

Gender Classification for Web Forums

Yulei Zhang, Yan Dang, and Hsinchun Chen, *Fellow, IEEE*

Abstract—More and more women are participating in and exchanging opinions through community-based online social media. Questions concerning gender differences in the new media have been raised. This paper proposes a feature-based text classification framework to examine online gender differences between Web forum posters by analyzing writing styles and topics of interest. Our experiment on an Islamic women's political forum shows that feature sets containing both content-free and content-specific features perform significantly better than those consisting of only content-free features, feature selection can improve the classification results significantly, and female and male participants have significantly different topics of interest.

Index Terms—Gender classification, online gender differences.

I. INTRODUCTION

THE RAPID development and evolution of the Internet have enabled people to access information whenever and wherever they want. Recently, with the advent of Web 2.0, the Internet has evolved toward multimedia-rich content delivery, end-user content generation, and community-based social interaction [44]. More and more Web forums, blogs, wikis, and other social media have been generated and become extremely popular. Such Web 2.0 social media help enhance information sharing, opinion generation, and community-based discussion for various emerging social and political topics.

Although it has a male-dominated history, the Internet is becoming a new medium for women to increasingly share their concerns and express opinions about personal, social, and political issues [26]. With this trend, the need for women to claim the Internet as an important space of their own has emerged. Women could gain equal presence or influence with men in the virtual community. In addition, their desire for gender equality continues to influence their Internet contributions and writings. Meanwhile, the increasing availability of the Internet offers marginalized groups and individuals a voice in the public sphere [27], [41]. For example, Harcourt [26] mentions the increasing voice of local Arab women on a global level through the Internet; Mitra [41] argues that the Internet has allowed women in South Asia to be heard by the outside world.

In many disciplines, questions concerning gender differences in the context of online communication have been raised [25]. Online gender differences (i.e., the digital gender gap in some studies), which refers

to the differences between women and men in Internet use, have been shown and studied in previous research [19], [20], [27]. Some studies point out that women are less likely to express political opinions and tend to have a less authoritative manner in their conversation style [45]. More research is critically needed to explain online gender differences in social, political, and even business (e.g., online shopping) activities.

Understanding online gender differences and why they occur could be important for Internet service providers, system developers, information analysts, and end users. Many domains, such as security and marketing, could benefit from such an understanding. The ability of security researchers and analysts to track individual contributors, analyze gender-specific trends and views, monitor certain opinion groups, and identify potential threats could be very useful. For the marketing domain, a better understanding of the different interests in various products between the two genders can help the sellers adopt and develop services and systems tailored for the two groups of people and thereby attract more customers.

In this paper, we adopted feature-based text classification techniques to identify and analyze online gender differences by examining the discrepancy between women's and men's writing styles. Most previous online text classification studies have focused on authorship and sentiment classification; relatively less effort has been put on gender classification. To improve classification performance, both the most recent authorship and sentiment classification studies incorporated all four types of features including lexical, syntactic, structural, and content-specific features. However, for gender classification, we have not seen a study using all four types of features. As to the context, previous gender classification studies have mainly focused on novels [28], nonfiction articles [32], e-mails [12], and Web blogs [43], [51]. Relatively less effort has been put on the Web forum context. In addition, even for those focusing on Web forum context, most studies used basic keyword-based analysis with a relatively small set of keywords to examine the topic differences between males and females [23], [53]. Few studies have investigated the gender differences in Web forums using feature-based text classification techniques. Therefore, we proposed a feature-based gender classification framework to analyze online gender differences for Web forums by examining the writing styles and contents (including different types of linguistic features) of female and male posters. Our experiment was conducted on an Islamic women's political forum, and we compared the performances of different feature sets. The best classification results were achieved by incorporating all four types of features and conducting feature selection, demonstrating the efficacy of this framework for gender classification for Web forums. We further analyzed the different topics preferred by women and men, respectively.

II. LITERATURE REVIEW

A. Online Gender Differences

With the increasing availability and popularity of the Internet, as well as the advent of Web 2.0, more and more women participate in community-based social media [11]. The Internet, therefore, has become a medium for women to share their political opinions and knowledge [26]. They are also creating their own online networks to exchange information and ideas [56].

The Internet is not only useful as a fast communication medium, it is also a very crucial channel of information on women's rights issues. Women use the Internet to fight against violence by building a strong layer of support through which their personal struggles can be discussed and solutions shared [26]. As an example, Harcourt [26] talks

Manuscript received July 7, 2009; revised February 16, 2010; accepted June 18, 2010. Date of publication March 3, 2011; date of current version June 21, 2011. This work is supported in part by the National Science Foundation's Computer and Network Systems (CNS) Program under Grant CNS-0709338. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This paper was recommended by Associate Editor C. Yang.

Y. Zhang and Y. Dang are with the W. A. Franke College of Business, Northern Arizona University, Flagstaff, AZ 86011 USA (e-mail: ylzhang@email.arizona.edu; ydang@email.arizona.edu).

H. Chen is with the Department of Management Information Systems, Eller College of Management, University of Arizona, Tucson, AZ 85721 USA (e-mail: hchen@eller.arizona.edu).

Color versions of one or more of the figures in this paper are available online at <http://icccxplora.ieee.org>.

Digital Object Identifier 10.1109/TSMCA.2010.2093886

about a case regarding a Muslim woman's right of choice of marriage in her study; she argues that "we could, within hours, receive case law on the issue from other Muslim countries, as well as legal and scholarly opinions and references, that prove critical in winning the case."

Researchers have also shown an increasing interest in studying online gender differences, which refers to the fact that there exist differences between women and men in Internet use [9]. Previously, the major online gender difference noted was that fewer women than men used the Internet. For example, the A. C. Nielsen CommerceNet consortium from 1999 showed that among U.S. and Canadian Internet users, 53% were men and 47% were women; among online shoppers, 62% were men and 38% were women; and among people who reported having used the Internet in the last twenty-four hours for any purpose, 68% were men and 32% were women [10]. In the realm of political activity, the National Election Studies data showed that visitors to Internet campaign sites during the 1998 election season were 60% male and 40% female [42]. However, with the rapid development and increasing availability of the Internet, more and more women are accessing the Internet to acquire information, express their ideas, and share common concerns. The May 2008 survey by the Pew Internet and American Life Project found that 73% of men and 73% of women use the Internet [48]. In contrast, its 2004 survey reported 66% and 61% Internet use for men and women, respectively.

Although access technology is not an issue today, women and men do have differences in Internet use depending on motivation and interest in the content being produced and consumed [27]. Jackson *et al.* [30] found that women are more likely to use the Internet as a communication tool, and men are more likely to use it as a means of information seeking. According to Ogan *et al.* [45], women are less likely to express political opinions and tend to have a less authoritative manner in their conversation style. Meanwhile, some studies [20], [64] observed that women's concerns tend to center around the private sphere of life, i.e., the domestic sphere of home, family, private relations, and sexual reproduction; on the other hand, men are more externally focused on the public sphere and political realm including government and commercial establishments.

As to online communication on Web forums, previous studies have used keyword analysis to show that women and men do have different topics that they are interested in and care about [53]. Seale *et al.* [53] analyzed cancer-related Web forums and found that women's discussions are more likely to lean toward the exchange of emotional support, including concern with the impact of illness on a wide range of other people; however, men are more likely to participate in threads related to treatment information, medical personnel, and procedures. Guiller and Durndell [23] analyzed an online course discussion board and found that women are more likely to explicitly agree and support others and make more personal and emotional contributions than men; on the other hand, men are more likely to use authoritative language and to respond negatively in interactions than women.

B. Online Text Classification

In this paper, we adopted online text classification techniques to study the online gender differences in Web forums by examining the writing style of the posted messages.

1) *Different Types of Online Text Classification Problems:* With the advent of Web 2.0, more and more automatic classification studies using online text-based social media data have appeared. In those studies, the investigated classification problems mainly include authorship, sentiment and gender classification. Unlike the classical topic-based classification problem in information retrieval, social media classification relies heavily on the information and fluid writing styles of authors in various online social media.

Authorship classification aims at determining which author produced which piece of writing by examining the styles and contents of writings produced by different authors. Previous studies have applied authorship classification to various online social media texts. De Vel and his collaborators [14], [15] applied the conventional text classification methods to identify the authors of e-mails. A recent comprehensive study conducted by Abbasi and Chen [2] tested their newly developed Writeprints technique with a rich set of features on various online data sets, including e-mails, instant messages, feedback comments, and program codes.

Sentiment classification for online texts aims to analyze direction-based texts (i.e., texts containing opinions and emotions) to determine whether a text is objective or subjective, or whether a subjective text contains positive or negative sentiments. The common two-class sentiment classification problem involves classifying sentiments as positive or negative [46], [58]. However, additional variations include classifying sentiments as opinionated/subjective or factual/objective [60], [61]. Instead of sentiments, other studies have attempted to classify emotions, including happiness, sadness, anger, horror, etc. [22], [57].

Gender classification aims to determine whether a piece of writing was produced by a female or male by examining the writing styles and contents of female and male authors. Gender classification is different from authorship classification in that authorship classification examines individual differences of people's writing styles no matter whether a person is a woman or a man, while gender classification is used to examine and identify the overall differences in writing styles between the two gender groups, in order to gain an understanding of gender-based differences. Previous gender classification studies using automatic text classification techniques have been done on both traditional articles (e.g., novels and nonfiction articles) and online social media texts (e.g., e-mails and Web blogs). As an example of gender classification on traditional articles, Koppel *et al.* [32] used the exponential gradient algorithm to classify genders for both fiction and nonfiction documents. By using a feature set combining function words and parts-of-speech (POS) tags, they achieved 79.5% accuracy for fiction documents and 82.6% accuracy for nonfiction documents. After feature selection, the accuracy increased to 98% for both fiction and nonfiction documents. Another study conducted by Hota *et al.* [28] classified the gender of Shakespeare's characters based on a collection of his plays. They achieved the highest accuracy of 74.28% using support vector machine (SVM) on the feature set consisting of both content-independent and content-based features. Argamon and his collaborators [5] analyzed writing styles and identified a set of lexical and syntactic features that differed significantly according to author gender in both fiction and nonfiction documents. In particular, they found that although the total number of nominals used by female and male authors was virtually identical, females used many more pronouns, and males used many more noun specifiers.

For online social media text, most previous gender classification studies focused on e-mails [12] and Web blogs [43], [51]. Corney *et al.* [12] used SVM to classify genders for e-mails and achieve the highest F-measure of 71.1% using a combination of lexical, structural, and selected gender-specific features. Nowson and Oberlander [43] used SVM to classify genders for Web blogs and achieved the highest accuracy of 91.5% using a combination of POS, bigrams, and trigrams as the features. Schler *et al.* [51] also conducted gender classification on Web blogs and emphasized the significant differences in writing styles and contents between female and male bloggers, as well as among authors of different ages.

2) *Features for Online Social Media Text Classification:* Features are very important for online social media text classification. Good feature sets can improve the performance of the classifier. There are four types of features that were often used in previous online social

media text classification studies, namely, lexical, syntactic, structural, and content-specific features. Among them, the first three types are content-free features; the fourth type contains features related to specific topics.

Lexical features refer to character- or word-based statistical measures of lexical variation. Lexical features mainly include the following: character-based lexical features [6], [21], vocabulary richness measures [65], and word-based lexical features [15], [66]. Examples of character-based lexical features are the total number of characters, the number of characters per sentence, the number of characters per word, and the usage frequency of individual letters. Examples of vocabulary richness measures include the number of words that occur once and twice and some other statistical measures defined by Yule [65]. Examples of word-based lexical features are the total number of words, the number of words per sentence, and word length distribution.

Syntactic features refer to the patterns used to form sentences. Commonly studied syntactic features are function words [31], [32], punctuation [7], and POS tags [4], [7], [21], [43]. These studies also demonstrated that syntactic features may be more reliable compared with lexical features. To study the writing style differences between females and males, Argamon and his collaborators [5] used over 1000 features including 467 function words and a set of POS tags.

Structural features show the text organization and layout. They are especially useful when studying online social media texts [15]. Traditional structural features include greetings, signatures, the number of paragraphs, and the average paragraph length [15], [66]. Other structural features include technical features, such as the use of various file extensions, font sizes, and font colors [1].

Different from the above content-free features, content-specific features are comprised of important keywords and phrases on certain topics [38], [66] such as word n -grams [1], [16], [43]. Usually, these features represent specific subject matter in a given domain. For example, content-specific features on a discussion of computers may include "laptop" and "notebook." Previous studies have showed that content-specific features can improve the performance of online text classification [1], [3], [51], [66].

3) *Different Types of Online Social Media Texts*: The major types of online social media texts include e-mails, online news, Web blogs, online reviews, and Web forums. Among them, e-mail and online news typically belong to Web 1.0, while Web blogs, Web forums, and online reviews are considered to be Web 2.0 media.

Different from the other types of online social media texts, e-mails can only be used to share information between the senders and receivers. Typically, the general public cannot access the content. In contrast, online news is always available for any Internet user. However, it enables only a one-way flow of information through static websites which contain "read-only" materials. Therefore, users can only passively acquire information instead of actively participating in the discussions.

As Web 2.0 media, Web blogs, online reviews, and Web forums contain a great deal of dynamic user-generated content. Different people can participate in and exchange opinions through these communication platforms. For Web blogs, the blog owner typically leads the discussion and others can follow up with their comments. Compared to Web blogs, online reviews and Web forums tend to have relatively more "balanced" discussions among participants. "Balanced" here refers to the number of participants and the number of discussion messages they posted. In forums, participants are generally free to initiate their own discussions on topics of their own choosing, as opposed to in blogs, topics are generally set by the blog owner. One major difference between online reviews and Web forums is that online reviews are more focused on a particular product or category of products, while the discussion topics in Web forums tend to be broader, meaning that in

addition to certain products, people also share their opinions on certain events or social and political issues.

Previous gender classification studies have mainly focused on either traditional articles (e.g., novels and nonfiction articles) or e-mails [12] and Web blogs [43], [51]. Relatively less effort has been put on the Web forum context. Considering its role as a major type of online social media with a balanced nature of discussions among participants and a relatively broader range of topics, it is important to understand the online gender differences in Web forums.

4) *Feature Selection for Text Classification*: To perform text classification, a real-world textual data set is usually represented by a set of features. When the number of features is large, not all the features are necessary to learning the concept of interest; instead, many of them may be noisy or redundant; feeding all these features into a model often results in overfitting and poor predictions [39]. In such cases, feature selection can be used to improve the classification performance by selecting an optimal subset of features [13], [24]. Previous text classification studies using n -gram features usually included some form of feature selection in order to extract the most important words or phrases [33]. The objective of feature selection is threefold as follows: to improve the prediction accuracy, provide faster and more cost-effective prediction, and provide a better understanding of the underlying process that generated the data [29], [35]. A feature selection method generates different candidates from the feature space and assesses them based on some evaluation criterion to find the best feature subset [35]. Graph-based search algorithms are often used to find the optimal features [35]. In general, when the feature set is large, using feature selection in text classification can improve the classification performance by offering more concise and precise feature representations of documents [52].

III. MOTIVATION

Understanding online gender differences and why they occur could be important for Internet service providers, system developers, information analysts, and end users. Many domains, such as security and marketing, could benefit from understanding gender-based differences. The ability of security researchers and analysts to track individual contributors, analyze gender-specific trends and views, monitor certain opinion groups, and identify potential threats could be very useful. For the marketing domain, a better understanding of the different interests in various products between the two genders could help sellers adopt and develop services and systems tailored for the two groups of people and thereby attract more customers.

According to our literature review, most previous online text classification studies focused on authorship classification and sentiment classification; relatively less effort has been put on gender classification. To improve the classification performance, the most recent authorship classification studies [1], [66] have incorporated all four types of features, i.e., lexical, syntactic, structural, and content-specific features. Although early sentiment classification studies often used one type of feature, later studies [21], [61] added other types of features to improve the classification performance. However, for gender classification, we have not seen a study using all four types of features. Although previous studies have shown the existence and evolution of online gender differences and the importance of gender role in political movements, most of them have focused on either traditional articles (e.g., novels and nonfiction articles) or e-mails [12] and Web blogs [43], [51]. For the relatively few studies in the Web forum context, most of them used basic keyword-based analysis [23], [53] instead of the more advanced feature-based text classification techniques.

Motivated by the earlier discussion, in this paper, we proposed a framework to investigate online gender differences in the context of

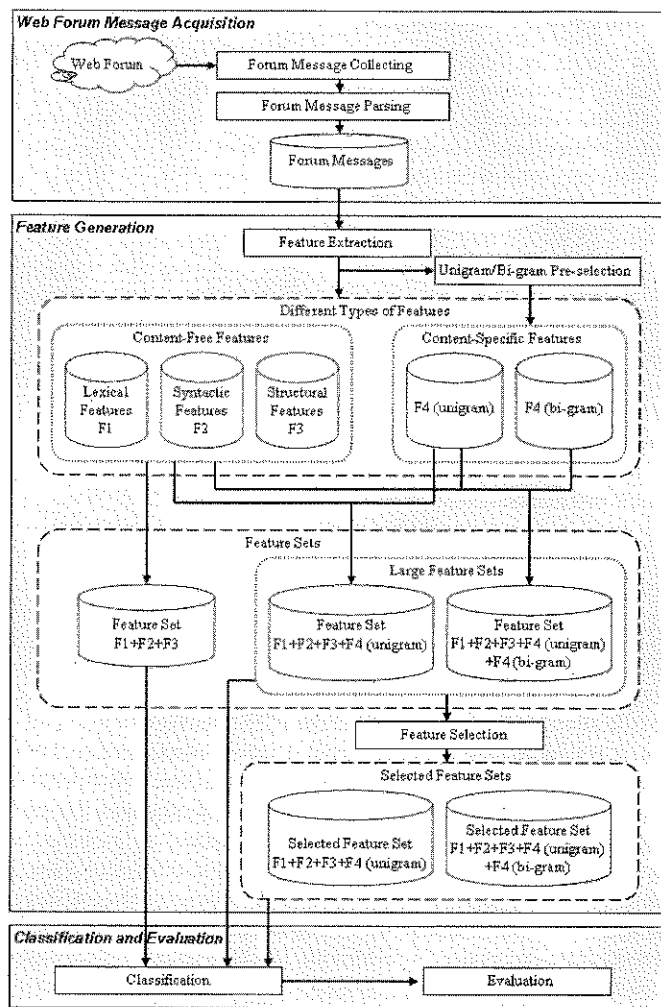


Fig. 1. Research framework.

Web forums using feature-based gender classification techniques by incorporating all four types of features. We intended to answer the following research questions raised from our literature review.

- 1) Can gender classification techniques be used to identify and analyze online gender differences in Web forums?
- 2) Will the use of both content-free features (i.e., lexical, syntactic, and structural features) and content-specific features improve gender classification performance for Web forums compared to using only content-free features?
- 3) For relatively large feature sets, will feature selection that returns a smaller number of the most important features improve the gender classification performance for Web forums?

The first question is more general, while the other two are more specific. Finding the answer to the first question is the primary motivation for this paper as a whole. The efficacy of the proposed gender classification framework is addressed later in this paper. In order to answer the remaining two questions, we developed the detailed hypotheses listed in Section V-B, which in turn drove the details of our design.

IV. RESEARCH DESIGN

In order to address these questions, we developed a framework of feature-based gender classification on Web forums (see Fig. 1).

A. Web Forum Message Acquisition

This component consisted of two steps, namely, forum message collecting and parsing. First, spidering programs were developed to collect all the messages in a given open source Web forum as HTML pages. After that, we built parsers to parse out the message information from the raw HTML pages and store the parsed data in a relational database.

B. Feature Generation

In this component, we generated different feature sets containing different types of features. By doing so, we could compare and evaluate the performance of different feature sets in gender classification for Web forums in order to answer research questions 2 and 3.

There were several steps in this component as follows: feature extraction, uni/bigram preselection, and feature selection. Each led to the generation of different feature sets.

Feature Extraction: Different types of features were extracted based on all messages collected from a given open source Web forum. In this study, we extracted the lexical (denoted by F1), syntactic (denoted by F2), and structural (denoted by F3) features as content-free features, and unigrams [denoted by F4(unigram)] and bigrams [denoted by F4(bigram)] as content-specific features.

For F1 features, we adopted the character-based lexical features [14], [18], [34]; the vocabulary richness features [59]; and the word-length frequency features [40], [15]. In total, we used 87 lexical features. For F2 features, we adopted a set of 150 function words used in Zheng *et al.* [66] since this study also focused on Web forum messages, although it is about authorship classification. In addition, we adopted 8 punctuation marks suggested by Baayen *et al.* [8]. Therefore, we used 158 syntactic features in total. For F3 features, we chose five of the most common ones from previous literature [1], [66] that could be applied to a broad number of general Web forums: the total number of sentences per message, the total number of paragraphs per message, the number of sentences per paragraph in a message, the number of characters per paragraph in a message, and the number of words per paragraph in a message. We did not use a large number of structural features related to technical structures (e.g., font colors and font sizes), because some Web forums may not have had the related characteristics. For example, some popular (but old) Web forums do not have functions which allow users to change the font colors and sizes.

Unigram/Bigram Pre-selection: Although content-free features are important for online text classification, content-specific features that consist of important keywords and phrases on certain topics could be more meaningful, thus leading to relatively high representative ability. Content-specific features used in previous online text classification studies are either a relatively small number of manually selected, domain-specific keywords [37], [66], or a relatively large number of n-grams automatically learned from the textual data collection [1], [3], [47], [51]. The large potential feature spaces of n-grams have been shown to be effective for online text classification [2]. Therefore, in this study we used n-grams as content-specific features. Specifically, we used unigrams [i.e., F4(unigram)] and bigrams [i.e., F4(bigram)]. The unigrams and bigrams were extracted from all the messages in the Web forum. After removing the stop-words, we kept the unigrams and bigrams that appeared more than ten times in the whole forum as our content-specific features.

By conducting feature extraction and unigram/bigram preselection, we obtained five types of features. Based on those different types of features, we built three feature sets in an

incremental way: 1) feature set $F1 + F2 + F3$; 2) feature set $F1 + F2 + F3 + F4(\text{unigram})$; and 3) feature set $F1 + F2 + F3 + F4(\text{unigram}) + F4(\text{bigram})$. This incremental order represents the evolutionary sequence of features used for online text classification [2], [66]. Studies [2], [66] have shown that lexical and syntactic features are the foundation for structural and content-specific features. In this study, we used feature set $F1 + F2 + F3$ as the baseline feature set to assess the performance of the other two proposed feature sets which also incorporates content-specific features.

Feature Selection: By adding unigrams and unigrams plus bigrams as content-specific features, respectively, feature sets $F1 + F2 + F3 + F4(\text{unigram})$ and $F1 + F2 + F3 + F4(\text{unigram}) + F4(\text{bigram})$ can be very large. In this study, we conducted feature selection on them using the Information Gain (IG) heuristic due to its reported effectiveness in previous online text classification research [2], [33], thus building the selected feature sets $F1 + F2 + F3 + F4(\text{unigram})$ and $F1 + F2 + F3 + F4(\text{unigram}) + F4(\text{bigram})$. In addition to IG, there are some other feature selection methods that have been reported to have effective performance. For example, Forman [17] found that in general Bi-Normal Separation (BNS) and IG are the two most effective methods compared to other feature selection methods. BNS sometimes performed “marginally” better than IG. However, the gap was barely visible in some cases. In terms of precision, “Information Gain yielded the best result most often” [17]. We did not further compare the performances among various feature selection methods since that was not the focus of this study.

As defined in the following formula, information gain $IG(C, A)$ measures the amount of entropy decrease on a class C when providing a feature A [50], [55]. The decreasing amount of entropy reflects the additional information gained by adding feature A . In the formula, $H(C)$ and $H(C|A)$ represent the entropies of class C before and after observing feature A , respectively. The Information Gain for each feature varies along the range 0–1 with higher values indicating more information gained by providing certain features

$$IG(C, A) = H(C) - H(C|A), \text{ where}$$

$$H(C) = - \sum_{c \in C} p(c) \log_2 p(c),$$

$$H(C|A) = - \sum_{a \in A} p(a) \sum_{c \in C} p(c|a) \log_2 p(c|a).$$

All features with an information gain greater than 0.0025 (i.e., $IG(C, A) > 0.0025$) are selected [3], [63]. The idea is to try to achieve the best classification performance by filtering out the features with less to contribution while keeping the ones with relatively higher discriminatory powers.

C. Classification and Evaluation

To assess the performance of each feature set, we adopted the standard classification performance metrics, i.e., accuracy, precision, recall, and F-measure. These metrics have been widely used in information retrieval and text classification studies [2], [3], [36], especially for data sets with balanced data points among different classes. In this paper, our testbed was quite balanced between the two gender groups, so we chose to use these performance measures. When data sets are unbalanced, the ROC curve can be used as another important performance measure [54]. We chose SVM as the classifier because of its often reported best performance in many previous online text classification studies [2], [3], [29], [37], [66].

Among the four standard measures, accuracy assesses the overall classification correctness; while precision, recall, and F-measure evaluate the correctness of each class, which is shown at the bottom of the page, with classes 1 and 2 being Web forum messages written by female and male authors, respectively.

V. EXPERIMENTAL STUDY

A. Testbed

We conducted our experiment on a large international Islamic women’s political forum to evaluate our proposed framework of gender classification for Web forums. We chose it for three reasons as follows: first, it is a large long-standing (about 4 years) international political forum and thus can be used to study the international cyber political movement; second, it has self-reported gender information¹ for each registered member, thus providing a gold standard to evaluate the performance of our automatic gender classifiers; third, since it is a women’s forum, more females participate, thus providing a larger number of messages written by female authors compared with other general male-dominated Web forums. We believe the international, political, and female-oriented nature of this large active forum makes it an ideal testbed² for our research.

We collected and parsed all the messages in the forum posted up to March, 2007. In total, we gathered 34 695 different messages in 4352 unique threads. The numbers of messages written by females and males were quite balanced, with 17 785 and 16 572 messages written

¹The use of self-reported gender information is a limitation of this paper. It is impossible to check the truthfulness of the self-reported gender information provided by each forum participant. However, because of the following two reasons, we believe that in most cases the self-reported gender information in our testbed forum is relatively reliable. First, the use of this testbed was suggested by a domain scientist who is an expert in women’s political and security studies. She provided a highly positive assessment of the reliability and importance of the content posted on this forum. Second, for this particular forum, there is no potential benefit to the participants for providing false information about their own gender.

$$\text{accuracy} = \frac{\text{number of all correctly classified Web forum messages}}{\text{total number of Web forum messages}}$$

$$\text{precision (i)} = \frac{\text{number of correctly classified Web forum messages for class i}}{\text{total number of Web forum messages classified as class i}}$$

$$\text{recall (i)} = \frac{\text{number of correctly classified Web forum messages for class i}}{\text{total number of Web forum messages in class i}}$$

$$\text{F-measure (i)} = \frac{2 \times \text{precision (i)} \times \text{recall (i)}}{\text{precision (i)} + \text{recall (i)}}, \text{ where } i = 1, 2$$

TABLE I
TESTBED DISTRIBUTION BETWEEN FEMALES AND MALES

| Number Range of Posted Messages | Female | Male |
|---------------------------------|--------|------|
| 5-20 messages | 16 | 18 |
| 21-50 messages | 10 | 16 |
| 51-100 messages | 6 | 4 |
| 101-200 messages | 8 | 4 |
| 201-500 messages | 5 | 6 |
| 501-1000 messages | 4 | 1 |
| 1000+ messages | 1 | 1 |
| Total | 50 | 50 |

by females and males, respectively. An additional 338 messages did not have gender information. The time span of the collected messages is from June 9, 2004 to March 13, 2007. Based on careful discussion with our political science collaborator, who has significant experience in studying women's political forums, we believe that this testbed is of high quality and has credible participant-specified gender information.

To test the performance of our classifiers, we randomly selected 100 authors, 50 females and 50 males. In total, there were 12 690 messages posted by those 100 authors. Table I shows the distribution of the numbers of messages written by females and males. In each number range, there were relatively balanced numbers of messages between females and males. On average, each female participant produced 142.26 messages, and each male participant wrote 111.54 messages.

B. Hypotheses

Drawing on the vast online social media classification literature we reviewed, we posited that adding content-specific features to the baseline content-free features would improve the performance of gender classification for Web forums (targeting research question 2), and that conducting feature selection on a relatively large number of features would improve the performance of gender classification for Web forums (targeting research question 3). The specific hypotheses tested are as follows:

- H1: Using the combination of content-free features and unigrams can achieve higher performance than using only content-free features.
- H2: Using the combination of content-free features, unigrams, and bigrams can achieve higher performance than using the combination of content-free features and unigrams.
- H3: Using the feature set generated by conducting feature selection on the combination of content-free features and unigrams can achieve higher performance than using the combination of content-free features and unigrams without feature selection.
- H4: Using the feature set generated by conducting feature selection on the combination of content-free features, unigrams, and bigrams can achieve higher performance than using the combination of content-free features, unigrams, and bigrams without feature selection.
- H5: Using the feature set generated by conducting feature selection on the combination of content-free features, unigrams, and bigrams can achieve higher performance than using the feature set created by conducting feature selection on the combination of content-free features and unigrams.

C. Experimental Results

Feature set $F1 + F2 + F3$ contained 250 content-free features. For the content-specific features, we obtained 6012 unigrams and 4022 bigrams. Therefore, there were 6262 and 10 284 features in feature sets $F1 + F2 + F3 + F4(\text{unigram})$ and $F1 + F2 + F3 + F4(\text{unigram}) + F4(\text{bigram})$, respectively. After feature selection, the two selected feature sets consisted of 351 and 640 features, respec-

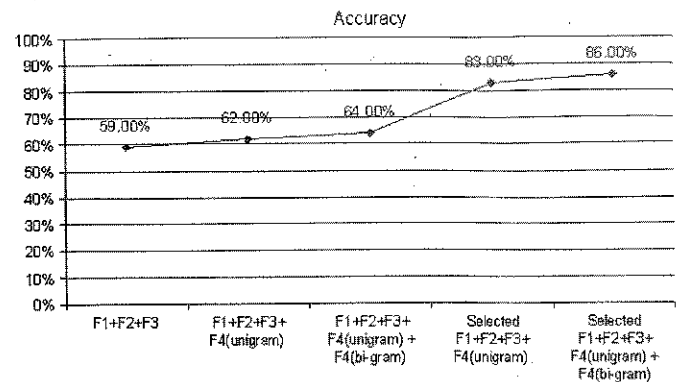


Fig. 2. Accuracy of gender classification on each feature set.

TABLE II
PERFORMANCE MEASURES USING DIFFERENT FEATURE SETS

| Feature Set | Class | Precision | Recall | F-measure |
|---|---------|-----------|--------|-----------|
| $F1 + F2 + F3$ | Female | 57.10% | 72.00% | 63.69% |
| | Male | 62.20% | 46.00% | 52.89% |
| | Average | 59.70% | 59.00% | 59.35% |
| $F1 + F2 + F3 + F4(\text{unigram})$ | Female | 63.00% | 58.00% | 60.40% |
| | Male | 61.10% | 66.00% | 63.46% |
| | Average | 62.10% | 62.00% | 62.05% |
| $F1 + F2 + F3 + F4(\text{unigram}) + F4(\text{bi-gram})$ | Female | 62.50% | 70.00% | 66.04% |
| | Male | 65.90% | 58.00% | 61.70% |
| | Average | 64.20% | 64.00% | 64.10% |
| Selected $F1 + F2 + F3 + F4(\text{unigram})$ | Female | 90.20% | 74.00% | 81.30% |
| | Male | 78.00% | 92.00% | 84.42% |
| | Average | 84.10% | 83.00% | 83.55% |
| Selected $F1 + F2 + F3 + F4(\text{unigram}) + F4(\text{bi-gram})$ | Female | 92.90% | 78.00% | 84.80% |
| | Male | 81.00% | 94.00% | 87.02% |
| | Average | 86.90% | 86.00% | 86.45% |

tively, each of which was much smaller than the corresponding one without feature selection. The feature selection was carried out by Weka's information gain attribute evaluator [62].

The classification was carried out by using a linear kernel with the Sequential Minimal Optimization algorithm [49] included in the Weka Data Mining Package [62]. Evaluation was done via tenfold cross validation. In each fold, 90% of the data was used as the training set and the remaining 10% as the testing set. Fig. 2 shows the accuracy of gender classification on each feature set. Accuracy increased as more types of features were incorporated. Specifically, feature set $F1 + F2 + F3 + F4(\text{unigram}) + F4(\text{bigram})$ outperformed $F1 + F2 + F3 + F4(\text{unigram})$, which in turn outperformed $F1 + F2 + F3$. In addition, feature selection improved the classification accuracies significantly. Particularly, after conducting feature selection, the classification accuracies on feature sets $F1 + F2 + F3 + F4(\text{unigram})$ and $F1 + F2 + F3 + F4(\text{unigram}) + F4(\text{bigram})$ increased from 62% to 83% and 64% to 86%, respectively. The highest classification accuracy was achieved on the selected feature set $F1 + F2 + F3 + F4(\text{unigram}) + F4(\text{bigram})$.

Table II shows the precision, recall, and F-measure of gender classification on each feature set. All three types of measurement values increased in the same way as the accuracy (see Fig. 2). The highest precision, recall, and F-measure for both classes (i.e., female and male) were achieved on the selected feature set $F1 + F2 + F3 + F4(\text{unigram}) + F4(\text{bigram})$. Compared to Corney *et al.*'s study [12] of gender classification on e-mails, the best F-measure (i.e., 86.45%) we have achieved was higher than that (i.e., 71.1%) reported in their study. However, the highest accuracy (i.e., 91.5%) reported in Nowson and Oberlander's work [43] on gender classification in the context of

TABLE III
RESULTS OF HYPOTHESES TESTING ON ACCURACY AND
F-MEASURE FOR HYPOTHESES 1 THROUGH 5

| No. | <i>p</i> value on accuracy | Result | <i>p</i> value on average F-measure | Result |
|-----|----------------------------|-----------|-------------------------------------|-----------|
| H1 | <0.0001** | Supported | <0.0001** | Supported |
| H2 | 0.0080** | Supported | 0.0400* | Supported |
| H3 | <0.0001** | Supported | <0.0001** | Supported |
| H4 | <0.0001** | Supported | <0.0001** | Supported |
| H5 | <0.0001** | Supported | <0.0001** | Supported |

Note. Significance levels * $\alpha = 0.05$ and ** $\alpha = 0.01$.

Web blog was higher than the highest accuracy (i.e., 86%) we achieved in this paper. This could be attributed to the shorter text in the Web forum messages compared with the Web blog corpus they used.

Pairwise *t*-tests on accuracy and F-measure were performed to test H1 through H5. The tests were conducted by randomly shuffling the data before performing tenfold cross validation; i.e., we reshuffled the data in a different way each time and then performed tenfold cross validation. We repeated this process 30 times. As summarized in Table III, the classification accuracy and F-measure were significantly higher ($p < 0.0001$ on both accuracy and F-measure) when using the combination of content-free features and unigrams than those using only content-free features. Thus, H1 is supported. Compared with the combination of content-free features and unigrams, the combination of content-free features, unigrams, and bigrams showed significantly higher classification accuracy ($p = 0.0080$) and F-measure ($p = 0.0400$). Therefore, H2 is supported. After feature selection, the selected feature set $F1 + F2 + F3 + F4(\text{unigram})$ outperformed $F1 + F2 + F3 + F4(\text{unigram})$ significantly ($p < 0.0001$ on both accuracy and F-measure), in support of H3; and the selected feature set $F1 + F2 + F3 + F4(\text{unigram}) + F4(\text{bigram})$ outperformed $F1 + F2 + F3 + F4(\text{unigram}) + F4(\text{bigram})$ significantly ($p < 0.0001$ on both accuracy and F-measure), in support of H4. In addition, the classification accuracy and F-measure were significantly higher ($p < 0.0001$ on both accuracy and F-measure) when using the selected feature set $F1 + F2 + F3 + F4(\text{unigram}) + F4(\text{bigram})$ than when using the selected feature set $F1 + F2 + F3 + F4(\text{unigram})$. Thus, H5 is supported.

D. Discussion

Feature set $F1 + F2 + F3$ achieved 59.00% accuracy, 59.70% average precision, 59.00% average recall, and 59.35% average F-measure. However, compared with the results reported in previous online text classification studies [1], [66], the performance results we achieved were worse. This could be attributed to several causes. First, as mentioned in previous studies, although they represent vocabulary richness, lexical features (F1) may not be very useful when the text length is short [66]. Since some Web forum messages in our data set are quite short, the lexical features would not have been, after all, very effective. Second, compared with the relatively small number of words in a Web forum message, the 150 function words we used as part of the syntactic features (F2) may be more than necessary. In their study, de Vel *et al.* [15] observed a decrease in performance when the number of function words increased from 122 to 320. Third, since we used only five structural features (F3) in this paper, we might have missed other important ones that may have had an impact in previous studies.

Feature set $F1 + F2 + F3 + F4(\text{unigram})$ was significantly better than $F1 + F2 + F3$. This result is consistent with the previous studies [2], [66] that point out the good discriminating capability of content-specific features. Those studies have noticed that Web forum participants were interested in different topics, thus providing the content-specific features with relatively high discriminatory power.

Feature set $F1 + F2 + F3 + F4(\text{unigram}) + F4(\text{bigram})$ was significantly better than $F1 + F2 + F3 + F4(\text{unigram})$ (Table III). This result indicates that although unigrams are very important content-specific features, they may not be sufficient to represent the content. By incorporating bigrams, we can capture more content information about the Web forum messages. However, we also noticed that the increases of both precision and F-measure, although significant, were not as great as the increases from feature set $F1 + F2 + F3 + F4(\text{unigram})$ to the baseline feature set $F1 + F2 + F3$. One possible reason could be that although the large number of bigrams captured more content information, it introduced noise as well [35], [39].

Koppel and Schler [33] have shown that conducting feature selection on *n*-grams can improve the text classification results. According to the *t*-test results of H3 and H4 (Table III), feature selection improved the classification performance significantly. As shown in Fig. 2 and Table II, all four performance measures increased significantly after conducting feature selection on feature sets $F1 + F2 + F3 + F4(\text{unigram})$ and $F1 + F2 + F3 + F4(\text{unigram}) + F4(\text{bigram})$, respectively. In addition, the numbers of features in the selected feature sets were reduced appreciably, thus leading to higher efficiency.

The testing between the two selected feature sets (see H5) showed that the selected feature set $F1 + F2 + F3 + F4(\text{unigram}) + F4(\text{bigram})$ outperformed the selected feature set $F1 + F2 + F3 + F4(\text{unigram})$ significantly (Table III). Similar to the comparison between feature sets $F1 + F2 + F3 + F4(\text{unigram}) + F4(\text{bigram})$ and $F1 + F2 + F3 + F4(\text{unigram})$, this result once again indicates that using both unigrams and bigrams can better improve the gender classification performance for Web forums than using only unigrams as content-specific features.

E. Different Topics of Interests: Females and Males

In this paper, we achieved the highest classification accuracy of 86% on the selected feature set $F1 + F2 + F3 + F4(\text{unigram}) + F4(\text{bigram})$. This indicates that gender differences do exist in Web forums and the features used for classification, especially the content-specific features, have a high discriminating capability of distinguishing the online gender differences between female and male posters, thus answering our research question 1.

By investigating the features in the selected feature set $F1 + F2 + F3 + F4(\text{unigram}) + F4(\text{bigram})$, we observed that females talked more about family members, God, peace, marriage, and good will; on the other hand, males talked more about extremism, holy men, and belief.

Table IV lists some examples of the unigrams and bigrams preferred by females and males, respectively from the selected feature set $F1 + F2 + F3 + F4(\text{unigram}) + F4(\text{bigram})$. They are among the features with the highest information gain values, therefore showing high discriminatory power. We conducted chi-square (χ^2) tests to examine the statistical significances of the differences in using those unigrams and bigrams between females and males. A domain expert from an Islamic country provided the meanings of some of those unigrams and bigrams.

As summarized in Table IV, significant terms/words in female conversations included the following: sis (i.e., sisters in Islam), sister, mother, husband, flower, amen, alhamdulillah (i.e., thank God), inshaallaah (i.e., in God's will), ahah kheir (i.e., God is good), and sexually defiled. Significant terms/words in male discussions included the following: Salafi (i.e., an extremist sect of Islam), Allah (i.e., Allah God of Muslims), army, deviant, ijtihaad (i.e., inferring or interpreting Islamic laws), e-mail, great scholar, Muslim intellectual, and imam Nawawi (i.e., Priest Nawawi). For the bigram "original Arabic," although men preferred to use it more frequently than women,

TABLE IV
EXAMPLES OF FEMALE AND MALE PREFERRED UNIGRAMS
AND BIGRAMS FROM THE SELECTED FEATURE SET
F1 + F2 + F3 + F4(unigram) + F4(bigram)

| <i>Female preferred unigrams and bi-grams</i> | | | |
|---|----------------|----------------|--|
| Keyword | χ^2 value | <i>p</i> value | Meaning |
| Sis | 456.07 | <0.0001** | Sisters in Islam |
| Sister | 165.08 | <0.0001** | |
| Mother | 123.88 | <0.0001** | |
| Husband | 51.87 | <0.0001** | |
| Flower | 9.00 | 0.0030** | |
| Amen | 166.64 | <0.0001** | |
| Alhamdulillah | 283.85 | <0.0001** | Thank God |
| Inshaallaah | 33.51 | <0.0001** | In God's will |
| Ahhah kheir | 15.16 | <0.0001** | God is good |
| Sexually defiled | 5.25 | 0.0220* | |
| <i>Male preferred unigrams and bi-grams</i> | | | |
| Keyword | χ^2 value | <i>p</i> value | Meaning |
| Salafi | 377.17 | <0.0001** | Extremist sect of Islam |
| Allah | 290.30 | <0.0001** | Allah God of Muslims |
| Army | 66.12 | <0.0001** | |
| Deviant | 35.79 | <0.0001** | |
| Ijtihad | 57.80 | <0.0001** | Inferring or interpreting Islamic laws |
| E-mail | 23.81 | <0.0001** | |
| Great scholar | 13.89 | 0.0002** | |
| Muslim intellectual | 11.27 | 0.0008** | |
| Imam Nawawi | 26.56 | <0.0001** | Priest Nawawi |
| Original Arabic | 3.52 | 0.0606 | |

Note. Significance levels * $\alpha = 0.05$ and ** $\alpha = 0.01$.

the difference was not statistically significant ($p = 0.0606 > 0.05$). This may be because the total number of its appearances in the whole forum was small and therefore could not show statistical significance.

The results of our experimental study show the importance of content-specific features in gender classification for Web forums and are consistent with previous gender classification studies for Web blogs [43], [51].

As an important type of social media, political Web forums have become a major communication channel for people to discuss and debate political, cultural, and social issues. More and more women are using this medium to share their political opinions and knowledge. Along with this trend, researchers have developed an increased interest in studying online gender differences. By analyzing writing styles and topics of interest, our experimental results indicate that female and male participants in political Web forums do have significantly different topics of interest.

VI. CONCLUSION AND FUTURE DIRECTIONS

With the rapid development and the increasing importance of the Internet in people's daily lives and work, understanding online gender differences and why they occur is becoming more and more important for Internet service providers, system developers, information analysts, and end users.

Nowadays, more and more women are participating in cyberspace. However, this does not diminish online gender differences. In contrast, discrepancies of motivation and interest in Internet use between females and males are becoming the focus of online gender difference research. In this paper, we have used feature-based online social media text classification techniques to investigate the online gender differences between female and male participants in Web forums, by examining their writing styles and topics of interest.

In the framework, we examined different types of features that have been widely used in previous online text classification studies, includ-

ing the following: lexical, syntactic, structural, and content-specific features. We built five different feature sets by adding content-specific features to the basic content-free features and conducting feature selection. In our experimental study on a large Islamic women's political forum, the feature sets combining both content-specific and content-free features performed significantly better than the ones consisting of only content-free features. In addition, feature selection on large feature sets improved the classification performance significantly. The results also indicated the existence of online gender differences in Web forums. Further investigation identified different topics of interest between females and males.

This research has made several contributions. First, we proposed a systematic framework of gender classification to analyze online gender differences in social media, an area which has received little investigative attention. The framework can be applied to study the gender differences in many other domains. Our approach provides an informative point of departure for continued research. Second, our empirical study demonstrated the effectiveness of the proposed framework, thus confirming the prevalence of online gender differences in Web forums. Third, we also make a research contribution by examining different feature sets and identifying the one with the best classification performance. The comparison of different feature sets also indicates the importance of incorporating content-specific features and conducting feature selection in automatic gender classification for online social media.

This paper also has some limitations that can be explored further. First, we used only unigrams and bigrams as content-specific features. Because of their computational complexity, we did not include n -grams with n greater than 2. However, those n -grams could capture more content information than unigrams and bigrams, thus potentially leading to higher classification performance. On the other hand, if n is too big, the n -grams may introduce additional noise and computational overhead. More systematic investigation may be needed. Second, to generate the selected feature sets, we adopted information gain which is one of the most widely used feature selection methods. However, there are other advanced feature selection methods, such as the wrapper model, the filter model, and the Markov blanket. Future research could explore and compare their performance in gender classification in the Web forum context. Third, we tested our proposed gender classification framework on only one English language forum. We believe the framework can be applied to Web forums in different languages, but feature representation and extraction research would need to be conducted to better develop a scalable, multilingual online gender classification model. Lastly, we plan to explore the gender differences in other important social media domains, such as marketing, e-commerce, health care, and education. These are all areas in which women may exhibit unique characteristics and exercise significant influence. Additional social, cultural, and psychological models would also then need to be considered in future research.

ACKNOWLEDGMENT

The authors would like to thank Dr. K. von Knop for her helpful suggestions and comments about our research testbed.

REFERENCES

- [1] A. Abbasi and H. Chen, "Applying authorship analysis to extremist-group web forum messages," *IEEE Intell. Syst.—Special Issue on Artificial Intelligence for National and Homeland Security*, vol. 20, no. 5, pp. 67–75, Sep./Oct. 2005.
- [2] A. Abbasi and H. Chen, "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace," *ACM Trans. Inf. Syst.*, vol. 26, no. 2, pp. 1–29, Mar. 2008.

- [3] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums," *ACM Trans. Inf. Syst.*, vol. 26, no. 3, pp. 1–34, Jun. 2008.
- [4] S. Argamon, M. Koppel, and G. Avneri, "Routing documents according to style," in *Proc. 1st Int. Workshop Innovative Inf.*, Pisa, Italy, 1988.
- [5] S. Argamon, M. Koppel, J. Fine, and A. Shimoni, "Gender, genre, and writing style in formal written texts," *Text*, vol. 23, no. 3, pp. 321–346, 2003.
- [6] S. Argamon, M. Saric, and S. S. Stein, "Style mining of electronic messages for multiple authorship discrimination," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2003, pp. 475–480.
- [7] R. H. Baayen, H. V. Halteren, A. Neijt, and F. J. Tweedie, "An experiment in authorship attribution," in *Proc. 6th Int. Conf. Stat. Anal. Textual Data*, 2002, pp. 69–75.
- [8] R. H. Baayen, H. V. Halteren, and F. J. Tweedie, "Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution," *Literary Linguistic Comput.*, vol. 11, no. 3, pp. 121–132, Sep. 1996.
- [9] B. Bimber, "Measuring the gender gap on the Internet," *Social Sci. Quart.*, vol. 81, no. 3, pp. 863–876, 2000.
- [10] CommerceNet, The CommerceNet/Nielsen Internet Demographic Survey, 1999. [Online]. Available: <http://www.commerce.net/>
- [11] M. Consalvo and S. Paasonen, *Women & Everyday Uses of the Internet: Agency & Identity*. New York: Peter Lang Publ., 2002.
- [12] M. Corney, O. deVel, A. Anderson, and G. Mohay, "Gender-preferential text mining of e-mail discourse," in *Proc. 18th ACSAC*, Las Vegas, NV, 2002, pp. 282–292.
- [13] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal.*, vol. 1, no. 3, pp. 131–156, 1997.
- [14] O. deVel, "Mining e-mail authorship," presented at the Workshop Text Mining, ACM Int. Conf. Knowledge Discovery Data Mining (KDD), Boston, MA, 2000.
- [15] O. deVel, A. Anderson, M. Corney, and G. Mohay, "Mining e-mail content for author identification forensics," *SIGMOD Rec.*, vol. 30, no. 4, pp. 55–64, 2001.
- [16] J. Diederich, J. Kindermann, E. Leopold, and G. Paass, "Authorship attribution with support vector machines," *Appl. Intell.*, vol. 19, no. 1/2, pp. 109–123, Jul. 2003.
- [17] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, 2003.
- [18] R. S. Forsyth and D. I. Holmes, "Feature finding for text classification," *Literary Linguistic Comput.*, vol. 11, no. 4, pp. 163–174, 1996.
- [19] J. E. Fountain, "Constructing the information society: Women, information technology, and design," *Technol. Soc.*, vol. 22, no. 1, pp. 45–62, Jan. 2000.
- [20] J. E. Fuller, "Equality in cyberdemocracy? Gauging gender gaps in on-line civic participation," *Social Sci. Quart.*, vol. 85, no. 4, pp. 938–957, Dec. 2004.
- [21] M. Gamon, "Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis," in *Proc. 20th Int. Conf. Comput. Linguistics*, 2004, pp. 841–847.
- [22] G. Grefenstette, Y. Qu, J. G. Shanahan, and D. A. Evans, "Coupling niche browsers and affect analysis for an opinion mining application," in *Proc. 12th Int. Conf. Recherche d'Information Assistee par Ordinateur*, 2004, pp. 186–194.
- [23] J. Guiller and A. Durndell, "Students' linguistic behaviour in online discussion groups: Does gender matter?," *Comput. Human Behavior*, vol. 23, no. 5, pp. 2240–2255, Sep. 2007.
- [24] B. Guo and M. S. Nixon, "Gait feature subset selection by mutual information," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 39, no. 1, pp. 36–46, Jan. 2009.
- [25] D. Halbert, "Shulamith firestone: Radical feminism and visions of the information society," *Inf. Commun. Soc.*, vol. 7, no. 1, pp. 115–136, 2004.
- [26] W. Harcourt, "The personal and the political: Women using the Internet," *CyberPsychology Behavior*, vol. 3, no. 5, pp. 693–697, Oct. 2000.
- [27] D. Harp and M. Tremayne, "The gendered blogosphere: Examining inequality using network and feminist theory," *Journalism Mass Commun. Quart.*, vol. 83, no. 2, pp. 247–264, 2006.
- [28] S. Hota, S. Argamon, M. Koppel, and I. Zigdon, "Performing gender: Automatic stylistic analysis of Shakespeare's characters," in *Proc. Digital Humanit. Conf. (Association for Computers in Humanities and the Association for Literary and Linguistic Computing)*, 2006, pp. 100–106.
- [29] P. J.-H. Hu, T.-H. Cheng, C.-P. Wei, C.-H. Yu, A. L. F. Chan, and H.-Y. Wang, "Managing clinical use of high-alert drugs: A supervised learning approach to pharmacokinetic data analysis," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 37, no. 4, pp. 481–492, Jul. 2007.
- [30] L. A. Jackson, K. S. Ervin, P. D. Gardner, and N. Schmitt, "Gender and the Internet: Women communicating and men searching," *Sex Roles: J. Res.*, vol. 44, no. 5/6, pp. 363–378, Mar. 2001.
- [31] M. Koppel, N. Akiva, and I. Dagan, "Feature instability as a criterion for selecting potential style markers," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 57, no. 11, pp. 1519–1525, Sep. 2006.
- [32] M. Koppel, S. Argamon, and A. Shimoni, "Automatically categorizing written texts by author gender," *Literary Linguistic Comput.*, vol. 17, no. 4, pp. 401–412, Nov. 2002.
- [33] M. Koppel and J. Schler, "Exploiting stylistic idiosyncrasies for authorship attribution," in *Proc. IJCAI Workshop Comput. Approaches Style Anal. Synthesis*, Acapulco, Mexico, 2003, pp. 69–72.
- [34] G. R. Ledger and T. V. N. Merriam, "Shakespeare, Fletcher, and the two Noble Kinsmen," *Literary Linguistic Comput.*, vol. 9, no. 3, pp. 235–248, 1994.
- [35] J. Li, H. Su, H. Chen, and B. W. Futscher, "Optimal search-based gene subset selection for gene array cancer classification," *IEEE Trans. Inf. Technol. Biomed.*, vol. 11, no. 4, pp. 398–405, Jul. 2007.
- [36] J. Li, Z. Zhang, X. Li, and H. Chen, "Kernel-based learning for biomedical relation extraction," *J. Amer. Soc. Inf. Sci. Technol. (JASIST)*, vol. 59, no. 5, pp. 756–769, Mar. 2008.
- [37] J. Li, R. Zheng, and H. Chen, "From fingerprint to writeprint," *Commun. ACM*, vol. 49, no. 4, pp. 76–82, Apr. 2006.
- [38] C. Martindale and D. Mckenzie, "On the utility of content analysis in author attribution: The federalist," *Comput. Humanit.*, vol. 29, no. 4, pp. 259–270, Aug. 1995.
- [39] R. Meiri and J. Zahavi, "Using simulated annealing to optimize the feature selection problem in marketing applications," *Eur. J. Oper. Res.*, vol. 171, no. 3, pp. 842–858, Jun. 2006.
- [40] T. C. Mendenhall, "The characteristic curves of composition," *Science*, vol. 11, no. 11, pp. 237–249, 1887.
- [41] A. Mitra, "Voices of the marginalized on the Internet: Examples from a website for women of South Asia," *J. Commun.*, vol. 54, no. 3, pp. 492–510, Sep. 2004.
- [42] NationalElectionStudy, "American National Election Study. 1998 Pre- and Post- Election Survey," Conducted by the Center for Political Studies of the Institute for Social Research, The University of Michigan, Ann Arbor, Inter-University Consortium for Political and Social Research 1998.
- [43] S. Nowson and J. Oberlander, "The identity of bloggers: Openness and gender in personal weblogs," in *Proc. AAAI Spring Symposia Comput. Approaches Analyzing Weblogs*, Stanford, CA, 2006.
- [44] T. O'Reilly, "What is Web 2.0? Design Patterns and Business Models for the Next Generation of Software, 2005." [Online]. Available: <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>
- [45] C. Ogan, F. Cicek, and M. Ozakca, "Letters to Sarah: Analysis of email responses to an online editorial," *New Media Soc.*, vol. 7, no. 4, pp. 533–557, Aug. 2005.
- [46] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2002, pp. 79–86.
- [47] F. Peng, D. Schuurmans, V. Kesselj, and S. Wang, "Automated authorship attribution with character level language models," in *Proc. 10th Conf. Eur. Chapter Assoc. Comput. Linguistics*, Budapest, Hungary, 2003.
- [48] Pew Internet and American Life Project, 2008. [Online]. Available: http://www.pewinternet.org/trends/User_Demo_7.22.08.htm
- [49] J. Platt, "Fast training on SVMs using sequential minimal optimization," in *Advances in Kernel Methods: Support Vector Learning*, B. Scholkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1999.
- [50] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.
- [51] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker, "Effects of age and gender on blogging," in *Proc. AAAI Spring Symp. Comput. Approaches Analyzing Weblogs*, Menlo Park, CA, 2006, pp. 199–205.
- [52] S. Scott and S. Matwin, "Feature engineering for text classification," in *Proc. 16th ICML*, 1999, pp. 379–388.
- [53] C. Seale, S. Ziebland, and J. Charteris-Black, "Gender, cancer experience and Internet use: A comparative keyword analysis of interviews and online cancer support groups," *Social Sci. Med.*, vol. 62, no. 10, pp. 2577–2590, May 2006.
- [54] C. Seiffert, T. M. Khoshgoftaar, J. V. Hulse, and A. Napolitano, "RUS-Boost: A hybrid approach to alleviating class imbalance," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 40, no. 1, pp. 185–197, Jan. 2010.
- [55] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 4, pp. 379–423, 1948.

- [56] A. P. Sherman, *Cybergrrl @ Work: Tips and Inspiration for the Professional You*. New York: Berkley Trade, 2001.
- [57] P. Subasic and A. Huettner, "Affect analysis of text using fuzzy semantic typing," *IEEE Trans. Fuzzy Syst.*, vol. 9, no. 4, pp. 483–496, Aug. 2001.
- [58] P. D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in *Proc. 40th Annu. Meetings Assoc. Comput. Linguistics*, Philadelphia, PA, 2002, pp. 417–424.
- [59] F. J. Tweedie and R. H. Baayen, "How variable may a constant be? Measures of lexical richness in perspective," *Comput. Humanit.*, vol. 32, no. 5, pp. 323–352, Sep. 1998.
- [60] J. Wiebe, T. Wilson, and M. Bell, "Identifying collocations for recognizing opinions," in *Proc. ACL/EACL Workshop Collocation*, Toulouse, France, 2001.
- [61] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin, "Learning subjective language," *Comput. Linguistics*, vol. 30, no. 3, pp. 277–308, Sep. 2004.
- [62] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005.
- [63] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proc. ICML*, 1997, pp. 412–420.
- [64] G. Youngs, "Cyberspace: The new feminist frontier," in *Women and Media: International Perspectives*, K. Ross and C. M. Byerly, Eds. Malden, MA: Wiley-Blackwell, 2004, pp. 185–208.
- [65] G. U. Yule, *The Statistical Study of Literary Vocabulary*. Cambridge, U.K.: Cambridge Univ. Press, 1944.
- [66] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *J. Amer. Soc. Inf. Sci. Technol. (JASIST)*, vol. 57, no. 3, pp. 378–393, Feb. 2006.