where the dotted line does not show, it overlaps with the solid one

All of these works draw upon a large number of features, including Part-of-Speech (PoS) tags, dictionaries, term frequencies, stylistic markers, etc. For instance, (Zhang & Zhang, 2010) achieve an accuracy of >70% with PoS tags and dictionaries; (Kucukyilmaz et al., 2006) achieve an accuracy of 84.2% with term- and style-based features. (Rosenthal & McKeown, 2011)'s approach is most similar to ours in that they also use syntactic dependencies as features—as we do; the accuracy they achieve is of 81.57%. However, as in other approaches, they also use a large quantity of further language-dependent features.

(Groom & Pennebaker, 2005) classifiy authors of online personal advertisements by their gender and sexual orientation. They analyze if the miss-classified instances match existing social stereotypes: are homosexual men confused by heterosexual women? This is an issue that to the best of our knowledge is addressed only in this work. Given its relevance, we plan to explore it in the future as well.

Few works consider in one way or the other the question of language background of authors. For instance, (Koppel et al., 2005) and (Wong & Dras, 2009) work with English learner texts, using idiosyncratic errors in these texts to determine the native tongue of their authors. However, none of them undertakes cross-language studies of the kind we did that would allow for an analysis of language-specific differences in the writings of the different genders. The recent and upcoming shared tasks organized in the field have and will hopefully continue to contribute to a change of this state of affairs; see,

e.g., the PAN 2014 challenge, in which the participants had to address the task of author gender and age identification in English and Spanish texts (Rangel et al., 2014).

Conclusions and Future Work

We used a set of language- and content-independent features that were normalized in order to avoid a bias resulting from the idiosyncratic syntactic, punctuation and writing style characteristics of a language. Compared to state-of-the-art proposals in the field, our set of features is very small. Nonetheless, the results are very competitive.

The conclusion that can be drawn from our work is that it is feasible to use the same set of features to determine the gender of the authors of texts written in different languages with high accuracy. The setup of the experiments that we carried out and their outcome make us furthermore hypothesize that if a set of language- and content-independent features could profile the writing of authors effectively, it might be possible to detect the native language of an author writing in a foreign language.

In the future, we also plan to explore how unsupervised or semi-supervised approaches can be used in author profiling problems. This possibility seems to be of high relevance in particular in forensic applications, where no training data of sufficient size as needed for supervised learning is available.

As already pointed out above, we also plan to compile a dataset tagged by gender and sexual orientation in order to explore not only automatic classification of texts by the sexual orientation of the authors, but also to analyze the misclassifications along the lines done in (Groom & Pennebaker, 2005).

References

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

Argamon, S., & Shimoni, A. R. (2003). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, *17*, 401–412.

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)* (pp. 89–97). Stroudsburg, PA, USA: Association for Computational Linguistics.

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1301–1309). Stroudsburg, PA, USA: Association for Computational Linguistics.

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

Estival, D., Gaustad, T., Pham, S. B., Radford, W., & Hutchinson, B. (2007). Author Profiling for English Emails. In *Proceedings of the Australasian Language Technology Workshop* (pp. 21–30).

Groom, C. J., & Pennebaker, J. W. (2005). The language of love: Sex, sexual orientation, and language use in online personal advertisements. *Sex Roles*, *52*(7-8), 447–461.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, *11*(1).

Koppel, M., Schler, J., & Zigdon, K. (2005). Determining an author's native language by mining a text for errors. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining – KDD '05.* New York, New York, USA: ACM Press. doi: 10.1145/1081870.1081947

Kucukyilmaz, T., Cambazoglu, B. B., Aykanat, C., & Can, F. (2006). Chat mining for gender prediction. In T. M. Yakhno & E. J. Neuhold (Eds.), *Advis* (Vol. 4243, p. 274-283). Springer.

Lorber, J. (2011). Believing is seeing: Biology as ideology. In M. S. Kimmel, A. Aronson, & A. Kaler (Eds.), *The gendered society reader* (pp. 11–18).

Mukherjee, A., & Liu, B. (2012). Improving gender classification of blog authors. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing (EMNLP)* (p. 207-217). ACL.

Pham, D. D., Tran, G. B., & Pham, S. B. (2009, December). Author Profiling for Vietnamese Blogs. *2009 International Conference on Asian Language Processing*, 190–194. doi: 10.1109/IALP.2009.47

Rangel, F., & Rosso, P. (2013). 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.*

Rangel, F., Rosso, P., Verhoeven, B., Potthast, M., Trenkmann, M., Stein, B., ... Daelemans, W. (2014, 2014/09/18). Overview of the 2nd author profiling task at PAN 2014. In *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers.* Sheffield, UK.

Rosenthal, S., & McKeown, K. (2011). Age Prediction in Blogs: A Study of Style, Content, and Online Behavior in Pre- and Post-Social Media Generations. In D. Lin, Y. Matsumoto, & R. Mihalcea (Eds.), *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (p. 763-772). The Association for Computational Linguistics.

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

Wong, S.-M. J., & Dras, M. (2009). Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Workshop.*

Zhang, C., & Zhang, P. (2010). Predicting gender from blog posts. *Technical Report. University of Massachusetts Amherst, USA*.