

Affect Analysis of Web Forums and Blogs Using Correlation Ensembles

Ahmed Abbasi, *Member, IEEE*, Hsinchun Chen, *Fellow, IEEE*, Sven Thoms, and Tianjun Fu

Abstract—Analysis of affective intensities in computer-mediated communication is important in order to allow a better understanding of online users' emotions and preferences. Despite considerable research on textual affect classification, it is unclear which features and techniques are most effective. In this study, we compared several feature representations for affect analysis, including learned n-grams and various automatically and manually crafted affect lexicons. We also proposed the support vector regression correlation ensemble (SVRCE) method for enhanced classification of affect intensities. SVRCE uses an ensemble of classifiers each trained using a feature subset tailored toward classifying a single affect class. The ensemble is combined with affect correlation information to enable better prediction of emotive intensities. Experiments were conducted on four test beds encompassing web forums, blogs, and online stories. The results revealed that learned n-grams were more effective than lexicon-based affect representations. The findings also indicated that SVRCE outperformed comparison techniques, including Pace regression, semantic orientation, and WordNet models. Ablation testing showed that the improved performance of SVRCE was attributable to its use of feature ensembles as well as affect correlation information. A brief case study was conducted to illustrate the utility of the features and techniques for affect analysis of large archives of online discourse.

Index Terms—Affective computing, discourse, emotion recognition, linguistic processing, machine learning, text mining.

1 INTRODUCTION

THE need for enhanced information retrieval and knowledge discovery from computer-mediated communication archives has been articulated by many in recent years. One suggested information access refinement has been to mine directional text: text containing emotions and opinions [11], [29]. Affects play an important role in influencing people's perceptions and decision making [21]. Analysis of sentiments and affects is particularly important for online discourse, where such information is often more pervasive than topical content [26], [18]. With the increased popularity of social computing, the presence and significance of affective text is likely to grow [14]. There has been considerable recent work on sentiment analysis of online forums and product reviews [27], [30]. However, research on analysis of affects (including emotions and moods) is still relatively sparse [5]. While recent studies have analyzed the presence of affects in blogs, online stories, chat dialog, transcripts, song lyrics, etc., it is unclear which features and techniques are most useful for affective computing of online texts. There is, therefore, a need to compare existing features for representing affective content as well as the techniques used for assigning emotive intensities.

In this study, we evaluate features and techniques for classification of affective intensities in online text. The

features investigated include a large set of learned n-grams as well as automatically and manually generated affect lexicons used in prior research. We also propose a support vector regression correlation ensemble (SVRCE) method for text-based affect classification. SVRCE combines feature subset ensembles with affect correlation information for improved affect classification performance. Evaluation of the various feature representations and the proposed method in comparison with existing affect analysis techniques found that the use of SVRCE with n-grams is highly effective for affect classification of online forums, blogs, and stories.

The remainder of this paper is organized as follows: Section 2 provides a review of related work on textual affect analysis. Section 3 outlines our research framework based on gaps and questions derived from the literature review. Section 4 presents an experimental evaluation of the various features and techniques incorporated in our framework. Section 5 features a brief case study illustrating how the proposed affect analysis methods can be applied to large CMC archives. Section 6 provides concluding remarks and describes future research directions.

2 RELATED WORK

Affect analysis is concerned with the analysis of text containing emotions [21], [26]. Emotional intelligence, the ability to effectively recognize emotions automatically, is crucial for learning preference related information and determining the importance of text content [22]. Affect analysis is associated with sentiment analysis, which looks at the directionality of text, i.e., whether a text segment is positively or negatively oriented [11]. However, there are two major differences between affect analysis and sentiment analysis. First, affect analysis involves a large number of potential emotions or affect classes [26]. These include

• The authors are with the Artificial Intelligence Lab, Department of Management Information Systems, University of Arizona, 1130 E. Helen Street, PO Box 210108, Tucson, AZ 85721.
E-mail: {aabbasi, sthoms, futj}@email.arizona.edu, hchen@eller.arizona.edu.

Manuscript received 14 Mar 2007; revised 22 Aug. 2007; accepted 20 Feb. 2008; published online 4 Mar. 2008.

For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2007-03-0107.

Digital Object Identifier no. 10.1109/TKDE.2008.51.

TABLE 1
Related Prior Affect Analysis Studies

Study	Features	Technique(s)	Analysis Level	Test Bed and Results
Donath et al., [7]	Manual lexicon, punctuation	Posting scoring	Posting	Greek USENET forums, visualization of anger intensities over time
Subasic & Hueitner, [26]	Manual lexicon (fuzzy semantic typing)	Word scoring	Word	Movie reviews and news stories; visualization of 83 affects
Liu et al., [14]	Language patterns derived from knowledge base	Sentence scoring	Sentence	User study on email browser
Chuang & Wu, [3]	Manual lexicon	Support vector machine (SVM)	Sentence	Drama broadcast transcripts; 76.44% accuracy for 7 class experiments
Grefenstette et al., [9]	Manual lexicon, semantic orientation	Manual tagging, point-wise mutual information (PMI)	Word	Candidate affect words; scored intensities across 86 affects
Grefenstette et al., [10]	Manual lexicon	Word scoring	Word	Political web pages; scored text relating to certain topic
Read, [23]	Semantic orientation	Point-wise mutual information (PMI)	Sentence	Short stories; 47.14% accuracy for 2 class experiments
Ma et al., [15]	Manual lexicon (WordNet-Affect database)	Word scoring	Sentence	Instant messaging chat data; no formal evaluation
Mishne, [16]	BOWs, POS tags, document length, emphasized words, semantic orientation, WordNet lexicon	Support vector machine (SVM)	Posting	LiveJournal blog postings; 60.25% accuracy for 2 class experiments
Cho & Lee, [5]	Manual lexicon, BOWs	Sentence scoring, Support vector machine (SVM)	Song	Korean song lyrics; 77.3% accuracy on 5 class experiments
Mishne & Rijke, [17]	Word n-grams	Pace regression	Posting	LiveJournal blog postings; average error of 52.53%, correlation coefficient of 0.827 for 2 class experiments
Wu et al., [32]	Emotion generation and association rules	Separable mixture models	Posting	Student chat dialog; 80.98% accuracy for 3 class experiments

happiness, sadness, anger, hate, violence, excitement, fear, etc. In contrast, sentiment analysis primarily deals with positive, negative, and neutral sentiment polarities. Second, while the sentiments associated with particular words or phrases are mutually exclusive, text segments can contain multiple affects [26], [10]. For example, the sentence "I can't stand you!" has a negative sentiment polarity but simultaneously contains hate and anger affects. Word level examples include the verb form of "alarm," which can be attributed to the fear, warning, and excitement affects [26], and the adjective "gleeful," which can be assigned to the happiness and excitement affect classes [10]. Additionally, certain affect classes may be correlated [26]. For instance, hate and anger often co-occur in text segments, resulting in a positive correlation. Similarly, happiness and sadness are opposing affects that are likely to have a negative correlation. In summary, affect analysis involves assigning text with emotive intensities across a set of mutually inclusive and possibly correlated affect classes. Important affect analysis characteristics include the features used to represent the presence of affects in text, techniques for assigning affective intensity scores, and the level of text granularity at which the analysis is performed. Table 1 presents a summary of the relevant prior studies based on these important affect analysis characteristics.

Based on the table, we can make several observations regarding the features and techniques used in previous affect analysis research:

1. Most prior research has used either manually generated lexicons, lexicons automatically created using WordNet (WNet) or semantic orientation (SO),

or generic feature representations such as word and part-of-speech (POS) tag n-grams. It is unclear which of these feature representations is most effective for affect analysis.

2. Techniques used for assigning affect intensities can be predominantly categorized as scoring methods or machine learning techniques. However, we are unaware of any prior work attempting to compare various techniques for affect classification.
3. Previous affect classification studies typically utilized between two and seven affect classes, applied at the word, sentence, or document levels. Despite the presence of multiple interrelated affects [26], [10], class correlation information was not leveraged for improved affect intensity assignment. Additionally, regression-based methods have seen limited usage despite their effectiveness in related application domains [20], [24].
4. Prior studies mainly focused on a single application domain, such as movie reviews, web forums, blogs, chat dialog, song lyrics, stories, etc. Given the differences in the degree of interaction, language usage, and communication structure across these domains, it is unclear if an approach suitable for classifying story affects will be applicable on web forums and blogs. The features and techniques used in prior research are expounded upon in the remainder of the section.

2.1 Features for Affect Analysis

The attributes used to represent affects can be classified as either lexicon-based features or generic n-gram-based

features. Considerable prior research has used manually or automatically generated lexicons. As previously stated, in affect lexicons, the same word/phrase can be assigned to multiple affect classes. The intensity score for an attribute is based on its degree of severity toward that particular affect class. Depending upon the semantic relation between affects, certain classes can have a positive or negative occurrence correlation [26].

Many studies have incorporated manually developed affect lexicons. Subasic and Huettner [26] used Fuzzy Semantic Typing where each feature was assigned to multiple affect categories with varying intensity and centrality scores depending upon the word and usage context. For example, the word "rat" was assigned to the disloyalty, horror, and repulsion affect categories with intensity scores of 0.9, 0.6, and 0.7, respectively (on a 0.0-1.0 scale where 1.0 was highest). In order to compensate for word-sense ambiguity, their approach also assigned each word-affect pair a centrality score indicating the likelihood of the word being used for that particular affect class. For example, the word "rat" was assigned a centrality score of 0.3 for the disloyalty affect and 0.6 for the repulsion affect (also on a 0.0-1.0 scale), since the usage of "rat" to convey disloyalty is not as common. Thus, while "rat" was more intense for the disloyalty affect, it was also less central to this class. In Subasic and Huettner's [26] approach, the intensity and centrality scores were both utilized for determining the affective composition of a text document. Although the accuracy for specific term affects may be inaccurate, the fuzzy logic approach is intended to capture the essence of a document's various affect intensities. A similar method for generating lexicons was employed in related work [9], [10]. Many other studies have also utilized manually constructed affect lexicons [3], [5]. Donath et al. [7] used a set of keywords relating to anger for analyzing USENET forums. Ma et al. [15] incorporated the WNet-Affect database created by Valitutti et al. [28]. This database is comprised of manually assigned affect intensities for words found in the WNet lexical resource [8]. Liu et al. [14] manually constructed sentence level language patterns for identification of six affect classes, including happiness, sadness, anger, fear, etc.

Although manually created affect lexicons can provide powerful insight, their construction can be time consuming and tedious. As a result, many studies have explored the use of automated lexicon generation methods such as SO [9], [23] and WNet lexicons [16]. These methods take a small set of manually generated seed/paradigm words which accurately reflect the particular affect class, and use automated methods for lexicon expansion.

Based on the work of Turney and Littman [27], the SO approach assesses the intensity of each word based on its frequency of co-occurrence with a set of core paradigm words reflective of that affect class [9]. The occurrence frequencies for the paradigm words and candidate words are derived from search engines such as AltaVista [9], [23] or Yahoo! [16]. The number of paradigm words used for a particular affect class is generally five to seven [9], [23]. For example, the paradigm words for the praise affect may include "acclaim, praise, congratulations, homage, approval," [9], and additional lexicon items generated automatically using SO include the words "award, honor, extol."

The SO approach is typically coupled with a point-wise mutual information (PMI) scoring mechanism for assigning candidate words' intensity scores [27]. Traditional PMI assigns each word a score based on how often it occurs in proximity with positive and negative paradigm words; however, it has been modified to be applicable with affect classes [23], [9]. The affect analysis rendition of PMI proposed by Grefenstette et al. [9] is as follows:

$$\text{PMI Score}(\text{word}, \text{Class}) = \log_2 \left(\frac{\prod_{\text{cword} \in \text{Class}} \text{hits}(\text{word Near cword})}{\prod_{\text{cword} \in \text{Class}} \log_2(\text{hits}(\text{cword}))} \right),$$

where *cword* is one of the paradigm words chosen for an affect class *Class* and *hits* is the number of pages found by Alta Vista.

The occurrence frequencies for the paradigm words and candidate words are derived from search engines such as AltaVista [9], [23] or Yahoo! [16]. Another automated affect lexicon generation method is WNet lexicons. Originally proposed by Kim and Hovy [13], this method is similar to SO. However, it uses WNet to expand the seed words associated with a particular affect class by comparing each candidate word's synset with the seed word list [16]. The intensity for a candidate word is proportional to the percentage of its synset also present in the seed word list for that particular affect class. Word scores are assigned using the following formula [13]:

$$\text{WordNet Score}(\text{word}, \text{Class}) = P(\text{Class}) \frac{1}{\text{count}(c)} \sum_{i=1}^n \text{count}(\text{syn}_i, \text{Class}),$$

where *Class* is an affect class, *syn_i* is one of the *n* synonyms of *word*, *P(Class)* is the number of words in *Class* divided by the total number of words considered.

In addition to lexicon-based affect representations, studies have also used generic n-gram features. Mishne [16] incorporated bag-of-words (BOWs) and POS tags in combination with automatically generated lexicons, while Mishne and de Rijke [17] used word n-grams for affect analysis of blog postings. Cho and Lee [5] used BOWs for classifying affects inherent in Korean song lyrics. N-grams have also been shown to be highly effective in the related area of sentiment classification [30], [1], especially when combined with machine learning methods capable of learning n-gram patterns conveying opinions and emotions. While prior research has used various n-gram and lexicon representations, we are unaware of any work done to evaluate the effectiveness of various potential affect analysis features.

2.2 Techniques for Assigning Affect Intensities

Prior research has utilized scoring and machine learning methods for assigning affect intensities. However, there has been no research done to investigate the effectiveness of these methods. Scoring-based methods, which are generally used in conjunction with lexicons, typically use the average intensity across lexicon items occurring in the text (i.e., word spotting) [26], [14], [5]. Sentence level

averaging has also been performed in combination with the word-level PMI scores generated using SO [27] as well as with WNet lexicons [13]. Studies utilizing manually developed lexicons comprised of sentence patterns obviously do not use averaging (at least at the sentence level), instead simply matching sentences with lexicon entries and assigning intensity scores accordingly [14], [5].

Machine learning techniques have also been utilized for assigning affect intensities. Many studies used support vector machine (SVM) for determining whether a text segment contains a particular affect class [3], [16], [5]. One shortcoming of using SVM is that it can only deal with discrete class labels, whereas affect intensities can vary along a continuum. Recent work has attempted to address this problem by using regression-based classifiers [20]. For example, Mishne and de Rijke [17] used word n-grams in unison with Pace regression [31] for assigning affect intensities in LiveJournal blogs. Pace regression overcomes many of the traditional problems associated with linear regression, such as lack of robustness against noise and redundancy [31]. Nevertheless, regression-based learning methods have seen limited usage despite their effectiveness in related application domains such as prediction of stock prices using financial text [24] or movie sales using blogs [34]. Grounded in Statistical Learning Theory [35], Support Vector Regression (SVR) is capable of predicting continuous affect intensities while still benefiting from the robustness of SVM [36]. It has been shown to work well on sparse data sets, which may be useful when dealing with affect occurrences [37].

3 RESEARCH DESIGN

In this section, we highlight affect analysis research gaps based on our review of the related work. Research questions are then posed based on the relevant gaps identified. Finally, a research framework is presented in order to address these research questions, along with some research hypotheses. The framework encompasses various feature representations and techniques for assigning affective intensities to sentences.

3.1 Gaps and Questions

Prior research has utilized manually or automatically generated lexicons as well as generic n-gram features for representing affective content in text. Since most studies used a single feature category and did not compare different alternatives, it is unclear which emotive representation is most effective. Furthermore, prior research has used scoring-based techniques and machine learning methods such as SVM. Regression-based methods capable of assigning continuous intensity scores have not been explored in great detail, with the exception of Mishne and de Rijke [17]. Leveraging the relationship between mutually inclusive affect classes in combination with powerful machine learning methods such as SVR could be highly effective for accurate assignment of affect intensities. Additionally, most prior affect analysis research was applied to a single domain (e.g., blogs, stories, etc.). Application across multiple domains could lend greater validity to the effectiveness of affect analysis features and

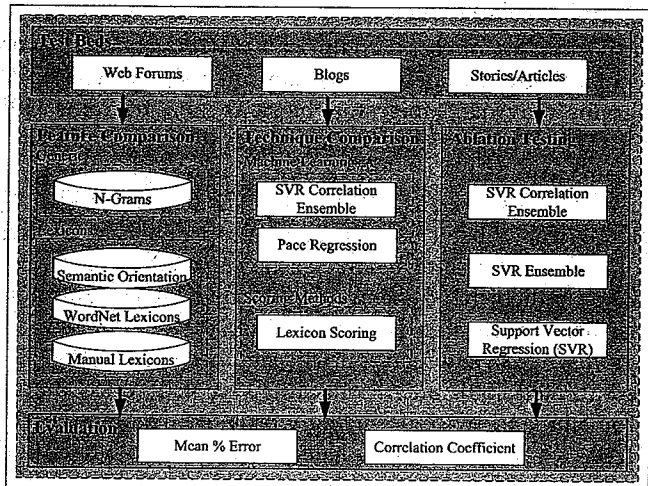


Fig. 1. Affect analysis research framework.

techniques. Based on these gaps, we present the following research questions:

- Which feature categories are best at accurately assigning affect intensities?
- Can the use of an extended feature set enhance affect analysis performance over individual generic and lexicon-based feature categories?
- Can a regression ensemble that incorporates affect correlation information outperform existing machine learning and scoring-based methods?
- What impact will the application domain have on affect intensity assignment?

3.2 Research Framework

Our research framework (shown in Fig. 1) relates to the features and techniques used for assigning affect intensity scores. We intend to compare generic n-gram features with automatically and manually generated lexicons. We also plan to assess the effectiveness of using an extended feature set encompassing all these attributes in comparison with individual feature categories. With respect to affect analysis techniques, we propose a support vector regression ensemble (SVRE) that considers affect correlation information when assigning emotive intensities to sentences. We intend to compare the SVRCE with other machine learning and scoring-based methods used in prior research. These include Pace regression [31], [17], SO [9], [23], WNet [13], and manual lexicon (ML) scoring [26]. We also plan to perform ablation testing to see how the different components of the proposed SVRCE method contribute to its overall performance. All testing will be performed on several test beds encompassing sentences derived from web forums, blogs, and stories. Features and techniques will be evaluated with respect to their percentage mean error and correlation coefficients in comparison with a human annotated gold standard. Further details about the features, techniques, ablation testing, and our research hypotheses are presented below, while the test bed and evaluation metrics are discussed in greater detail in the ensuing evaluation section.

3.3 Affect Analysis Features

The n-gram feature set is comprised of word, character, and POS tag n-grams. For each n-gram category, we use up to trigrams only, i.e., unigrams, bigrams, and trigrams [19], [30]. Word n-grams, including unigrams (e.g., "LIKE"), bigrams (e.g., "I LIKE", "LIKE YOU"), and trigrams (e.g., "I LIKE YOU") as well as POS tag n-grams (e.g., "NP VB", "JJ NP VB") have been used in prior affect analysis research [16]. We also include character n-grams (e.g., "li", "ik", "ike"), which have been useful in related sentiment classification studies [1]. In addition to standard word n-grams, hapax legomena and dis legomena collocations are incorporated [30]. Such collocations replace once (hapax legomena) and twice occurring words (dis legomena) with "HAPAX" and "DIS" tags. Hence, the trigram "I hate Jim" would be replaced with "I hate HAPAX" provided "Jim" only occurs once in the corpus. The intuition behind such collocations is to remove sparsely occurring words with tags that will allow the extracted n-grams to be more generalizable and, hence, more useful [30]. For instance, in the above example, the fact that the writer hates is more important from an affect analysis perspective than the specific person the hate is directed toward.

The lexicons employed are comprised of automated lexicons derived using SO and WNet models [9], [16]. We selected seven paradigm words for each affect class for input into the SO algorithm, as described in Section 2.1. For the WNet models, up to 50 words are used as seeds for each affect class, following the guidelines described by Kim and Hovy [13].

Our feature set also includes a manually crafted word level lexicon. The lexicon is comprised of over 1,000 affect words for several emotive classes (e.g., happiness, sadness, anger, hate, violence, etc.). Each word is assigned an intensity and ambiguity score between 0 and 1. The intensities are assigned based on the word's degree of severity or valence for its particular affect category (with 1 being highest). This approach is consistent with the intensity score assignment methods incorporated in previous studies that utilized manually crafted lexicons [7], [26], [10], [3]. Each affect word is also assigned an ambiguity score that signifies the probability of an instance of the word having semantic congruence with the affect class represented by that word. This score is determined by taking a sample set of instances of the word's occurrence and coding each as relevant or irrelevant; the percentage of relevant samples represents the word's ambiguity score. Hence, an ambiguity value of one suggests that the term always appears in the appropriate affective connotation. A maximum of 20 samples is used per word. Using more instances would be exhaustive, and we observed that the size used is sufficient to accurately capture the probability of a word being relevant to a particular affect class. The intensity and ambiguity assignment was done by two independent coders. Each coder initially assigned values without consulting the other. The coders then consulted one another in order to resolve tagging differences. The intercoder reliability tests revealed a kappa statistic of 0.78 prior to coder discussions and 0.89 after discrepancy resolution. For situations where the disparity could not be

TABLE 2
Manual Lexicon Examples for the Happiness Affect

Term	Intensity	Ambiguity	Weight
please	0.350	0.100	0.035
happy	0.600	1.000	0.600
overjoyed	0.950	1.000	0.950
elated	0.875	1.000	0.875

resolved even after discussions, the two coders' values were averaged. Table 2 shows examples from the happiness affect lexicon. The weight for each term is the product of its intensity and ambiguity value. This is the value assigned to each occurrence of the term in the text being analyzed. For example, "overjoyed" and "elated" were considered more severe by the coders than "happy". Although all three terms convey happiness with little ambiguity, elation and joy signify greater levels of happiness (i.e., higher intensity).

3.4 Affect Analysis Techniques

Ensemble classifiers use multiple classifiers with each built using different techniques, training instances, or feature subsets [6]. Particularly, the feature subset classifier approach has been shown to be effective for analysis of text patterns. Stamatakos and Widmer [25] used an SVM ensemble for music performer recognition. They used multiple SVMs each trained using different feature subsets. Similarly, Cherkauer [4] used a Neural Network ensemble for imagery analysis. Their ensemble consisted of 32 neural networks trained on eight different feature subsets. The intuition behind using a feature ensemble is that it allows each classifier to act as an "expert" on its particular subset of features [4], [25], thereby improving performance over simply using a single classifier. We propose the use of an SVRE that incorporates the relationship between various affect classes in order to enhance affect classification performance. Our ensemble includes multiple SVR models; each trained using a subset of features most effective for differentiating emotive intensities for a single affect class. We use the information gain (IG) heuristic to select the features for each SVR classifier. Since affect intensities are continuous, discretization is performed before IG can be applied. We use 5 and 10 class bins (e.g., an intensity value of 0.15 would be placed into class 1 of 5 and 2 of 10 using 5 and 10 class bins). All features with an average IG greater than a threshold t are selected [33].

The SVRCE adjusts the affect intensity prediction for a particular sentence based on the predicted intensities of other affects. The amount of adjustment is proportional to the level of correlation between affect classes (i.e., the affect class being predicted and the ones being used to make the adjustment) as derived from the training data. The SVRCE formulation is shown in Fig. 2. The rationale behind SVRCE is that, in certain situations, a particular sentence may get misclassified by a trained model due to a lack of prior exposure to the affective cues inherent in its text. In such circumstances, leveraging the relationship between affect classes may help alleviate the magnitude of such erroneous classifications.

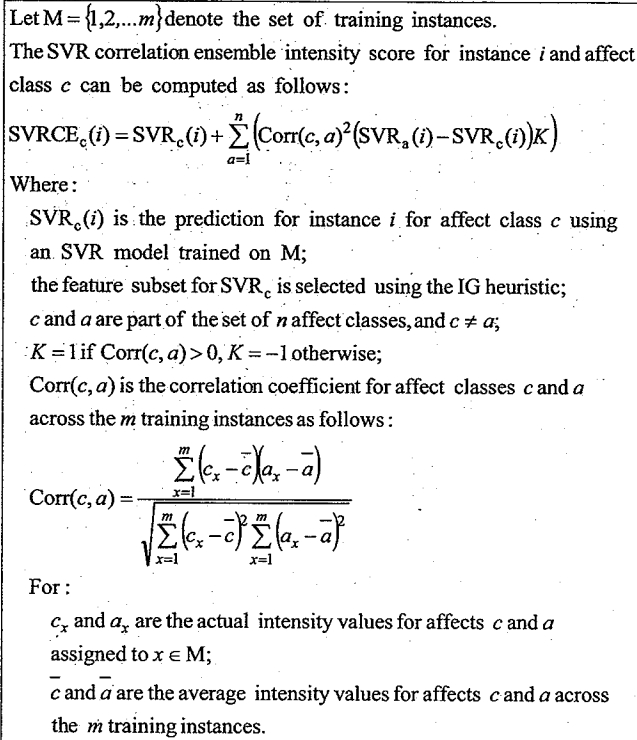


Fig. 2. SVRCE for assigning affect intensities.

We intend to compare the proposed SVRCE method against machine learning and scoring-based methods used in prior affect analysis research. These include the Pace regression technique proposed by Witten and Frank [31], which was used to analyze affect intensities in weblogs [17], as well as the SO, WNet model, and ML scoring approaches. In addition to comparing the proposed SVRCE against other affect analysis techniques, we also intend to perform ablation testing to better understand the impact different components of our proposed method have on classification performance. Since SVRCE uses correlation information and feature subset-based ensembles, we plan to compare it against an SVRE that does not use correlation information as well as an SVR trained using a single feature set for all affect classes. We also intend to compare it with an SVR stack; a classifier that combines the prediction values from an ensemble of underlying classifiers [38]. The potential benefit of using a stack is that it can combine low-level classifiers' prediction scores in a nonlinear manner using a higher level classifier [38], [39]. The SVR stack will use the SVRE classifiers' prediction scores as input features. The hypotheses associated with our research framework are presented below.

3.5 Research Hypotheses

H1: Features. The use of learned generic n-gram features will outperform manually and automatically crafted affect lexicons. Additionally, using an extended feature set encompassing all features will outperform individual feature sets.

- H1a: N-Grams > ML, SO, WNet models.
- H1b: All features > n-grams, MLs, SO, WNet models.

H2: Techniques. The proposed SVRCE method will outperform comparison techniques used in prior studies for affect analysis.

- H2: SVRCE > Pace regression, SO scores, WNet model scores, ML scores.

H3: Ablation Testing. The SVRCE method will outperform an SVRE not using correlation information, an SVR stack, as well as SVR run using a single feature set. Furthermore, the SVRE will also significantly outperform SVR run using a single feature set.

- H3a: SVRCE > SVRE, SVR, SVR stack.
- H3b: SVRE > SVR.
- H3c: SVR stack > SVR.

4 EVALUATION

We conducted experiments to evaluate various affective feature representations along with different affect analysis techniques, including the proposed SVRCE. The experiments were conducted on four test beds comprised of sentences taken from web forums, blogs, and short stories. This section encompasses a description of the test beds, experimental design, experimental results, and outcomes of the hypotheses testing.

4.1 Test Bed

Analyzing affect intensities across application domains is important in order to get a better sense of the effectiveness and generalizability of different features and techniques. As a result, our test bed consisted of sentences taken from four corpora (shown in Table 2). The first test bed was a set of supremacist web forums discussing issues relating to Nazi and socialist ideologies. The second was comprised of 1,000 sentences taken from a couple of Middle Eastern forums discussing issues relating to the war in Iraq. Analysis of such forums is important to better understand Cyberactivism, social movements, and people's political sentiments. Additionally, sentences were extracted from LiveJournal weblogs; a test bed used in prior research [16], [17]. The fourth test bed, which was also used in prior affect analysis research, consisted of sentences taken from Fifty Word Fiction, a website that posts short stories [23].

Two independent coders tagged the sentences for intensities across the four affect classes used for each test bed (shown in Table 3). Each sentence was tagged with an intensity score between 0 and 1 (with 1 being most intense) for each of the affects. The tagging followed the same format as the one used for the ML creation. Each coder initially assigned values without consulting the other. The coders then consulted one another in order to resolve tagging differences. For situations where the disparity could not be resolved even after discussions, the two coders' values were averaged. The intercoder reliability kappa values shown in Table 3 are from after discrepancy resolution (prior to averaging). For the Middle Eastern forums, the coders were unable to meet to resolve coding differences. For this test bed, the kappa value shown is for the two coders' initial tagging.

4.2 Experimental Design

Based on our research framework and hypotheses presented in Section 3, three experiments were conducted. The

TABLE 3
Test Bed Description

Test Bed	Source URL(s)	# Sentences	Affect Classes	Inter-coder
Fifty Word Short Stories (FWF)	tan-gents.co.uk/50words	758	Happiness, sadness, pleasantness, excitement	0.91
LiveJournal weblogs (LJ)	livejournal.com	1,000	Happiness, sadness, anger, hate	0.93
Supremacist Forums (SF)	stormfront.org nazi.org	1,000	Violence, anger, hate, racism	0.89
Middle Eastern Forums (MEF)	montada.com alfirdaws.com	1,000	Violence, anger, hate, racism	0.79*

first was intended to compare the performance of learned n-grams against manually and automatically crafted lexicons. We also investigated the effectiveness of an extended feature set comprised of n-grams and lexicons versus individual feature groups. The second experiment compared different affect analysis techniques, including the proposed SVRCE, Pace regression, and scoring methods. The final experiment pertained to ablation analysis of the major components of SVRCE, including the use of correlation information, stacking, and an ensemble approach to affect classification. In order to allow statistical testing of results, we ran 50 bootstrap instances for each condition across all three experiments. In each bootstrap run, 95 percent of the sentences were randomly selected for training, while the remaining 5 percent were used for testing [2]. The average results across the 50 bootstrap runs were reported for each experimental condition. Performance was evaluated using standard metrics for affect analysis, which include the mean percentage error and the correlation coefficient [17]:

$$\text{Mean Percentage Error} = \frac{100}{n} \sum |x - y|$$

$$\text{Corr}(X, Y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

where x and y are the actual and predicted intensity values for one of the n testing instances denoted by the vectors X and Y .

4.3 Experiment 1: Comparison of Feature Sets

In this experiment, we compared generic n-grams with SO, WNet model, and the ML. We also constructed an extended feature set comprised of n-grams, SO, WNet, and ML (labeled "All"). All feature sets were evaluated using the SVRCE. SVRCE was run using a linear kernel. N-grams were selected using the IG heuristic applied at the affect level. IG was applied to the 95 percent training data during each of the 50 bootstrap instances. These features were then used to train the SVRCE classifiers used on the testing data. This resulted in 16 n-gram feature subsets (one for each affect

TABLE 4
Overall Results for Various Feature Sets

Features/ Test Bed	Mean % Error			
	FWF	LJ	SF	MEF
N-Grams	4.9527	6.6472	4.6360	3.8066
SO	6.4928	7.1601	5.0725	4.4742
WNet	5.9080	7.1019	4.9646	4.5507
ML	5.5881	7.3417	4.9767	4.6147
All	5.0184	6.9265	4.8176	4.3522
	Correlation Coefficient			
N-Grams	0.4547	0.4367	0.6627	0.7455
SO	0.2809	0.2389	0.4558	0.5308
WNet	0.3147	0.1993	0.4952	0.5122
ML	0.3448	0.1810	0.5388	0.4121
All	0.4422	0.3577	0.6238	0.6036

class across the four test beds), and a corresponding SVRCE model for each feature subset. SO and WNet were run using the formulas described in Section 2.1. For SO, WNet, and ML, the word level scores were computed for each sentence, resulting in a vector of scores for each sentence. Since different paradigm/seed words were used for each affect across all four test beds, the lexicon methods also generated 16 sets of sentence vectors each. Consistent with Mishne [16], these vectors were treated as features input into the SVRCE. For the "All" feature set, the lexicon sentence vectors were merged with the n-gram frequency vectors.

Table 4 and Fig. 3 show the macrolevel experimental results for the mean percentage error and correlation coefficients across the five feature sets applied to all four test beds. The values shown were averaged across the four affect classes used within each test bed. The test bed labels correspond to the abbreviations presented in Table 3 under the column "Test Bed Name." The n-gram features appeared to have the best performance, with the lowest mean percentage error and highest correlation coefficient for all four test beds. The automated (i.e., SO and WNet) and MLs all had fairly similar performance, with mean errors typically in the 5 percent-7 percent range and correlation coefficients between 0.2 and 0.5. As anticipated, the use of all features performed well, outperforming the use of individual lexicons. Surprisingly however, using all features (i.e., n-grams in conjunction with lexicons) did not outperform the use of n-grams alone. N-grams outperformed the extended feature set by as much as 0.5 percent and 0.14 on mean error and correlation coefficient, respectively. This suggests that the learned n-grams were able to effectively represent affective patterns in the text. Adding lexicon features introduced redundancy, and in some instances, noise. Further elaboration regarding the performance of n-grams in comparison with other feature sets is provided in the hypotheses testing section (Section 4.6).

Fig. 4 shows the microlevel results for mean percentage error and correlation coefficient across the 16 classes incorporated (4 affects \times 4 test beds). Each class is labeled with its test bed and the first letter of its affect. The microlevel results indicate that the performance differences for various feature sets were fairly consistent across classes. N-grams had the lowest class-level mean error and the highest correlation coefficients, followed by the extended

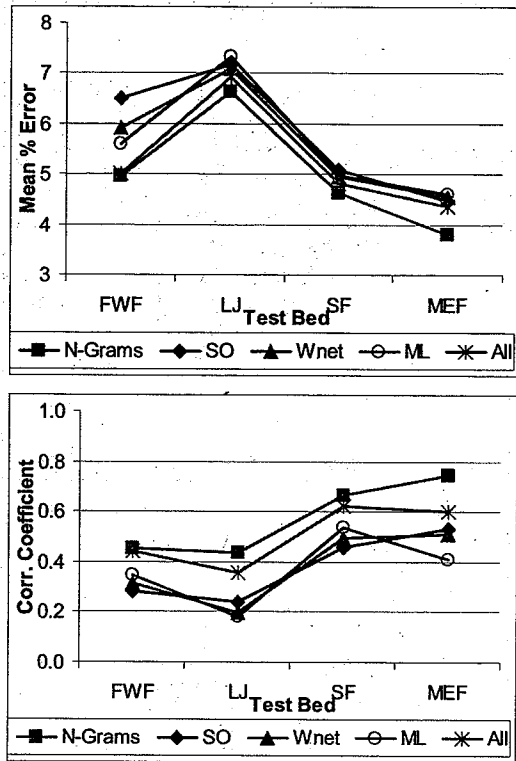


Fig. 3. Macrolevel mean percentage error and correlation coefficients for various feature sets.

feature set. Generally, the highest mean errors occurred on the sadness and hate affects on the short story and blog test beds, respectively (FWF-S and LJ-HT). The SO features had the worst performance, with especially low correlation coefficients on the supremacist forum test bed when analyzing the racism affect class (SF-R).

4.4 Experiment 2: Comparison of Techniques

The SVRCE method was compared against scoring and machine learning methods used in prior studies. The comparison techniques included Pace regression [17], WNet scores [13], [16], the PMI scores from the SO approach, and the scores from our ML. For SO, WNet, and ML, the average word level intensities were used as the sentence level scores as done in prior affect analysis research [26], [9], [23], [5]. SVRCE and Pace regression were both run using the n-gram features. N-grams were used since they had the best performance in Experiment 1. Both techniques (i.e., SVRCE and Pace) were run using identical features; with each using 16 feature subsets selected using the IG heuristic as described in Experiment 1. Any scores outside the 0-1 range were adjusted to fit the possible range of intensities (this was done in order to avoid inflated errors stemming from values well outside the feasible range).

Fig. 5 and Table 5 show the macrolevel experimental results for the mean percentage error and correlation coefficients across the five techniques. The SVRCE method had the best performance, with the lowest mean percentage error and highest correlation coefficient for all four test beds. Pace regression, WNet models, and the ML scoring methods were all in the middle, while the SO scoring method had the worst performance. The results are consistent with prior research that has also observed large differences between the word level scores assigned using

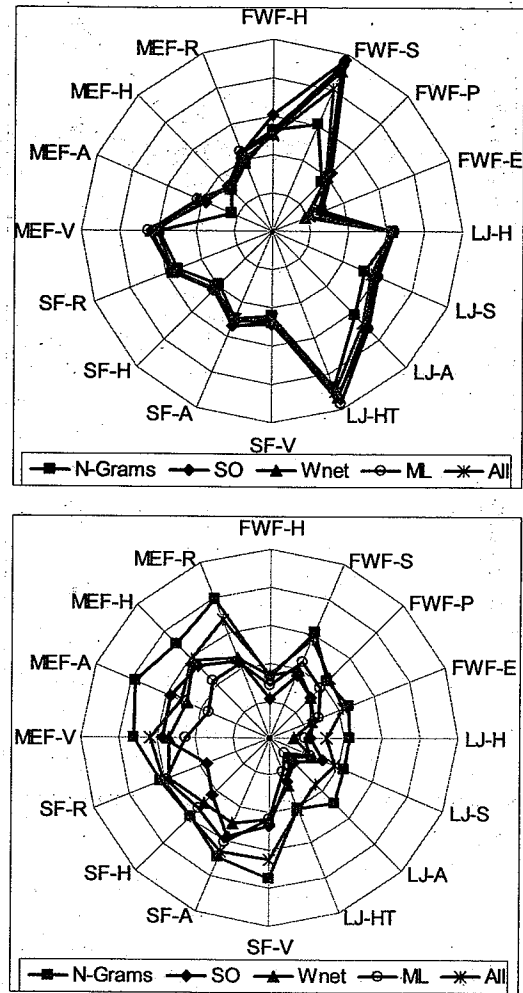


Fig. 4. Microlevel mean percentage error and correlation coefficients for various feature sets.

WNet and SO [16]. The machine learning methods (SVRCE and Pace) both fared well with respect to their correlation coefficients. Pace also performed well on the supremacist and Middle Eastern forums in terms of mean percentage error, but not on the blogs test bed (LJ).

Fig. 6 shows the microlevel results for mean percentage error and correlation coefficient across the 16 classes. The microlevel results indicate that the performance differences for various techniques were fairly consistent across classes. SVRCE had the lowest mean percentage error and the highest correlation coefficient for almost each class. SO fared especially poorly on the Middle Eastern forums for the racism, hate, and violence affects (MEF-R, MEF-H, MEF-V), with very high error percentages and low correlation coefficients. The WNet models and ML scoring methods were fairly close to one another in terms of error and correlation values across the 16 classes.

4.5 Experiment 3: Ablation Testing

Ablation testing was performed to evaluate the effectiveness of the different SVRCE components. The SVRCE was compared against an SVRE that does not utilize correlation information, an SVR stack, as well as an SVR classifier using only a single feature set (SVR). The SVR was trained using a single feature set (for each test bed) selected by using all n-grams occurring at least five times in the corpus [12]. The

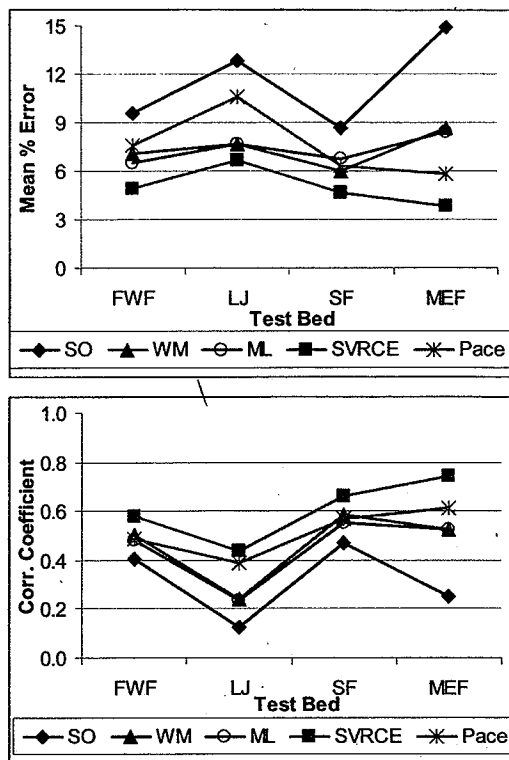


Fig. 5. Macrolevel mean percentage error and correlation coefficients for various techniques.

SVRE and SVRCE were both run using IG on the training data to select the 16 feature subsets most representative of each affect class. The SVR stack used the predictions from the 16 SVRE classifiers as input features. The experiment was intended to evaluate the two core components of SVRCE: 1) its use of feature ensembles to better represent affective content and 2) the use of correlation information for enhanced affect classification.

Table 6 and Fig. 7 show the macrolevel results for the mean percentage error and correlation coefficients for SVRCE, SVRE, SVR stack, and SVR. The SVRCE method had the best performance, with the lowest mean percentage error and highest correlation coefficient for all four test beds. SVRCE marginally outperformed SVRE and SVR

TABLE 5
Overall Results for Various Techniques

Features/ Test Bed	Mean % Error			
	FWF	LJ	SF	MEF
N-Grams	9.5634	12.8245	8.6590	14.8759
SO	7.0981	7.6321	5.9899	8.6639
WNet	6.4866	7.7012	6.7270	8.3860
ML	4.9527	6.6472	4.6360	3.8066
All	7.5748	10.6183	6.3038	5.8473
Correlation Coefficient				
N-Grams	0.4044	0.1271	0.4673	0.2530
SO	0.5005	0.2396	0.5837	0.5224
WNet	0.4805	0.2352	0.5500	0.5251
ML	0.5797	0.4367	0.6627	0.7455
All	0.4878	0.3856	0.5692	0.6124

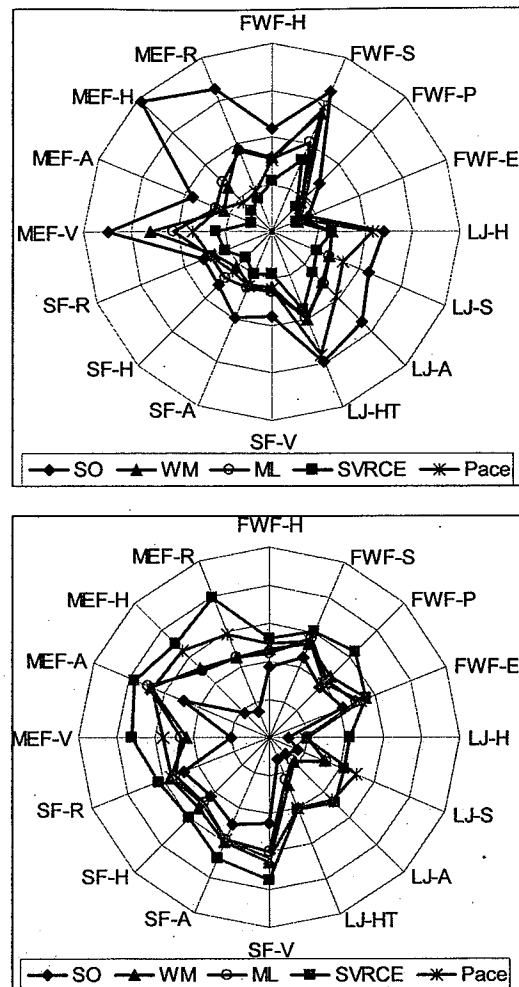


Fig. 6. Microlevel mean percentage error and correlation coefficients for various techniques.

stack, while all three techniques outperformed SVR. The results suggest that use of feature ensembles and correlation information are both useful for classifying affective intensities.

4.6 Hypotheses Results

We conducted pairwise t-tests on the 50 bootstrap runs for all three experiments. Given the large number of comparison conditions, a Bonferroni correction was performed to avoid spurious positive results. All p-values less than 0.0005 were considered significant at alpha = 0.01.

4.6.1 H1: Feature Comparison

Pairwise t-tests were conducted to compare the effectiveness of the extended and n-gram feature sets with other feature categories. N-grams and the extended feature set both significantly outperformed the lexicon-based representations on all test beds with respect to mean error and correlation (all p-values < 0.0001). Surprisingly, the extended feature set did not outperform n-grams. In contrast, the n-gram feature set significantly outperformed the use of all features (n-grams plus the three lexicons), with all p-values significant at alpha = 0.01 except the correlation coefficient on the FWF test bed (p-value = 0.0034).

Table 7 provides examples of learned n-grams taken from the LiveJournal test bed for the hate affect. It also

TABLE 6
Overall Results for Ablation Testing

Features/ Test Bed	Mean % Error			
	FWF	LJ	SF	MEF
SVRCE	4.9527	6.6472	4.6360	3.8066
SVRE	5.0351	6.6501	5.0776	4.0667
SVR	5.2379	7.7871	5.7676	5.0460
SVR Stack	4.9969	6.8799	5.0312	4.0699
	Correlation Coefficient			
SVRCE	0.4547	0.4367	0.6627	0.7455
SVRE	0.4271	0.4098	0.5990	0.7231
SVR	0.3896	0.3267	0.5631	0.5757
SVR Stack	0.4402	0.3770	0.6015	0.7173

TABLE 7
Sample Learned N-Grams for the Hate Affect

Learned N-Grams		Lexicon Items
Category	N-Gram	
Character N-Grams	uck, ck, fuc	awful, stupid, terrible,
Word N-Grams	terribly, suck, the stupid, the s**t, the f**k	sicken, s**t, f**k
Hapax and Dis Legomena Collocations	HAPAX so awful	
POS Tag N-Grams	PERSON_SG, WEEKDAY_NNP, TIME_SG	

shows some related hateful items from the ML. The n-grams were able to learn many of the concepts conveyed in the lexicon. Furthermore, the n-grams were able to provide better context for some features and also learn deeper patterns in several instances. For example, the hate in LiveJournal blogs is often directed toward specific people and frequently involves places and times. This pattern is captured by the POS tag n-grams. In contrast, word lexicons cannot accurately represent such complex patterns. The example illustrates how the n-gram features learned were more effective than the lexicons employed in this study.

4.6.2 H2: Technique Comparison

Based on pairwise t-tests performed on the techniques, the SVRCE method significantly outperformed all four comparison techniques on mean percentage error and correlation

coefficient across all four test beds. All p-values were less than 0.0005 and, therefore, significant at alpha = 0.01. The results indicate that the SVRCE method's use of ensembles of learned n-gram features combined with affect correlation information allows the classifier to assign affect intensities with greater effectiveness than comparison approaches used in prior research.

4.6.3 H3: Ablation Tests

Pairwise t-tests were also conducted to assess the contribution of the major components of the SVRCE method. The results of SVRCE versus SVRE and SVR stack revealed that the use of correlation information significantly enhanced performance in most cases, significant for three out of four test beds on mean error and correlation. The results were not significant for mean error on the LiveJournal blog test bed (p-value = 0.3452) as well as for correlation on the Middle Eastern forum data set (p-value = 0.0013) when for SVRCE compared to SVRE. SVRCE also did not significantly outperform the SVR stack on the Fifty Word Fiction data set (p-value = 0.2943 on mean error, p-value = 0.1798 on correlation). SVRCE, SVRE, and SVR stack all significantly outperformed SVR, indicating that the use of feature ensembles is effective for classifying affect intensities (all p-values less than 0.0001, significant at alpha = 0.01).

Table 8 shows example sentences taken from the LiveJournal blog data set. The anger affect intensity scores assigned by the coders (gold standard) as well as the SVRE and SVRCE scores are shown. SVRCE was more accurate for these sentences because it was able to leverage information from the closely correlated hate affect. The hate affect had a correlation coefficient of over 0.6 with anger on the training data. Hence, although the SVRE anger intensity scores were low for these three sentences, the higher hate scores boosted the anger intensity values assigned by SVRCE. The example illustrates how the use of correlation information between affect classes can improve affect intensity classification in text where appropriate intensity cues/patterns may not be readily apparent.

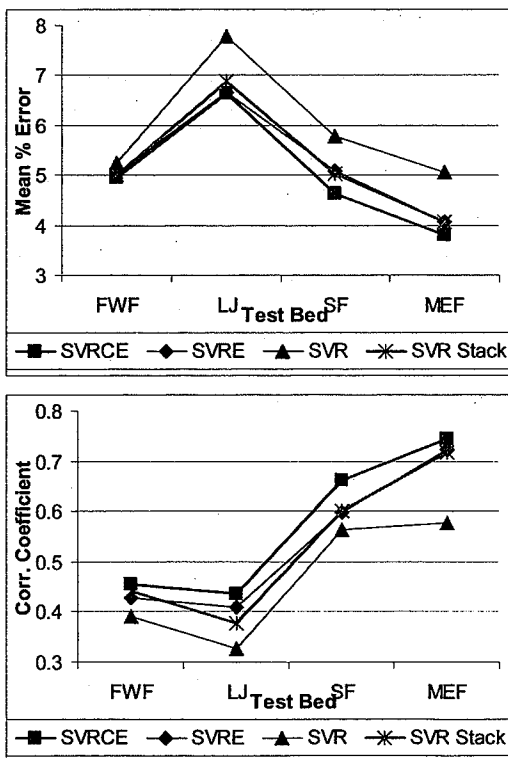


Fig. 7. Microlevel mean percentage error and correlation coefficients for ablation testing.

5 CASE STUDY

Many prior studies have used brief case studies to illustrate the utility of their proposed affect analysis

TABLE 8
Example Sentences where Correlation Information Improved
Anger Affect Classification Performance

Sentence	Coders	SVRE	SVRCE
We're so apathetic as a nation and it just makes me sick!	0.900	0.109	0.484
Damn those confusing streets with no street name signs.	0.750	0.123	0.348
Which part of "personal space" don't you understand?	0.600	0.000	0.218

methods [26], [17]. In order to demonstrate the usefulness of the SVRCE method coupled with a rich set of learned n-grams, we analyzed the affective intensities in two popular Middle Eastern Web forums: www.alfirdaws.org/vb and www.montada.com. Analysis of affects in such forums is important for sociopolitical reasons and to better our understanding of social phenomena in online communities. Firdaws is considered a more extreme forum by domain experts, with considerable content dedicated to support the Iraqi insurgency and Al-Qaeda. In contrast, Montada is a general discussion forum with content and discussion pertaining to various social matters. We hypothesized that our SVRCE method would be able to effectively depict the likely intensity differences for appropriate affect classes, between these two Web forums.

We used spidering programs to collect the content in both Web forums. Table 9 shows summary statistics for the content collected from the two forums. The Montada forum was considerably larger, with over 31,000 authors and a large number of threads and postings, partially because it had been around for approximately 7 years. Firdaws was a relatively newer forum, beginning in 2005. Due to the nature of its content and time duration of existence, this forum had fewer authors and postings.

Fig. 8 shows number of posts for each month the forums have been active. Montada was very active in 2002 and 2005, with over 20,000 posts in some months, yet appears to be in a down phase in 2007 (similar to 2004). Firdaws consistently had between 2,500 and 3,000 posts per month since the second half of 2006.

The SVRCE classifier was employed in conjunction with the n-gram feature set to analyze affect intensities in the two web forums. Analysis was performed on violence, hate, racism, and anger affects. We computed the average posting level intensities (averaged across all sentences in a posting) as well as the total intensity per post (the summation of sentence intensities in each posting). The analysis was performed on all postings in each forum (approximately

TABLE 9
Summary Statistics for Two Web Forums Collected

Forum	Firdaws	Montada
# Authors	2,189	31,692
# Threads	14,526	114,965
# Messages	39,775	869,264
# Sentences	244,917	2,052,511
Duration	1/2005 - 7/2007	9/2000 - 7/2007

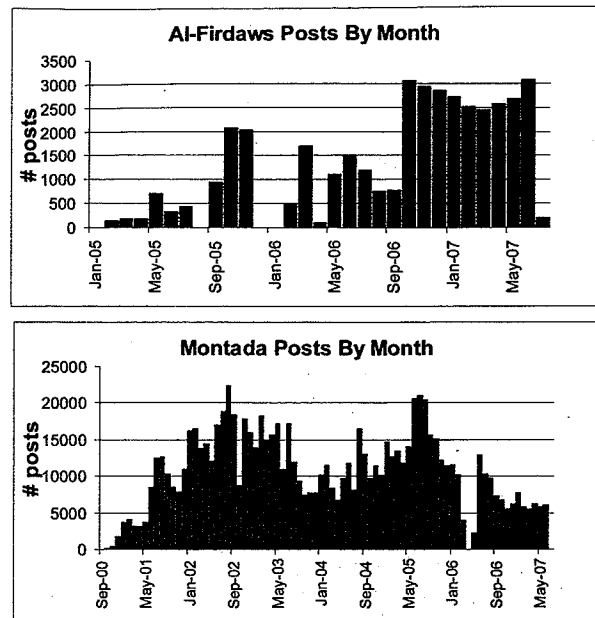


Fig. 8. Posting frequency for two Web forums.

900,000 postings and 2.3 million sentences). As shown in Table 10, the Al-Firdaws forum had considerably higher affect intensities for all four affect classes, usually two to three times greater than Montada.

Fig. 9 depicts the average message violence and hate intensities over time for all postings in each of the two web forums. The x -axis indicates time, while the y -axis shows the intensities (on a scale of 0 to 1). Each point represents a single message, while areas with greater message concentrations are darker. The blank periods in the diagrams correspond to periods of posting inactivity in forums (see Fig. 8 for correspondence). Based on the diagram, we can see that Firdaws has considerably higher violence and also greater hate intensity across time. Firdaws also appears to have increasing violence intensity in 2007 (based on the concentration of postings), possibly attributable to the increased activity in this forum. In contrast, violence and hate intensities are consistently low in Montada. The results generated using SVRCE and n-gram features are consistent with existing knowledge regarding these two forums. The case study illustrates how the proposed features and techniques can be successfully applied toward affect analysis of computer-mediated communication text.

6 CONCLUSIONS

In this study, we evaluated various features and techniques for affect analysis of online texts. In addition, the SVRCE

TABLE 10
Affect Intensities per Posting Across Two Forums

Intensity	Forum	Violence	Anger	Hate	Racism
Average	Firdaws	0.084	0.018	0.037	0.032
	Montada	0.027	0.012	0.010	0.014
Total	Firdaws	0.523	0.127	0.178	0.191
	Montada	0.246	0.105	0.092	0.134

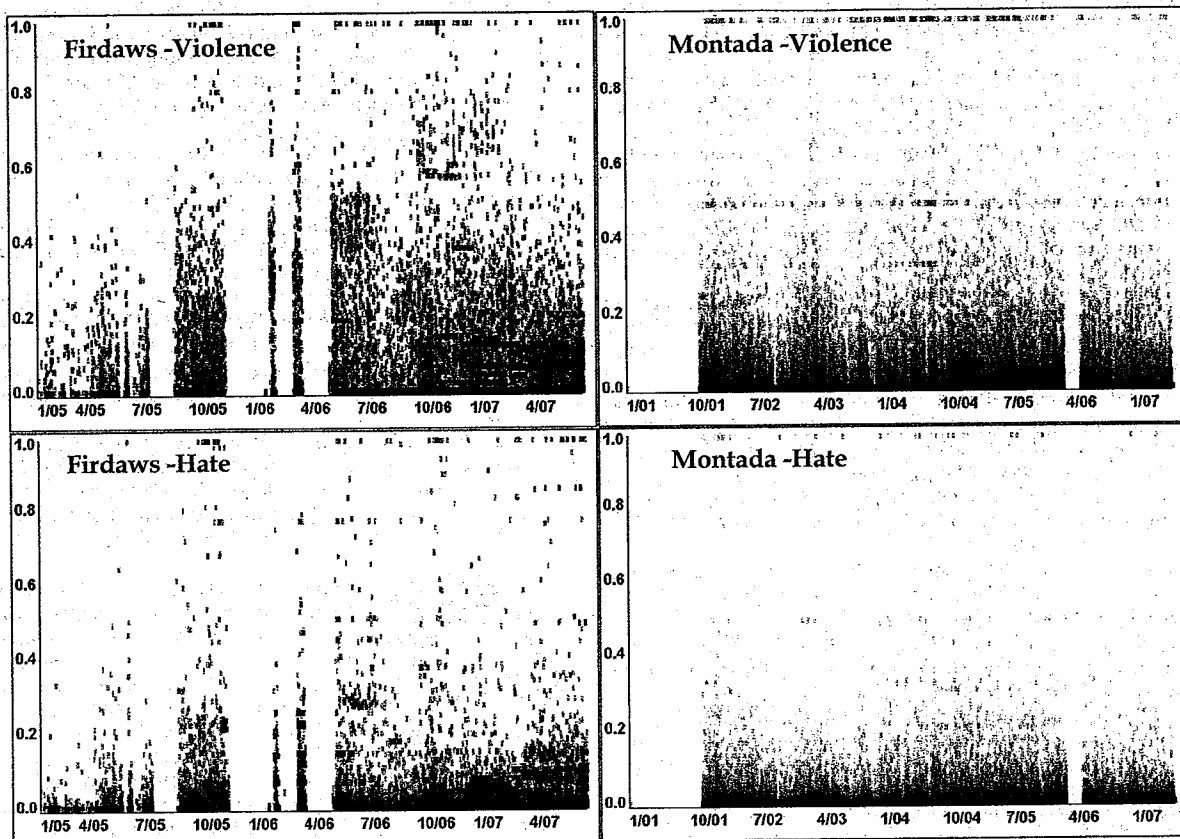


Fig. 9. Temporal view of intensities in two Web forums.

was proposed. This method leverages an ensemble of SVR classifiers with each constructed for a separate affect class. The ensemble of predictions combined with the correlation between affect classes is leveraged for enhanced affect classification performance. Experimental results on test beds derived from online forums, blogs, and stories revealed that the proposed method outperformed existing affect analysis techniques. The results also suggested that learned n-grams can improve affect classification performance in comparison with lexicon-based representations. However, combining n-gram and lexicon features did not improve performance due to increased amounts of noise and redundancy in the extended feature set. A case study was also performed to illustrate how the proposed features and techniques can be applied to large cyber communities in order to reveal affective tendencies inherent in these communities' discourse.

To the best of our knowledge, the experiments conducted in this study are the first to evaluate features and techniques for affect analysis. Furthermore, we are also unaware of prior research applied to such a vast array of domains and test beds.

We believe this study provides an important stepping stone for future work intended to further enhance the feature representations and techniques used for classifying affects. Based on this work, we have identified several future research directions. We intend to apply the techniques across a larger set of affect classes (e.g., 10-12 affects per test bed). We are also interested in exploring additional feature representations, such as the use of richer

learned n-grams (e.g., semantic collocations, variable n-gram patterns, etc.). We also plan to evaluate the effectiveness of real world knowledge bases such as those employed by Liu et al. [14].

ACKNOWLEDGMENTS

The authors wish to thank the associate editor and reviewers for their invaluable feedback. This research has been supported in part by the following grant: US NSF/ITR, "COPLINK Center for Intelligence and Security Informatics—A Crime Data Mining Approach to Developing Border Safe Research," EIA-0326348, September 2003-August 2005.

REFERENCES

- [1] A. Abbasi, H. Chen, and A. Salem, "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums," *ACM Trans. Information Systems*, vol. 26, no. 3, article 12, July 2008.
- [2] S. Argamon, C. Whitelaw, P. Chase, S.R. Hota, N. Garg, and S. Levitan, "Stylistic Text Classification Using Functional Lexical Features," *J. Am. Soc. for Information Science and Technology*, vol. 58, no. 6, pp. 802-822, 2007.
- [3] Z. Chuang and C. Wu, "Multi-Modal Emotion Recognition from Speech and Text," *Computational Linguistics and Chinese Language Processing*, vol. 9, no. 2, pp. 45-62, 2004.
- [4] K.J. Cherkauer, "Human Expert-Level Performance on a Scientific Image Analysis Task by a System Using Combined Artificial Neural Networks," *Working Notes of the AAAI Workshop Integrating Multiple Learned Models*, P. Chan, ed., pp. 15-21, 1996.
- [5] Y.H. Chó and K.J. Lee, "Automatic Affect Recognition Using Natural Language Processing Techniques and Manually Built Affect Lexicon," *IEICE Trans. Information Systems*, vol. E89, no. 12, pp. 2964-2971, 2006.

- [6] T.G. Dietterich, "Ensemble Methods in Machine Learning," *Proc. First Int'l Workshop Multiple Classifier Systems (MCS '00)*, pp. 1-15, 2000.
- [7] J. Donath, K. Karahalio, and F. Viegas, "Visualizing Conversation," *Proc. 32nd Conf. Computer-Human Interaction (CHI)*, 1999.
- [8] *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
- [9] G. Grefenstette, Y. Qu, D.A. Evans, and J.G. Shanahan, "Validating the Coverage of Lexical Resources for Affect Analysis and Automatically Classifying New Words Along Semantic Axes," *Proc. AAAI Spring Symp. Exploring Attitude and Affect in Text: Theories and Applications (AAAI-EAAT '04)*, Y. Qu, J. Shanahan, and J. Wiebe, eds., pp. 71-78, 2004.
- [10] G. Grefenstette, Y. Qu, J.G. Shanahan, and D.A. Evans, "Coupling Niche Browsers and Affect Analysis for an Opinion Mining Application," *Proc. 12th Int'l Conf. Recherche d'Information Assistee par Ordinateur (RIAO '04)*, pp. 186-194, 2004.
- [11] M.A. Hearst, "Direction-Based Text Interpretation as an Information Access Refinement," *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, P. Jacobs, ed., Lawrence Erlbaum Assoc., 1992.
- [12] M. Jiang, E. Jensen, S. Beitzel, and S. Argamon, "Choosing the Right Bigrams for Information Retrieval," *Proc. Meeting of the Int'l Federation of Classification Soc.*, 2004.
- [13] S. Kim and E. Hovy, "Determining the Sentiment of Opinions," *Proc. 20th Int'l Conf. Computational Linguistics (COLING '04)*, pp. 1367-1373, 2004.
- [14] H. Liu, H. Lieberman, and T. Selker, "A Model of Textual Affect Sensing Using Real-World Knowledge," *Proc. Eighth Int'l Conf. Intelligent User Interfaces*, 2003.
- [15] C. Ma, H. Prendinger, and M. Ishizuka, "Emotion Estimation and Reasoning Based on Affective Textual Interaction," *Proc. First Int'l Conf. Affective Computing and Intelligent Interaction (ACII '05)*, pp. 622-628, 2005.
- [16] G. Mishne, "Experiments with Mood Classification," *Proc. First Workshop Stylistic Analysis of Text for Information Access Workshop (Style)*, 2005.
- [17] G. Mishne and M. de Rijke, "Capturing Global Mood Levels Using Blog Posts," *Proc. AAAI Spring Symp. Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, 2006.
- [18] K. Nigam and M. Hurst, "Towards a Robust Metric of Opinion," *Proc. AAAI Spring Symp. Exploring Attitude and Affect in Text (AAAI-EAAT)*, 2004.
- [19] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up? Sentiment Classification Using Machine Learning Techniques," *Proc. Empirical Methods in Natural Language Processing (EMNLP '02)*, pp. 79-86, 2002.
- [20] B. Pang and L. Lee, "Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales," *Proc. Ann. Meeting on Assoc. for Computational Linguistics (ACL '05)*, pp. 115-124, 2005.
- [21] R.W. Picard, *Affective Computing*. MIT Press, 1997.
- [22] R.W. Picard, E. Vyzas, and J. Healey, "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1179-1191, Oct. 2001.
- [23] J. Read, "Recognizing Affect in Text Using Point-Wise Mutual Information," master's thesis, 2004.
- [24] R. Schumaker and H. Chen, "Textual Analysis of Stock Market Prediction Using Financial News Articles," *Proc. 11th Americas Conf. Information System (AMCIS)*, 2006.
- [25] E. Stamatatos and G. Widmer, "Music Performer Recognition Using an Ensemble of Simple Classifiers," *Proc. 15th European Conf. Artificial Intelligence (ECAI)*, 2002.
- [26] P. Subasic and A. Huettner, "Affect Analysis of Text Using Fuzzy Semantic Typing," *IEEE Trans. Fuzzy Systems*, vol. 9, no. 4, pp. 483-496, 2001.
- [27] P.D. Turney and M.L. Littman, "Measuring Praise and Criticism: Inference of Semantic Orientation from Association," *ACM Trans. Information Systems*, vol. 21, no. 4, pp. 315-346, 2003.
- [28] A. Valitutti, C. Strapparava, and O. Stock, "Developing Affective Lexical Resources," *PsychNology J.*, vol. 2, no. 1, pp. 61-83, 2004.
- [29] J. Wiebe, "Tracking Point of View in Narrative," *Computational Linguistics*, vol. 20, no. 2, pp. 233-287, 1994.
- [30] J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin, "Learning Subjective Language," *Computational Linguistics*, vol. 30, no. 3, pp. 277-308, 2004.
- [31] I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, second ed. Morgan Kaufman, 2005.
- [32] C. Wu, Z. Chuang, and Y. Lin, "Emotion Recognition from Text Using Semantic Labels and Separable Mixture Models," *ACM Trans. Asian Language Information Processing*, vol. 5, no. 2, pp. 165-182, 2006.
- [33] Y. Yang and J.O. Pederson, "A Comparative Study on Feature Selection in Text Categorization," *Proc. 14th Int'l Conf. Machine Learning (ICML '97)*, pp. 412-420, 1997.
- [34] G. Mishne and N. Glance, "Predicting Movie Sales from Blogger Sentiment," *Proc. AAAI Spring Symp. Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, 2006.
- [35] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.
- [36] A. Webb, *Statistical Pattern Recognition*. John Wiley & Sons, 2002.
- [37] K.R. Muller, A.J. Smola, G. Ratsch, B. Scholkopf, J. Kohlmorgen, and V. Vapnik, "Predicting Time Series with Support Vector Machines," *Proc. 15th Int'l Joint Conf. Artificial Intelligence (IJCAI)*, 1997.
- [38] A.L. Prodromidis and S.J. Stolfo, "A Comparative Evaluation of Meta-Learning Strategies over Large and Distributed Data Sets," *Proc. 16th Int'l Conf. Machine Learning (ICML '99)*, pp. 18-27, 1999.
- [39] K.M. Ting and I.H. Witten, "Stacked Generalization: When Does It Work?" *Proc. 15th Int'l Joint Conf. Artificial Intelligence (IJCAI)*, 1997.



Ahmed Abbasi received the BS and MBA degrees in information technology from Virginia Tech and the PhD degree in information systems from the University of Arizona. He is currently a professor of information systems at the University of Wisconsin-Milwaukee. He has published several peer-reviewed articles on computer-mediated communication, text mining, electronic commerce, and information visualization. He is a member of the IEEE, the Association for Information Systems (AIS), and the Decision Sciences Institute (DSI).



Hsinchun Chen received the BS degree from the National Chiao-Tung University, Hsinchu, Taiwan, R.O.C., the MBA degree from SUNY Buffalo, and the PhD degree in information systems from New York University. He is a professor of information systems and the director of the Artificial Intelligence Lab, University of Arizona. He has authored/edited 13 books, 17 book chapters, and more than 150 SCI journal articles covering digital library, intelligence analysis, biomedical informatics, data/text/Web mining, knowledge management, and Web computing. He serves on 10 editorial boards and has been an advisor for major US National Science Foundation, US Department of Justice, US National Library of Medicine, US Department of Defense, US Department of Homeland Security, and other international research programs. He received the IEEE Computer Society 2006 Technical Achievement Award. He is a fellow of the IEEE and the AAAS.



Sven Thoms received the BS degree from Washington State University and the MS degree from the University of Arizona. He is currently a doctoral student in the Department of Management Information Systems, University of Arizona.



Tianjun Fu received the BS degree from Shanghai Jiao Tong University. He is currently working toward the PhD degree in information systems and is also a research associate in the Artificial Intelligence Lab, University of Arizona. His research interests include text and social network analysis of computer-mediated communication.