Paul D Farrell. Automating Author Gender Identification from Blogs. A Master's Paper for the M.S. in IS degree. April, 2015. 27 pages. Advisor: Stephanie Haas

The rapid growth of public blogging on the Internet has opened up a vast trove of information that can be text mined for potential insights. This study explores the potential of automating blog author gender based on differences in lexical expressions. The results of this study were mixed, and further refinement is needed.

Headings:

Text Mining

Natural Language Processing

# AUTOMATING AUTHOR GENDER INDENTIFICATION FROM BLOGS

by
Paul David Farrell

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April, 2015

Approved by:

_____

Stephanie W. Haas

## TABLE OF CONTENTS

**Introduction:**

The demand for online blogging services has grown to be a popular medium for people to express their views on the Internet. Blogs are typically noted for their casual and informal writing style. The rapid rise of blogging poses interesting new ways to study informal online communication. One area of active research is trying to identify the gender identification of blog writers. Do men and women express themselves lexically differently on blogs, and if so what are the differences in forms of expression? The vast growth of public blogs has provided a trove of information that can be text mined for potential patterns, but does pose a challenge due to the high volume of data. *If a highly accurate classifier that can detect author gender identification from blog postings could be created then you could quickly automate new blog postings based on differences in lexical expression.* One use case for exploring a single classifier approach is for marketing. Burger, Henderson, Kim, and Zarrella (2011) state that "accurate prediction of these features would be useful for marketing and personalization concerns, as well as for legal investigation" (pg. 1). The potential for creating an automatized classifier for author gender classification is one area actively being developed for commercialization. Mukherjee and Liu (2010) point out:

> It can help the user find what topics or products are most talked about by males and females, and what products and services are liked or disliked by men and women. Knowing this information is crucial for market intelligence because the information can be exploited in targeted advertising and also product development. (p.207)

The current trend in blogging suggests that many of the active bloggers are young adults according to Newson and Oberlander (2006), “while each gender accounts for about half of all weblogs, personal blogs are dominated by females of teen age and preferred by females in general” (p.2). This young adult demographic is a prime area for marketers to target their advertisements. Indeed having such a system would be a powerful tool that would allow businesses and information professionals to track their online presence with men or women, and potentially structure their messages to each gender respectively.

**Literature Review:**

There has been a large research interest in the natural language processing (NLP) and text mining communities about how to leverage the rich information found in blogs. Zhang and Zhang (2010), state that “from a research perspective, blog author classification is also an interesting problem. Blogs differ tremendously from formal texts, since they have many informal sentences, grammar errors, slang words, and phrases, and wrong spellings” (p.1). All of these characteristics make automated author gender classification a challenging endeavor.

Researchers have shown that there are semantic differences between how men and women write in blogs. According to Goswami, Sarkar, and Rustagi (2009), “female authors tend to use pronouns with high frequency, and male authors tend to use numeral and representation related numbers with high frequency” (p. 214). Herring and Paolillo (2006) add that “determiners are more common among male writers” (pg. 444). However, despite these syntactic clues of author gender, Mohtasseb and Ahmed highlight

that “scaling existing solutions with the huge, and increasing number of authors is a challenge” (pg. 2).

Researchers have tried tackling the problem of automating blog author gender detection in several ways. Schler, Koppel, Argamon, and Pennebaker (2005), frame the problem that “broadly speaking, two different kinds of potential distinguishing features can be considered: style-related and content-related” (p. 1). Style related features are “selected parts-of-speech, function words and blogging specific features such as “blog words” and hyperlinks.” (Schler, Koppel, Argamon, and Pennebaker, 2005, p.1). “Blog words” refers to common abbreviations or slang found in many blogs informal language usage. Another style feature is average sentence length. However, Rustagi, Prasath, Goswami, and Sarkar,(2009) found that “average sentence length itself is not a good feature to predict the variation as there is a wide variation in sentence length in informal writing”(pg. 211). Variation in this context refers to the fact that average sentence length alone is not a sufficiently distinguishing feature for a model, but one of many feature indicators for an author gender model. With content features you are measuring the frequency of specific terms in the corpus by each gender**.**

Current popular classifiers utilized for blog author gender classification fall into either Naive Bayes or Support Vector Machine (SVM), along with 10 fold cross validation. Yan and Yan (2006) created a promising classifier using “Naive Bayes classification approach to identify genders of weblog authors. In addition to features employed in traditional text categorization, we use weblog-specific features such as web page background colors and emoticons”(p.1). Vel, Corney, Anderson, Mohay (2002) write, “one advantage of SVM’s is that they do not require a reduction in the number of

features in order to avoid the problem of overfitting, which is useful when dealing with large dimensions as encountered in the area of text mining"(pg. 5). The common metrics used for evaluation of accuracy in the literature are precision, recall, and F-measure.

Another critical component is using POS n-grams to try and identify common words that may be used by males and females. POS stands for parts of speech and refers to the grammatical classes of words such as verbs, nouns, etc. POS n-grams help statistically track the frequency and distribution of these grammatical classes in a text. The methods and approaches listed above have all provided possible solutions to automating blog author identification. However, Mukherjee and Liu (2010) note that "although there have been several existing papers studying the problem, the current accuracy is still far from ideal" (p. 216). Mohtasseb and Ahmed point out that "the complexity of text and the high percentage of new words motivate us to focus more on these new or misspelling words that could appear in the text"(pg. 2). Adding an additional layer of complexity, Perseus asserts "the majority of blogs started are dissolving into static, abandoned web pages"(as cited in Pendersen and Macafee, 2007). If any such single classifier is feasible it will face persistent needs for refinement due to the mix of both current blogs emerging with new words balanced against potentially outdated lexical expressions found in stagnant older blogs.

Preparing a balanced gender dataset across blog domains is a difficult endeavor. Mukherjee and Liu (2010) outline their data preparation for testing in the following manner:

> To keep the problem of gender classification of informal text as general as possible, we collected blog posts from many blog hosting sites and blog search

> engines, e.g., blogger.com, technorati.com, etc. The data set consists of 3100 blogs. Each blog is labeled with the gender of its author. The gender of the author was determined by visiting the profile of the author. Profile pictures or avatars associated with the profile were also helpful in confirming the gender especially when the gender information was not available explicitly. To ensure quality of the labels, one group of students collected the blogs and did the initial labeling, and the other group double-checked the labels by visiting the actual blog pages. Out of 3100 posts, 1588 (51.2%) were written by men and 1512 (48.8%) were written by women. The average post length is 250 words for men and 330 words for women.(p. 214)

Mukherjee and Liu crafted the corpus to be " as general as possible". This poses a potential problem since the dataset does not take into consideration domain balancing. These 3100 blog posts were drawn from various blogging sites that covered different domains. The domains in the corpus ranged from travel blogs to technology. In order to have a highly accurate classifier that covers various domains, it is imperative to have a sizable amount of balanced instances.

Another limitation to Mukherjee and Liu's corpus is that the blog author age was not taken into consideration. The potential for capturing a disproportionate amount of younger bloggers is high. To craft a truly generalizable classifier for detecting blog authors genders it is imperative to have a balanced amount of instances across all age demographics. Kobayashi, Matsumura, Ishizuka (2006) describe "trends for men and for women are definitely different in most domains, and the same can be said for bloggers' age, residential area etc. The problem is derived from the lack of bloggers' personal information because such information is not opened to the public in general"(pg. 1). Argamon, Koppel, Pennebaker, and Schler(2007) research found "that a number of stylistic and content–based indicators are significantly affected by both age and gender, and that the main difference between older and younger bloggers, and between male and female bloggers, lies in the extent to which their discourse is outer– or inner–

directed"(pg. 1). Not having the blog authors' ages eliminates any possibility of identifying gender author nuances across age ranges.

## 3. Methodology:

### 3.1 *Data Preparation*

The corpus of blogs that I used for my research was a publicly provided dataset provided by Bing Liu and Arjun Mukherjee. The dataset was used in their paper *Improving Gender Classification of Blog Authors*. Each instance from the dataset was manually created with two column identifiers. The first column is text, which is designated for the actual prose of the blog. The second column is the specific gender class associated for each text entry as either a male author or female author. In the original dataset, there were more posts authored by males than females. I removed 76 randomly selected posts authored by men, to create a set of 3,024 posts, half of which were authored by men, and half by women. Next, I randomly distributed 80% of the corpus as training data and 20% as test data into two separate csv files. This left 2,420 posts for training and 604 posts for testing. Both training and test data sets had equal distributions of male and female blog posts of 1210 and 302 posts in each respectively. Finally, I repeated the same process above, to create a 90% training and 10% testing set. The training data had 2,722 posts and the test data 302 posts.

**Table 1: Number of posts in each corpus before and after gender balancing**

| Original Corpus Size | Gender Balanced Corpus |
|---|---|
| 3,100 | 3,024 |

**Table 2: Number of posts in the training and testing partitions**

| Corpus Partition | Training Posts | Testing Posts |
|---|---|---|
| **80% Train 20% Test** | 2,420 | 604 |
| **90% Train 10% Test** | 2,722 | 302 |

### 3.2 *Data Mining Tools*

The main tool I used was a free open source product called LightSIDE (Mayfield & Rose, 2012). LightSide supports a variety of features, such as Words/POS, and line length, it supports the most common algorithms, as well as a variable number of folds for cross-validation. Additionally, I selected Weka, a popular text mining program for my testing. I selected Weka to provide as a contrast to LightSIDE to see if there was any difference in model performance by algorithm.

I used the following for performance measures for my classifier evaluation: Accuracy, Kappa, Recall, Precision, F-Score, and Correlation. Accuracy is the measurement of how close the measured value is to the (true) value. Kappa measures the strength of inter rater agreement. The agreement here is between the system and the class associated with the post in the original corpus. A Kappa of 1 would be perfect agreement. The closer the Kappa score is to 1 the higher the agreement. Recall is the percentage of

correct items that are selected, for example, the proportion of actual female-authored posts identified as such by the system. Conversely, Precision is the percentage of selected items that are correct, for example, the proportion of positively predicted male-author posts that were identified as correct by the model. F-Score is a balanced average between precision and recall. An F-Score of 1 is the highest score and 0 is the lowest. The higher the F-Score the better the predictive power of the classification procedure, for example, an F-Score of 1 for male-author posts means that the classification procedure is perfect. Finally, correlation is a measure of relation between two or more variables. A correlation of 1 is the best score, and indicates one variable can be predicted to be the value of the other variable, for example a classifier identifies the word "mother" has a correlation of 1 for female-author posts. That would mean every time the word "mother" appeared in a text the system would identify the text as being a female-author post. A correlation of -1 is equally important, it indicates an inverse relationship between variables. A correlation of 0 indicates absolutely no correlation between variables.

I used Naive Bayes and Support Vector Machine (SVM) classifiers for my experiment. The literature review noted that these are the two most common classifiers used by researchers. I also tried different configuration options to try to glean the best performance.

- Configuration 1: unigrams, Words/POS, Include Punctuation, Ignore All Stopword in N grams, and Track Feature Hits.
- Configuration 2: bigrams, Words/POS, Include Punctuation, and Track Feature Hits.

- Configuration 3: trigrams, Words/POS, Include Punctuation, and Track Feature Hits.

Additionally, I used 10 fold Cross Validation to test the model performance and set a benchmark for comparison. The features were the individual words in the corpuses and I used a word frequency threshold of five. In other words, if a particular word occurred less than five times in the training corpus, it would not be counted as a feature when creating the classification model.

## 4. Results:

### *4.1 Test Results Overview*

The following tests were run from February 3, 2015 to February 6, 2015. Configuration 1 which included unigrams, Words/POS, Include Punctuation, Ignore All Stopword in N grams, and Track Feature Hits gave the best performance. Naive Bayes was the best performing classifier overall configurations with the 90% training and 10% testing performing the best. Below are the results for each configuration. The first set of tables gives Accuracy and Kappa scores for each configuration, broken out by training/testing set and classifier. The second set of tables give recall, precision, and F measures for each configuration, training/testing set and classifier, broken out by male and female. In other words, these tables show differences in classification performance in identifying male and female language. The third set of tables shows the most highly correlated words for male and female language for the highest performing classifier.

### 4.2 *Test Results Accuracy and Kappa*

**Table 3:** Configuration 1, (unigrams, Words/POS, Include Punctuation, Ignore All Stopword in N grams, and Track Feature Hits)

| **90% Train 10% Test** | **Accuracy** | **Kappa** |
|---|---|---|
| Naive Bayes | **.8113** | **.6225** |
| SVM | .6887 | .3775 |
| Weka Naive Bayes | .7185 | .4371 |
| Weka SVM | .6589 | .3179 |
| **80% Train 10% Test** | | |
| Naive Bayes | .654 | .3079 |
| SVM | .5894 | .1788 |
| Weka Naive Bayes | .654 | .3079 |
| Weka SVM | .5861 | .1722 |

**Table 4:** Configuration 2, (bigrams, Words/POS, Include Punctuation, Ignore All Stopword in N grams, and Track Feature Hits)

| **90% Train 10% Test** | **Accuracy** | **Kappa** |
|---|---|---|
| Naive Bayes | .8046 | .6093 |
| SVM | .7285 | .457 |
| Weka Naive Bayes | .8046 | .6093 |
| Weka SVM | .6722 | .3444 |
| **80% Train 20% Test** | | |
| Naive Bayes | .6093 | .2185 |
| SVM | .6026 | .2053 |

| | | |
|---|---|---|
| Weka Naive Bayes | .6093 | .2185 |
| Weka SVM | .5844 | .1689 |

**Table 5:** Configuration 3, (trigrams, Words/POS, Include Punctuation, Ignore All Stopword in N grams, and Track Feature Hits)

| **90% Train 10% Test** | **Accuracy** | **Kappa** |
|---|---|---|
| Naive Bayes | .798 | .596 |
| SVM | .702 | .404 |
| Weka Naive Bayes | .798 | .596 |
| Weka SVM | .6523 | .3046 |
| **80% Train 10% Test** | | |
| Naive Bayes | .5977 | .1954 |
| SVM | .5894 | .1788 |
| Weka Naive Bayes | .5977 | .1954 |
| Weka SVM | .5844 | .1689 |

### 4.3 *Recall, Precision, and F-Score Performance by Gender*

**Table 6:** Configuration 1, (unigrams, Words/POS, Include Punctuation, Ignore All Stopword in N grams, and Track Feature Hits)

| **90% Train 10% Test** | **Recall** | **Precision** | **F-Score** |
|---|---|---|---|
| Naive Bayes Female | **.83** | **.79** | **.81** |

| | | | |
|---|---|---|---|
| Naive Bayes Male | **.79** | **.83** | **.81** |
| SVM Female | .72 | .62 | .67 |
| SVM Male | .67 | .75 | .71 |
| Weka Naive Bayes Female | .74 | .68 | .71 |
| Weka Naive Bayes Male | .70 | .76 | .75 |
| Weka SVM Female | .69 | .58 | .63 |
| Weka SVM Male | .64 | .74 | .69 |
| **80% Train 20% Test** | | | |
| Naive Bayes Female | .73 | .48 | .58 |
| Naive Bayes Male | .61 | .82 | .70 |
| SVM Female | .61 | .49 | .54 |
| SVM Male | .57 | .69 | .62 |
| Weka Naive Bayes Female | .73 | .48 | .58 |
| Weka Naive Bayes Male | .61 | .82 | .70 |
| Weka SVM Female | .61 | .48 | .54 |
| Weka SVM Male | .57 | .69 | .62 |

**Table 7:** Configuration 2, (bigrams, Words/POS, Include Punctuation, Ignore All Stopword in N grams, and Track Feature Hits)

| **90% Train 10% Test** | **Recall** | **Precision** | **F-Score** |
|---|---|---|---|
| Naive Bayes Female | .82 | .74 | .78 |
| Naive Bayes Male | .77 | .82 | .79 |

| | | | |
|---|---|---|---|
| SVM Female | .76 | .69 | .72 |
| SVM Male | .70 | .79 | .74 |
| Weka Naive Bayes Female | .82 | .74 | .78 |
| Weka Naive Bayes Male | .77 | .82 | .79 |
| Weka SVM Female | .87 | .40 | .55 |
| Weka SVM Male | .61 | .94 | .74 |
| **80% Train 20% Test** | | | |
| Naive Bayes Female | .71 | .37 | .49 |
| Naive Bayes Male | .57 | .84 | .68 |
| SVM Female | .63 | .49 | .55 |
| SVM Male | .58 | .72 | .64 |
| Weka Naive Bayes Female | .71 | .37 | .49 |
| Weka Naive Bayes Male | .57 | .84 | .68 |
| Weka SVM Female | .61 | .48 | .54 |
| Weka SVM Male | .57 | .69 | .62 |

**Table 8:** Configuration 3, (trigrams, Words/POS, Include Punctuation, Ignore All Stopword in N grams, and Track Feature Hits)

| **90% Train 10% Test** | **Recall** | **Precision** | **F-Score** |
|---|---|---|---|
| Naive Bayes Female | .85 | .73 | .79 |

| | | | |
|---|---|---|---|
| Naive Bayes Male | .86 | .76 | .80 |
| SVM Female | .73 | .65 | .69 |
| SVM Male | .68 | .75 | .71 |
| Weka Naive Bayes Female | .85 | .73 | .79 |
| Weka Naive Bayes Male | .86 | .76 | .80 |
| Weka SVM Female | .88 | .35 | .50 |
| Weka SVM Male | .60 | .95 | .74 |
| **80% Train 20% Test** | | | |
| Naive Bayes Female | .70 | .34 | .46 |
| Naive Bayes Male | .56 | .86 | .68 |
| SVM Female | .62 | .47 | .53 |
| SVM Male | .57 | .71 | .63 |
| Weka Naive Bayes Female | .70 | .34 | .46 |
| Weka Naive Bayes Male | .56 | .86 | .68 |
| Weka SVM Female | .61 | .48 | .54 |
| Weka SVM Male | .57 | .69 | .62 |

### 4.4 Top Correlated Words for each Gender

**Table 9: Top 10 Positively Correlated Words for each Gender (Configuration 1):**

| Male | | Female | |
|---|---|---|---|
| **Feature** | **Correlation** | **Feature** | **Correlation** |
| similar | .1036 | love | .1439 |
| game | .1011 | husband | .1406 |
| data | .0949 | mom | .1337 |
| bill | .0947 | ! | .124 |
| users | .0916 | lunch | .124 |
| rsb | .091 | food | .115 |
| john | .0905 | lovely | .1119 |
| web | .0903 | !! | .111 |
| link | .0902 | eating | .1038 |
| lsb | .0895 | little | .1007 |

## 5. Discussion:

### *5.1 Results Discussion*

The results from my testing were mixed. The best accuracy score was with LightSide, which achieved an accuracy of 81% using Naive Bayes in configuration 1 using the 90% training and 10% testing corpora. However, this score of 81% has to be taken with some amount of caution. Using 90% training and 10% testing can bias the classification results and the results may not be well suited for generalization to other datasets. The 10 fold cross validation I applied is a well-accepted approach to attempt smoothing out this classification bias by randomizing the data and partitioning it into folds, but it still has the potential for overfitting.

Interestingly, the best classifier for Mukherjee and Liu was not Naive Bayes, but SVM regression. In fact, Naive Bayes performed the worst in their testing. However, Mukherjee and Liu used an algorithm called ensemble feature selection (EFS) for their initial extraction of features that I did not have available for my testing. Mukherjee and Liu were able to achieve a higher best accuracy of 88.56% with SVM regression. However, the accuracies I achieved from configuration 1 90% training and 10% testing roughly match Mukherjee and Liu's results.

The number of misclassified female posts is higher than the number of misclassified male posts throughout all of the testing results. The travel blog subject domain was one area that was responsible for a lot of the misclassification errors. One possibility is that travel is a relatively gender neutral domain that provides few strong author gender clues. Instead travel blogs are more interested in discussing landmarks,

travel accommodations, etc. This neutrality reduces the classifier to using a guessing approach.

The intended marketing use case for a highly accurate single broad domain classifier appears elusive. The costs of using a one size fit all classifier could hinder an organization's marketing efforts more than help. There are too many lexical nuances between genders from one domain to another to draw clear-cut conclusions. An organization would be allocating their finite resources based on unreliable information. In fact, using a single broad classifier that misclassifies the gender could alienate the very audience an organization is intending to target. It would reflect poorly on the professionalism of an organization that advertised messages intended for men to women and vice versa.

The classification problem is exacerbated by the fact that the n-gram correlations are not very strong for either male or female classes. This highlights the linguistic difficulty in distinguishing n-grams that are predominantly in a male or female domain. Sarawgi, Gajulapalli, and Choi (2011) contend that "despite strong evidence for deep syntactic structure that characterizes gender-specific language styles, such deep patterns are not as robust as shallow morphology-level when faced with topic and genre change"(pg. 79). However, the positive n-gram correlations that exist for females fits stereotyped ideas in contrast to the male n-grams. The top three positively correlated n-grams for females are *love, husband, and mom*. The top three positively correlated n-grams for males are *similar, game*, and *data*.

One interesting finding in the top correlated n-grams is that the female class had a high use of the exclamation mark for positive correlation. Some of the top positively female correlated n-grams ! and !! were ranked 4th and 8th respectively. The inverse was true for the negative correlation for ! and !! in the male class. The high use of asserting positive feelings by females reflects previous studies. Wolf writes, “women have added dimensions including solidarity, support, assertion of positive feelings, and thanks, which were absent from male-created definition of emoticons and their use”(pg. 1).

## 6. Conclusion:

My results are inconclusive that creating a single cross domain classifier for automating gender identification is feasible. Past research has shown that men and women do differ semantically when they write. Further research is needed to refine automated classification of author blogs.

Future research needs to be broadened from individuals identifying as either male or female. For example, are there any distinctive lexical patterns for individuals that identify as transgender? Additionally, how can demographic features such as age, education level, and socioeconomic level be incorporated as features?

One of the key benefits of this research is to study an increasingly popular way people express themselves online. The potential for using this information is that information professionals can craft their messages for each gender. This information can take on commercial implications for target advertising or public awareness campaigns.

The current trend of young adults bloggers is a traditionally very ripe demographic for target advertising. The rapid rise and popularity of the blogging platform calls for a better way to quickly sift and sort pertinent information by gender that cannot efficiently be studied manually.

## Works Cited

1) Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2007). Mining the Blogosphere: Age, gender and the varieties of self-expression. *First Monday*,*12*(9).

2) Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *To appear in Text*, *23*, 3.

3) Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011, July). Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1301-1309). Association for Computational Linguistics.

4)de Vel, O. Y., Corney, M. W., Anderson, A. M., & Mohay, G. M. (2002). Language and gender author cohort analysis of e-mail for computer forensics

5) Gehrke, G. T. (2008). *Authorship discovery in blogs using Bayesian classification with corrective scaling* (Doctoral dissertation, Monterey, California. Naval Postgraduate School)

6) Goswami, S., Sarkar, S., & Rustagi, M. (2009, March). Stylometric analysis of bloggers' age and gender. In *Third International AAAI Conference on Weblogs and Social Media*.

7) Herring, S. C., & Paolillo, J. C. (2006). Gender and genre variation in weblogs.*Journal of Sociolinguistics*, *10*(4), 439-459.

8) Huffaker, D. A., & Calvert, S. L. (2005). Gender, identity, and language use in teenage blogs. *Journal of Computer-Mediated Communication*, *10*(2), 00-00.

9) Kobayashi, D., Matsumura, N., & Ishizuka, M. (2007, March). Automatic Estimation of Bloggers' Gender. In *ICWSM*.

10) Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, *17*(4), 401-412.

11) Madigan, D., Genkin, A., Lewis, D. D., Argamon, S., Fradkin, D., & Ye, L. (2005, June). Author identification on the large scale. In *Proc. of the Meeting of the Classification Society of North America*.

12)Mayfield, E., & Rosé, C. P. (2012). LightSIDE: Open source machine learning for text accessible to non-experts. *Invited chapter in the Handbook of Automated Essay Grading*.

13)Mohtasseb, H., & Ahmed, A. (2009). Mining online diaries for blogger identification.

14) Mohtasseb, H., & Ahmed, A. (2009). More blogging features for author identification.

15) Mukherjee, A., & Liu, B. (2010, October). Improving gender classification of blog authors. In *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing* (pp. 207-217). Association for Computational Linguistics.

16) Newson, S., & Oberlander, J. (2006, March). The Identity of Bloggers: Openness and Gender in Personal Weblogs. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (pp. 163-167)

17) Pedersen, S., & Macafee, C. (2007). Gender differences in British blogging.*Journal of Computer-Mediated Communication*, *12*(4), 1472-1492.

18) Rustagi, M., Prasath, R. R., Goswami, S., & Sarkar, S. (2009). Learning age and gender of blogger from stylistic variation. In *Pattern Recognition and Machine Intelligence* (pp. 205-212). Springer Berlin Heidelberg.

19) Santosh, K., Bansal, R., Shekhar, M., & Varma, V. Author Profiling: Predicting Age and Gender from Blogs.

20) Sarawgi, R., Gajulapalli, K., & Choi, Y. (2011, June). Gender attribution: tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning* (pp. 78-86). Association for Computational Linguistics.

21) Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006, March). Effects of Age and Gender on Blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (Vol. 6, pp. 199-205).

22) Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational linguistics*, *26*(4), 471-495.

23) Thomson, R., & Murachver, T. (2001). Predicting gender from electronic discourse. *British Journal of Social Psychology*, *40*(2), 193-208.

24) Wolf, A. (2000). Emotional expression online: Gender differences in emoticon use. *CyberPsychology & Behavior*, *3*(5), 827-833.

25) Yan, X., & Yan, L. (2006, April). Gender Classification of Weblog Authors. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*(pp. 228-230).

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