Using Sociolinguistic Inspired Features for Gender Classification of Web Authors

Vasiliki Simaki[1], Christina Aravantinou[1], Iosif Mporas[2] and Vasileios Megalooikonomou[1]

Multidimentional Data Analysis and Knowledge Management Laboratory,
Dept. of Computer Engineering and Informatics, University of Patras, 26500-Rion, Greece

1{simaki, aravantino, vassilis}@ceid.upatras.gr
2imporas@upatras.gr

Abstract. In this article we present a methodology for classification of text from web authors, using sociolinguistic inspired text features. The proposed methodology uses a baseline text mining based feature set, which is combined with text features that quantify results from theoretical and sociolinguistic studies. Two combination approaches were evaluated and the evaluation results indicated a significant improvement in both combination cases. For the best performing combination approach the accuracy was 84.36%, in terms of percentage of correctly classified web posts.

Keywords: text classification algorithms, sociolinguistics, gender identification

1 Introduction

The expansion of text-based social media is impressive and the need of classifying the provided information into sub categories is an important task. This categorization can be made in terms of topic, genre, author, gender, age, etc. according to the informational need and the purpose of the users. This is implemented by identifying differential features characterizing the demanded purpose. Every social media user leaves his digital fingerprints on the web, not only by declaring personal information, but unconsciously through his writing style. One of the most important issues on this field is the identification of the user's gender and the classification of documents according to this specification. It is a challenging task, given that in the typical case the gender is identified without taking into account the personal information the user provides, but estimated only using the content of his/her texts.

Gender classification is an important field of text mining with many commercial applications. The knowledge of the user's gender is important to companies in order to promote a product or a service, if it is preferable mostly by women or men. Market analysis and advertising professionals are interested in which product or service is more talked or liked between the two groups, and should be addressed to women or men. Gender classification is also considerable in e-government services and social science studies. Useful conclusions can be extracted about the different trends among women and men, different topics of interests, political views, social concerns, world theories, and many other issues. Since it is quite difficult for social scientists to manually go through large volumes of data, computer-based solutions supported by the recent advances in natural language processing and machine learning have techniques been proposed. In parallel to computer-based solutions sociolinguists have offered essential knowledge in support of the task of gender identification from written language.

Sociolinguistics is the specific scientific domain of linguistics which studies the influence of social factors into the written and spoken language. Factors as gender, age, education, etc., delimitate the linguistic diversity and variation, the linguistic choices that people and social groups make in everyday life. The differences between men and women’s language can be detected in their texts, due to the separate linguistic choices they make. These choices can be identified in all levels of linguistic analysis (from the phonetic to the pragmatic one) and they may be conscious or not, differentiating the speaker’s attitude from the standard language in a given communicative occasion [1].

In our study, an interdisciplinary methodology for the detection of the author’s gender is proposed, based on features derived from two different disciplines, the gender linguistic variation and the gender classification. These two kinds of features are fused in order to achieve higher accuracy and prove that linguistics and text mining, when combined, can contribute to better gender identification results.

The remainder of this paper is organized as follows. In Section 2 we present the background work in the field of gender identification, after theoretical, empirical and computational studies. Section 3 describes our methodology and in Section 4 the experimental part of our work is presented. Finally, in Section 5 we conclude this work.

2 Related Work

Several studies related to author's gender discrimination have been reported in the literature, both based on computer-based methods (text mining) and theoretical models (sociolinguistic). The first ones concentrate on efficient computational algorithms while the latter ones on social cues expressed through linguistic expressions on written text.

As considers text mining based approaches, they typically consider author’s gender identification as a text classification issue [2, 3]. Koppel et al. [4] propose text classification methods to extract the author’s gender from formal texts, using features such as n-grams and function words that are more usual in authorship attribution. This research combines stylometric and text classification techniques, in order to extract the author’s gender. Argamon et al. [5] have applied factor analysis for gender and age classification in texts mined from the blogosphere. Ansari et al. [6]have used frequency counting of tokens, tf-idf and POS-tags to find the gender of blog authors. In Burger et al. [7] a study on gender recognition of texts from Twitter was presented, where the content of the tweet combined with the username and other information related to the user we used. Many recent studies around gender classificationdeal with social media and they propose methods that identify the gender [8, 9, 10] and in some cases the age of the web users [11]. Most of the reported approaches implement their experiments, taking into account features, such as gender-polarized words, POS tags and sentence length, in order to obtain best classification results. In Sarawgi et al. [12] a comparative study of gender attribution, without taking into account the topic or the genre of the selected text is presented. Holgrem and Shyu[13] applied machine learning techniques using a feature vector containing word counts, in order to detect the author's gender of Facebook statuses. In Rangel and Rosso [14] a set of stylistic features to extract the gender and age of authors using a large set of documents from the social web written in Spanish was presented. Marquardt et al.[15] evaluated the appropriateness of several feature setsfor age and gender classification in social media.

Except the text-mining approaches, sociolinguistic studies offer valuable information about the gender characterization of a text. The basic concept of sociolinguistics, and more specifically the gender linguistic variation, is perceived as a socially different but linguistically equal way to say the same thing [8]. A general opinion about the women’s language is that women tend to make a more conservative use of language by using more standard types than men [16]. Women use non-normative forms only when they adapt socially prestigious changes, local linguistic elements, communicative indirection, and under specific communicative situations [17, 18]. Under standard conditions, they have a smaller vocabulary than men, using a narrower range of different lexical types. Compared to men discourse, women tend to use more complex syntactic structures by forming many explanatory secondary phrases in the period. The use of "empty" adjectives which have the sense of admiration and/or approval is also frequent in women’s language, as well the use of questions in place of statements [19, 20, 21]. Moreover, specific lexical choices that women do unlike men (use of norm types, avoid bad words, etc.), researchers observe their effort in many cases to decline the illocutionary force of their utterances. This phenomenon is achieved by using palliative forms like tag-questions, interrogative intonation instead of affirmations, extension of requests and hedges of uncertainty. As considers women's language, theyuse different politeness, agreement and disagreement strategies than men and more sentimental expressions, indirect requests and hypercorrected grammar types [22, 23]. Men on the other hand, tend to use more bad words, slang types and coarse language. They insert in their vocabulary non-norm forms and neologisms. In Alami et al.[24] study of the lexical density in male and female discourseand comparison of the relationship to the discourse length is performed. Eckert [25] merged existing and traditional theories, in order to create patterns about the gender-specific variation, and analyzed the meaning and the social context around a given linguistic attitude. In recent studies [26, 27, 28] researchers discuss the social factor and the stylistic information in different communicative situation in order to explain the specific linguistic choice of speakers.

3 Gender Classification Methodology

Most of the previous studies in the field of gender identification are based either in theoretical analysis and empirical findings or in computational approaches. The first kind of research, conducted by expert sociolinguists, can reveal frequent but also rare differential characteristics after empirical studies. These studies confirm existing theories and they create new rules. However, theoretical studies are time consuming, since working with large and different data collections is tedious, especially when need to verify rare discriminative rules which will probably appear only in large volumes of text data. On the other hand, computational approaches based on data mining algorithms can perform efficient and fast process of large data collections; however, the results are frequently biased to the specifications of the dataset used. Moreover, infrequent discriminative rules either will not appear in the evaluation text or they will be considered by the algorithm as outliers rather than newly discovered patterns.

The objective of the present approach for author gender identification is to exploit existing knowledge from the sociolinguistics domain in order to enhance the performance of the dominating text mining solutions. Thus, we combine sociolinguistic characteristics and data-driven features for gender classification. Specifically, a number of well-known and widely used in text mining methods features for text, author and gender classification are used to build a baseline feature vector [29]. This feature vector is combined with features inspired from sociolinguistic studies in order to enhance the gender discriminative ability of a classification engine.The sociolinguistic characteristics of gender variation may be summarized as: ‘syntacticcomplexity’, ‘use of adjectives’, ‘sentence length’, ‘different politeness and agreement/disagreement strategies’, ‘tag questions’, ‘slang types’, ‘bad words’, ‘sentimental language’, ‘lexical density’, ‘interrogative intonation’ and ‘vocabulary richness’.

The baseline (BASE) feature set and the features inspired from sociolinguistics (SLING) are presented in Table 1. The baseline feature vector has length equal to 24 and the sociolinguistic-inspired list of features has length equal to 11.

Table 1. The BASE and SLING features used in author’s gender classification.

BASE features	SLING features
# of characters per web post	normalized # of the sentence verbs
normalized # of alphabetic characters	normalized # of adjectives per comment
normalized # of upper case characters	normalized # of the text’s words
# of occurrence of each alphabetic character	# of standard polite, agreement/disagreement phrases
normalized # of digit characters	# of tag question phrases
normalized # of tab ('\t') characters	# of slang types
normalized # of space characters	# of bad words
normalized # of special characters ("@", "#", "$", "%", "&", "*", "~", "^", "-", "=", "+", ">", "<", "[", "]", "{", "}", "|", "\", "/")	normalized # of sentimentally polarized words of the comment, according to SentiWordNet[30]
total # of words	normalized # of the document’s content

	words
normalized # of words with length less than 4 characters	normalized # of the question marks to the total # of the document’s punctuation
# of punctuation symbols (".", ",", "!", "?", ":", ";", "'", "\"")	normalized # of different words per comment
average word length	
# of lines	
average # of characters per sentence	
# of sentences	
normalized # of unique words	
# of paragraphs	
average # of words per sentence	
# of "hapax legomena"	
# of "hapax dislegomena"	
normalized # of characters per word	
# of function words	
average # of sentences per paragraph	
average # of characters per paragraph	

For the combination of the baseline (BASE) and sociolinguistic-inspired (SLING) features we relied on two fusion approaches. In the first approach (early combination), the SLING features are appended to the BASE vector and the concatenated feature vector is processed by a classification algorithm. In the second approach (late combination), the data-driven (BASE) and the knowledge-based (SLING) vectors are separately processed by classification engines and the results are fused by a second-stage classifier. In both early and late fusion scenarios both data-based (from data mining) and sociolinguistic-inspired knowledge is utilized in the classification procedure.

4 Experimental Setup and Results

The text mining based and sociolinguistic-inspired combination methodology described in Section 3 was evaluated using a dataset collection of users’ comments on web. Our dataset consists of user comments in English about various topics extracted from forums and web sites. It contains comments from different sources, covering various thematic areas both from gender-preferential sites and forums, like fashion (typically preferred by women) or cars (typically preferred by men) and neutral web sources (like news, health etc). The size of the corpus is 326,736 words. The number of the characters is equal to 1,643,547. The gender division between men and women is 42% and 58% respectively.

For the classification stage, we relied on several dissimilar machine learning algorithms, which have extensively been reported in the literature. In particular, we used a multilayer perceptron neural network (MLP) and support vector machines (SVMs), using radial basis kernel (RBF) and polynomial kernel (poly). Furthermore, we employed Adaboost.M1, which is a boosting algorithm combined with decision trees (AdaBoost) and a bagging algorithm using decision trees (Bagging). Finally, we used three decision tree algorithms, namely the random tree (RandTree), the random forest (RandForest) and thefast decision tree learner (RepTree). All classifiers were implemented using WEKA toolkit [31]. In order to avoid overlap between training and test subsets a 10-fold cross validation evaluation protocol was followed. The performance results in terms of percentages of correctly classified web posts are tabulated in Table 2. The best performance per setup is indicated in bold.

Table 2. Gender classification results using different combination setups and algorithms.

	BASE	**SLING**	**BASE+SLING (early fusion)**	**BASE+SLING (late fusion)**
MLP	82.31	66.87	82.51	**84.36**
SVM(rbf)	67.49	50.00	68.31	83.13
SVM(poly)	82.72	63.17	**84.16**	82.92
Bagging	82.72	69.35	83.54	82.30
Boosting	82.10	69.14	82.51	81.07
RepTree	**82.92**	67.08	80.86	81.48
RandForest	82.72	**69.34**	82.72	79.84
RandTree	79.84	66.05	81.07	75.51

As can be seen in Table 1, the use of SLING features improves gender classification accuracy by almost 1,5% comparing to the best BASE alone. Specifically, the best BASE performance was 82.92% using the RepTree classifier, while the overall best performance was 84.36%, which was achieved with the late combination approach and the MLP classification algorithm. The SLING approach standalone does not offer competitive performance comparing to the BASE setup, however in both fusion setups there is an increase of performance which shows the importance of the sociolinguistic-inspired features. As considers the evaluated classification algorithms for the case of early combination where the fusion feature vector is of length equal to 24+11=35, the SVM algorithm outperforms all others, probably to the fact that it does not suffer from the curse of dimensionality. In the late fusion case, where the fusion vector consists of the probability of being male/female from BASE and SLING (i.e. 2+2=4 length) the MLP classifier performs better than SVM.

5 Conclusions

The exploitation of the existing knowledge extracted from theoretical and sociolinguistic studies and the transformation of this qualitative information to quantitative metrics can improve text-based gender classification accuracy. The use of sociolinguistic-inspired text features is not essential only for combination with typical text mining features, as demonstrated in this article, but can also be used to fine-tune computational algorithms by supporting the training of statistical-based models through definition initialization values and restriction of range of values of free parameters which will protect from models biased to specific data.

References

1. Archakis, A. and Kondyli, M. (2004). *Introduction to sociolinguistic issues*. Athens: Nisos(in Greek)
2. Cheng, N., Chandramouli, R.andSubbalakshmi, K.P. (201)1. Author gender identification from text. In *The International Journal of Digital Forensics & Incident Response,* Volume 8 Issue 1, July 2011, pp. 78-88.
3. Soler, J. andWanner, L. (2014). How to Use Less Features and Reach Better Performance in Author Gender Identification. In *Proceedings of LREC 2014.*
4. Koppel, M., Argamon, S. and Shimoni, A. R. (2003). Automatically categorizing written texts by author gender. In *Literary and Linguistic Computing,* 17: 401–12
5. Argamon, S., Koppel, M., Pennebaker, W. and Schler, J. (2007). Mining the Blogosphere: Age, gender and the varieties of self-expression. First Monday. 12, 9 (September 2007). DOI=Http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2003.
6. Ansari, Y.Z., Azad, S.A.and Akhtar, H. (2013). Gender Classification of Blog Authors. International Journal of Sustainable Development and Green Economic, ISSN No.: 2315-4721, V-2, I-1, 2, 2013.
7. Burger, J., Henderson, J., Kim, G. andZarrella, G. (2011). Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP '11). Association for Computational Linguistics, Stroudsburg, PA, USA, 1301-1309.
8. Kobayashi, D., Matsumura, N. and Ishizuka, M. (2007). Automatic Estimation of Bloggers' Gender. In *Proceedings of International Conference on Weblogs and Social Media*. Boulder: Omnipress.
9. Zhang, C., and Zhang, P. (2010). *Predicting gender from blog posts*. Technical Report. University of Massachusetts Amherst, USA.
10. 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. DOI=http:/www.aclweb.org/anthology/D10-1021.
11. Peersman, C., Daelemans, W. and Van Vaerenbergh, L. (2011). Predicting Age and Gender in Online Social Networks. In *Proceedings of the 3rd Workshop on Search and Mining User-Generated Contents*. SMUC'11, Glasgow, UK.
12. Sarawgi, R., Gajulapalli, K. and Choi, Y. (2011). Gender Attribution: Tracing Stylometric Evidence Beyond Topic and Genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning* (Portland, USA, 9 – 24 June, 2011). Association for Computational Linguistics, Stroudsburg, PA, USA, 78 – 86.
13. Holgrem, J., Shyu, E. (2013). Gender Classification of Facebook Posts.
14. Rangel, F., Rosso, P. (2013). Use of Language and Author Profiling: Identification of Gender and Age. In *Proceedings of the Tenth International Workshop on Natural Language Processing and Cognitive Science* (Marseille, France, October 2013).

15. Marquardt, J., Farnadi, G., Vasudevan, G., Moens, M., Davalos, S., Teredesai, A. and De Cock, M. (2014). Age and Gender Identification in Social Media. *Author Profiling Task at PAN 2014.*

16. Gordon, E. (1997). Sex, speech and stereotypes. Why women use prestige speech forma than men. In *Language in society*, 26 (1). pp. 47-63.

17. Cameron, D. (1998). Gender, Language and discourse: A review. In *Journal of women, Culture and Society,* 23(4). pp. 945-60.

18. Cameron, D. (2005). Language, Gender and Sexuality: Current Issues and New Directions. In *Applied Linguistics*, 26(4).pp.482-502. DOI=10.1093/applin/ami027.

19. Bucholtz, M. (1999). You da man: Narrating the racial other in production of white masculinity. In *Journal of Sociolinguistics,* 3: 443-460.

20. Bucholtz, M., Liang, A. & Sutton, L. (1999).*Reinventing identities: The gendered Self in Discourse*. New York: Oxford University Press.

21. Fishman, P. (1983). Interaction: The work women do. In *Language, Gender and Society*. pp. 89-102. Rowley, Mass.: Newbury House.

22. Lakoff, R. (1990). Talking Power: The Politics of Language in Our Lives. New York: Basic Books.

23. Lakoff, R. (1975). *Language and Women's Place*. New York: Harper and Row.

24. Alami, M., Sabbah, M. &Iranmanesh, M. (2013).Male-Female Discourse Difference in Terms of Lexical Density. In *Research Journal Of Applied Sciences*, Engineering and Technology, 5(23). pp. 5365-5369.

25. Eckert, P. (2012). Three waves of variation study: the emergence of meaning in the study of sociolinguistic variation. In *The Annual Review of Anthropology,* 2012(41). pp. 87-100.

26. Moore E. &Podesva RJ.(2009). Style, indexicality and the social meaning of tag questions. In *Language in Society*,38. pp. 447–85.

27. Bucholtz, M. (2002).From 'Sex Differences' to Gender Variation in Sociolinguistics. In *Papers from NWAV 30* (University of Pennsylvania Working Papers in Linguistics 8.2).University of Pennsylvania, Department of Linguistics. pp. 33-45.

28. Bucholtz, M. (2003). Theories of Discourse as Theories of Gender: Discourse Analysis in Language and Gender Studies. In *The Handbook of Language and Gender*, ed. Holmes, J. &Meyerhoff, M.. pp. 43-68. Oxford: Blackwell.

29. Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology, 57*(3), 378-393.

30. Esuli, A. &Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of Language Resources and Evaluation* (LREC).

31. Witten, I.H., & Frank, E. (2005). Data mining: practical machine learning tools and techniques (2nd ed, Morgan-Kaufman Series of Data Management Systems). San Francisco: Elsevier.