Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums

AHMED ABBASI, HSINCHUN CHEN, and ARAB SALEM
The University of Arizona

The Internet is frequently used as a medium for exchange of information and opinions, as well as propaganda dissemination. In this study the use of sentiment analysis methodologies is proposed for classification of Web forum opinions in multiple languages. The utility of stylistic and syntactic features is evaluated for sentiment classification of English and Arabic content. Specific feature extraction components are integrated to account for the linguistic characteristics of Arabic. The entropy weighted genetic algorithm (EWGA) is also developed, which is a hybridized genetic algorithm that incorporates the information-gain heuristic for feature selection. EWGA is designed to improve performance and get a better assessment of key features. The proposed features and techniques are evaluated on a benchmark movie review dataset and U.S. and Middle Eastern Web forum postings. The experimental results using EWGA with SVM indicate high performance levels, with accuracies of over 91% on the benchmark dataset as well as the U.S. and Middle Eastern forums. Stylistic features significantly enhanced performance across all testbeds while EWGA also outperformed other feature selection methods, indicating the utility of these features and techniques for document-level classification of sentiments.

Categories and Subject Descriptors: I.5.3 [**Pattern Recognition**]: Clustering—*Algorithms*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Sentiment analysis, opinion mining, feature selection, text classification

ACM Reference Format:

Abbasi, A., Chen, H., and Salem, A. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Trans. Inform. Syst. 26, 3, Article 12 (June 2008), 34 pages. DOI = 10.1145/1361684.1361685 http://doi.acm.org/10.1145/1361684.1361685

Authors' addresses: A. Abbasi, H. Chen, and A. Salem, Department of Management Information Systems, University of Arizona, 1130 E. Helen St., Tucson, AZ 85721; email: {aabbasi, hchen, asalem}@email.arizona.edu.

1. INTRODUCTION

Analysis of Web content is becoming increasingly important due to augmented communication via computer mediated communication (CMC) Internet sources such as email, Web sites, forums, and chat rooms. The numerous benefits of the Internet and CMC have been coupled with the realization of some vices, including cybercrime. In addition to misuse in the form of deception, identity theft, and the sales and distribution of pirated software, the Internet has also become a popular communication medium and haven for extremist and hate groups. This problematic facet of the Internet is often referred to as the Dark Web [Chen 2006].

Stormfront, what many consider to be the first hate-group Web site [Kaplan and Weinberg 1998], was created around 1996. Since then, researchers and hate watch-organizations have begun to focus their attention towards studying and monitoring such online groups [Leets 2001]. Despite the increased focus on analysis of such groups' Web content, there has been limited evaluation of forum postings, with the majority of studies focusing on Web sites. Burris et al. [2000] acknowledged a need to evaluate forum and chat-room discussion content. Schafer [2002] also stated that it was unclear as to how much and what kind of forum activity was going on with respect to hateful cyberactivist groups. Due to the lack of understanding and current ambiguity associated with the content of such groups' forum postings, analysis of extremist-group forum archives is an important endeavor.

Sentiment analysis attempts to identify and analyze opinions and emotions. Hearst [1992] and Wiebe [1994] originally proposed the idea of mining direction-based text, namely, text containing opinions, sentiments, affects, and biases. Traditional forms of content analysis such as topical analysis may not be effective for forums. Nigam and Hurst [2004] found that only 3% of USENET sentences contained topical information. In contrast, Web discourse is rich in sentiment-related information [Subasic and Huettner 2001]. Consequently, in recent years, sentiment analysis has been applied to various forms of Web-based discourse [Agarwal et al. 2003; Efron 2004]. Application to extremist-group forums can provide insight into important discussion and trends.

In this study we propose the application of sentiment analysis techniques to hate/extremist-group forum postings. Our analysis encompasses classification of sentiments on a benchmark movie review dataset and two forums: a U.S. supremacist and a Middle Eastern extremist group. We evaluate different feature sets consisting of syntactic and stylistic features. We also develop the entropy weighted genetic algorithm (EWGA) for feature selection. The features and techniques result in the creation of a sentiment analysis approach geared towards classification of Web discourse sentiments in multiple languages. The results, using support vector machines (SVM) indicate a high level of classification accuracy, demonstrating the efficacy of this approach for classifying and analyzing sentiments in extremist forums.

The remainder of this article is organized as follows. Section 2 presents a review of current research on sentiment classification. Section 3 describes

research gaps and questions, while Section 4 presents our research design. Section 5 describes the EWGA algorithm and our proposed feature set. Section 6 presents experiments used to evaluate the effectiveness of the proposed approach and discussion of the results. Section 7 concludes with closing remarks and future directions.

2. RELATED WORK

Extremist groups often use the Internet to promote hatred and violence [Glaser et al. 2002]. The Internet offers a ubiquitous, quick, inexpensive, and anonymous means of communication for such groups [Crilley 2001]. Zhou et al. [2005] did an in-depth analysis of U.S. hate-group Web sites and found significant evidence of fund raising-, propaganda-, and recruitment-related content. Abbasi and Chen [2005] also corroborated signs of Web usage as a medium for propaganda by U.S. supremacist and Middle Eastern extremist groups. These findings provide insight into extremist-group Web usage tendencies; however, there has been little analysis of Web forums. Burris et al. [2000] acknowledged the need to evaluate forum and chat-room discussion content. Schafer [2002] was also unclear as to how much and what kind of forum activity was going on with respect to extremist groups. Automated analysis of Web forums can be an arduous endeavor due to the large volumes of noisy information contained in CMC archives. Consequently, previous studies have predominantly incorporated manual or semiautomated methods [Zhou et al. 2005]. Manual examination of thousands of messages can be an extremely tedious effort when applied across thousands of forum postings. With increasing usage of CMC, the need for automated text classification and analysis techniques has grown in recent years. While numerous forms of text classification exist, we focus primarily on sentiment analysis for two reasons. Firstly, Web discourse is rich in opinion- and emotion-related content. Secondly, analysis of this type of text is highly relevant to propaganda usage on the Web, since directional/opinionated text plays an important role in influencing people's perceptions and decision making [Picard 1997].

2.1 Sentiment Classification

Sentiment analysis is concerned with analysis of direction-based text, that is, text containing opinions and emotions. We focus on sentiment classification studies which attempt to determine whether a text is objective or subjective, or whether a subjective text contains positive or negative sentiments. Sentiment classification has several important characteristics, including various tasks, features, techniques, and application domains. These are summarized in the taxonomy presented in Table I.

We are concerned with classifying sentiments in extremist-group forums. Based on the proposed taxonomy, Table II shows selected previous studies dealing with sentiment classification. We discuss the taxonomy and related studies in detail next.

Table I. A Taxonomy of Sentiment Polarity Classification

Tasks		
Category	Description	Label
Classes	Positive/negative sentiments or objective/subjective texts	C1
Level	Document or sentence/phrase-level classification	C2
Source/Target	Whether source/target of sentiment is known or extracted	C3
Features		
Category	Examples	Label
Syntactic	Word/POS tag n-grams, phrase patterns, punctuation	F1
Semantic	Polarity tags, appraisal groups, semantic orientation	F2
Link Based	Web links, send/reply patterns, and document citations	F3
Stylistic	Lexical and structural measures of style	F4
Techniques		
Category	Examples	Label
Machine Learning	Techniques such as SVM, naïve Bayes, etc.	T1
Link Analysis	Citation analysis and message send/reply patterns	T2
Similarity Score	Phrase pattern matching, frequency counts, etc.	T3
Domains		
Category	Description	Label
Reviews	Product, movie, and music reviews	D1
Web Discourse	Web forums and blogs	D2
News Articles	Online news articles and Web pages	D3

2.2 Sentiment Analysis Tasks

There have been several sentiment polarity classification tasks. Three important characteristics of the various sentiment polarity classification tasks are the classes, classification levels, and assumptions about sentiment source and target (topic). The common two-class problem involves classifying sentiments as positive or negative [Pang et al. 2002; Turney 2002]. Additional variations include classifying messages as opinionated/subjective or factual/objective [Wiebe et al. 2004, 2001]. A closely related problem is affect classification, which attempts to classify emotions instead of sentiments. Example affect classes include happiness, sadness, anger, horror, etc. [Subasic and Huettner 2001; Grefenstette et al. 2004; Mishne 2005].

Sentiment polarity classification can be conducted at document-, sentence-, or phrase- (part of sentence) level. Document-level polarity categorization attempts to classify sentiments in movie reviews, news articles, or Web forum postings [Wiebe et al. 2001; Pang et al. 2002; Mullen and Collier 2004; Pang and Lee 2004; Whitelaw et al. 2005]. Sentence-level polarity categorization attempts to classify positive and negative sentiments for each sentence [Yi et al. 2003; Mullen and Collier 2004; Pang and Lee 2004], or whether a sentence is subjective or objective [Riloff et al. 2003]. There has also been work on phrase-level categorization in order to capture multiple sentiments that may be present within a single sentence [Wilson et al. 2005].

In addition to sentiment classes and categorization levels, different assumptions have also been made about the sentiment sources and targets [Yi et al. 2003]. In this study we focus on document-level sentiment polarity categorization (i.e., distinguishing positive- and negative-sentiment texts). However, we

Table II. Selected Previous Studies in Sentiment Polarity Classification

Study	Features				Reduce Feats.	Techniques			Domains			No. Lang.
	F1	F2	F3	F4	Yes/No	T1	T2	T3	D1	D2	D3	1-n
Subasic & Huettner, 2001	√	√			No			√			√	1
Tong, 2001	√	√			No			√	√			1
Morinaga et al., 2002	√				Yes			√	√			1
Pang et al., 2002	√				No	√			√			1
Turney, 2002	√	√			No			√	√			1
Agrawal et al., 2003	√		√		No	√	√			√		1
Dave et al., 2003	√				No	√		√	√			1
Nasukawa & Yi, 2003	√	√			No			√	√			1
Riloff et al., 2003		√		√	No	√					√	1
Yi et al., 2003	√	√			Yes			√	√		√	1
Yu & Hatzivassiloglou, 2003	√	√			No	√		√			√	1
Beineke et al., 2004		√			No	√		√	√			1
Efron, 2004	√		√		No	√	√			√		1
Fei et al., 2004		√			No			√	√			1
Gamon, 2004	√			√	Yes	√			√			1
Grefenstette et al., 2004	√	√			No			√		√		1
Hu & Liu, 2004	√	√			No			√	√			1
Kanayama et al., 2004	√	√			No			√	√			1
Kim & Hovy, 2004		√			No			√		√		1
Pang & Lee, 2004	√	√			No	√		√	√			1
Mullen & Collier, 2004	√	√			No	√			√			1
Nigam & Hurst, 2004	√	√			No	√				√		1
Wiebe et al., 2004	√			√	Yes	√		√		√	√	1
Liu et al., 2005	√	√			No			√	√			1
Mishne, 2005	√	√		√	No	√				√		1
Whitelaw et al., 2005	√	√			No	√			√			1
Wilson et al., 2005	√	√			No	√					√	1
Ng et al., 2006	√	√			Yes	√			√			1
Riloff et al., 2006	√				Yes	√			√		√	1

also review related sentence-level and subjectivity classification studies due to the relevance of the features and techniques utilized and the application domains.

2.3 Sentiment Analysis Features

There are four feature categories that have been used in previous sentiment analysis studies. These include syntactic, semantic, link-based, and stylistic features. Along with semantic features, syntactic attributes are the most commonly used set of features for sentiment analysis. These include word *n*-grams [Pang et al. 2002; Gamon 2004], part-of-speech (POS) tags [Pang et al. 2002; Yi et al. 2003; Gamon 2004], and punctuation. Additional syntactic features include phrase patterns, which make use of POS tag *n*-gram patterns [Nasukawa and Yi 2003; Yi et al. 2003; Fei et al. 2004]. The cited authors noted that phrase patterns such as "n+aj" (noun followed by positive adjective) typically represent positive sentiment orientation, while "n+dj" (noun followed by negative adjective) often express negative sentiment [Fei et al. 2004]. Wiebe et al. [2004]

used collocations where certain parts of fixed-word *n*-grams were replaced with general word tags, thereby also creating *n*-gram phrase patterns. For example, the pattern "U-adj as-prep" would be used to signify all bigrams containing a unique (once-occurring) adjective followed by the preposition "as". Whitelaw et al. [2005] used a set of modifier features (e.g., very, mostly, not); the presence of these features transformed appraisal attributes for lexicon items.

Semantic features incorporate manual/semiautomatic or fully automatic annotation techniques to add polarity- or affect intensity-related scores to words and phrases. Hatzivassiloglou and McKeown [1997] proposed a semantic orientation (SO) method, later extended by Turney [2002], that uses a mutual information calculation to automatically compute the SO score for each word/phrase. The score is computed by taking the mutual information between a phrase and the word "excellent" and subtracting the mutual information between the same phrase and the word "poor". In addition to pointwise mutual information, the SO approach was later also evaluated using latent semantic analysis [Turney and Littman 2003].

Manual- or semiautomatically generated sentiment lexicons (e.g., Tong [2001], Fei et al. [2004], Wilson et al. [2005]) typically use an initial set of automatically generated terms which are manually filtered and coded with polarity and intensity information. The user-defined tags are incorporated to indicate whether certain phrases convey positive or negative sentiment. Riloff et al. [2003] used semiautomatic lexicon generation tools to construct sets of strong subjectivity, weak subjectivity, and objective nouns. Their approach outperformed the use of other features, including bag-of-words, for classification of objective versus subjective English documents. Appraisal groups [Whitelaw et al. 2005] is another effective method for annotating semantics to words/phrases. Initial term lists are generated using WordNet, which are then filtered manually to construct the lexicon. Developed based on appraisal theory [Martin and White 2005], each expression is manually classified into various appraisal classes. These classes include attitude, orientation, graduation, and polarity of phrases. Whitelaw et al. [2005] were able to get very good accuracy using appraisal groups on a movie review corpus, outperforming several previous studies (e.g., Mullen and Collier [2004]), the automated mutual-information-based approach [Turney 2002], as well as the use of syntactic features [Pang et al. 2002]. Manually crafted lexicons have also been used for affect analysis. Subasic and Huettner [2001] used affect lexicons along with fuzzy semantic typing for affect analysis of news articles and movie reviews. Abbasi and Chen [2007a, 2007b] used manually constructed affect lexicons for analysis of hate and violence in extremist Web forums.

Other semantic attributes include contextual features representing the semantic orientation of surrounding text, which have been useful for sentence-level sentiment classification. Riloff et al. [2003] utilized semantic features that considered the subjectivity and objectivity of text surrounding a sentence. Their attributes measured the level of subjective and objective clues in the sentences prior to and following the sentence of interest. Pang and Lee [2004] also leveraged coherence in discourse by considering the level of subjectivity of sentences in close proximity to the sentence of interest.

Link-based features use link/citation analysis to determine sentiments for Web artifacts and documents. Efron [2004] found that opinion Web pages heavily linking to each other often share similar sentiments. Agarwal et al. [2003] observed the exact opposite for USENET newsgroups discussing issues such as abortion and gun control. They noticed that forum replies tended to be antagonistic. Due to the limited usage of link-based features, it is unclear how effective they may be for sentiment classification. Furthermore, unlike Web pages and USENET, other forums may not have a clear message-link structure and some forums are serial (no threads).

Stylistic attributes include lexical and structural attributes incorporated in numerous prior stylometric/authorship studies (e.g., De Vel et al. [2001], Zheng et al. [2006]). However, lexical and structural style markers have seen limited usage in sentiment analysis research. Wiebe et al. [2004] used hapax legomena (unique/once-occurring words) effectively for subjectivity and opinion discrimination. They observed a noticeably higher presence of unique words in subjective texts as compared to objective documents across a *Wall Street Journal* corpus and noted that "apparently, people are creative when they are being opinionated" [Wiebe et al. 2004, p. 286]. Gamon [2004] used lexical features such as sentence length for sentiment classification of feedback surveys. Mishne [2005] used lexical style markers, such as words per message and words per sentence, for affect analysis of Web blogs. While it is unclear whether stylistic features are effective sentiment discriminators for movie/product reviews, style markers have been shown highly prevalent in Web discourse [Abbasi and Chen 2005; Zheng et al. 2006; Schler et al. 2006].

2.4 Sentiment Classification Techniques

Previously used techniques for sentiment classification can be classified into three categories. These include machine learning algorithms, link analysis methods, and score-based approaches.

Many studies have used machine learning algorithms, with support vector machines (SVM) and naïve Bayes (NB) being the most commonly used. SVM have been used extensively for movie reviews [Pang et al. 2002; Pang and Lee 2004; Whitelaw et al. 2005], while naïve Bayes has been applied to reviews and Web discourse [Pang et al. 2002; Pang and Lee 2004; Efron 2004]. In comparisons, SVM have outperformed other classifiers such as NB [Pang et al. 2002]. While SVM have become a dominant technique for text classification, other algorithms such as Winnow [Nigam and Hurst 2004] and AdaBoost [Wilson et al. 2005] have also been used in previous sentiment classification studies.

Studies using link-based features and metrics for sentiment classification have often used link analysis. Efron [2004] used cocitation analysis for sentiment classification of Web-site opinions, while Agarwal et al. [2003] used message-reply link structures to classify sentiments in USENET newsgroups. An obvious limitation of link analysis methods is that they are not effective where link structure is not clear or where links are sparse [Efron 2004].

Score-based methods are typically used in conjunction with semantic features. These techniques generally classify message sentiments based on the

total sum of comprised positive or negative sentiment features. Phrase pattern matching [Nasukawa and Yi 2003; Yi et al. 2003; Fei et al. 2004] requires checking text for manually created, polarized phrase tags (positive and negative). Positive phrases are assigned a +1 while negative phrases are assigned a −1. All messages with a positive sum are assigned positive sentiment while negative messages are assigned to the negative-sentiment class. The semantic orientation approach [Hatzivassiloglou and McKeown 1997; Turney 2002] uses a similar method to score the automatically generated, polarized phrase tags. Score-based methods have also been used for affect analysis, where the affect features present within a message/document are scored based on their degree of intensity for a particular emotion class [Subasic and Huettner 2001].

2.5 Sentiment Analysis Domains

Previously used techniques for sentiment classification can be grouped into three categories. These include machine learning. Sentiment analysis has been applied to numerous domains, including reviews, Web discourse, and news articles and documents. Reviews include movie, product, and music reviews [Morinaga et al. 2002; Pang et al. 2002; Turney 2002]. Sentiment analysis of movie reviews is considered very challenging, since movie reviewers often present lengthy plot summaries and also use complex literary devices such as rhetoric and sarcasm. Product reviews are also fairly complex, since a single review can feature positive and negative sentiments about particular facets of the product.

Web discourse sentiment analysis includes evaluation of Web forums, newsgroups, and blogs. These studies assess sentiments about specific issues/topics. Sentiment topics include abortion, gun control, and politics [Agarwal et al. 2003; Efron 2004]. Robinson [2005] evaluated sentiments about the World Trade Center on 9/11 in three forums in the United States, Brazil, and France. Wiebe et al. [2004] performed subjectivity classification of USENET newsgroup postings.

Sentiment analysis has also been applied to news articles [Yi et al. 2003; Wilson et al. 2005]. Henley et al. [2004] analyzed newspaper articles for biases pertaining to violence-related reports. They found a significant difference between the manner in which the *Washington Post* and the *San Francisco Chronicle* reported news stories relating to anti-gay attacks, with the reporting style reflecting newspaper sentiments. Wiebe et al. [2004] classified objective and subjective news articles in a *Wall Street Journal* corpus.

Some general conclusions can be drawn from Table II and the literature review. Most studies have used syntactic and semantic features. There has also been little use of feature reduction/selection techniques which may improve classification accuracy. In addition, most previous studies have focused on English data, predominantly in the review domain.

3. RESEARCH GAPS AND QUESTIONS

Based on our review of previous literature and conclusions, we have identified several important research gaps. Firstly, there has been limited previous

sentiment analysis work on Web forums, and most studies have focused on sentiment classification of a single language. Secondly, there has been almost no usage of stylistic feature categories. Finally, little emphasis has been placed on feature reduction/selection techniques.

3.1 Web Forums in Multiple Languages

Most previous sentiment classification of Web discourse has focused on USENET and financial forums. Applying such methods to extremist forums is important in order to develop a viable set of features for assessing the presence of propaganda, anger, and hate in these online communities. Furthermore, there has been little evaluation of non-English content, with the exception of Kanayama et al. [2004] performing sentiment classification on Japanese text. Even in that study, machine translation software was used to convert the text to English. Thus, multiple-language features have not been used for sentiment classification. The globalized nature of the Internet necessitates more sentiment analysis across languages.

3.2 Stylistic Features

Previous work has focused on syntactic and semantic features. There has been little use of stylistic features such as word-length distributions, vocabulary richness measures, character- and word-level lexical features, and special-character frequencies. Gamon [2004] and Pang et al. [2002] pointed out that many important features may not seem intuitively obvious at first. Thus, while prior emphasis has been on adjectives, stylistic features may uncover latent patterns that can improve classification performance of sentiments. This may be especially true for Web forum discourse, which is rich in stylistic variation [Abbasi and Chen 2005; Zheng et al. 2006]. Stylistic features have also been shown highly prevalent in other forms of computer-mediated communication, including Web blogs [Herring and Paolillo 2006].

3.3 Feature Reduction for Sentiment Classification

Different automated and manual approaches have been used to craft sentiment classification feature sets. Little emphasis has been given to feature subset selection techniques. Gamon [2004] and Yi et al. [2003] used log likelihood to select important attributes from a large initial feature space. Wiebe et al. [2004] evaluated the effectiveness of various potential subjective elements (PSEs) for subjectivity classification based on their occurrence distribution across classes. However, many powerful techniques have not been explored. Feature reduction/selection techniques have two important benefits [Li et al. 2006]. They can potentially improve classification accuracy and also provide greater insight into important class attributes, resulting in a better understanding of sentiment arguments and characteristics [Guyon and Elisseeff 2003]. Using feature reduction, Gamon [2004] was able to improve accuracy and focus in on a key feature subset of sentiment discriminators.

3.4 Research Questions

We propose the following research questions:

(1) Can sentiment analysis be applied to Web forums in multiple languages?
(2) Can stylistic features provide further sentiment insight and classification power?
(3) How can feature selection improve classification accuracy and identify key sentiment attributes?

4. RESEARCH DESIGN

In order to address these questions, we propose the use of a sentiment classification feature set consisting of syntactic and stylistic features. Furthermore, the utilization of feature selection techniques such as genetic algorithms [Holland 1975] and information gain [Shannon 1948; Quinlan 1986] is also included to improve classification accuracy and get insight into the important features for each sentiment class.

Based on the prevalence of stylistic variation in Web discourse, we believe that lexical and structural style markers can improve the ability to classify Web forum sentiments. Integrated stylistic features include attributes such as word-length distributions, vocabulary richness measures, letter usage frequencies, use of greetings, presence of requoted content, use of URLs, etc.

We also propose the use of an entropy weighted genetic algorithm (EWGA) that incorporates the information-gain (IG) heuristic with a genetic algorithm (GA) to improve feature selection performance. GA is an evolutionary computing search method [Holland 1975] that has been used in numerous feature selection applications [Siedlecki and Sklansky 1989; Yang and Honavar 1998; Li et al. 2007, 2006]. Oliveira et al. [2002] successfully applied GA to feature selection for handwritten digit recognition. Vafiaie and Imam [1994] showed that GA outperformed other heuristics such as greedy search for image recognition feature selection. Like most random-search feature selection methods [Dash and Liu 1997], it uses a wrapper model where performance accuracy is used as the evaluation criterion to improve the feature subset in future generations.

In contrast, IG is a heuristic based on information theory [Shannon 1948]. It uses a filter model for ranking features, which makes it computationally more efficient than GA. IG has outperformed numerous feature selection techniques in head-to-head comparisons [Forman 2003]. Since our experiments will use the SVM classifier, we also plan to compare the proposed EWGA technique against the use of SVM weights for feature selection. In this method, the SVM weights are used to iteratively reduce the feature space, thereby improving performance [Koppel et al. 2002]. SVM weights have been shown effective for text categorization [Koppel et al. 2002; Mladenic et al. 2004] and gene selection for cancer classification [Guyon et al. 2002]. GA, IG, and SVM weights have been used in several previous text classification studies, as shown in Table III. A review of feature selection for text classification can be found in Sebastiani [2002].

A consequence of using an optimal search method such as GA in a wrapper model is that convergence towards an ideal solution can be slow when dealing

Table III. Text Classification Studies using GA, IG, and SVM Weights

Technique	Task	Study
GA	Stylometric Analysis	Li et al. [2006]
IG	Topic Classification	Efron et al. [2003]
	Stylometric Analysis	Juola and Baayen [2003]
		Koppel and Schler [2003]
		Abbasi and Chen [2006]
SVM Weights	Topic Classification	Mladenic et al. [2004]
	Gender Categorization	Koppel et al. [2002]

with very large solution spaces. However, as previous researchers have argued, feature selection is considered an offline task that does not need to be repeated constantly [Jain and Zongker 1997]. This is why wrapper-based techniques using genetic algorithms have been used for gene selection with feature spaces consisting of tens of thousands of genes [Li et al. 2007]. Furthermore, hybrid GAs have previously been used for product design optimization [Alexouda and Paparrizos 2001; Balakrishnan et al. 2004] and scheduling problems [Levine 1996] to facilitate improved accuracy and convergence efficiency [Balakrishnan et al. 2004]. We developed the EWGA hybrid GA that utilizes the information-gain (IG) heuristic with the intention of improving feature selection quality. More algorithmic details are provided in the next section.

5. SYSTEM DESIGN

We propose the following system design (shown in Figure 1). Our design has two major steps: extracting an initial set of features and performing feature selection. These steps are used to carry out sentiment classification of forum messages.

5.1 Feature Extraction

We incorporated syntactic and stylistic features in our sentiment classification attribute set. These features are more generic and applicable across languages. For instance, syntactic, lexical, and structural features have been successfully used in stylometric analysis studies applied to English, Chinese [Peng et al. 2003; Zheng et al. 2006], Greek [Stamamatos et al. 2003], and Arabic [Abbasi and Chen 2006, 2005]. Link-based features were not included, since our messages were not in sequential order (insufficient cross-message references). These types of features are only effective where the testbed consists of entire threads of messages and message referencing information is available. Semantic features were not used since these attributes are heavily context dependent [Pang et al. 2002]. Such features are topic- and language specific. For example, the set of positive-polarity words describing a good movie may not be applicable to discussions about racism. Unlike stylistic and syntactic features, semantic features such as manually crafted lexicons incorporate an inherent feature selection element via human involvement. Such human involvement makes semantic features (e.g., lexicons and dictionaries) very powerful for sentiment analysis. Lexicon developers will only include features that are considered important and will weight these features based on their significance, thereby reducing the need for feature selection. For example, Whitelaw et al.

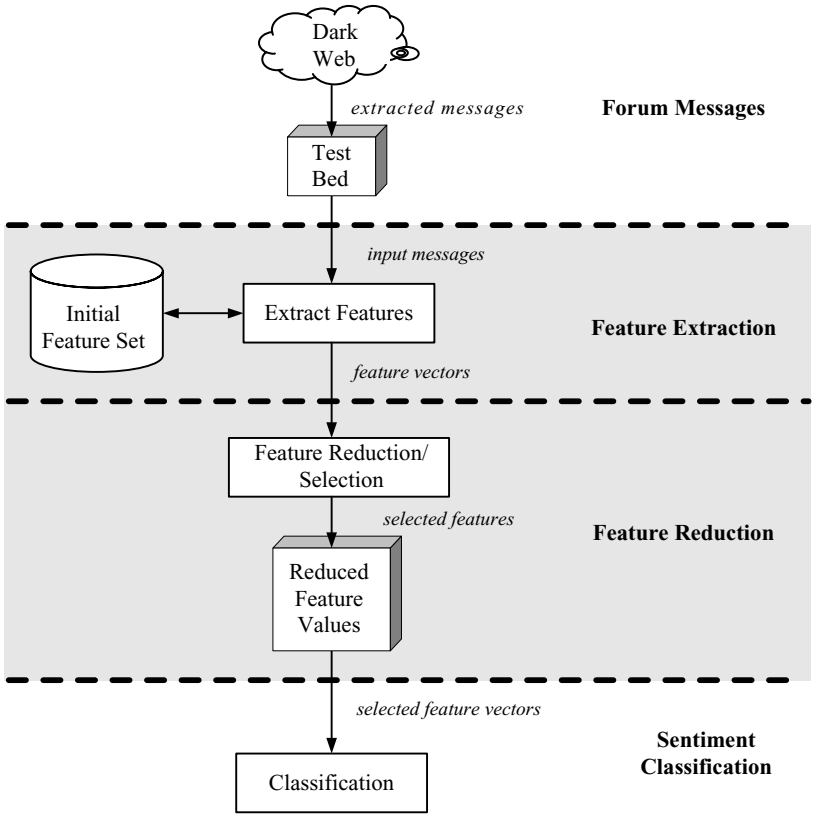


Fig. 1. Sentiment classification system design.

[2005] used WordNet to construct an initial set of features which were manually filtered and weighted to create the lexicon. Unfortunately, the language specificity of semantic features is particularly problematic for application to the Dark Web, which contains text in dozens of languages [Chen 2006]. We hope to overcome the lack of semantic features by incorporating feature selection methods intended to isolate the important subset of stylistic and syntactic features and remove noise.

5.1.1 *Determining Size of Initial Feature Set.* Our initial feature set consisted of 14 different feature categories which included POS tag *n*-grams (for English), word roots (for Arabic), word *n*-grams, and punctuation for syntactic features. Style markers included word- and character-level lexical features, word-length distributions, special characters, letters, character *n*-grams, structural features, vocabulary richness measures, digit *n*-grams, and function words. The word-length distribution includes the frequency of 1- to 20-letter words. Word-level lexical features include total words per document, average word length, average number of words per sentence, average number of words per paragraph, total number of short words (i.e., ones less than 4 letters), etc. Character-level lexical features include total characters per document,

Table IV. English and Arabic Feature Sets

Category	Feature Group	English	Arabic	Examples
Syntactic	POS N-grams	varies	—	frequency of part-of-speech tags (e.g., NP_VB)
	Word Roots	—	varies	frequency of roots (e.g., slm, ktb)
	Word N-grams	varies	varies	word n-grams (e.g. senior editor, editor in chief)
	Punctuation	8	12	occurrence of punctuation marks (e.g., !;:,.?)
Stylistic	Letter N-Grams	26	36	frequency of letters (e.g., a, b, c)
	Char. N-grams	varies	varies	character n-grams (e.g., abo, out, ut, ab)
	Word Lexical	8	8	total words, % char. per word
	Char. Lexical	8	8	total char., % char. per message
	Word Length	20	20	frequency distribution of 1–20-letter words
	Vocab. Richness	8	8	richness (e.g., hapax legomena, Yule's K)
	Special Char.	20	21	occurrence of special char. (e.g., @#$% ^&*+)
	Digit N-Grams	varies	varies	frequency of digits (e.g., 100, 17, 5)
	Structural	14	14	has greeting, has url, requoted content, etc.
	Function Words	250	200	frequency of function words (e.g., of, for, to)

average number of characters per sentence, average number of characters per paragraph, percentage of all characters that are in words, and the percentage of alphabetic, digit, and space characters. Vocabulary richness features include the total number of unique words used, hapax legomena (number of once-occurring words), dis legomena (number of twice-occurring words), and various previously defined statistical measures of richness such as Yule's K, Honore's R, Sichel's S, Simpson's D, and Brunet's W measures. The structural features encompass the total number of lines, sentences, and paragraphs, as well as whether the document has a greeting or signature. Additional structural attributes include whether there is a separation between paragraphs, whether the paragraphs are indented, the presence and position of quoted and forwarded content, and whether the document includes email, URL, and telephone contact information. Further descriptions of the lexical, vocabulary richness, and structural attributes can be found in De Vel et al. [2001], Zheng et al. [2006], and Abbasi and Chen [2005]. The Arabic function words were Arabic words translated from the English function-word list, as in previous research (e.g., Chen and Gey [2002]). Only words were considered; for convenience no affixes were included.

Many feature categories are predefined in terms of the number of potential features. For example, there are only a certain number of possible punctuation- and stylistic lexical features (e.g., words per sentence, words per paragraph, etc.). In contrast, there are countless potential *n*-gram-based features. Consequently, some shallow selection criterion is typically incorporated to reduce the feature space for *n*-grams. A common approach is to select features with a minimum usage frequency [Mitra et al. 1997; Jiang et al. 2004]. We used a minimum frequency threshold of 10 for *n*-gram-based features. Less common features are sparse and likely to cause overfitting. In addition, we only used unigrams, bigrams, and trigrams, as these higher-level *n*-grams tend to be redundant. Using only up to trigrams has been shown effective for stylometric analysis [Kjell et al. 1994] and sentiment classification [Pang et al. 2002; Wiebe et al. 2004]. Based on this criterion for *n*-gram features, Table IV shows the English and Arabic feature sets.

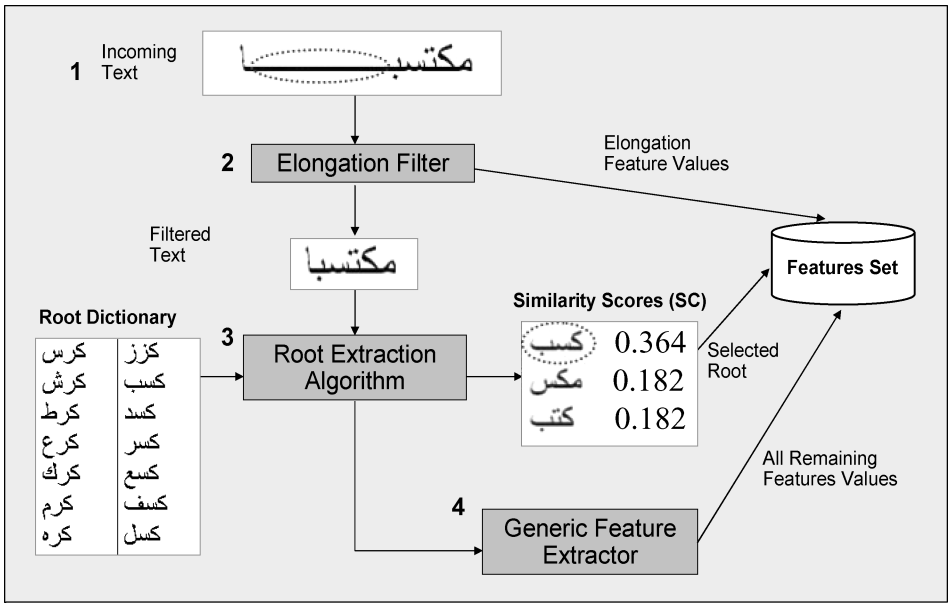


Fig. 2. Arabic extraction component.

5.1.2 *Feature Extraction Component.* Due to the challenging morphological characteristics of Arabic, our attribute extraction process features a component for tracking elongation, as well as a root extraction algorithm (illustrated in Figure 2). Elongation is the process of using a dash-like "kashida" character for stylistic word stretching (shown in step 1 in Figure 2). The use of elongation is very prevalent in Arabic Web forum discourse [Abbasi and Chen 2005]. In addition to tracking the presence and extent of elongation, we filter out these "kashida" characters in order to ensure reliable extraction of the remaining features (step 2 in Figure 2). The filtered words are then passed through a root extraction algorithm [Abbasi and Chen 2005] that compares each word against a root dictionary to determine the appropriate word-root match (step 3). Root frequencies are tracked in order to account for the highly inflective nature of Arabic, which reduces the effectiveness of standard bag-of-words features. The remaining stylistic and syntactic features are then extracted in a similar manner for English and Arabic (step 4).

5.2 Feature Selection: Entropy Weighted Genetic Algorithm (EWGA)

Most previous hybrid GA variations combine GA with other search heuristics such as beam-search, where the beam-search output is used as part of the initial GA population [Alexouda and Paparrizos 2001; Balakrishnan et al. 2004]. Additional hybridizations include modification of the GA's crossover [Aggarwal et al. 1997] and mutation operators [Balakrishnan et al. 2004]. The entropy weighted genetic algorithm (EWGA) uses the information-gain (IG) heuristic to weight the various sentiment attributes. These weights are then incorporated into the GA's initial population as well as crossover and mutation operators. The major steps for the EWGA are as follows.

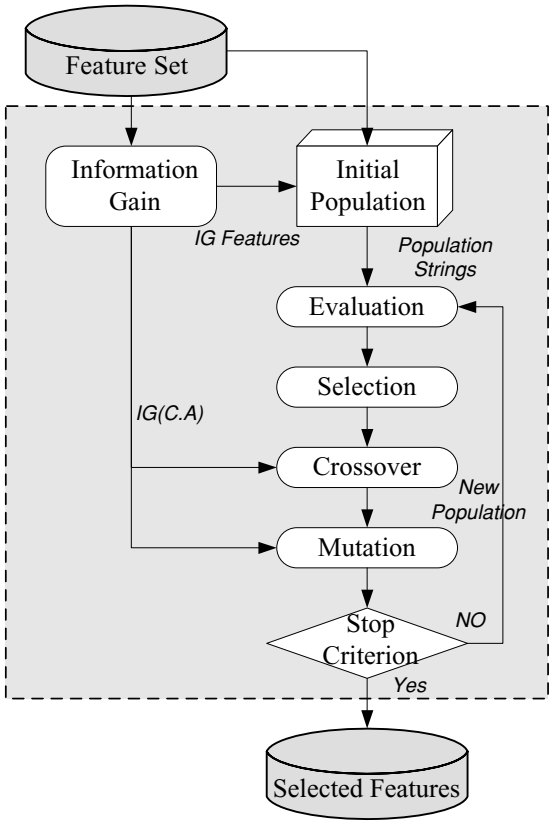


Fig. 3. EWGA illustration.

Algorithm. EWGA

1) Derive feature weights using IG.
2) Include IG selected features as part of initial GA solution population.
3) Evaluate and select solutions based on fitness function.
4) Crossover solution pairs at point that maximizes total IG difference between the two solutions.
5) Mutate solutions based on feature IG weights.
   Repeat steps 3-5 until stopping criterion is satisfied.

Figure 3 shows an illustration of the EWGA process. A detailed description of the IG, initial population, evaluation, selection, crossover, and mutation steps is presented next.

5.2.1 *Information Gain.* For information gain (IG) we used the Shannon entropy measure [Shannon 1948] in which

$$IG(C, A) = H(C) - H(C|A), \tag{1}$$

where

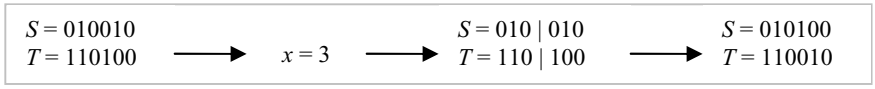
$IG(C, A)$	information gain for feature A;
$H(C) = -\sum_{i=1}^{n} p(C = i) \log_2 p(C = i)$	entropy across sentiment classes C;
$H(C\|A) = -\sum_{i=1}^{n} p(C = i\|A) \log_2 p(C = i\|A)$	specific feature conditional entropy;
n	total number of sentiment classes.

If the number of positive- and negative-sentiment messages is equal, $H(C)$ is 1. Furthermore, the information gain for each attribute A will vary along the range 0 to 1 with higher values indicating greater information gain. All features with an information gain greater than 0.0025 (i.e., $IG(C, A) > 0.0025$) are selected. The use of such a threshold is consistent with prior work using IG for text feature selection [Yang and Pedersen 1997].

5.2.2 *Solution Structure and Initial Population.* We represent each solution in the population using a binary string of length equal to the total number of features, with each binary-string character representing a single feature. Specifically, 1 represents a selected feature while 0 represents a discarded one. For example, a solution string representing five candidate features, "10011", means that the first, fourth, and fifth, features are selected, while the other two are discarded [Li et al. 2006]. In the standard GA, the initial population of n strings is randomly generated. In the EWGA, $n - 1$ solution strings are randomly generated while the IG solution features are used as the final solution string in the initial population.

5.2.3 *Evaluation and Selection.* We use the classification accuracy as the fitness function used to evaluate the quality of each solution. Hence, for each genome in the population, tenfold cross-validation with SVM is used to assess the fitness of that particular solution. Solutions for the next iteration are selected probabilistically, with better solutions having higher probability of selection. While several population replacement strategies exist, we use the generational replacement method originally defined by Holland [1976] in which the entire population is replaced every generation. Other replacement alternatives include steady-state methods where only a fraction of the population is replaced every iteration, while the majority is passed over to the next generation [Levine 1996]. Generational replacement is used in order to maintain solution diversity and to prevent premature convergence attributable to the IG seed solution dominating the other solutions [Bentley 1990; Aggarwal et al. 1997; Balakrishnan et al. 2004].

5.2.4 *Crossover.* From the n solution strings in the population (i.e., $n/2$ pairs), certain adjacent string pairs are randomly selected for crossover based on a crossover probability P_c. In the standard GA, we use single-point crossover by selecting a pair of strings and swapping substrings at a randomly determined crossover point x.



The IG heuristic is utilized in the EWGA crossover procedure in order to improve the quality of the newly generated solutions. Given a pair of solution strings S and T, the EWGA crossover method selects a crossover point x that maximizes the difference in cumulative information gain across strings S and T. Such an approach is intended to create a more diverse solution population: those with heavier concentrations of features with higher IG values and those with fewer IG features. The crossover-point selection procedure can be formulated as

$$\arg\max_{x} \left| \sum_{A=1}^{x} IG(C, A)(S_A - T_A) + \sum_{A=x}^{m} IG(C, A)(T_A - S_A) \right|,$$

where

$IG(C, A)$	information gain for feature A;
S_A	Ath character in solution string S;
T_A	Ath character in solution string T;
m	total number of features;
x	crossover point in solution pair S and T, where $1 < x < m$.

Maximizing the IG differential between solution pairs in the crossover process allows the creation of potentially better solutions. Solutions with higher IG contain attributes that may have greater discriminatory potential, while the lower IG solutions help maintain diversity balance in the solution population. Such balance is important to avoid premature convergence of solution populations towards local maxima [Aggarwal et al. 1997].

5.2.5 *Mutation.* The traditional GA mutation operator randomly mutates individual feature characters in a solution string based on a mutation probability constant P_m. The EWGA mutation operator factors the attribute information gain into the mutation probability as shown in the following. This is done in order to improve the likelihood of inclusion into the solution string for features with higher information gain, while decreasing the probability of features with lower information gain. Our mutation operator sets the probability of a bit to mutate from 0 to 1 based on the feature's information gain, whereas the probability to mutate from 1 to 0 is set to the value 1 minus the feature's information gain. Balakrishnan et al. [2004] demonstrated the potential for modified mutation operators that favor features with higher weights in their hybrid genetic algorithm geared towards product design optimization.

$$P_m(A) = \begin{cases} B[IG(C, A)], & \text{if } S_A = 0 \\ B[1 - IG(C, A)], & \text{if } S_A = 1, \end{cases}$$

where

$P_m(A)$	probability of mutation for feature A;
$IG(C, A)$	information gain for feature A;
S_A	Ath character in solution string S;
B	constant in the range 0 to 1.

5.3 Classification

Because our research focus is on sentiment feature extraction and selection, in all experiments a support vector machines (SVM) is used with tenfold cross-validation and bootstrapping to classify sentiments. We chose an SVM in our experiments because it has outperformed other machine learning algorithms for various text classification tasks [Pang et al. 2002; Abbasi and Chen 2005; Zheng et al. 2006]. We use a linear kernel with the sequential minimal optimization (SMO) algorithm [Platt 1999] included in the Weka data mining package [Witten and Frank 2005].

6. EVALUATION

Experiments were conducted on a benchmark movie review dataset (Experiment 1) and on English and Arabic Web forums (Experiment 2). The purpose of Experiment 1 was to evaluate the effectiveness of the proposed features and selection technique (EWGA) in comparison with previous document-level sentiment classification approaches. Experiment 2 was concerned with evaluating the system on English and Arabic Web forums. The overall accuracy was the average classification accuracy across all ten folds where the classification accuracy was computed as

$$\text{Classification Accuracy} = \frac{\text{Number of Correctly Classified Documents}}{\text{Total Number of Documents}}.$$

In addition to tenfold cross-validation, bootstrapping was used to randomly select 50 samples for statistical testing, as done in previous research (e.g., Whitelaw et al. [2005]). For each sample, we used 5% of the instances for testing and the other 95% for training. Pairwise t-tests were performed on the bootstrap values to assess statistical significance.

6.1 Experiment 1: Benchmark Testbed

In Experiment 1, we conducted two experiments to evaluate the effectiveness of our features as well as feature selection methods for document-level sentiment polarity classification on a benchmark movie review dataset [Pang et al. 2002; Pang and Lee 2004]. This dataset has been used for document-level sentiment categorization in several previous studies (e.g., Pang et al. [2002], Mullen and Collier [2003], Pang and Lee [2004], Whitelaw et al. [2005]). The testbed consists of 2000 movie reviews (1000 positive and 1000 negative) taken from the IMDb movie review archives. The positive reviews are comprised of four- and five-star reviews while the negative ones are those receiving one or two stars. For all experiments, an SVM was run using tenfold cross-validation, with 1800 reviews used for training and 200 for testing in each fold. Bootstrapping was performed

Table V. Experiment 1(a) Results

Movie Reviews				
Features	10-Fold CV	Bootstrap	Standard Dev.	# Features
Stylistic	73.65%	73.26%	2.832	1,017
Syntactic	83.80%	83.74%	1.593	25,853
Stylistic + Syntactic	**87.95%**	**88.04%**	1.133	26,870

Table VI. P-Values for Pairwise T-Tests on Accuracy ($n = 50$)

Features / Test Bed	Movie Reviews
Sty. vs. Syn.	**<0.0001***
Sty. vs. Syn + Sty.	**<0.0001***
Syn. vs. Syn. + Sty.	**<0.0001***

* P-values significant at alpha = 0.05.

by randomly selecting 100 reviews for testing and the remaining 1900 for training, 50 times. In Experiment 1(a) we evaluated the effectiveness of syntactic and stylistic features for sentiment polarity classification. Experiment 1(b) focused on evaluating the effectiveness of EWGA for feature selection.

6.1.1 *Experiment 1(a): Evaluation of Features.* In order to evaluate the effectiveness of syntactic and stylistic features for movie review classification, we used a feature set permutation approach (e.g., stylistic, syntactic, stylistic + syntactic). Stylistic features are unlikely to effectively classify sentiments on their own. Syntactic features have been used in most previous studies and we suspect that these are most important. However, stylistic features may be able to supplement syntactic features; nevertheless, this set of features has not been tested sufficiently. Table V shows the results for the three feature sets. The bootstrap accuracy and standard deviation were computed across the 50 samples.

The best classification accuracy result using an SVM was achieved when using both syntactic and stylistic features. The combined feature set outperformed the use of only syntactic or stylistic features. As expected, the results using only syntactic features were considerably better than those using just style markers. In addition to improved accuracy, the results using stylistic and syntactic features had less variation based on the lower standard deviation. This suggests that using both feature categories in conjunction results in more consistent performance. In contrast, stylistic features had considerably higher standard deviation, indicating that their effectiveness varies across messages.

Table VI shows the pairwise t-tests conducted on the 50 bootstrap samples to evaluate the statistical significance of the improved results using stylistic and syntactic features. As expected, syntactic features outperformed stylistic features when both were used alone. However, using both feature categories significantly outperformed the use of either category individually. The results suggest that stylistic features are prevalent in movie reviews and may be useful for document-level sentiment polarity classification.

6.1.2 *Experiment 2(a): Evaluation of Feature Selection Techniques.* This experiment was concerned with evaluating the effectiveness of feature selection

Table VII. Experiment 1(b) Results

Movie Reviews				
Techniques	10-Fold CV	Bootstrap	Std. Dev.	# Features
Base	87.95%	88.04%	4.133	26,870
IG	89.85%	89.60%	2.631	2,314
GA	90.05%	89.84%	2.783	2,011
SVMW	90.20%	89.96%	2.124	2,000
EWGA	**91.70%**	**91.52%**	2.843	1,748
Ng et al., 2006	90.50%	-	-	25,000
Whitelaw et al. 2005	90.20%	-	-	49,911
Pang and Lee 2004	87.20%	-	-	-
Mullen and Collier 2004*	86.00%	-	-	-
Pang et al., 2002*	82.90%	-	-	-

*Applied to earlier version of data set containing 1,300 reviews.

for sentiment classification. The feature set consisted of all features (syntactic and stylistic), since Experiment 1(a) had already demonstrated the superior performance of using syntactic and stylistic features in unison. We compared the EWGA feature selection approach to no selection/reduction (baseline), feature selection using information gain (IG), the genetic algorithm (GA), and the SVM weights. Feature selection was performed on the 1800 training reviews for each fold, while the remaining 200 were used to evaluate the accuracy for that fold. Thus, the ideal set of features chosen using each selection technique on the 1800 training reviews was used on the testing messages. Thus, IG was applied to the training messages for each fold in order to rank and select those features for that particular fold that would be used on the testing messages. For the GA and EWGA wrappers, this meant that they were run using an SVM with ten-fold cross-validation on the 1800 reviews from each fold. The selected feature subset was then used for evaluating the messages from that particular fold. The overall accuracy was computed as the average accuracy across all ten folds (as standard when using cross-validation). Once again, an SVM was used to classify the message sentiments. The GA and EWGA were each run for 200 iterations, with a population size of 50 for each iteration, using a crossover probability of 0.6 ($= 0.6$) and a mutation probability of 0.01($= 0.01$). These parameter settings are consistent with prior GA research [Alexouda and Paparizzos 2001; Balakrishnan et al. 2004]. The EWGA mutation operator constant was set to 0.1 ($B = 0.1$). For the SVM weight (SVMW) approach, we used the method proposed by Koppel et al. [2002]. We iteratively reduced the number of features for each class from 5000 to 250 in increments of 250 (i.e., decreased overall feature set from 10000 to 500). For each iteration, features were ranked based on the product of their average occurrence frequency per document and the absolute value of their SVM weight. For all experiments, the number of features yielding the best result was reported for the SVMW feature selection method.

Table VII shows the results for the four feature reduction methods and the no feature selection baseline applied to the movie reviews. The bottom half of Table VII also provides the results from prior document-level sentiment classification studies conducted on the same testbed. All four feature selection techniques improved the classification accuracy over the baseline. Consistent with

Table VIII. P-values for Pair Wise T-Tests on Accuracy ($n = 50$)

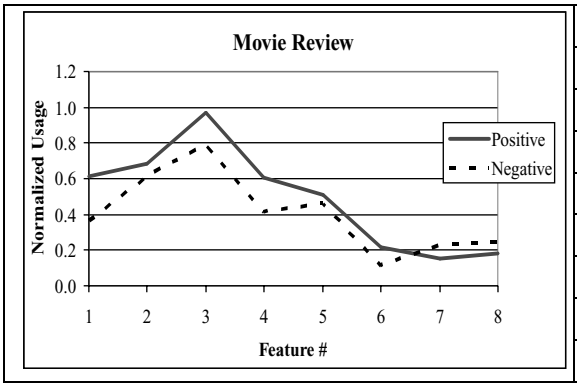
Technique/Test Bed	Movie Reviews
Base vs. IG	<**0.0001***
Base vs. GA	<**0.0001***
Base vs. EWGA	<**0.0001***
Base vs. SVMW	<**0.0001***
IG vs. GA	0.0834
IG vs. EWGA	<**0.0001***
IG vs. SVMW	0.1562
GA vs. EWGA	<**0.0001***
GA vs. SVMW	0.2063
SVMW vs. EWGA	<**0.0001***

*P-values significant at alpha = 0.05.

previous research (e.g., Koppel et al. [2002], Mladenic et al. [2004]), the SVM weights approach also performed well, outperforming IG and GA. The EWGA had the best performance in terms of overall accuracy, resulting in a 4% improvement in accuracy over the no feature selection baseline and a 1.5% to 2% improvement over the other feature selection methods. Furthermore, the EWGA was also the most efficient in terms of the number of features used, improving accuracy while utilizing a smaller subset of the initial feature set. EWGA-based feature selection was able to identify a more concise set of key features as compared to other selection methods.

In comparison with prior work, the results indicate that we were able to achieve higher accuracy than many previous studies on the movie review dataset. Most previous work has had accuracy in the 80–90% range [Pang et al. 2002; Whitelaw et al. 2005] while our performance was over 91% when using stylistic and syntactic features in conjunction with EWGA for feature selection. This is attributable to the prevalence of varying style markers across sentiment classes, as well as the use of feature selection to remove noise and isolate the most effective sentiment discriminators. As noted by Whitelaw et al. [2005], the studies by Pang et al. [2002] and Mullen and Collier [2004] used an earlier, smaller version of the testbed, and are therefore not directly comparable to ours. Table VIII shows the pairwise t-tests conducted to evaluate the statistical significance of the improved results using feature selection ($n = 50$, $df = 49$). EWGA significantly outperformed all other techniques, including the no feature selection baseline, IG, GA, and SVMW.

6.1.3 *Results Discussion.* Figure 4 shows some of the important stylistic features for the movie review dataset. The diagram to the left shows the normalized average feature usage across all positive and negative reviews. The table to the right shows the description for each feature as well as its IG and SVM weight. The positive movie reviews in our dataset tend to be longer in terms of total number of characters and words (features 1 and 2 in the aforesaid table). These reviews also have higher vocabulary richness, based on the various richness formulas that measure the uniqueness of words in a document, such as Simpson's D, Brunet's W, Honore's R, and Yule's K (features 3 through 6). The negative reviews have greater occurrence of the function words "no" and "if".



No.	Feature	IG	SVM
1	total char.	0.014	-0.163
2	total words	0.012	-0.164
3	Simpson	0.028	-0.430
4	Brunet	0.012	-0.121
5	Honore	0.012	-0.161
6	Yule	0.016	-0.116
7	no	0.025	0.252
8	if	0.018	0.249

Fig. 4. Key stylistic features for movie review dataset.

6.2 Experiment 2: English and Arabic Web Forums

We conducted two experiments to evaluate the effectiveness of our features as well as feature selection methods for sentiment classification of messages from English and Arabic extremist Web forums. Once again, an SVM was run using tenfold cross-validation, with 900 messages used for training and 100 for testing in each fold. Bootstrapping was performed by randomly selecting 50 messages for testing and the remaining 950 for training, 50 times. In Experiment 2(a) we evaluated the effectiveness of syntactic and stylistic features. Experiment 2(b) focused on evaluating the effectiveness of feature selection for sentiment analysis across English and Arabic forums.

6.2.1 *Testbed.* Our testbed consists of messages from two major extremist forums (one U.S. and one Middle Eastern) collected as part of the Dark Web project [Chen 2006]. This project involves spidering the Web and collecting Web sites and forums relating to hate and extremist groups. The initial list of group URLs is collected from authoritative sources such as government agencies and the United Nations. These URLs are then used to gather additional relevant forums and Web sites.

The U.S. forum www.nazi.org is an English-language forum that belongs to the Libertarian National Socialist Green Party (LNSG). This is an Aryan supremacist group that gained notoriety when a forum member was involved in a school shooting in 2004. The Middle Eastern forum www.la7odood.com is a major Arabic-speaking partisan forum discussing the war in Iraq and support for the insurgency. The forum's content includes numerous Al-Qaeda speeches and beheading videos.

We randomly selected 1000 polar messages from each forum, which were manually tagged. The polarized messages represented those in favor of (agonists) and against (antagonists) a particular topic. The number of messages used is consistent with previous classification studies [Pang et al. 2002]. In accordance with previous sentiment classification experiments, a maximum of 30 messages were used from any single author. This was done in order to ensure that sentiments were being classified, as opposed to authors. For the U.S. forum,

Table IX. Characteristics of English and Arabic Testbeds

Forum	Messages	Authors	Average Length (char.)	Data Range
U.S.	1000	114	854	3/2004–9/2005
Middle Eastern	1000	126	1126	11/2005–3/2006

we selected messages relating to racial issues. Agonistic-sentiment messages were considered to be those in favor of racial diversity. In contrast, antagonistic-sentiment messages had content denouncing racial diversity, integration, interracial marriage, and race mixing. For the Middle Eastern forum we selected messages relating to the insurgency in Iraq. Agonistic-messages were considered to be those opposed to the insurgency. These messages had positive sentiments about the Iraqi government and U.S. troops in Iraq. Antagonistic-sentiment messages were those in favor of the insurgents and against the current Iraqi government and U.S. forces. These messages had negative sentiments about the Iraqi government and U.S. troops. The occurrence of messages with opposing sentiments is attributable to the presence of agitators (also referred to as trolls) and debaters in these forums [Donath et al. 1999; Herring et al. 2002; Viegas and Smith 2004]. Thus, while the majority of the forum membership may have negative sentiments about a topic, a subset has opposing sentiment polarity. For the sake of simplicity, from here on we will refer to agonistic messages as "positive" and antagonistic messages as "negative" as these terms are more commonly used to represent the two sides in most previous sentiment analysis research. Here, we use the terms positive and negative as indicators of semantic orientation with respect to the specific topic; however the "positive" messages may also contain sentiments about other topics (which may be positive or negative), as described by Wiebe et al. [2005]. This is similar to the document-level annotations used for product and movie reviews [Pang et al. 2002; Yi et al. 2003]. Using two human annotators, 500 positive (agonistic)- and 500 negative (antagonistic)- sentiment messages were incorporated from each forum. Both annotators/coders were bilingual, fluent in English and Arabic. The message annotation task by the independent coders had a Kappa (k) value of 0.90 for English and 0.88 for Arabic, which is considered reliable, suggesting sufficient intercoder reliability. Table IX shows some summary statistics for our English and Arabic Web forum testbeds.

6.2.2 *Experiment 2(a): Evaluation of Features.* In our first experiment, we repeated the feature set tests previously performed on the movie review dataset in Experiment 1(a). Once again, the three permutations of stylistic and syntactic features were used. Table X shows the results for the three feature sets across the U.S. and Middle Eastern forum message datasets.

The best classification accuracy results using SVM were achieved when using both syntactic and stylistic features. The combined feature set statistically outperformed the use of only syntactic or stylistic features across both datasets. The increase was more prevalent in the Middle Eastern forum messages, where the use of stylistic and syntactic features resulted in a 5% improvement in accuracy over the use of syntactic features alone. Surprisingly, stylistic features alone were able to attain over 80% accuracy for the Middle Eastern messages,

Table X. Characteristics of English and Arabic Test Bed

U.S. Forum				
Features	10-Fold CV	Bootstrap	Standard Dev.	# Features
Stylistic	71.40%	71.08%	3.324	867
Syntactic	87.00%	87.13%	2.439	12,014
Stylistic + Syntactic	**90.60%**	**90.56%**	2.042	12,881
Middle Eastern Forum				
Features	Accuracy	Bootstrap	Standard Dev.	# Features
Stylistic	80.20%	80.01%	4.145	1,166
Syntactic	85.40%	85.23%	2.457	12,645
Stylistic + Syntactic	**90.80%**	**90.52%**	2.093	13,811

Table XI. P-Values for Pairwise T-tests on Accuracy ($n = 50$)

Features / Test Bed	U.S.	Middle Eastern
Sty. vs. Syn.	**<0.0001***	**<0.0001***
Sty. vs. Syn + Sty.	**<0.0001***	**<0.0001***
Syn. vs. Syn. + Sty.	**<0.0001***	**<0.0001***

*P-values significant at alpha = 0.05.

nearly a 9% improvement in the effectiveness of these features as compared to the English forum messages. This finding is consistent with previous stylometric analysis studies that have also found significant stylistic usage in Middle Eastern forums, including heavy usage of fonts, colors, elongation, numbers, and punctuation [Abbasi and Chen 2005].

Table XI shows the pairwise t-tests conducted on the bootstrap samples to evaluate the statistical significance of the improved results using stylistic and syntactic features. As expected, syntactic features outperformed stylistic features when both were used alone. However, using both feature categories significantly outperformed the use of either category individually. The results suggest that stylistic features are prevalent and important in Web discourse, even when applied to sentiment classification.

6.2.3 *Experiment 2(b): Evaluation of Feature Selection Techniques.* This experiment was concerned with evaluating the effectiveness of feature selection for sentiment classification of Web forums. The same experimental settings as in Experiment 1(b) were used for all techniques. Table XII shows the results for the four feature reduction methods and the no feature selection baseline applied across the U.S. and Middle Eastern forum messages. All four feature selection techniques improved the classification accuracy over the baseline. The EWGA had the best performance across both testbeds in terms of overall accuracy, resulting in a 2 to 3% improvement in accuracy over the no feature selection baseline. Furthermore, the EWGA was also the most efficient in terms of number of features used, improving accuracy while utilizing a smaller subset of the initial feature sets. EWGA-based feature selection was able to identify a more concise set of key features that was 50% to 70% smaller than those IG and SVMW and 75% to 90% smaller than that of the baseline. GA also used a smaller number of features; however, the use of EWGA resulted in considerably improved accuracy.

Table XII. Experiment 2(b) Results

U.S. Forum				
Technique	10-Fold CV	Bootstrap	Standard Dev.	# Features
Base	90.60%	90.56%	2.042	12,881
IG	91.10%	91.16%	1.564	1,055
GA	90.90%	90.64%	1.453	505
SVMW	91.10%	91.20%	1.656	1,000
EWGA	**92.80%**	**92.84%**	1.458	508
Middle Eastern Forum				
Technique	10-Fold CV	Bootstrap	Standard Dev.	# Features
Base	90.80%	90.52%	2.093	13,811
IG	93.40%	93.36%	1.665	1,045
GA	92.10%	92.24%	1.438	462
SVMW	93.30%	93.28%	1.337	1,000
EWGA	**93.60%**	**93.84%**	2.831	338

Table XIII. P-Values for Pair Wise T-tests on Accuracy ($n = 50$)

Technique / Test Bed	U.S.	Middle Eastern
Base vs. IG	<**0.0384***	<**0.0001***
Base vs. GA	0.1245	**0.0134***
Base vs. EWGA	<**0.0001***	<**0.0001***
Base vs. SVMW	<**0.0369***	<**0.0001***
IG vs. GA	**0.0485***	0.0685
IG vs. EWGA	<**0.0001***	0.2783
IG vs. SVMW	0.2934	0.4130
GA vs. EWGA	<**0.0001***	**0.0456***
GA vs. SVMW	**0.0461***	0.0728
SVMW vs. EWGA	<**0.0001***	0.2025

*P-values significant at alpha = 0.05.

Table XIII shows the pairwise t-tests conducted on the bootstrap values to evaluate the statistical significance of the improved results using feature selection. EWGA significantly outperformed the baseline and GA for both datasets. In addition, EWGA provided significantly better performance than IG and SVMW on the English Web forum messages. EWGA also outperformed IG and SVMW on the Middle Eastern forum dataset, though the improved performance was not statistically significant.

6.3 Results Discussion

Figure 5 shows the selection accuracy and number of features selected (out of over 12800 potential features) for the U.S. forum using EWGA as compared to GA across the 200 iterations (average of ten folds). The Middle Eastern forum graphs looked similar to the U.S. forum and hence were not included. The EWGA accuracy declines initially, despite being seeded with the IG solution. This is due to the use of generation replacement, which prevents the IG solution from dominating the other solutions and creating a stagnant solution population. As intended, the IG solution features are gradually disseminated to the remaining solutions in the population until the new solutions begin to improve in accuracy at around the 20th iteration. Overall, the EWGA is able to converge on an

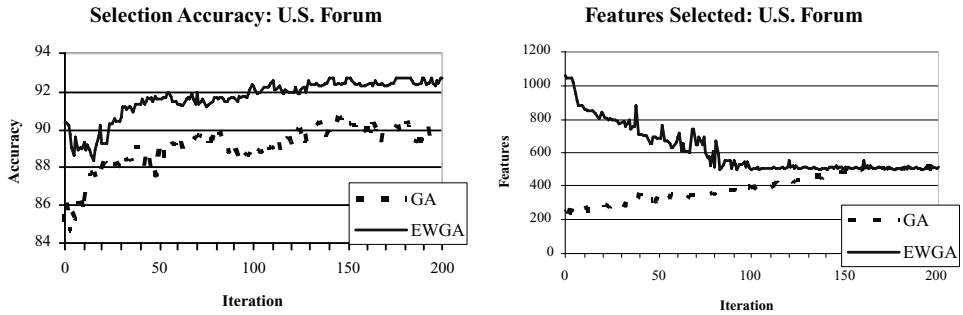


Fig. 5. U.S. forum results using EWGA and GA.

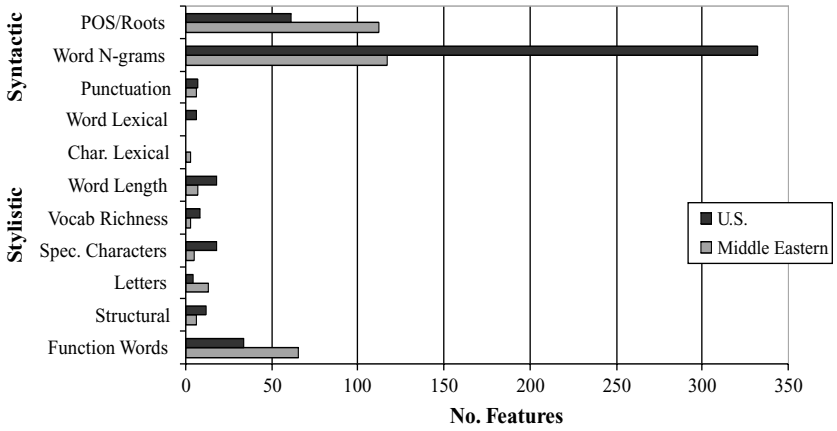


Fig. 6. Key feature usage frequencies by category.

improved solution while only using half of the features originally transferred from IG. It is interesting to note that EWGA and GA both converge to a similar number of features when applied to the U.S. forum; however, EWGA is better able to isolate the more effective sentiment discriminators.

6.3.1 *Analysis of Key Sentiment Features.* We chose to analyze the EWGA features, since they provided the highest performance with the most concise set of features. Thus, the EWGA-selected features are likely the most significant discriminators with the least redundancy. Figure 6 shows the number of each feature category selected by the EWGA for the English and Arabic feature sets. As expected, more syntactic features (POS tags, *n*-grams, word roots) were used, since considerably more of these features were included.

While Figure 6 shows the number of features selected by the EWGA for each feature category, Figure 7 shows the percentage of the overall number of features in each category that were selected. For example, the EWGA selected 12 structural features from the U.S. (English-language) feature set; however, this represents 86% percent of the structural features as shown in Figure 7. Looking at the percentage of usage, stylistic features were more efficient than word *n*-grams and POS tags/roots, also as shown in Figure 7. Many of the

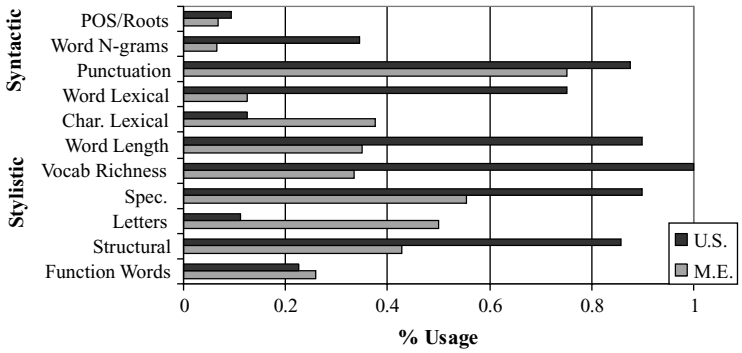
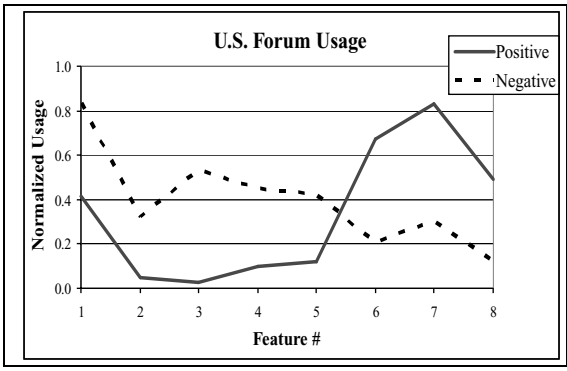


Fig. 7. Key feature usage percentage by category.



No.	Feature	IG	SVM
1	total char.	0.027	0.243
2	$	0.029	0.130
3	&	0.017	0.141
4	{	0.022	0.126
5	digit count	0.015	0.316
6	therefore	0.021	-0.104
7	however	0.017	-0.120
8	nevertheless	0.014	-0.119

Fig. 8. Key stylistic features for U.S. forum.

stylistic feature groups had over 40% usage, whereas syntactic features rarely had such high usage, with the exception of punctuation. For the U.S. feature set, some categories such as word length, vocabulary richness, special characters, and structural features had well over 80% representation in the final feature subset. Comparing across regions, U.S. features had higher usage rates than the Middle Eastern feature set. Approximately 10% of Middle Eastern features were used by the EWGA versus 25% of the U.S. attributes.

6.3.2 *Key Stylistic Features.* Figure 8 shows some of the important stylistic features for the U.S. forum. The diagram on the left side of the figure shows the normalized average feature usage across all positive- and negative-sentiment messages. The table on the right shows the description for each feature as well as its IG and SVM weight.

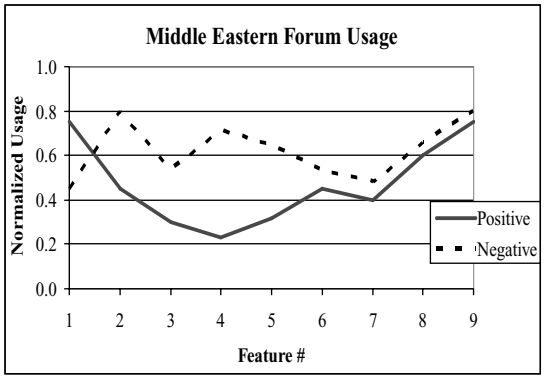
The positive-sentiment messages (agonists, in favor of racial diversity) tend to be considerably shorter (feature 1), containing a few long sentences. These messages also feature heavier usage of conjunctive function words such as “however”, “therefore”, and “nevertheless” (features 6 through 8). In contrast, the negative-sentiment messages are nearly twice as long and contain lots of digits (feature 5) and special characters (features 2 through 4). Higher digit

usage in the negative messages is due to references to news articles that are used to stereotype. Article snippets begin with a date, resulting in the higher digit count. The negative messages also feature shorter sentences. The stylistic feature usage statistics suggest that positive-sentiment messages follow more of a debating style, with shorter, well-structured arguments. In contrast, the negative-sentiment messages tend to contain greater signs of emotion. The following verbal joust between two members in the U.S. forum exemplifies the stylistic differences across sentiment classes. It should be noted that some of the content in the messages has been sanitized for vulgar word usage; however, the stylistic tendencies that are meant to be illustrated remain unchanged.

Negative: You're a total %#$*@ idiot!!! You walk around thinking you're doing humanity a favor. Sympathizing with such barbaric slime. They use your sympathy as an excuse to fail. They are a burden to us all!!! Your opinion means nothing.
Positive: Neither does yours. But at least my opinion is an educated and informed one backed by well-reasoned arguments and careful skepticism about my assumptions. Race is nothing more than a social classification. What have you done for society that allows you to deem others a burden?

Figure 9 shows some of the important stylistic features for the Middle Eastern forum. There are a few interesting similarities between the U.S. and Middle Eastern forum feature usage tendencies across sentiment lines. The positive-sentiment messages in the Middle Eastern forum (agonists, opposed to the insurgency) also tend to be considerably shorter than the negative-sentiment messages in terms of total number of characters (feature 2). Additionally, like their U.S. forum counterparts, negative Arabic messages contain heavy digit usage attributable to news article snippets (feature 5). The negative-sentiment messages make greater use of stylistic word stretching (elongation) which is done in order to emphasize key words (feature 3). Consequently, the negative messages include greater use of words longer than 10 characters (feature 4) while the positive messages are more likely to use shorter words of less than 4 characters in length (feature 1). The negative-sentiment messages also have higher vocabulary richness (features 6 through 9), various vocabulary richness formulas).

6.3.3 *Key Syntactic Features.* Table XIV shows the keyword *n*-grams for each sentiment class selected by the EWGA. Many of the terms and phrases comprised racist content that was not included in the table, but rather represented using a description label. Items in quotes indicate actual terms (e.g., "criminals") while nonquoted items signify term descriptions (e.g., racist terms). For the Middle Eastern forum, sentiments seem to be drawn along sectarian lines. In contrast, U.S. forum sentiments are not clearly separated along racial lines. While the majority of the negative sentiments towards racial issues are generated by white supremacists, many of the positive sentiments are also presented by those with the same self-proclaimed racial affiliations. This



No.	Feature	IG	SVM
1	short words	0.026	-0.020
2	total char.	0.039	0.210
3	elongation	0.037	0.319
4	long words	0.035	0.362
5	digits	0.014	0.086
6	Simpson	0.029	0.135
7	Yule	0.023	0.137
8	Brunet	0.027	0.109
9	Honore	0.024	0.140

Fig. 9. Key stylistic features for middle eastern forum.

Table XIV. Key N-Grams for Various Sentiment Classes

U.S. Forum		Middle Eastern Forum	
Positive (agonist)	Negative (antagonist)	Positive (agonist)	Negative (antagonist)
Racist terms "racism" "subhuman racist" "anti-Semitism" "ignorant slime"	Racist terms "criminals" "whites" "Americans" "get a job" "lmwao" "Odin's rage" "urban jungle"	Racist Shia terms "terrorists" – "ارهابيين" "Shia" – "شيعة" "Shiite" – "شيعي"	Racist Sunni terms "freedom fighters" – "مجاهدين" "martyrdom" – "استشهاد" "Zarqawi" – "الزرقاوي" "Sunni"– "سني" "American"– "امريكي" "Iraq"– "العراق" " international forces"– "دولية قوات"

reduced the amount of racial name-calling across sentiments in the U.S. forums, resulting in the need for considerably larger numbers of *n*-grams to effectively discern sentiment classes. Consequently, the number of *n*-grams used for the U.S. feature set (i.e., 332) is nearly threefold those used for the Middle Eastern sentiment classification (i.e., 117).

7. CONCLUSIONS AND FUTURE DIRECTIONS

In this study we applied sentiment classification methodologies to English and Arabic Web forum postings. In addition to syntactic features, a wide array of English and Arabic stylistic attributes, including lexical, structural, and function-word style markers, were included. We also developed the entropy weighted genetic algorithm (EWGA) for efficient feature selection in order to improve accuracy and identify key features for each sentiment class. EWGA significantly outperformed the no feature selection baseline and GA on all testbeds. It also outperformed IG and SVMW on all three datasets (statistically significant for the movie review and U.S. forum datasets) while isolating a smaller subset of key features. EWGA demonstrated the utility of these key features in terms of

classification performance and for content analysis. Analysis of EWGA-selected stylistic and syntactic features allowed greater insight into writing-style and content differences across sentiment classes in the two Web forums. Our approach of using stylistic and syntactic features in conjunction with the EWGA feature selection method achieved a high level of accuracy, suggesting that these features and techniques may be used in the future to perform sentiment classification and content analysis of Web forums discourse. Applying sentiment analysis to Web forums is an important endeavor and the current accuracy is promising for effective analysis of forum conversation sentiments. Such analysis can help provide a better understanding of extremist-group usage of the Web for information and propaganda dissemination.

In the future we would like to evaluate the effectiveness of the proposed sentiment classification features and techniques for other tasks, such as sentence- and phrase-level sentiment classification. We also intend to apply the technique to other sentiment domains (e.g., news articles and product reviews). Moreover, we believe the suggested feature selection technique may also be appropriate for other forms of text categorization, and plan to apply our technique to topic, style, and genre classification. We also plan to investigate the effectiveness of other forms of GA hybridization, such as using SVM weights instead of the IG heuristic.

REFERENCES

ABBASI, A. AND CHEN, H. 2005. Identification and comparison of extremist-group Web forum messages using authorship analysis. *IEEE Intell. Syst. 20*, 5, 67–75.

ABBASI, A. AND CHEN, H. 2006. Visualizing authorship for identification. In *Proceedings of the 4th IEEE International Conference on Intelligence and Security Informatics*, San Diego, CA, 60–71.

ABBASI, A. AND CHEN, H. 2007a. Affect intensity analysis of Dark Web forums. In *Proceedings of the 5th IEEE International Conference on Intelligence and Security Informatics*, New Brunswick, NJ, 282–288.

ABBASI, A. AND CHEN, H. 2007b. Analysis of affect intensities in extremist group forums. In *Intelligence and Security Informatics*. E. Reid and H. Chen, Eds. Springer (forthcoming).

ALEXOUDA, G. AND PAPPARRIZOS, K. 2001. A genetic algorithm approach to the product line design problem using the seller's return criterion: An extensive comparative computational study. *Eur. J. Oper. Res. 134*, 165–178.

AGGARWAL, C. C., ORLIN, J., AND TAI, R. P. 1997. Optimized crossover for the independent set problem. *Oper. Res. 45*, 2, 226–234.

AGRAWAL, R., RAJAGOPALAN, S., SRIKANT, R., AND XU, Y. 2003. Mining newsgroups using networks arising from social behavior. In *Proceedings of the 12th International World Wide Web Conference (WWW)*, 529–535.

BALAKRISHNAN, P. V., GUPTA, R., AND JACOB, V. S. 2004. Development of hybrid genetic algorithms for product line designs. *IEEE Trans. Syst. Man Cybernet. 34,* 1, 468–483.

BEINEKE, P., HASTIE, T., AND VAITHYANATHAN, S. 2004. The sentimental factor: Improving review classification via human-provided information. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 263.

BURRIS, V., SMITH, E., AND STRAHM, A. 2000. White supremacist networks on the Internet. *Sociol. Focus 33,* 2, 215–235.

CHEN, A. AND GEY, F. 2002. Building an Arabic stemmer for information retrieval. In *Proceedings of the 11th Text Retrieval Conference (TREC)*, Gaithersburg, MD, 631–639.

CHEN, H. 2006. *Intelligence and Security Informatics for International Security: Information Sharing and Data Mining*. Springer, London.

CRILLEY, K. 2001. Information warfare: New battle fields, terrorists, propaganda, and the Internet. *Aslib Proc. 53,* 7, 250–264.

DASH, M. AND LIU, H. 1997. Feature selection for classification. *Intell. Data Anal. 1*, 131–156.

DAVE, K., LAWRENCE, S., AND PENNOCK, D. M. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on the World Wide Web (WWW)*, 519–528.

DE VEL, O., ANDERSON, A., CORNEY, M., AND MOHAY, G. 2001. Mining e-mail content for author identification forensics. *ACM SIGMOD Rec. 30*, 4, 55–64.

DONATH, J. 1999. Identity and deception in the virtual community. In *Communities in Cyberspace*, Routledge Press, London.

EFRON, M. 2004. Cultural orientations: Classifying subjective documents by cocitation analysis. *In Proceedings of the AAAI Fall Symposium Series on Style and Meaning in Language, Art, Music, and Design*, 41–48.

EFRON, M., MARCHIONINI, G., AND ZHIANG, J. 2004. Implications of the recursive representation problem for automatic concept identification in on-line government information. In *Proceedings of the Annual Meeting of the American Society for Information Science and Technology (ASIST) SIG-CR Workshop*.

FEI, Z., LIU, J., AND WU, G. 2004. Sentiment classification using phrase patterns. In *Proceedings of the 4th IEEE International Conference on Computer Information Technology*, 1147–1152.

FORMAN, G. 2003. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res. 3*, 1289–1305.

GAMON, M. 2004. Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th International Conference on Computational Linguistics*, 841.

GLASER, J., DIXIT, J., AND GREEN, D. P. 2002. Studying hate crime with the Internet: What makes racists advocate racial violence? *J. Social Issues* 58, 1, 177–193.

GREFENSTETTE, G.., QU, Y., SHANAHAN, J. G., AND EVANS, D. A. 2004. Coupling niche browsers and affect analysis for an opinion mining application. In *Proceedings of the 12th International Conference Recherche d'Information Assistee par Ordinateur*, 186–194.

GUYON, I., WESTON, J., BARNHILL, S., AND VAPNIK, V. 2002. Gene selection for cancer classification using support vector machines. *Mach. Learn. 46*, 389–422.

GUYON, I. AND ELISSEEFF, A. 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res. 3*, 1157–1182.

HATZIVASSILOGLOU, V. AND MCKEOWN, K. R. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics*, 174–181.

HEARST, M. A. 1992. Direction-Based text interpretation as an information access refinement. In *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, P. Jacobs, Ed. Lawrence Erlbaum Associates, Mahwah, NJ.

HENLEY, N. M., MILLER, M. D., BEAZLEY, J. A., NGUYEN, D. N., KAMINSKY, D., AND SANDERS, R. 2002. Frequency and specificity of referents to violence in news reports of anti-gay attacks. *Discourse Soc. 13*, 1, 75–104.

HERRING, S., JOB-SLUDER, K., SCHECKLER, R., AND BARAB, S. 2002. Searching for safety online: Managing "trolling" in a feminist forum. *The Inf. Soc. 18*, 5, 371–384.

HERRING, S. AND PAOLILLO, J. C. 2006. Gender and genre variations in Weblogs. *J. Sociolinguist. 10*, 4, 439.

HOLLAND, J. 1975. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI.

HU, M. AND LIU, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168–177.

JAIN, A. AND ZONGKER, D. 1997. Feature selection: Evaluation, application, and small sample performance. *IEEE Trans. Pattern Anal. Mach. Intell. 19*, 2, 153–158.

JIANG, M., JENSEN, E., BEITZEL, S. AND ARGAMON, S. 2004. Choosing the right bigrams for information retrieval. In *Proceedings of the Meeting of the International Federation of Classification Societies*.

JUOLA, P. AND BAAYEN, H. 2005. A controlled-corpus experiment in authorship identification by cross-entropy. *Literar. Linguist. Comput. 20*, 59–67.

KANAYAMA, H., NASUKAWA, T., AND WATANABE, H. 2004. Deeper sentiment analysis using machine translation technology. In *Proceedings of the 20th International Conference on Computational Linguistics*, 494–500.

KAPLAN, J. AND WEINBERG, L. 1998. *The Emergence of a Euro-American Radical Right*., Rutgers University Press, New Brunswick, NJ.

KIM, S. AND HOVY, E. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, 1367–1373.

KJELL, B. WOODS, W. A., AND FRIEDER, O. 1994. Discrimination of authorship using visualization. *Inf. Process. Manage. 30*, 1, 141–150.

KOPPEL, M., ARGAMON, S., AND SHIMONI, A. R. 2002. Automatically categorizing written texts by author gender. *Literar. Linguis. Comput. 17*, 4, 401–412.

KOPPEL, M. AND SCHLER, J. 2003. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of the IJCAI Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico.

LEVINE, D. 1996. Application of a hybrid genetic algorithm to airline crew scheduling. *Comput. Oper. Res. 23,* 6, 547–558.

LEETS, L. 2001. Responses to Internet hate sites: Is speech too free in cyberspace? *Commun. Law Policy* 6, 2, 287–317.

LI, J., ZHENG, R., AND CHEN, H. 2006. From fingerprint to writeprint. *Commun. ACM 49*, 4, 76–82.

LI, J., SU, H., CHEN, H., AND FUTSCHER, B. 2007. Optimal search-based gene subset selection for gene array cancer classification. *IEEE Trans. Inf. Technol. Biomed* (to appear).

LIU, B., HU, M., AND CHENG, J. 2005. Opinion observer: Analyzing and comparing opinions on the Web. In *Proceedings of the 14th International World Wide Web Conference (WWW)*, 342–351.

MARTIN, J. R. AND WHITE, P. R. R. 2005. *The Language of Evaluation: Appraisal in English*. Palgrave, London.

MISHNE, G. 2005. Experiments with mood classification. In *Proceedings of the 1st Workshop on Stylistic Analysis of Text for Information Access*, Salvador, Brazil.

MITRA, M., BUCKLEY, C., SINGHAL, A., AND CARDIE, C. 1997. An analysis of statistical and syntactic phrases. In *Proceedings of the 5th International Conference Recherche d'Information Assistee par Ordinateur,* Montreal, Canada, 200–214.

MLADENIC, D., BRANK, J., GROBELNIK, M., AND MILIC-FRAYLING, N. 2004. Feature selection using linear classifier weights: Interaction with classification models. In *Proceedings of the 27th ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, UK, 234–241.

MORINAGA, S., YAMANISHI, K., TATEISHI, K., AND FUKUSHIMA, T. 2002. Mining product reputations on the Web. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Canada, 341–349.

MULLEN, T. AND COLLIER, N. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP) Conference*, Barcelona, Spain, 412–418.

NASUKAWA, T. AND YI, J. 2003. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge Capture*, Sanibel Island, FL, 70–77.

NIGAM, K. AND HURST, M. 2004. Towards a robust metric of opinion. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text.*

NG, V., DASGUPTA, S., AND ARIFIN, S. M. N. 2006. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of the COLING/ACL Conference.* Sydney, Australia, 611–618.

OLIVEIRA, L. S., SABOURIN, R., BORTOLOZZI, F., AND SUEN, C. Y. 2002. Feature selection using multi-objective genetic algorithms for handwritten digit recognition. In *Proceedings of the 16th International Conference on Pattern Recognition*, 568–571.

PANG, B., LEE, L., AND VAITHYANATHAIN, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 79–86.

PANG, B. AND LEE, L. 2004. A sentimental education: Sentimental analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 271–278.

PICARD, R. W. 1997. *Affective Computing*. MIT Press, Cambridge, MA.

PLATT, J. 1999. Fast training on SVMs using sequential minimal optimization. In *Advances in Kernel Methods: Support Vector Learning*. B. Scholkopf et al. Eds., MIT Press, Cambridge, MA, 185–208.

QUINLAN, J. R. 1986. Induction of decision trees. *Mach. Learn. 1*, 1, 81–106.

RILOFF, E., PATWARDHAN, S., AND WIEBE, J. 2006. Feature subsumption for opinion analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* Sydney, Australia, 440–448.

RILOFF, E., WIEBE, J., AND WILSON, T. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the 7th Conference on Natural Language Learning*, Edmonton, Canada, 25–32.

ROBINSON, L. 2005. Debating the events of September 11th: Discursive and interactional dynamics in three online for a. *J. Comput. Mediat. Commun. 10,* 4.

SCHAFER, J. 2002. Spinning the web of hate: Web-based hate propagation by extremist organizations. *J. Criminal Just. Popular Culture 9,* 2, 69–88.

SCHLER, J., KOPPEL, M., ARGAMON, S., AND PENNEBAKER, J. 2006. Effects of age and gender on blogging. In *Proceedings of the AAAI Spring Symposium Computational Approaches to Analyzing Weblogs*, Menlo Park, CA, 191–197.

SEBASTIANI, F. 2002. Machine learning in automated text categorization. *ACM Comput. Surv. 34,* 1, 1–47.

SHANNON, C. E. 1948. A mathematical theory of communication. *Bell Syst. Tech. J. 27,* 4, 379–423.

SIEDLECKI, W. AND SKLANSKY, J. 1989. A note on genetic algorithms for large-scale feature selection. *Pattern Recogn. Lett. 10,* 5, 335–347.

SUBASIC, P. AND HUETTNER, A. 2001. Affect analysis of text using fuzzy semantic typing. *IEEE Trans. Fuzzy Syst. 9,* 4, 483–496.

TONG, R. 2001. An operational system for detecting and tracking opinions in on-line discussion. In *Proceedings of the ACM SIGIR Workshop on Operational Text Classification*, 1–6.

TURNEY, P. D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meetings of the Association for Computational Linguistics*, Philadelphia, PA, 417–424.

TURNEY, P, D. AND LITTMAN, M. L. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst. 21,* 4, 315–346.

VAFAIE, H. AND IMAM, I. F. 1994. Feature selection methods: Genetic algorithms vs. greedy-like search. In *Proceedings of the International Conference on Fuzzy and Intelligent Control Systems*.

VIEGAS, F. B. AND SMITH, M. 2004. Newsgroup crowds and AuthorLines: Visualizing the activity of individuals in conversational cyberspaces. In *Proceedings of the 37th Hawaii International Conference on System Sciences*, Hawaii, USA.

WHITELAW, C., GARG, N., AND ARGAMON, S. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM Conference on Information and Knowledge Management*, 625–631.

WIEBE, J. 1994. Tracking point of view in narrative. *Comput. Linguist. 20,* 2, 233–287.

WIEBE, J., WILSON, T., AND BELL, M. 2001. Identifying collocations for recognizing opinions. In *Proceedings of the ACL/EACL Workshop on Collocation*, Toulouse, France.

WIEBE, J., WILSON, T., BRUCE, R., BELL, M., AND MARTIN, M. 2004. Learning subjective language. *Comput. Linguist. 30,* 3, 277–308.

WIEBE, J., WILSON, T., AND CARDIE, C. 2005. Annotating expressions of opinions and emotions in language. *Lang. Resources Eval. 1,* 2,165–210.

WITTEN, I. H. AND FRANK, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques,* 2nd ed. Morgan Kaufmann, San Francisco, CA.

WILSON, T., WIEBE, J., AND HOFFMAN, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, British Columbia, Canada, 347–354.

YANG, Y. AND PEDERSON, J. O. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning,* 412–420.

YANG, J. AND HONAVAR, V. 1998. Feature subset selection using a genetic algorithm. *IEEE Intell. Syst. 13,* 2, 44–49.

YI, J., NASUKAWA, T., BUNESCU, R., AND NIBLACK, W. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of the 3rd IEEE International Conference on Data Mining*, 427–434.

YI, J. AND NIBLACK, W. 2005. Sentiment mining in WebFountain. In *Proceedings of the 21st International Conference on Data Engineering*, 1073–1083.

YU, H. AND HATZIVASSILOGLOU, V. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 129–136.

ZHENG, R., LI, J., HUANG, Z., AND CHEN, H. 2006. A framework for authorship analysis of online messages: Writing-Style features and techniques. *J. Amer. Soc. Inf. Sci. Technol. 57*, 3, 378–393.

ZHOU, Y., REID, E., QIN, J., CHEN, H., AND LAI, G. 2005. U.S. extremist groups on the Web: Link and content analysis. *IEEE Intell. Syst. 20,* 5, 44–51.

Received December 2006; revised June 2007; accepted July 2007