# Authorship Attribution and Gender Identification in Greek Blogs

George K. Mikros

Department of Italian Language and Literature, School of Philosophy,
National and Kapodistrian University of Athens, Athens, Greece
gmikros@isll.uoa.gr

**Abstract.** The aim of this study is to obtain authorship attribution and author's gender identification in a corpus of blogs written in Modern Greek language. More specifically, the corpus used contains 20 bloggers equally divided by gender (10 males & 10 females) with 50 blog posts from each author (1,000 posts in total and an overall size of 406,460 words). From this corpus we calculated a number of standard stylometric variables (e.g. word length statistics and various vocabulary "richness" indices) and 300 most frequent word and character n-grams (character and word unigrams, bigrams, trigrams). Support Vector Machines (SVM) were trained on this data, and the author's gender prediction accuracy in 10-fold cross-validation experiment reached 82.6% accuracy, a result that is comparable to current state-of-the-art author profiling systems. Authorship attribution accuracy reached 85.4%, an equally satisfying result given the large number of candidate authors (n=20).

**Keywords:** Authorship Attribution, Author profiling, Blogs, Machine Learning, Support Vector Machines, Gender Identification, Stylometry

## 1 Introduction

Over the last two decades Automatic Authorship Identification (AAI) has been evolved in a highly dynamic research strand exploiting recent advances in a number of fields like Artificial Intelligence, Linguistics and Computing. Furthermore, AAI research now is concerned not only with problems of authorship in the broad field of the Humanities (Literature, History, Theology), but also with applications in various law-enforcement tasks such as Intelligence, Forensics e.g. [1-4] etc. The major application areas are described below:

1. Authorship Attribution: This is the most common authorship identification analysis with the study of the *Federalist Papers* by Mosteller & Wallace [5] being a typical example. In this case we are trying to find who is the author of one or more disputed texts among a closed set of 2, 3 … n known authors. This scenario assumes that we are certain that at least one of the possible authors is actually the author of the disputed texts and that an adequate corpus in size and quality for every possible author is available [6].
2. Author verification: In this case we are investigating whether certain text(s) were written by a specific author. We are assuming an open set of authors and each du-

bious document must be attributed to the specific author without reference to corpora from other authors [7-9].
3. Author profiling: In some applications related to Information Retrieval or Opinion Mining and Sentiment Analysis we are interested in identifying the author's gender [10-12], age [13] or psychological type [14-16].

In this paper we will focus in the first and the third type of research, namely authorship attribution and author profiling. More specifically, we will try to detect automatically the author of a blog and his/her gender training machine learning algorithms using data from a Modern Greek blog corpus.

## 2 Language Usage in Blogs

During the last decade the Internet has evolved from a static field of simple information provision into a digital carrier of language production characterized by interactivity and dynamic configuration of the online textual content.

Blogs are among the best known Web tools that have transformed Web communication and overcame the unidirectionality of standard online communication. Up to 2011 approximately 181 million blogs have been created worldwide, producing 900,000 posts every day which are being read by 77% of internet users (Source: NM Incite). Since many blogs are important information nodes and attract many more readers than most of the traditional printed media, they can exert influence in language usage and produce linguistic innovations accelerating linguistic change. For this reason, blog language usage has started to attract attention and become a challenging research subject in the linguistic community.

Blogs represent a new text genre with interesting characteristics. They combine personal views, news and reporting on current events [17]. Their structure is a hybrid containing both monologue and dialogue features. At the same time they are both log entries reflecting personal opinions and open calls for public discussion [18]. Mishne [17] studied in detail various properties of linguistic usage in English blogs and showed that they present increased usage of personal pronouns and words relating to personal surroundings emerging from personal experience. Furthermore, he examined the linguistic complexity of the blogs using the perplexity measure [19] and the out-of-vocabulary rate (OOV) and found that their linguistic structure was more complex than most of the similar written genres (e.g. personal correspondence). Increased perplexity, according to Mishne, equates with increased irregularity in linguistic usage (i.e. free-form sentences, decreased compliance with grammatical rules etc.). In addition, blogs presented increased OOV rates, meaning that blog texts exhibit a topical diffused vocabulary, with many neologisms, possible typographical errors and increased level of references to named entities from the blogger's personal environment.

Another interesting characteristic of the blog's linguistic structure is its equilibrium between spoken and written language. Sentence construction in blogs is highly variable using selectively structures from both spoken and written norms [20]. An equally important effect in language usage in blogs is the age of the bloggers. Half of the

them are aged 18-34 (Source: The Social Media Report: Q3 2011, MN Incite, Nielsen). For this reason, formality in language usage is decreased, with shorter that average sentence lengths and lower readability scores in the best-known readability formulas (Gunning-Fog, Flesch-Kincaid, SMOG).

## 3 Gender Identification in Blogs. A Literature Review

Blogs' textual production is increasing rapidly. At the same time anonymous posting often covers illegal acts ranging from copyright infringement to criminal offences. AAI methods can be effectively employed in the framework of Forensic Linguistics. Due to their special linguistic structure described in the previous section, anonymous blog posts represent a serious challenge for both the stylometric features and the machine learning methods used to reveal a malicious blogger's identity [10, 11, 21, 22].

The detection of the blogger's gender is an equally important research issue with many possible applications including forensics, online audience identification for targeted advertisement and socio(linguistic) analysis on gender identity issues.

Schler et al. [10] used a large blog corpus (37,475 blog posts totaling 300 million words) and tried to predict both the authors' gender and age. The specific study used 1,502 features including specific content words, selected parts-of-speech, function words and blogs specific features such as "blog words" - lol, haha, ur etc. - and hyperlinks. The machine learning algorithm used was Multi-Class Real Winnow and the prediction accuracy for the author's gender reached 80.1%. Interestingly, the authors noted that despite the great diversity found among stereotyped word content usage between men and women, the most important gender distinctive features were semantically neutral (such as frequent functional words and Parts of Speech).

Argamon et al. [11] have also examined how age and gender affect writing style and topic in blog postings. They presented an analysis based on 140 million words mined from 46,747 English language blogs. They extracted the 1,000 most frequent words from this corpus and recorded their frequency in each blog. Using these data they performed a factor analysis in order to find groups of related words that tend to occur in similar documents. Results indicated that women bloggers prefer personal pronouns, conjunctions and auxiliary verbs while male bloggers use more articles and prepositions. Prediction accuracy of the bloggers' gender using the 1,000 most frequent words reached 80.5%. The researchers, however, warn that style and content effects are highly correlated and it may be that the choice of content determined particular style preferences, or both content and style may be influenced by a single underlying variable such as genre preference.

In another study [23], 73 Vietnamese bloggers' gender was predicted using a variety of machine learning algorithms and stylometric features based on character and word units. The classification accuracy for gender reached 83.3% with the word-based features to contribute more to the gender identification than the character-based features.

Mohtasseb and Ahmed [24] studied a large number of demographic characteristics of authors including gender in blog texts. They trained Support Vector Machines

(SVMs) using various standard stylometric indices and 88 features from the Linguistic Inquiry Word Count (LIWC) [25], a special psycholinguistic lexical database that groups words into specific psychological categories. Results indicated that men's posts could be recognized more accurately that women's under all experimental conditions.

Mukherjee and Liu [26] also studied author gender classification in blog posts. They proposed a new class of features which are POS sequence patterns that are able to capture complex stylistic regularities of male and female authors. Furthermore, they proposed an ensemble feature selection method which takes advantage of many different types of feature selection criteria. These methods were tested in 3,100 blog posts and compared against known public domain gender detection systems (Gender Genie, Gender Guesser) and relative published algorithms [10, 11, 27]. In all cases their proposed methodology proved considerably more accurate.

Sarawgi et al. [28] studied the effect of text topic and genre in the accuracy of automatic gender identification methods. Using a sophisticated experimental design and multiple datasets (mostly blogs of different topics), they compared multiple machine learning methods controlling for genre and topic bias. They noticed that the most robust approach was based on character-level language models which used morphological patterns, rather than token-level language models that learned shallow lexico-syntactic patterns. In addition, they traced statistical evidence of gender-specific language styles beyond topics and genre, and even in modern scientific papers.

## 4 Research Methodology

### 4.1 The Greek Blog Corpus (GBC)

In order to explore authorship attribution and gender identification in Greek blogs we had to develop from scratch a Greek blog corpus (GBC). For this reason we harvested the Greek blogosphere from September 2010 till August 2011 and manually collected 100 Greek blogs equally divided to 50 male and 50 female bloggers. Since topic can induce significant bias into stylometric measurements [29], we decided to explore only a part of the collected corpus, using blogs that share the same topic. In this study we used 20 blogs (10 male and 10 female authors) with a common topic (Personal affairs), with a total of 1,000 blog posts counting 406,460 words. For each author we collected the 50 most recent blog posts.

A close examination of the word length descriptive statistics reveals that male and female bloggers produce texts that vary considerably in size even when the topic is roughly the same. Female (fm) bloggers produce longer posts with less variation in size (Mfm=423.4 words, SDfm=243.6) than male (ma) bloggers (Mma=389.5, SDma=351.1).

### 4.2 Stylometric Features and Classification Algorithm

Authorship attribution has a long history of using a large variety of textual features in order to correlate them with a specific author's style. In the present study we will use

a wide set of stylometric features in order observe their association with authorship and author gender. The feature list we mined is extensive and contains both "classic" stylometric features such as lexical "richness" and word length measures, and "modern" features borrowed from Information Retrieval and Language Modeling such as character and word n-grams. The detailed list of the features used in this study is the following:

"Classic" stylometric features

- Vocabulary "richness"
  - Yule's K, [30]
  - Functional Density, [31]
  - Percentage of Hapax and Dis-legomena
  - Ratio of Dis to Hapax-legomena, [32]
  - Lexical Entropy and Redundancy, [33]
- Word Length
  - Average Word Length – AWL (in characters)
  - Standard Deviation of Average Word Length – sd AWL
  - Word Length Spectrum: Normalized frequency of 1, 2, 3 … 14-letter words.
- Letter frequencies
  - Normalized frequencies of each letter.

"Modern" features

- Character bigrams
- Character trigrams
- Unigrams (words)
- Word bigrams
- Word trigrams

For each character and word n-gram feature group described above we counted the 300 most frequent features and normalized their frequency in 100-word text size. Feature counting was performed using customized PERL scripts and the total vector size produced was 1,356 features.

This vector fed the Sequential Minimal Optimization (SMO) algorithm [34], an optimized version of the Support Vector Machines (SVMs). SVMs represent the state-of-the-art in machine learning methods regarding text classification and have been used extensively in authorship attribution research [35-38]. They are suited for solving binary classification problems, though there are many extension methods that make them appropriate also for multi-class problems. They project the points of the training sample to a higher dimension area and find a hyperplane that separates with the best possible way the points of the two classes. Points from the testing sample are classified according to the side of the hyperplane in which they are located. Vectors which define the hyperplane are called support vectors.

Evaluation of the classification performance was obtained using accuracy, i.e. percentage of the texts that were attributed correctly to their author, or author's gender.

In order to avoid random fluctuations in algorithm performance we used 10-fold cross-validation methodology, i.e. we took the mean accuracy of 10 different complementing training and testing cycles with each cycle to use 90% of the data as training sample and 10% as validation sample.

## 5 Results

Using the features and the algorithm described in the previous section we had 85.4 accuracy in authorship attribution and 82.6 in gender identification. Both reported accuracies can be considered as excellent regarding the data size and the number of candidate authors (n=20). This last parameter is very important since two-class authorship attribution problems are less demanding and most stylometric methods can successfully deal with them.

In order to understand better the impact of the number of candidate authors on the evolution of the authorship attribution accuracy we created a controlled experiment. We segmented our data into 4 size groups (2 authors, 4 authors, 8 authors, 16 authors). For each size group we selected 10 different author combinations using stratified random sampling. Reported accuracy measures are based on the mean of these 10 different author combinations in each size group. This method minimizes systematic errors which can intervene due to an unusual stylistic (dis)similarity between specific authors. In total we ran 40 (4 x 10) classification experiments using SMO in 10-fold cross-validation scheme. The mean accuracies are displayed in figure 1.

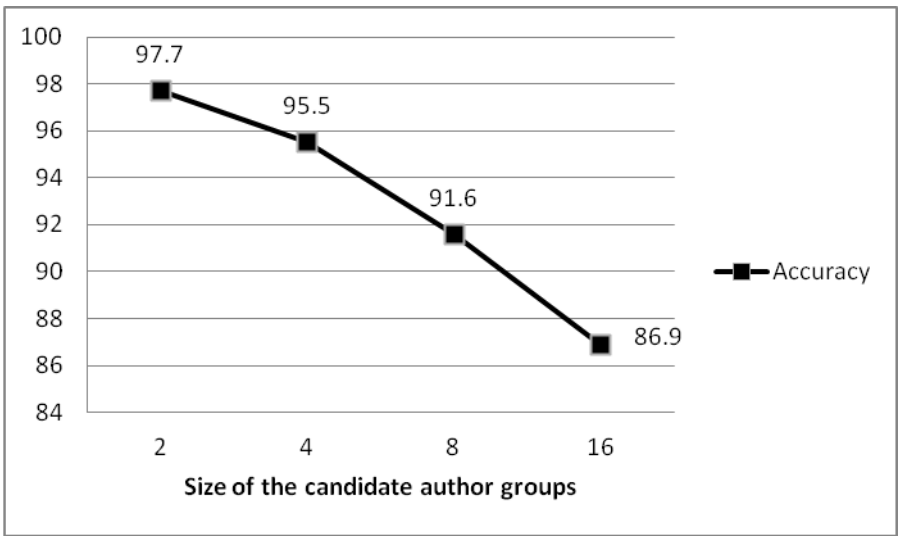


**Fig. 1.** Influence of the number of candidate authors

Classification performance is directly related to the number of candidate authors. When we examine 2 candidate authors the obtained classification accuracy is very

high (97.7%). Accuracies drop as the number of candidate authors increases with the lowest accuracy reported in the 16-author group (86.9%). In order to evaluate further the impact of candidate group size to the classification accuracy, we examined the full experimental data using one-way ANOVA with dependent variable the obtained accuracies and independent variable the group size. Results were statistically significant at the 0.05 level ($F(3, 36)=53.4$, $p<0.05$) indicating that overall means in the different group sizes are indeed different. In order to further explore which specific group sizes differentiate, we applied the Tukey post hoc test. Results indicated that all group sizes differ statistically significantly between themselves except the 2 and 4-groups. This means that authorship attribution accuracy using the above mentioned combination of features and algorithm performs its best up to 4 candidate authors and then its performance drops linearly as the number of the authors is increasing exponentially.

Another important research question we confronted was related to the influence of the stylometric features in each type of attribution, i.e. authorship and gender. We applied Information Gain [39], a well-known feature selection algorithm for text classification tasks and recorded the 10 most influential features in authorship attribution and gender identification task. The relative importance of each feature in the two classifications is displayed in figure 2.

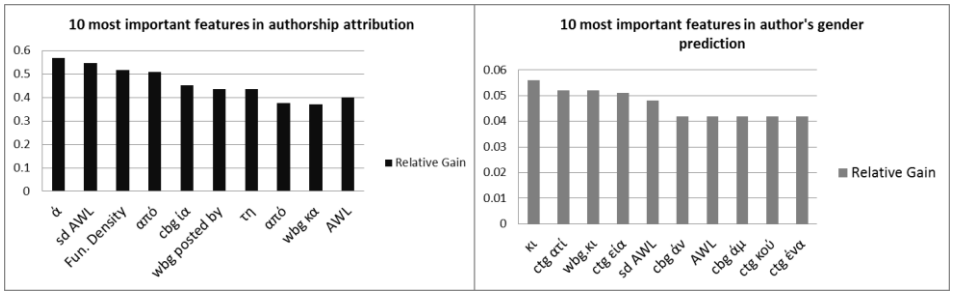


**Fig. 2.** 10 most important stylometric features for authorship attribution and gender identification.

A general conclusion that can be drawn from examining feature importance in the two classification tasks is that specific character n-grams carry significant authorship information while specific word n-grams have increased importance in author gender identification. Another finding that deserves comment is that word length measures (AWL, sd AWL) convey both authorship and gender evidence.

In order to explore further the way n-grams reveal authorship and gender patterns we performed authorship and gender classification using only these as features. We recorded the classification accuracy first using all n-gram features and in a second step we performed classification without a specific n-gram feature group. We subtracted the new accuracy from the one that was based on all n-grams, resulting in a relative difference that could be explained as the importance of the feature group that was missing, i.e. the larger the difference, the larger the importance of this feature group in the classification. We calculated all these differences by removing sequentially all n-gram feature groups one at a time for both classification tasks (authorship

and gender). N-gram importance in relation to the classification task is displayed in figure 3.

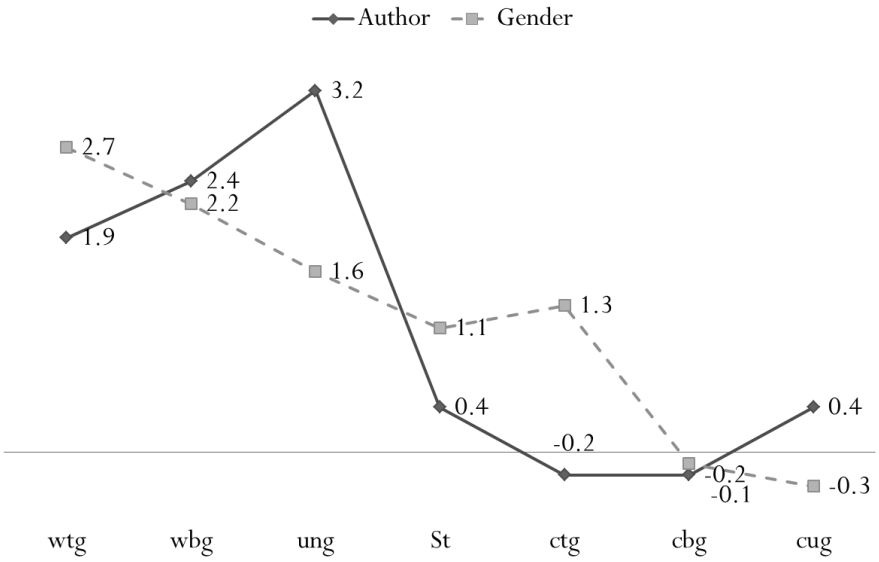


**Fig. 3.** Relative importance of n-gram feature groups in the authorship attribution and the gender identification task.

In the above chart we observe two different tensions regarding word n-grams. As we move from words (ung) to word bigrams (wbg) and trigrams (wtg) gender identification is getting more accurate. The exact opposite trend can be observed in authorship identification, where increasing the size of word sequences leads to a linear drop in accuracy. This trend could be related to the way the syntax is connected to these tasks. It seems that gender distinctions are associated with specific syntactic patterns while authorship is based more on most frequent words usage.

Character n-grams follow a similar trend. As we move towards longer sequences, gender identification becomes more accurate, meaning that morphological information is highly relevant to the way gender is manifested in a text. On the other hand, simple character frequency is the most productive feature group among sub-word features in authorship attribution. These trends in word and character n-grams reveal that authorship and gender classifications are quite different tasks which utilize complementary linguistic means. Author gender finds expression using specific morpho-syntactical patters which differentiate male from female authors. Authorship on the other hand, is based on the selection of high frequency words and their idiosyncratic usage by each author. This phenomenon is partly reflected in the representation of specific characters and their derived usefulness in authorship attribution, since specific very frequent words increase the frequency of their constituent characters. This complementarity, however, is not absolute. N-grams function as markers of both authorship and gender, and their increased discriminatory power in each of these tasks is

just an indication that gender and authorship exploits more or less specific elements of the grammatical spectrum.

## 6 Conclusions

The present study has investigated methods for authorship attribution and gender identification in Greek blogs using state-of-the-art machine learning algorithm (SVM) and a large variety of stylometric features. Authorship attribution and author gender prediction in blog posts reached reasonable accuracy (85.4% & 82.6%) with many candidates (n=20).

Furthermore, the relation of authorship attribution accuracy to the number of candidate authors was examined. Using a controlled experimental design our methodology performed optimally up to 4 candidate authors. From this point, authorship attribution accuracy dropped linearly as the number of candidate authors was increased exponentially.

Another finding of this study was that author identification and gender detection are two different tasks with distinct patterns of stylometric feature interaction. As we moved towards longer lexical chains (bigger word n-grams), we noticed an increase in the author's gender identification accuracy. An opposite trend was spotted in the authorship classification task. As we moved towards single words (unigrams), we noticed an increase in author identification accuracy. The same trend was detected in the character n-grams. Longer sequences of character n-grams led to a better accuracy rate in gender identification while shorter n-grams and single characters boosted accuracy in authorship attribution. These observations lead us to the conclusion that author gender is conveyed through specific syntactical and morphological patterns while authorship seems to rely on over- or under-representation of specific high frequency words.

## 7 References


1. de Vel, O., Anderson, A., Corney, M.W., Mohay, G.: Multi - Topic E-mail Authorship Attribution Forensics. Proceedings of ACM Conference on Computer Security - Workshop on Data Mining for Security Applications. Philadelphia, PA, USA (2001).
2. Chaski, C.E.: Who's at the keyboard? Authorship attribution in digital evidence investigations. International Journal of Digital Evidence 4(1), pp. 1-13 (2005).
3. Iqbal, F., Binsalleeh, H., Fung, B.C.M., Debbabi, M.: Mining writeprints from anonymous e-mails for forensic investigation. Digital Investigation 7(1-2), pp. 56-64 (2010).
4. Li, J., Zheng, R., Chen, H.: From fingerprint to writeprint. Communications of the ACM 49(4), pp. 76-82 (2006).
5. Mosteller, F., Wallace, D.L.: Applied bayesian and classical inference. The case of The Federalist Papers. 2nd ed. Springer-Verlag, New York (1984).
6. Juola, P.: Authorship attribution. Foundations and Trends® in Information Retrieval 1(3), pp. 233-334 (2008).

7. Koppel, M., Schler, J.: Authorship verification as a one-class classification problem. Proceedings of 21st International Conference on Machine Learning, July 2004, pp. 489-495. Banff, Canada (2004).
8. Van Halteren, H.: Author verification by linguistic profiling: An exploration of the parameter space. ACM Transactions on Speech and Language Processing (TSLP) 4(1), pp. 1-17 (2007).
9. Iqbal, F., Khan, L.A., Fung, B.C.M., Debbabi, M.: E-mail Authorship Verification for Forensic Investigation. Proceedings of the 2010 ACM Symposium on Applied Computing (SAC '10), March 22-26, 2010, Sierre, Switzerland, pp. 1591-1598. ACM, New York (2010).
10. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of age and gender on blogging. Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, 27-29 March 2006, Stanford, California, pp. 199-205. (2006).
11. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Mining the blogosphere: Age, gender and the varieties of self–expression. First Monday, 12(9). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2003/1878 (2007).
12. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), pp. 401-412 (2002).
13. Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, genre, and writing style in formal written texts. Text 23(3), pp. 321-346 (2003).
14. Luyckx, K., Daelemans, W.: Personae: A corpus for author and personality prediction from text. In: Calzolari, N., et al. (eds.). Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), 28-30 May 2008. Marrakech, Morocco (2008).
15. Luyckx, K., Daelemans, W.: Using syntactic features to predict author personality from text. Proceedings of Digital Humanities 2008 (DH 2008), pp. 146-149. (2008).
16. Argamon, S., Dhawle, S., Koppel, M., Pennebaker, J.: Lexical predictors of personality type. Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America: Theme: Clustering and Classification, 8-12 Jun 2005. St. Louis, MO (2005).
17. Mishne, G.: Applied Text Analytics for Blogs. University of Amsterdam, Amsterdam (2007).
18. Nilsson, S.: The function of language to facilitate and maintain social networks in research weblogs. Umeå Universitet (2003).
19. Brown, P.F., deSouza, P.V., Mercer, R.L., Della Pietra, V.J., Lai, J.C.: Class-based *n*-gram models of natural language. Computational Linguistics 18(4), pp. 467-479 (1992).
20. Chafe, W., Danielewicz, J.: Properties of spoken and written language. In: Horowitz, R., Samuels, J.S. (eds.). Comprehending oral and written language, pp. 83-113. Academic Press, New York (1987).
21. Mohtasseb, H., Ahmed, A.: Two-layered Blogger identification model integrating profile and instance-based methods. Knowledge and Information Systems, pp. 1-21 (2011).
22. Mohtasseb, H., Ahmed, A.: More blogging features for author identification. Proceedings of the International Conference on Knowledge Discovery (ICKD'09), 6-8 June 2009, Manila, Philippines pp. 534-539. (2009).
23. Dang Duc, P., Giang Binh, T., Son Bao, P.: Author profiling for vietnamese blogs. Asian Language Processing, 2009 (IALP '09), pp. 190-194. (2009).
24. Mohtasseb, H., Ahmed, A.: The Affects of Demographics Differentiations on Authorship Identification. In: Ao, S.-I., Gelman, L. (eds.). Electronic Engineering and Computing Technology, Vol. 60, pp. 409-417. Springer, Heidelberg (2010).

25. Pennebaker, J.W., Francis, M.E.: Linguistic Inquiry and Word Count: LIWC2001. Erlbaum Publishers Mahwah, NJ (2001).
26. Mukherjee, A., Liu, B.: Improving gender classification of blog authors. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 207-217. Association for Computational Linguistics, Cambridge, Massachusetts (2010).
27. Yan, X., Yan, L.: Gender classification of weblog authors. Computational Approaches to Analyzing Weblogs, 27-29 March 2006, Stanford University, California, USA, pp. 228-230. American Association of Artificial Intelligence, (2006).
28. Sarawgi, R., Gajulapalli, K., Choi, Y.: Gender attribution: tracing stylometric evidence beyond topic and genre. Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pp. 78-86. Association for Computational Linguistics, Portland, Oregon (2011).
29. Mikros, G.K., Argiri, E.K.: Investigating topic influence in authorship attribution. In: Stein, B., Koppel, M., Stamatatos, E. (eds.). Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, Vol. 276, pp. 29-35. CEUR, Amsterdam, Netherlands (2007).
30. Tweedie, F.J., Baayen, H.R.: How variable may a constant be? Measures of lexical richness in perspective. Computers and the Humanities 32(5), pp. 323-352 (1998).
31. García, A.M., Martín, J.C.: Function Words in Authorship Attribution Studies. Literary and Linguistic Computing 22(1), pp. 49-66 (2007).
32. Hoover, D.L.: Another Perspective on Vocabulary Richness. Computers and the Humanities 37(2), pp. 151-178 (2003).
33. Oakes, M.P.: Statistics for corpus linguistics. Edinburgh University Press, Edinburgh (1998).
34. Platt, J.C.: Fast training of support vector machines using sequential minimal optimization. In: Schölkopf, B., Burges, C.J.C., Smola, A.J. (eds.). Advances in kernel methods, pp. 185-208. MIT Press, Cambridge (1999).
35. Diederich, J., Kindermann, J., Leopold, E., Paass, G.: Authorship Attribution with Support Vector Machines. Applied Intelligence 19(1), pp. 109-123 (2003).
36. Escalante, H.J., Solorio, T., Montes-y-Gómez, M.: Local histograms of character N-grams for authorship attribution. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, pp. 288-298. Association for Computational Linguistics, Portland, Oregon (2011).
37. Houvardas, J., Stamatatos, E.: N-Gram Feature Selection for Authorship Identification. In: Euzenat, J., Domingue, J. (eds.). Artificial Intelligence: Methodology, Systems, and Applications, Vol. 4183, pp. 77-86. Springer Heidelberg (2006).
38. Zheng, R., Li, J., Chen, H., Huang, Z.: 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), pp. 378-393 (2006).
39. Yang, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. In: Fisher, D.H. (ed.). Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97), pp. 412-420. Morgan Kaufmann Publishers Inc., San Francisco, CA. (1997).