Gender Identification of Blog Authors: Do Men and Women Prefer DifferentCharacter Unigrams?

Vasiliki Simaki[1], Athanasia Koumpouri[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, koumpour, vasilis}@ceid.upatras.gr
[2]imporas@upatras.gr

Abstract. In this paper we present a studyon the topic of gender identification of web userswhich, to the best of our knowledge for first time,investigates the potential existence of character unigrams preferred by men or women. In our study we used the 'Blog author gender classification data set', a well-knowngender-annotated US English corpus. We examined 57 unigrams, in order to examine the existence of gender preferential letters and special characters, after the statistical analysis of the male and female corpus. The discovery of 25 statistically different character unigrams is evidence that, among other, women and men make different linguistic choices in character level.

Keywords: gender identification, n-grams, text processing.

1 Introduction

The expansion of social media is impressive and the information provided is of high volume and continuously updated. It is thus important to classify this information in terms of topic, author, as well as, gender, age, and other demographic characteristic. Users are able to express opinions, make online transactions, and communicate with other users. They leave their digital fingerprints on the web, not only by declaring personal information, but unconsciously through the writing style of the text they produce, their comments, etc. It is a challenging task to identify demographic information on users without taking into account their profile statement,but estimate clues only by using the content of the users' posts/texts. This requires an extensive analysis of the textual information provided and substantial knowledge of the background research on authorship profiling both in theoretical and computational level.

Gender identification is an important field of text mining with scientific interest and commercial applications. The detection of the user's gender is of high importance to companies in order to promote a product or a service if it is more preferable by women or men. Market analysts and advertising professionals are interested in finding

which product or service is more talked or liked between the two groups, and hence, which should be addressed to women or men. Gender identification is also important in e-government services and social science studies. Useful conclusions can be derived about the different trends among women and men, different topics of interest, political views, social concerns, world theories, etc. 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 techniques have been proposed based on theoretical sociolinguistic studies.

In this article we present a study on gender identification of web users which, to the best of our knowledge for the first time, investigates the existence of character unigramsthat are mostly preferred by men and women in written language. We performed a statistical analysis in all character unigrams of the 'Blog author gender classification data set' and we present the statistically significant character unigrams in the male and female data set. The remaining of this article is organized as follows: Section 2 presents the previous work in theoretical and automatic gender identification. Section 3 describes the motivation for a study in gender preferential character, the related/background work in gender identification and the methodology of our study. Section 4 presents the gender preferential character unigrams investigation and the results of the statistical analysis. Finally, Section 5 discusses the results and the conclusions.

2 Previous Work in Gender Identification

As considers sociolinguistics, essential knowledge in support of the task of gender identification has been reported. The research in language and gender is a very old task in the history of sociolinguistics.There are many approaches about the linguistic variation based on gender distinction, and the empirical findings on the field are quite indicative for the existence of differentiated linguistic choices between genders.A general opinion about the women's language is that, statistically, women tend to make a more conservative use of language and they use more standard types than men [1].Compared to men discourse, women tend to use more analytical ways to describe a specific colourtone, they use more frequently "empty" adjectives than men, which carry a metaphorical sense of admiration and/or approval [2, 3, 4]. Women also prefer a more "gentle" way of conversation, by using questions in place of statements. Besides specific lexical choices that women, unlike men, do (use of norm types, avoiding 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 [5, 6]. Women have different politeness strategies than men and different ways to agree/disagree. They also use more sentimental expressions, indirect requests and hypercorrected grammar types [7, 8]. Men on the other hand, tend to use more bad words and slang types, in general coarse language than women, and in case of disagreement, they use strong and explicit expressions[9, 10]. They insert in their vocabulary non-standard forms and neologisms.

Considering the information technology field of study, most researchers perceive author's gender identification as a text classification issue [11, 12]. Koppel et al. [13] proposedclassification 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 combinedstylometric and text classification techniques, in order to extract the author's gender. Argamon et al. [14] applied factor analysis for gender and age classification in texts mined from the blogosphere. Ansari et al. [15]used frequency counting of tokens, tf-idf and POS-tags to find the gender of blog authors. In Burger et al. [16] 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 was used. Many recent studies around gender classificationdeal with social media and propose methods that identify the gender [17, 18, 19] and in some cases the age of the web users [20]. 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. [21] a comparative study of gender attribution, without taking into account the topic or the genre of the selected text is presented. Holgrem and Shyu[22] applied machine learning techniques using a feature vector containing word counts, in order to detect the author's gender of FB status. In Rangel and Rosso [23] 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. [24] evaluated the appropriateness of several feature setsfor age and gender classification in social media.

3 Motivation and Methodology

The scientific interest on gender identification is high and researchers have proposed various approaches on the task. Differential characteristics between male and female speech have been derived and used in severalclassification experiments. Most linguistic characteristics are based on words, syntacticand stylistic information, n-grams, and their occurrence and frequency in texts are calculated in studies we describe in greater detail below. However, none of the previously reported studies has focused on the investigation of characteristics which might be preferential to men and women in an implicit/subconscious way rather than in an explicit way.

Our study aims to examine if gender preferential choices may be made by men and women, even in character level.It is possible that men and women may prefer utterances constituted by gender preferable characters, when it is feasible. This attitude should not be a conscious one, and the existence of such an attitude may be decisive about a hypothesis of differentiated language acquisition and discourse analysis between the genders.

Several studies in gender identification, as described in Section 1, use as features dictionary based types, and observe that women and men make different lexical choices and prefer to use different words. This has been explained in terms of interests and conversation topics, which differ among women and men. Character-level n-grams have been used mostly in authorship attribution studies, in order to prove that

n-grams carry stylistic and lexical information [25]. The length of the n-gram and the definition of the *n*is essential. A large *n* captures best lexical and contextual information, while a small *n* performs better in representing the subword information. Bi-grams and trigrams are considered as small n-grams, while up to 4-grams achieved best results in short English texts [26]. In other studies, character n-gram language modeling techniques are used in order to perform a language independent authorship attribution [27]. Character unigrams were used as featuresonly when they formed words (uni-character words) [28]. To the best of our knowledge, this article is the first effort for an aggregated analysis of all character-level unigrams and their investigation without any other linguistic knowledge.

In our study we calculated all character unigrams, letters and special characters. We statistically analyzed all alphabetic letters –upper and lower case- ("a", "b", "c", "d", "e", "f", "g", "h", "i", "j", "k", "l", "m", "n", "o", "p", "q", "r", "s", "t", "u", "v", "w", "x", "y", "z") and special characters including punctuation ("!", """, "#", "$", "%", "&", "(", ")", "*", "+", ",", "-", ".", "/", ":", ";", "<", "=", ">", "?", "@", "[", "\", "]", "^", "_", "`", "{", "|", "}", "~") constituted a set of 57 unigrams.

4 Gender Preferential Character Unigrams Investigation

The corpus we used in our study is the 'Blog author gender classification data set' [29] which consists of a collection of 3232 blog posts from many blog hosting sites and blog search engines. The researchers extracted the author's gender by using all the provided information, the blogger's profile information, his/hers profile pictures or avatars. The collected posts are equally distributed (half male, half female posts).This is the only corpus publicly available about gender distribution and it is created from a new text type resulted by the expansion of social media. The posts may contain a unique sentence, but in most cases they contain a longer text, covering exhaustively a thematic area. An interesting element is that the dataset covers a large thematic and stylistic range, and it might be useful to extract information gender-associated according to the topic and the style of the posts.

In this study we perform a statistical analysis in order to investigate the male and female preferable characters. The mean calculates the average value for every unigram and the STD the distance outliers from the average value. In order to strengthen our results we perform thereafter a statistical test, thet-statistic test, which is a ratio of the departure of an estimated parameter from its notional value and its standard error [30]. With the t-test, we compare the data of one group to the data from another group. In our case, the first group is the set of documents having a male author and the second is the set of documents having a female author. The t-statistic is calculated as below, where X_n and Y_m arethe mean values of the two samples and, S_X^2 and S_Y^2 are the variations of the two samples accordingly:

$$t = \frac{\overline{X} - \overline{Y}}{S\sqrt{\frac{1}{n} - \frac{1}{m}}}$$

If the t value is positive the first group has a higher value than the second and if the t value is negative, the second group has a higher value than the first one. We performed the unpaired t-test, which compares two unrelated samples (male and female measurements for each character unigram). We are interested in whether the weights of the two samples are different and use the threshold $p < 0.05$, and in that case the sample values are different. The return value is a tuple containing the t-statistic and the p-value and these are the results of a two-sided test.

Table 1. The results for the statistically significant character unigrams

Characters	Male corpus		Female corpus		t test	p value
	Mean	STD	Mean	STD		
y	0.02054	0.00753	0.02287	0.00739	-8.461	0
r	0.05462	0.01	0.05177	0.01	7.6489	0
!	0.00115	0.0042	0.00248	0.006	-6.892	0
w	0.02046	0.00713	0.02224	0.007	-6.689	0
c	0.026102	0.00849	0.024	0.007846	6.4998	0
p	0.018926	0.00702	0.01736	0.006435	6.2569	0
d	0.034991	0.00859	0.03685	0.008784	-5.812	0
s	0.06276	0.0108	0.06058	0.0108	5.4568	0
]	0.000036	0.00024	0	0	5.4523	0
(	0.00357	0.0043	0.0043	0.0042	-5.17	0
k	0.009428	0.00504	0.01028	0.004946	-4.638	0
x	0.00187	0.00208	0.00159	0.00179	3.9131	0.0001
m	0.026255	0.00764	0.02734	0.00774	-3.816	0.0001
h	0.048811	0.01014	0.05010	0.010738	-3.358	0.0008
q	0.00075	0.00125	0.00061	0.001	3.1809	0.0015
.	0.0024	0.0042	0.00199	0.0034	2.9843	0.0029
`	0.000001	0.00002	0.00001	0.000115	-2.932	0.0034
n	0.0658	0.0105	0.06469	0.0102166	2.8899	0.0039
f	0.02041	0.00617	0.01975	0.00622	2.8837	0.004
\	0.000034	0.0002	0.00001	0.00012	2.8606	0.0043
:	0.0004	0.0016	0.00026	0.0011	2.8526	0.0044
^	0.000003	0.00007	0.00001	0.00012	-2.75	0.006
@	0.00076	0.00158	0.0009	0.0017	-2.392	0.0168
,	0.000045	0.00042	0.00001	0.0002	2.2495	0.0246
_	0.00004	0.00075	0.00000	0.000051	1.9639	0.0496

In Table 1, we present the character unigrams that can differentiate women and men, based on the statistical test. From the 57 unigrams investigated, 25 proved to be statistically significant. 14 of them are alphabetic letters and 11 are special characters and punctuation. After the first observation of the differential character, it does not appear that gender preferential letters or special characters come of an optical choice made by women or men. The exclamation mark is highly preferred by women, and that confirms the theoretical hypothesis that women tend to used more exclamatory utterances than men [2]. Men, on the other hand, prefer to make neutral statements which is confirmed by the differential "." character [2]. Concerning the alphabetical characters, we observe that apart from the semivowel [y], all other differential letters are consonants. In Table 2 we sort the significant unigrams according to the t-test results, in terms of male and female preferable unigrams. Concerning the alphabetic unigrams, there is no indication though, that women and men prefer letters after their place and manner of articulation. Both women and men use plosive or fricative, glottal or alveolar letters, ie. [n] and [m], preferred by men and women respectively, are both nasals.

Table 2. Gender preferential characters

Men-preferable unigrams	Women-preferable unigrams
,	!
.	(
:	@
\	^
]	`
_	d
c	h
f	k
n	m
p	w
q	y
r	
s	
x	

5 Conclusions

Classification from text has been dominated by methods which are based on text mining features. In this study, for the first time, we performed a statistical analysis of character-level unigrams in the 'Blog author gender classification data set' corpus. We investigated the existence of differential letters and special characters and we measured 57 character unigrams. It is calculated that 25 unigrams are statistically

significant, 14 men-preferable and 11 preferred by women. 14 alphabetic letters and 11 special characters and punctuation are gender preferential features, and diversify the male and the female corpus. After the results of this study we assume that even in character level, a gender identification research is possible and contributes to the scientific community.

References

1. Gordon, E. (1997). Sex, speech and stereotypes. Why women use prestige speech forma than men. In *Language in society*, 26 (1). pp. 47-63.
2. Makri-Tsilipakou, M. (2010).'Women's language' and the language of women. In *Gender and Social Sciences in Modern Greece*, ed. Kantsas, V., Moutafis, V. &Papataxiarchis, E.. pp. 119-146. Athens: Alexandreia Editions (in greek).
3. Lakoff, R. (1973). Language and women's place.*In Language in Society*, 2. pp. 45-80.
4. Lakoff, R. (1975). *Language and Women's Place*. New York: Harper and Row.
5. 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.
6. Bucholtz, M. and Hall, K. (2005). Identity and interaction: a sociocultural linguistic approach. In *Discourse Stud*, 7.pp. 585– 614.
7. Fishman, P. (1983). Interaction: The work women do. In *Language, Gender and Society*. pp. 89-102. Rowley, Mass.: Newbury House.
8. Lakoff, R. (1990). Talking Power: The Politics of Language in Our Lives. New York: Basic Books.
9. Cameron, D. (1998). Gender, Language and discourse: A review. In *Journal of women, Culture and Society*, 23(4). pp. 945-60.
10. 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.
11. 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.
12. Soler, J. andWanner, L. (2014). How to Use Less Features and Reach Better Performance in Author Gender Identification. In *Proceedings of LREC 2014.*
13. Koppel, M., Argamon, S. and Shimoni, A. R. (2003). Automatically categorizing written texts by author gender. In *Literary and Linguistic Computing*, 17: 401–12
14. 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.
15. 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.
16. 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.

17. 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.
18. Zhang, C., and Zhang, P. (2010). *Predicting gender from blog posts*. Technical Report. University of Massachusetts Amherst, USA.
19. 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.
20. 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.
21. 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.
22. Holgrem, J., Shyu, E. (2013). Gender Classification of Facebook Posts.
23. 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).
24. 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.*
25. Stamatatos, E. (2009). A survey of modern authorship attribution methods.*Journal of the American Society for information Science and Technology,60(3), 538-556.*
26. Sanderson, C., & Guenter, S. (2006, July). Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (pp. 482-491).* Association for Computational Linguistics.
27. Peng, F., Schuurmans, D., Wang, S., & Keselj, V. (2003, April). Language independent authorship attribution using character level language models. In*Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1* (pp. 267-274). Association for Computational Linguistics.
28. Grieve, J. W. (2005). *Quantitative authorship attribution: A history and an evaluation of techniques* (Doctoral dissertation, Department of Linguistics-Simon Fraser University).
29. Mukherjee, A. & 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.
30. Welch, B. L. (1947). "The generalization of "Student's" problem when several different population variances are involved".Biometrika 34 (1–2): 28–35. doi:10.1093/biomet/34.1-2.28. MR 19277.