Journal of Intelligent Learning Systems and Applications, **2012, 4, 169-175**
http://dx.doi.org/10.4236/jilsa.2012.43017 Published Online August 2012 (http://www.SciRP.org/journal/jilsa)

Author Gender Prediction in an Email Stream Using Neural Networks

William Deitrick, Zachary Miller, Benjamin Valyou, Brian Dickinson, Timothy Munson, Wei Hu*

Department of Computer Science, Houghton College, Houghton, USA.
Email: *Wei.Hu@houghton.edu

Received March 28th, 2012; revised June 8th, 2012; accepted June 15th, 2012

ABSTRACT

With the rapid growth of the Internet in recent years, the ability to analyze and identify its users has become increasingly important. Authorship analysis provides a means to glean information about the author of a document originating from the internet or elsewhere, including but not limited to the author's gender. There are well-known linguistic differences between the writing of men and women, and these differences can be effectively used to predict the gender of a document's author. Capitalizing on these linguistic nuances, this study uses a set of stylometric features and a set of word count features to facilitate automatic gender discrimination on emails from the popular Enron email dataset. These features are used in conjunction with the Modified Balanced Winnow Neural Network proposed by Carvalho and Cohen, an improvement on the original Balanced Winnow created by Littlestone. Experiments with the Modified Balanced Winnow show that it is effectively able to discriminate gender using both stylometric and word count features, with the word count features providing superior results.

Keywords: 1-Gram Word Counts; Balanced Winnow; Enron Email; Gender Prediction; Neural Network; Stream Mining, Stylometric Features

1. Introduction

The Internet allows its users to more quickly and effectively share information than ever before. In recent times, the rapid growth of the Internet has been furthered by developments such as e-commerce, social networking, and Internet newsgroups [1]. These services are quickly bringing more users to the Internet, and as of 2010 the number of Internet users worldwide eclipsed two billion [2].

With the Internet's rapid growth, it has increasingly been used for malicious and illegal activities. The Annual Report of the Internet Crime Complaint Center stated that there was a 33.1% increase in online crime in the year 2008 alone [1]. This Internet crime, exemplified by junk email, viruses, identity fraud, and child pornography, presents a complicated problem since Internet users are generally shrouded in anonymity [3]. The fact that many online resources do not require user identification necessitates the design of effective methods for finding cyber criminals.

Discovering criminal users can be facilitated by authorship analysis. Authorship analysis examines a document in order to determine general information about the author, such as his or her identity, gender, and era. The 18th century logician Augustus De Morgan was the first to suggest that an author could be identified by the characteristics of their writing [4], and modern stylometry has descended from his ideas. Over a thousand stylometric features have been proposed to date, including word and character based features [1]. Machine learning techniques have proven invaluable to the field of authorship analysis, with statistical methods such as the Bayesian classifier [1] and principle component analysis [5] being applied to problems of author identification. Recently, more powerful algorithms such as the decision tree [6] and support vector machine [1] have also been used to facilitate effective authorship analysis.

1.1. Gender Identification

It is widely accepted that there are significant linguistic differences between men and women. Previous studies have analyzed the forms of these linguistic distinctions and their links to social roles [7]. Multiple linguistic features have been determined, such as character usage, writing syntax, functional words, and word frequency. Other features have been also been examined, such as those contained in the Media Research Center (MRC) Psycholinguistic database and the Linguistic Inquiry and Word Count (LIWC) software [8]. Some studies have

*Corresponding author.

also partitioned text into n-grams for analysis. N-grams are combinations of n words or characters used in sequence to capture the structure of an author's unique writing style [9].

Many studies have come to similar conclusions regarding which of these features best distinguish male and female authors. It has been reported that women tend to use more emotionally charged language as well as more adjectives and adverbs, and apologize more frequently than men. On the other hand, men use more references to quantity and commit a greater number of grammatical errors [10,11]. It has also been observed that gender-specific language is more prevalent in conversations consisting of only one gender when compared to pairs or groups of both genders [12].

Using the various methods and features, researchers have automated prediction of an author's gender with accuracies ranging from 80% to 90% [1,8]. The expected accuracy is often limited by the type of text being evaluated, with certain types of documents proving more difficult than others. For instance, business-related emails have less gender-preferential language than blogs, making business emails harder to classify than blogs and lowering the expected accuracy.

1.2. Data Stream Mining

Previous studies have primarily used batch learning methods in their research on gender identification. However, for some data sets, (especially those that are incredibly large or constantly growing) the batch technique is simply not practical. In these cases a stream algorithm may be used to evaluate the data in real time. The streaming paradigm utilizes the natural flow of data, as instances are analyzed and classified in real time as they are encountered [13]. This of course requires that the instances be evaluated at a pace equal to or greater than the rate at which they arrive [14]. Due to the restrictions on processing time, a single pass over the data using stream mining becomes more appropriate as data sets increase in size [13]. While making a single pass over data will often cause a decrease in accuracy, the ability of stream mining techniques to process exceedingly large datasets often makes this technique preferable to batch-mining. Thus, we have chosen to use a stream-based learning algorithm to facilitate our research.

2. Materials

2.1. The Enron Corpus

The popular Enron email corpus [15] was used in this study. This dataset was first made public during the investigation of the Houston-based energy company's 2001 bankruptcy by the Federal Energy Regulatory Commission. Initial integrity issues in the dataset were overcome by Melinda Gervasio of SRI for the CALO (a Cognative Assistant that Learns and Organizes) project, and several emails from the dataset have been removed as per the requests of Enron employees [16]. The version of the dataset used in our study, provided by William Cohen of CMU, was last updated in April of 2011 [15].

2.2. Email Parsing

To permit gender identification using a streaming algorithm, emails from users' sent folders ("sent", "sent_ items" and "_sent_mail") were extracted. Each user was annotated with respect to gender, and any emails extracted from their mail directory were assigned the appropriate gender annotation. Email header information and reply text was removed from each email, leaving behind only the body of each sent email with an associated time stamp and message id. This extraction process created a dataset containing 125,495 emails. Additionally, any duplicate emails were removed, and emails less than 50 words or greater than 1000 words in length were excluded from the dataset. The resulting dataset was then sorted according to message timestamps in order to create a dataset as representative as possible of a streaming environment. This extraction process resulted in a final dataset containing body text from 17,893 emails with associated gender labels further described in **Table 1**.

2.3. Feature Extraction

To test our learning algorithm, two different sets of features were extracted from the parsed emails. The first set of features contained 436 stylometric features computed from the text of each email. These features were comprised of four types of stylometric features used in previous studies on gender identification [1]. The four extracted feature types included: character-based features, syntactic features, word-based features, and function words. A more detailed description of each of our features is included in **Table 2**.

Of the four types of features used, character-based and syntactic features are the most straightforward. Character-based features such as the number of tabs, number of uppercase letters, etc. are among those most commonly used in authorship analysis [1]. Syntactic features capture differences in punctuation use between authors. Studies have shown differences in punctuation usage between men and women, and syntactic features are able to encapsulate these stylistic differences [1].

Table 1. Enron dataset description.

Gender	Number of Users	Number of Messages	Avg. Messages Per User
Male	96	8,969	93.42
Female	48	8,924	185.92
Total	144	17,893	124.26

Table 2. Description of stylometric features.

Character-Based Features	
F0	Number of characters
F1	Number of letters/F1
F2	Number of uppercase characters/F1
F3	Number of digital characters/F1
F4	Number of whitespace characters/F1
F5	Number of tab characters/F1
F6-F27	Number of special character/F1
Syntactic Features	
F28	Number of single quotes
F29	Number of commas
F30	Number of periods
F31	Number of colons
F32	Number of semicolons
F33	Number of question marks
F34	Number of repeated question marks
F35	Number of exclamation points
F36	Number of repeated exclamation points
F37	Number of ellipsis
Word-Based Features	
F38	Number of words
F39	Average word length (characters)
F40	Vocabulary richness (total unique words)/F38
F41	Number of long words (more than 6 characters)/F38
F42	Number of short words (1 - 3 characters)/F38
F43	Hapax legomena/F38
F44	Hapax dislegomena/F38
F45	Yule's K
F46	Simpson's D
F47	Honores' R
F48	Entropy
Function Words	
F49-F51	Number of articles/F40
F52-F55	Number of pro-sentence words/F40
F56-F132	Number of pronoun words/F40
F133-F177	Number of auxiliary verbs/F40
F178-F311	Number of conjunction words/F40
F199-F307	Number of interjection words/F40
F308-F435	Number of adposition words

Word-based features and functional features are slightly more complex to extract, but provide information useful in distinguishing between genders [1]. Word-based features provide statistical measurements including Simpson's D, Yule's K, and vocabulary richness. Formulas for several of these features are given in appendix A of [1]. Function words are those words which express connections between words in a sentence or indicate a speaker's mood. Due to stylistic differences between the writing of men and women, these words are able to capture differences between the genders. The function words used in this study are similar to those used in [1].

The second group of features extracted from each email in the Enron dataset was a set of word counts. To compile these word count features, we first parsed through each sent email in the dataset to find the number of times every word appearing in this set of emails was used. Then, we eliminated any words used less than 50 times and did not consider these words as features for our dataset. This procedure resulted in a master list of 3925 words to be used as features. Each email was then parsed again, and a feature vector was generated from each email consisting of counts for every word in the feature set.

3. Methods

3.1. Mistake Driven Online Learner

This algorithm is an example of a Mistake Driven Online Learning Neural Network that first generates a learning rule then updates this learning rule if a mistake occurs in classification [17]. In other words, a Mistake-Driven Online Learner creates a model w_0, generally represented as a matrix, then classifies individual instances and updates w_0 if the classification is incorrect. All learning and classification is done in an online fashion as new instances come in. As each instance arrives, the learner will make a prediction using a predefined neural network function, then compare the prediction with the actual class of the instance. If the prediction is incorrect, the algorithm will update its model and then wait for more instances. The pseudo-code for this type of neural network is found below.

Pseudo-Code for the Basic Mistake-Driven Online Learner.

(1) Initialize model w_0. Define function: $f(w_i, x_t)$
(2) For $t = 1,2,\cdots,T$;
  a. Retrieve new example x_t
  b. Predict $\widehat{y_t} = f(w_i, x_t)$ and compare it with actual class y_t.
  c. If $\widehat{y_t} \neq y_t$:
    i. Update model $w_i \rightarrow w_{i+1}$.
  d. Else:
    i. Prediction was correct.

3.2. Balanced Winnow Neural Network

The Balanced Winnow Neural Network is an example of a Mistake-Driven Online Learner. The Balanced Winnow algorithm defines three parameters used for classification and updating the weight models. The α parameter is used to promote the model through multiplication and is set such that $\alpha > 1$ so that the values in the model will be

increased. Similarly, β is the demotion parameter set $0 < \beta < 1$ and will always decrease the values stored by the model. These parameters are most critical to the operation of the Modified Balanced Winnow, as they are chosen to control the rate at which the model learns. Appropriate values in the ranges given above are dependent on the data, and a trial and error approach often facilitates effective discovery. The third and final parameter for the Balanced Winnow algorithm is the threshold, θ_{th} which is responsible for biasing the prediction. The bias is subtracted from the difference between the inner products of the instance and the models to help the models better fit the data. These three parameters are necessitated by the Balanced Winnow algorithm's departure from the general Mistake-Driven Online Learner by introducing two models that combine to form w_i. Because the Balanced Winnow has both a positive and a negative model u_i and v_i, which need to change at different rates, two weight parameters α and β must be used to increase the effectiveness of the accurate model while decreasing the influence of the mistaken model. In this way, the Balanced Winnow is able to create two effective models that can easily determine the correct class of the instance [18,19].

Two important features of the Balanced Winnow algorithm are its scoring function and its update rule. Let (x_t, w_i) denote the inner product between the current instance x_t and the current weight vector w_i. The scoring function for a Balanced Winnow algorithm is defined as

$$f = \text{sign}\left((x_t, u_i) - (x_t, v_i) - \theta_{th}\right),$$

where sign(x) is the signum function. This scoring function is compared to the actual class of the instance, and if a mistake occurs, the update rules are called. The update rules for the Balanced Winnow are defined below.

Update Rules for the Balanced Winnow Algorithm.
Function:updateModels()

Given: Models u_i, v_i where u_i is the positive model and v_i is the negative, true class y_t

(1) If ($y_t < 0$):
  a. $u_{i+1} = u_i \times \beta$,
  b. $v_{i+1} = v_i \times \alpha$,
(2) Else:
  c. $u_{i+1} = u_i \times \alpha$,
  d. $v_{i+1} = v_i \times \beta$.

3.3. Modified Balanced Winnow Neural Network

The Modified Balanced Winnow algorithm modifies the Balanced Winnow to improve the learning process. Although the Modified Balanced Winnow algorithm is similar to the Balanced Winnow, two important additions are included [17]. First, a margin M is created such that a prediction generated by the Modified Balanced Winnow is a mistake if the true class multiplied by the score function is less than or equal to M. This ensures that updates only occur when the prediction is absolutely correct. The second addition is the modified Model Update function. Instead of just multiplying the current models by either α or β, the model is also multiplied by the incoming instance. This change is seen below.

Modified Update Rule: Allows the Instance to Have an Effect on the Model.
Function:updateModels2()

Given: Models u_i, v_i where u_i is the positive model and v_i is the negative, instance x_t having true class y_t

(1) If ($y_t < 0$):
  a. $u_{i+1} = u_i \times \beta \times (1 - x_t)$,
  b. $v_{i+1} = v_i \times \alpha \times (1 + x_t)$,
(2) Else:
  c. $u_{i+1} = u_i \times \alpha \times (1 + x_t)$,
  d. $v_{i+1} = v_i \times \beta \times (1 - x_t)$.

The pseudo-code for the entire Modified Balanced Winnow is seen below. As each instance comes in, the Modified Balanced Winnow Algorithm adds a bias to each feature and normalizes the instance. The score function is then calculated then multiplied by the true class of the instance. If this is less than or equal to the M parameter, then the prediction is considered to be wrong. If the prediction is wrong, the updateModels2 function is called and the models are updated. This process continues as new instances arrive.

Full Pseudo-Code for the Modified Balanced Winnow Algorithm.

(1) Initialize models u_0 and v_0.
(2) For $t = 1,2,\cdots,T$:
  a. Receive new example x_t and add the bias.
  b. Normalize x_t.
  c. Calculate score function:
    i. $score = (x_t, u_i) - (x_t, v_i)\ \theta_{th}$.
  d. Retrieve true class y_t.
  e. If ($score \times y_t$) $\leq M$: //prediction is wrong
    i. Update Models2():
  f. Else
    i. Continue:

4. Results and Discussion

4.1. Setup and Results

Using the features collected from the Enron email dataset, we attempted to differentiate the gender of email authors using the Modified Balanced Winnow algorithm. Since the Modified Balanced Winnow runs relatively quickly, we decided to tune the parameters α and β using the full dataset. To automate this process, we wrote a threaded tuning program capable of running several hundred experiments simultaneously. Because α is the promotion

parameter of the Modified Balanced Winnow algorithm, we first roughly scanned through some acceptable values ranging from 1 to 2. We found that a promotion rate of 1.5 is beneficial as it gives the current instance sufficient influence over the model. Then for the second and third tuning tests we fixed α at 1.5 and varied the β rate for both the word count and the stylometric feature sets. This approach allows us to see how a particular parameter will affect the gender classification. The results from these tests are displayed in **Tables 3** and **4**.

Table 3. Performance metrics with stylometric features.

Alpha	Beta	Acc	Prec	Sens	Spec	F-M
	0.1	0.51	0.69	0.07	0.97	0.13
	0.3	0.52	0.68	0.11	0.95	0.19
1	0.5	0.53	0.67	0.14	0.93	0.24
	0.7	0.53	0.63	0.18	0.89	0.28
	0.9	0.90	0.96	0.84	0.96	0.89
	0.1	0.58	0.62	0.25	0.85	0.35
	0.3	0.89	0.94	0.84	0.94	0.89
1.5	0.5	0.94	0.96	0.94	0.96	0.95
	0.7	0.95	0.95	0.95	0.95	0.95
	0.9	0.94	0.95	0.95	0.95	0.95
	0.1	0.55	0.61	0.33	0.79	0.42
	0.3	0.98	0.96	0.93	0.96	0.94
2	0.5	0.93	0.94	0.94	0.94	0.94
	0.7	0.94	0.95	0.95	0.95	0.95
	0.9	0.95	0.95	0.95	0.95	0.95

Accuracy, Precision, Sensitivity, Specificity, F-Measure.

Table 4. Performance metrics with word-based features.

Alpha	Beta	Acc	Prec	Sens	Spec	F-M
	0.1	0.53	0.60	0.16	0.89	0.26
	0.3	0.54	0.62	0.23	0.86	0.33
1	0.5	0.54	0.60	0.26	0.83	0.36
	0.7	0.54	0.59	0.28	0.80	0.38
	0.9	0.55	0.59	0.33	0.76	0.43
	0.65	0.56	0.56	0.55	0.57	0.55
	0.7	0.66	0.66	0.66	0.66	0.66
1.5	0.77	0.89	0.89	0.89	0.89	0.89
	0.79	0.56	0.56	0.56	0.56	0.56
	0.8	0.56	0.56	0.57	0.55	0.56
	0.1	0.55	0.59	0.32	0.78	0.42
	0.3	0.56	0.58	0.45	0.68	0.51
2	0.5	0.56	0.56	0.56	0.56	0.56
	0.7	0.54	0.53	0.77	0.32	0.63
	0.9	0.53	0.52	0.87	0.20	0.65

As can be seen from **Tables 3** and **4**, the Modified Balanced Winnow performed better with word based features than with stylometric features. In the best case for the word based data (α = 1.5 and β = 0.7) we see that our accuracy is above 95% with 96% precision. These very high results are surprising given the one-pass nature of the Modified Balanced Winnow Neural Network, but show the network's potential for fast and accurate performance. Similarly, the stylometric features also exhibited up to 88% accuracy and precision when the parameters were tuned optimally. Our results can be further supported by data analysis.

4.2. Discussion

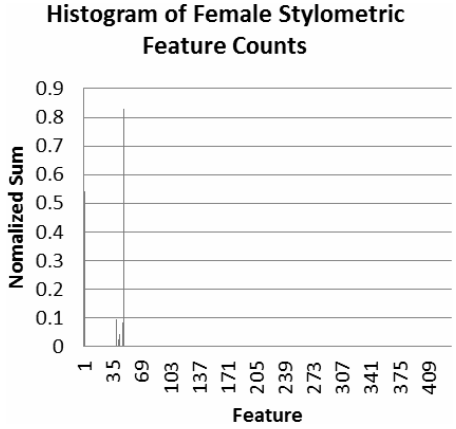
Using two different feature sets, our Modified Balanced Winnow algorithm was able to accurately differentiate between emails written by male and female Enron employees. This method shows an improvement in accuracy over previous studies using stylometric features by several percentage points, and shows the effectiveness of the Modified Balanced Winnow as a mistake-driven learner.

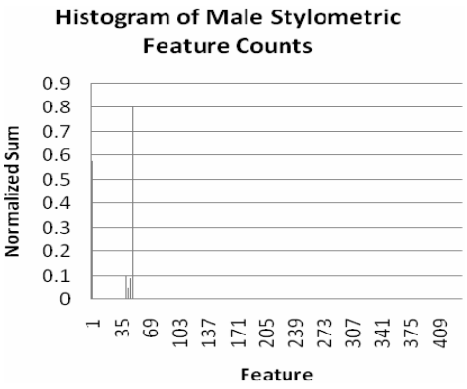
To further observe the improvements made algorithmically, we ran both the modified and the traditional Balanced Winnow on each feature set and computed the resulting performance metrics. The Modified Balanced Winnow algorithm shows improvement over the traditional Balanced Winnow when performing gender classification. In **Table 5**, we display the best results achieved by the Balanced Winnow Neural Network which is roughly 56% accuracy for both the stylometric and word-based features. It is interesting to note the fact that both feature representations performed equally well in comparison to the Modified Balanced Winnow Neural Network favoring the word-based feature representation. Also, the modifications made to the update rule of the Balanced Winnow algorithm allowed the Modified Balanced Winnow Neural Network to perform substantially better.

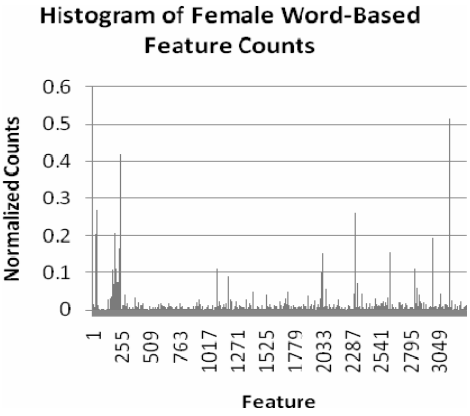
Table 5. Performance metrics for the balanced winnow algorithm.

Stylometric Features						
α	β	Acc	Prec	Sens	Spec	F-M
1.5	0.59	0.5629	0.5715	0.5113	0.6147	0.5398
1.5	0.62	0.5626	0.5683	0.53	0.5955	0.5485
1.5	0.63	0.5626	0.5682	0.5315	0.594	0.5492
Word-Based Features						
α	β	Acc	Prec	Sens	Spec	F-M
3.5	0.2	0.5613	0.5701	0.508	0.615	0.5372
4.5	0.3	0.561	0.5598	0.5816	0.5405	0.5705
3	0.5	0.5609	0.5558	0.6171	0.5044	0.5849

Interestingly, a significant improvement in accuracy was observed with word count features over stylometric features. While there are various reasons that could be cited for this difference, we believe that this is primarily due to the large range of values exhibited by the stylometric features. Though the algorithm normalizes feature values, the significant differences between values nevertheless causes feature vectors to less clearly differentiate male and female authors. In contrast, the word-based features, while showing similarities between male and female writing, also more clearly demonstrate differences between gender word usage. This fact is illustrated in the histograms shown in **Figure 1**. Each histogram displays the normalized value of each feature on the y-axis for each of the features for the given feature type (displayed on the x-axis). From these histograms, we can see that the word-based features more clearly show differences between male and female authors. This explains why greater accuracy was achieved using word-based features.







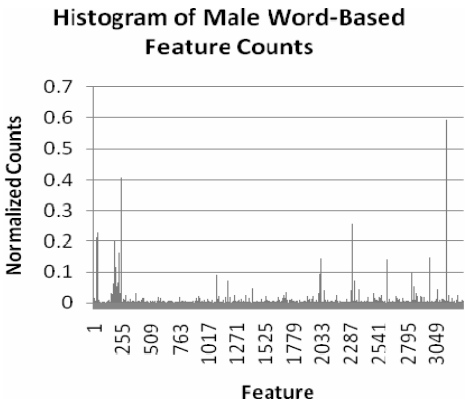


Figure 1. Histograms displaying normalized counts of stylometric and word-based features for males and females.

5. Conclusions

Gender prediction through text classification has multiple uses in today's communication-driven world. To improve upon previous research, we applied a simple Neural Network to this type of classification. Data for this study was extracted from the Enron email dataset, provided by William Cohen of CMU. Users within the Enron corpus were then labeled according to gender, and emails from the users' sent mail folders were extracted and converted into stylometric and word-based feature sets.

Both the Modified Balanced Winnow and the traditional Balanced Winnow Algorithms were used to discriminate gender on the Enron dataset. The best results were achieved by the Modified Balanced Winnow, with a range of values for the α and β parameters being tested. When the best parameters were found, the stylometric features achieved 88% accuracy and the word based features achieved around 95% in comparison to approximately 56% accuracy for the traditional Balanced Winnow using both feature sets. To further understand our results, we generated four normalized frequency counts for the different features. We noticed that because the feature-space of the stylometric features was more diverse, the distribution was very similar. The word based features, however, displayed a clear difference between the two genders.

6. Acknowledgements

We would like to thank Houghton College for its financial support and for providing this research opportunity.

REFERENCES

[1] N. Cheng, R. Chandramouli and K. P. Subbalakshmi, "Author Gender Identification from Text," *Digital Investigation*, Vol. 8, No. 1, 2011, pp. 78-88. doi:10.1016/j.diin.2011.04.002

[2] H. Touré, "Brief Remarks to Media," *Commission on Information and Accountability for Women's and Children's Health*, 2011.

[3] R. Zheng, Y. Qin, Z. Huang and H. Chen, "Authorship Analysis in Cybercrime Investigation," *Proceedings from the* 1*st NSF/NIJ Symposium*, 2003, pp. 59-73.

[4] R. Zheng, J. Li, H. Chen and Z. Huang, "A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques," *Journal of the American Society for Information Science and Technology*, Vol. 57, No. 3, 2006, pp. 378-393. doi:10.1002/asi.20316

[5] J. Burrows, "Word Patterns and Story Shapes: The Statistical Analysis of Narrative Style," *Literary and Linguistic Computing*, Vol. 2, No. 2, 1987, pp. 61-67. doi:10.1093/llc/2.2.61

[6] A. F. Damerau and A. F. S. Weiss, "Text Mining with Decision Trees and Decision Rules," *Conference on Automated Learning and Discovery*, Carnegie-Mellon University, Pittsburgh, 1998.

[7] F. R. Bilous and R. M. Krauss, "Dominance and Accommodation in the Conversation Behaviours of Same- and Mixed-Gender Dyads," *Language and Communicaion,* Vol. 8, No. 3-4, 1988, pp. 183-194. doi:10.1016/0271-5309(88)90016-X

[8] S. Nowson and J. Oberlander, "The Identity of Bloggers: Openness and Gender in Personal Weblogs," *Proceedings of the AAAI Spring Symposia on Computational Approaches to Analyzing Weblogs*, California, 2006.

[9] J. D. Burger, J. Henderson, G. Kim and G. Zarella, "Discriminating Gender on Twitter," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Edinburgh, 27-31 July 2011, pp. 1301-1309. http://www.mitre.org/work/tech_papers/2011/11_0170/11_0170.pdf

[10] M. W. Corney, "Analyzing E-Mail Text Authorship for Forensic Purposes," Masters Thesis, Queensland University of Technology, Queensland, 2003.

[11] O. Y. de Vel, M. W. Corney, A. M. Anderson and G. M. Mohay, "Language and Gender Author Cohort Analysis of E-Mail for Computer Forensics," *Proceedings Digital Forensics Research Workshop*, Syracuse, 6-8 August 2002.

[12] R. Thomson and T. Murachver, "Predicting Gender from Electronic Discourse," *The British Journal of Social Psychology*, Vol. 40, No. 2, 2001, pp. 193-208. doi:10.1348/014466601164812

[13] W. Fan, H. Wang and P. S. Yu, "Active Mining of Data Streams," *Proceedings of the Fourth SIAM International Conference on Data Mining*, Lake Buena Vista, 22-24 April 2004, pp. 457-461.

[14] P. Domingos and G. Hulten, "Mining High Speed Data Streams," University of Washington, Seattle, 2000.

[15] L. Kaelbling, "Enron Email Dataset," CALO Project, 2011. http://www.cs.cmu.edu/~enron

[16] J. Shetty and J. Adibi, "Enron Email Dataset Database Schema and Brief Statistical Report," Information Sciences Institute Technical Report, University of Southern, California, 2004.

[17] V. R. Carvalho and W. W. Cohen, "Single-Pass Online Learning: Performance, Voting Schemes and Online Feature Selection," *Proceedings of the* 12*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, 20-23 August 2006, pp. 548-553.

[18] I. Dagan, Y. Karov and D. Roth, "Mistake-Driven Learning in Text Categorization," *Conference on Empirical Methods on Natural Language Processing*, Providence 1-2 August 1997, pp. 55-63.

[19] N. Littlestone, "Learning Quickly When Irrelevant Attributes Abound: A New Linear-Threshold Algorithm," *Machine Learning*, Vol. 2, No. 4, 1988, pp. 285-318. doi:10.1007/BF00116827

