Using Verbs and Adjectives to Automatically Classify Blog Sentiment

Paula Chesley
Department of Linguistics
University at Buffalo
pchesley@buffalo.edu

Bruce Vincent and Li Xu
Department of Computer Science
and Engineering
University at Buffalo
{bvincent,lixu}@buffalo.edu

Rohini K. Srihari
Center of Excellence for Document
Analysis and Recognition (CEDAR)
University at Buffalo
rohini@cedar.buffalo.edu

Abstract

This paper presents experiments on subjectivity and polarity classification of blog posts, making novel use of a linguistic feature, verb class information, and of an online resource, the Wikipedia dictionary. The verb classes we use express objectivity and polarity, i.e., a positive or negative opinion. We derive the polarity of adjectives from their entries in the online dictionary. Each post from a blog is classified as *objective*, *subjective-positive*, or *subjective-negative*. Our approach is topic- and genre-independent, and our method of determining the polarity of adjectives has an accuracy rate of 90.9%. Accuracy rates of two verb classes demonstrating polarity are 89.3% and 91.2%. Initial classifier results show accuracies of 72.4% on objective posts, 84.2% for positive posts, and 80.3% for negative posts. These accuracies all represent substantial increases above the established baseline classification.

Introduction

As the blogging phenomenon continues its exponential growth, its increasingly influential role in the global marketplace of ideas and opinions is becoming widely acknowledged. In the *Harvard Business Review* of February 2005 Mohanbir Sawhney notes that "The 'blogosphere' is gaining the power to influence what people think and do... Bloggers are driven by a desire to share their ideas and opinions with anyone who cares to tune in" (Sawney 2005). Consequently, new techniques to automatically extract and analyze sentiment expressed in blogs constitute lively research topics.

In this paper we describe textual and linguistic features we extract and a classifier we develop to categorize blog posts with respect to sentiment. We ask whether a given blog post expresses subjectivity (vs. objectivity), and whether the sentiment a post expresses represents positive ("good") or negative ("bad") polarity.

Although the expression of subjectivity vs. objectivity is ideally expressed on a continuous scale, we simplify by using a binary classification. Similarly, we aim for a binary classification of positive vs. negative sentiment. In this research we address the following issues related to blog sentiment analysis:

1. How effectively can classes of verbs in blog posts categorize sentiment?
2. How can we utilize online lexical resources, such as Wikipedia's dictionary, to automatically categorize adjectives expressing polarity?
3. What is the best way to propagate up sentiment information from the lexical level to the post level?
4. Are blogs different from other genres expressing sentiment and if so, how?

Having manually examined hundreds of blogs, the answer to (4) seems to reside precisely in the diverse rhetorical structure of blog posts and large-scale references to other media. An op-ed piece from a newspaper often states the opinion in the first paragraph of the article and reiterates this opinion when closing (so much so that examining simply the first and last paragraphs for sentiment would potentially yield robust results). However, the majority of a blog post is often a quoted excerpt of an article with which the blogger agrees or disagrees, followed by a short comment expressing his or her agreement or disagreement. We believe such *nested* speech events (Wiebe, Wilson, & Cardie 2004) are a common occurrence in blogs and represent a significant challenge when classifying their sentiment.

To aid in the sentiment classification task, we make use of verb-class information, since exploiting lexical information contained in verbs has shown to be a successful technique for classifying documents (Klavans & Kan 1998). To obtain this information we use InfoXtract (Srihari *et al.* 2003), an automatic text analyzer which groups verbs according to classes that often correspond to our polarity classification. Additionally, we utilize Wikipedia's online dictionary, the Wiktionary[1], to determine the polarity of adjectives seen throughout the posts. With adjective accuracy at 90.9% and initial evaluations on verb classes to show 89.3% and 91.2% accuracy, these methods for extracting lexical polarity prove very robust.

We then propagate this lexical classification information up to the post level. This approach could then be extended to an entire blog, by categorizing a blog as subjective if the majority of its posts are subjective. Presumably, an author

[1]This dictionary is available at http://en.wiktionary.org/wiki/.

writes a post about one topic in order to preserve the coherence of the text. A reader generally expects a blogger's opinion within a post to stay consistent, unless a post expresses mixed sentiment, at which point the sentiment should be classified as mixed. Given this topic-independent approach, our sentiment classifier is well-suited to integration with a blog topic/subtopic classifier and a blog search engine. We anticipate such a search engine, which we are currently prototyping, to be useful not only for corporations interested in public-opinion information but also for the general public.

After examining related work, this paper describes our experiments in the classification of blog posts. We first describe our training data. We then discuss the features we use in the classification experiments, followed by a discussion of our results. Finally, we conclude and offer new directions for research.

Related Work

This work can be positioned within the fields of text classification and web mining, genre detection, and sentiment analysis. The majority of work on text classification is topic-related, and, to a lesser extent, event-related ((Klavans & Kan 1998), MUC conferences, etc.). Genre detection is related to sentiment analysis in that some genres, such as editorials, are inherently more subjective than others, e.g., "breaking news" reportage.

Recently there has been rapid expansion of work on sentiment analysis and online sentiment analysis in particular, in many cases for commercial purpose. Since many consumer-oriented websites include fields for user feedback, it is quite interesting to examine consumer reviews of products. In this light (Hu & Liu 2004) query online documents for positive and negative opinions of various product features, such as the size of a digital camera. The polarity of adjectives is determined using a seed list of adjectives having known orientation, and assigning unknown adjectives the orientation of their synonyms with known orientation. (Glance *et al.* 2005) examine online message boards and blogs to see what opinions consumers have on a given product. These authors determine sentiment orientation via techniques used in (Hurst & Nigam 2004), which make use of manual determination of polarized adjectives for a specific domain. They too examine digital cameras and note that the adjective *blurry* is most likely negative, while *crisp* is in all likelihood positive.

Work on sentiment analysis can be divided into approaches working at the sentence or clause level ((Wilson, Wiebe, & Hoffmann 2005), (Glance *et al.* 2005), (Hu & Liu 2004)) and those examining the document level ((Turney 2002), (Dave, Lawrence, & Pennock 2003), (Mishne 2005)). Working at the document level represents a challenge in that, for multiple sentence-documents, polarity information is propagated up from individual sentences. (Riloff & Wiebe 2003) remark that it is difficult to obtain clearly subjective or objective sentences; we find this challenge at least equally difficult at the document level.

Important aspects distinguishing our work from previous sentiment analysis studies are:

1. Performing our analysis on blogs at the post level. (Mishne 2005) takes up this task but uses different data and features;
2. The verb class information we take into account. Most experiments determining polarity restrict themselves to polar adjectives;
3. The resources we use to determine polarity. We use InfoXtract to obtain verb polarity information and Wikipedia's online dictionary, the Wiktionary, to obtain adjective polarity information;
4. A topic- and genre-independent classification. Of the document-level analyses, Turney as well as Dave *et al.* examine one genre, reviews.

Building training and test datasets

Our classifier was trained using a dataset of objective, subjective-positive, and subjective-negative documents. Training text from the web was obtained by a semi-automated, manually verified approach comprising two steps:

1. Obtain web documents via RSS and Atom web syndication feeds. A customized aggregator was developed using APIs of the Apache Software Foundation's ROME (Rss and atOM utilitiEs) project[2]. Compared to possible alternatives of web crawling or "page scraping", XML-based syndication formats of RSS (Really Simple Syndication) and Atom can capture web content in a more straightforward manner from a very large, fast-growing number of websites. The websites strategically selected for each training category provide diverse content of objective or subjective categories. Although the test dataset was manually chosen strictly from a set of blog posts of diverse topics, the training data were obtained from traditional websites as well as blogs. Many non-blog sites were prolific sources of syndicated news and topical writing that consistently met our criteria of an objective text. Objective feeds were from sites providing content such as world and national news (CNN, NPR, etc.), local news (Atlanta Journal and Constitution, Seattle Post-Intelligencer, etc.), and various sites focused on topics such as health, science, business, and technology. For subjective categories, traditional websites were also key sources of positive and negative text meeting our criteria that documents be categorically positive or negative. Subjective feeds included content from newspaper columns (Charles Krauthammer, E. J. Dionne, etc.), letters to the editor (Washington Post, Boston Globe, etc.), reviews (DvdVerdict.com, RottenTomatoes.com, etc.), and political blogs (Powerline, Huffington Post, etc.).
2. Manually verify each document written by the aggregator to confirm it strongly corresponds to an objective vs. subjective categorization. In subjective cases, a further judgment is made to verify a sufficiently positive or negative polarization. Documents not in one of these categories were discarded, and retained documents were grouped

[2] `http://rome.dev.java.net`.

	Objective	Positive	Negative	Total
Training	580	263	233	1076
Test	29	25	22	76

Table 1: A breakdown of our training and test datasets.

into the three data sets. While documents from the objective feeds were almost always verified as such, this was unsurprisingly much less the case for positive and negative documents from subjective feeds. The average rate of document retention from subjective feeds was approximately 19%, a figure which illustrates the difficulty of obtaining quality subjective data.

Given varying degrees of sentiment expression in the feeds, we decided to include only documents of very clear polarity in the training data. That is, we considered documents that are categorically positive, negative, or objective. For subjective texts, this means that the posts only express one opinion (e.g. "I was delighted today by your little plug for Futurama. . . Futurama, like early-90's Simpsons, was genius, and I love it more with every new episode I faithfully TIVO now"). For objective texts, this meant that the only goal of the author was to inform, and not to offer any kind of opinion. In so doing we obtained 263 positively oriented documents, 233 negatively oriented documents, and 580 objective documents. These proportions reflect the stringent criteria for subjective documents. Relatively similar proportions of each category were used in the test dataset. Table 1 shows a breakdown of our datasets.

It remains to be seen if blog-only training data could significantly improve classifier performance. However, obtaining a comparable amount of training exclusively from blogs without compromising our criteria for each classification would have proved inefficient: if blog posts do express overt sentiment, the sentiment is very often mixed. Our intuition is that a mix of blog and non-blog training text is beneficial to the classifier.

Sentiment agreement

Classifying posts for sentiment proved challenging for our raters, despite clear, formal criteria for such a task. These criteria concerned the author intent (to persuade? to categorically state his or her opinion on an issue? to inform?) as well as the genre of the post (review? editorial? an informational post?). Nevertheless, inter-rater agreement is not often discussed in the literature, with the exception of a series of papers by Wiebe and Wilson ((Wiebe, Wilson, & Cardie 2004), (Wilson, Wiebe, & Hoffmann 2005), etc.). That is, many studies operate on one sentence which is usually categorical in its sentiment orientation, perhaps rendering an agreement study superfluous (Hu & Liu 2004), or they automate this process by classifying according to a mood indicator on LiveJournal posts (Mishne 2005). We had also explored an automatic classification according to LiveJournal mood indicators. However, after examining LiveJournal posts and observing that these mood indicators seem at times enigmatic given the post, we felt a manually verified approach to be the judicious choice.

Our agreement study is on manual classification at the document level, where multiple sentences express sentiment. We checked the agreement of our two raters on 75 randomly chosen documents in our training data, equally distributed across the three categories. Inter-rater agreement for judging a document objective, subjective-positive, or subjective-negative was K = .77.

This K-value is slightly above that of (Wilson, Wiebe, & Hwa 2004), who originally obtain a K-value of .72. After proposing to exclude 18% of rater data marked as *uncertain* these authors obtain a K-value of .84. We asked our raters to make categorical judgments (i.e., no possibility for uncertainty, unless they wished to comment about individual posts) for two reasons: (1) we did not want to eliminate data from our training set; and (2) sometimes the *uncertain* choice becomes too easy for raters. Nevertheless, both raters did say there were documents on which they would have liked to mark their uncertainty. In all likelihood it is these documents that lower this K-value.

To ensure no differences in our test data due to inter-rater discord, we only used blog posts in our test set upon which both raters marked objective, subjective-positive, or subjective-negative.

Feature-based Classification

In this work we make use of several features of various types for our classification task. Textual features do not make use of linguistic information, while part-of-speech features require at least a tagger. A priori these two feature types contribute solely to determining subjectivity vs. objectivity and not to determining post polarity. We obtained these features with InfoXtract. Potentially polar verbs and adjective features could aid in not only a subjective vs. objective classification but also in positive vs. negative subjective classification. Verb-class information requires an InfoXtract analysis or some similar verb class information. Finally, our use of the online Wiktionary requires a substantial amount of adjectives to be listed in that resource. The Wiktionary currently provides a sufficient amount of adjectives in English, although if our approach is ever extended to other languages it is likely an alternative adjective source, along with an alternative to InfoXtract for verb class tagging, would be needed. Conceptually, however, we believe our approach could be so extended.

Textual features

Textual features without linguistic analysis have proven robust clues in automatic genre detection. (Kessler, Numberg, & Schuetze 1997) use a combination of lexical cues, such as the presence of Latinate affixes of nouns, graphemic cues, e.g. punctuation, and derivative cues, or ratios of lexical and graphemic cues. These surface-level cues proved quite effective, with over 90% of the documents correctly identified in four of the six genres studied. While textual features – other than word tokens – have not yet proven useful in the analysis of polarity, we include them for their utility in determining subjective vs. objective posts. We use the number of exclamation points and question marks: presumably, subjec-

tive posts will have more exclamation points, as this punctuation marker serves to emphasize the content expressed in a sentence. It would certainly seem odd to see exclamation marks in "breaking news" reportage. Additionally, question marks can be used for expressing irony and doubts, and barring FAQ documents, which we do not include in training or test data, they too may be more likely to be seen in subjective texts.

Part-of-speech features

Part-of-speech features have been used effectively in sentiment classification (Wilson, Wiebe, & Hoffmann 2005). Intuitively, a document with higher numbers of adjectives and adverbs is more likely to be subjective than a document with little amounts of these parts of speech: the fundamental function of adjectives and adverbs is to denote qualities of entities and events. We presume that subjective texts mention such qualities more frequently than objective objective texts, which tend more to recount events and relay information. We also posit that a document containing high amounts of first-person subject and object pronouns will prove more subjective than a document with lower amounts of these pronouns. By definition, subjectivity expresses reality from an individual's point of view, and a natural way to express one's own point of view when writing is to use a first-person perspective. We also included second-person subject and object pronouns, since our hunch was that they too would be instrumental in teasing out subjective texts. A full list of the part-of-speech features we utilize is listed in table 3.

Polarity features

To determine lexical-level polarity, we make use of two resources, the verb classes identified by InfoXtract, and the online Wikipedia dictionary. In so doing we aim to determine the overall polarity of the post, which is the polarity the blogger wants to convey. In these preliminary experiments, we assume: (1) that the distribution of polarity-expressing verbs and adjectives will show regularities across the training data for all posts within a given category, and (2) that the number of polarity-expressing verbs and adjectives oriented with the overall post polarity will outnumber those parts of speech with orientation opposite that of the blogger.

An innovative aspect of our approach is the integration of verbs and verb class information into sentiment analysis. As mentioned above, most studies of which we are aware examine only adjectives in polarity experiments. While adjectives may be a logical choice since their function is to describe entities or concepts, there is no reason why other parts of speech, such as nouns, verbs, and adverbs, could not be studied for their polarity content. Examples of these parts of speech expressing polarity include *greatness*, *mediocrity*, *like*, *despise*, *unfortunately*, *terribly*. Thus our experiments extend beyond examining polarity only expressed in adjectives.

We aim to tease out how verbs express sentiment in using the verb classes in InfoXtract. Here we do not mean "verb class" to be taken in the typical sense in NLP, which often refers to the Levin verb classes (Levin 1993) or to Levin-compatible classes. Our verb-class information is often more fine-grained and can systematically incorporate polarity into a class. For example, *approving* verbs most often show positive orientation, while *doubting* verbs tend to show negative orientation. In this way we can capture sentiment being expressed by statements like "I'm not one of those conservatives who has turned on Bush. I wholeheartedly *support* him for the same reason I did when I wrote the above posts." Verbs such as *agree*, *support*, and *appreciate* are included in the *approving* verb class. Some verb class information, such as *approving* and *doubting* verbs are included to elicit sentiment polarity, while others, such as *answering* and *suggesting* verbs, are meant to tease out subjective/objective factors. Table 3 lists the full set of verb classes we employ in this experiment. These verb classes were manually constructed and are discussed in greater detail in (Srihari *et al.* 2003) and (Li *et al.* 2003).

We use the online Wikipedia dictionary to determine the polarity of adjectives in the text. We decided to use this dictionary for its coarse-grained content. That is, adjectives in this resource are usually tersely defined, often by a list of synonyms. In brief, we find this resource similar to WordNet with its concepts of glosses and synsets, but the Wiktionary is not as specific. For example, the adjective *good* has 24 senses in WordNet and over 80 potential synonyms. In contrast, the Wiktionary has 12 senses for a total of 14 synonyms for this adjective. While (Hu & Liu 2004) yield positive results using WordNet for sentiment analysis, (Klavans & Kan 1998) note that their results are degraded by ambiguity of the synsets and that the same phenomenon occurs with nouns. Given that many of the synonyms for *good*, and other frequent adjectives, did not necessarily show positive orientation, we did not wish to overgenerate orientation where there is none. The Wiktionary thus provides an ideal tool. Although some rarer adjectives currently have no entry (*unspeakable* is currently not in the dictionary, but surprisingly, *vertiginous* is) since the dictionary is under development, we do not think this aspect adversely affects our results.

To exploit this resource, we queried the Wiktionary page for each adjective in a given post. We limited our search to the adjectival portion of an entry, further excluding the antonyms of the adjective and the example usages of the adjective. While antonym information could potentially be used in future research, we found example usages to lead too often to spurious results. We then looked for the number of adjectives in two small, manually constructed lists of known polarity in this entry. We count the number of adjectives of known polarity and assign the adjective in question the polarity with the greater number of adjectives from our manually constructed lists. If adjectives in the definition were negated (since it often the case that adjectival dictionary entries with negative orientation are defined in terms of a negated positive adjective), we consider the orientation to be the opposite of its manually established polarity. Table 2 shows an example classification of an adjective using the Wiktionary.

The adjectives in our lists were chosen for their lack of part-of-speech and polarity ambiguity. The adjective *great*,

"lucky"	
positive	negative
fortunate	–
good	–
good	–
good	–
fortunate	–
$5 > 0$	
output: positive	

Table 2: Classification of the polar adjective *lucky* using the Wiktionary. The adjectives *fortunate* and *good* figure among our list of manually constructed adjectives of positive polarity and were respectively seen two and three times in the entry for *lucky*. Five sightings of adjectives of known positive polarity outweigh zero sightings of adjectives of known negative polarity, and so the adjective is classified as positive.

for example, is also an adverb, modifying other adjectives, so it may appear in contexts not expressing positive polarity ("great fear"). Thus it cannot be used as an adjective of known positive polarity. We have initially opted for smaller lists of 34 adjectives for both positive and negative polarities. This number can be changed in order to maximize recall and precision rates of adjective polarity.

For each of these polarity features, we determine a sentiment to be negated if a verb or adjective expressing polarity is within a context window of five words of the verb or adjective (Hu & Liu 2004). However, given the complex sentences and writing styles in our training data, we do not know if taking the opposite sentiment of the would prove useful information. Thus in these first experiments we thought it wise simply to leave out of the post's sentiment classificaton any verb or adjective in a five-word window of a negation (within, of course, the same sentence).

Classifier

We use a Support Vector Machine (SVM) classifier for the classification tasks, since this technique is robust at classifying for sentiment (Mishne 2005) and it can handle noisy data well. Our LIBSVM implementation[3] gives the following binary classifications for all posts: objective/non-objective (i.e., subjective), positive/non-positive, and negative/non-negative. In our classification tasks we maximize the parameters of the SVM for maximal accuracy.

Results

We present results of our classification experiments both at the lexical level and at the post level.

Lexical-Level Results

In the 76 test files there were a total of 3,460 adjective tokens and 836 adjective types. Of these 836 types, 88, or 10.5%, were assigned a polarity by our program. We consider type accuracy of our Wiktionary method to be the number of adjectives assigned a plausible polarity for that adjective, given the Wiktionary entry for each adjective; this figure is 90.9%. In future experiments we hope to improve considerably on recall and not lose much precision by simply increasing the number of adjectives on our manually collected lists of adjectives of known polarity.

To get an estimation of the accuracy of our verb classes, we looked at two classes to see whether or not the verbs in the class do represent the a priori polarity we attributed them. We compute accuracy to be the number of verbs having a primary sense expressing the a priori polarity; for *doubting* verbs in our training data, accuracy is 25/28, or 89.3%. Accuracy on *positive mental affecting* verbs was higher at 31/34, or 91.2%. Given that these classes were not constructed with our task in mind, we are pleased with these initial results.

Post-Level Results

Classifier results on all three categories show clear improvements over baseline accuracy. Given our test data of 29 objective posts, 25 positive posts, and 22 negative posts, the baseline accuracy for each classification is guessing the majority category, which is the opposite of the category elicited. For example, in the positive/non-positive classification we assume a post is non-positive, a classification comprising both objective and negative documents, for a total of 51/76 posts, a 67.1% baseline. For the negative/non-negative classification, the baseline is 71.0% (54/76 posts), and for the objective/non-objective task this figure is 61.8% (47/76 posts).

Using all the textual and linguistic features yields a test accuracy rate of 72.4% for objective posts, 84.2% for positive posts, and 80.3% for negative posts. Every classification using our textual and linguistic features yields results that represent a substantial increase above the baseline figures. A summary of these findings, as well as precision and recall rates, is given in table 4.

To get an idea of which individual features were contributing to the accuracy of our classifier, we next effectuated experiments in holding out each feature. Holding out features such as the positive adjectives we obtained from the Wiktionary typically decreases accuracy, which entails that this feature plays a large role in the correct classification of posts. Holding out some features can actually increase accuracy, which would imply that these features have a detrimental effect on proper classification. However, all these increases are small, and currently we cannot say if they are significant. The effects of holding out individual features are given in table 5.

Discussion

The preliminary results in table 4 are encouraging. In every classification, our system yields results that are well over the baseline accuracy for that task. Our system scores both the best accuracy and F-measure on the positive sentiment classification task. This is despite a high precision on the negative sentiment classification task: negative sentiment recall was the lowest of all three classification tasks. Only nine

[3] `http:www.csie.ntu.edu.tw/cjlin/libsvm`.

Subjective vs. Objective Features			Polarity Features
Objective Verb Classes answering, asking, asserting, explaining **Subjective Verb Classes** believing suggesting mental sensing	**Textual Features** exclamation points question marks	**Parts of speech** first-person pronouns second-person pronouns # of adjectives # of adverbs	**Positive Verb Classes** positive mental affecting, approving, praising **Negative Verb Classes** abusing, accusing, arguing, doubting, negative mental affecting **Adjectives (Wiktionary)** positive adjectives negative adjectives

Table 3: Features used in the sentiment classification task.

	Classifications								
	Objective			Positive			Negative		
Baseline Accuracy	61.8%			67.1%			71.1%		
Accuracy	72.4%			84.2%			80.3%		
	Prec	Rec	F-meas	Prec	Rec	F-meas	Prec	Rec	F-meas
All features	68.2%	51.7%	59.2%	84.2%	64%	72.7%	88.9%	36.4%	51.7%

Table 4: Results of classification experiments.

Feature group	Held-out feature	Objective	Positive	Negative
All features	–	**72.4%**	**84.2%**	**80.3%**
Textual Features	exclamation points	69.7%	82.9%	80.3%
	question marks	72.4%	84.2%	80.3%
Parts of Speech	first-person pronouns	72.4%	85.5%	80.3%
	second-person pronouns	72.4%	85.5%	80.3%
	# of adjectives	72.4%	84.2%	80.3%
	# of adverbs	72.4%	84.2%	80.3%
Objective Verb Classes	answering	72.4%	84.2%	80.3%
	asking	72.4%	84.2%	80.3%
	asserting	72.4%	**76.3%**	80.3%
	explaining	76.3%	86.8%	80.3%
Subjective Verb Classes	believing	72.4%	84.2%	80.3%
	suggesting	76.3%	85.5%	80.3%
	mental sensing	72.4%	84.2%	80.3%
Positive Verb Classes	positive mental affecting	75%	84.2%	80.3%
	approving	73.7%	**77.6%**	80.3%
	praising	73.7%	84.2%	81.6%
Negative Verb Classes	abusing	73.7%	84.2 %	80.3%
	accusing	72.4%	84.2%	80.3%
	arguing	72.4%	81.6%	80.3%
	doubting	71.1%	84.2%	80.3%
	negative mental affecting	71.1%	82.9%	81.6%
Adjectives	positive adjectives	**61.8%**	**67.2%**	80.3%
	negative adjectives	75%	82.8%	78.9%

Table 5: Effects of holding out an individual feature on accuracy, against accuracy for all features.

documents were classified as negative, eight of which were in fact negative.

The hold-out experiments show initial results of what features prove influential in a given classification. Currently we have not tested these results for significance: thus we cannot notice if the features that change only slightly will be consistent over other test datasets. There are, however, some features that seem to affect greatly the objective and positive classifications; the figures for these features on a given classification task are given in bold. For objective and positive posts, positive adjectives acquired from Wikipedia's Wictionary seem to play a key role in increasing overall accuracy. Thus our strategy for using an online dictionary in a sentiment classification task appears to have been successful. Also, for the positive classification two verb classes appear to play a key role in improving accuracy, namely, the asserting and approving verb classes. These results, if consistently duplicated, would show that verb classes can improve results on sentiment classification of blogs. However, it is important to note that these preliminary results, with little change if any several features are individually held out, would seem to indicate that we can reduce our feature space for the classification task. For example, we could probably reduce significantly the number of verb classes we use to classify posts.

Since precision for the negative documents is high (there is only one misclassified post) and recall is low, there is little room in the test data for seeing changes in this classification. We do however remark that the one hold-out that decreases negative accuracy is the amount of negative adjectives we obtained from the Wiktionary. Hence it is possible that this feature plays a significant role in correctly classifying negative documents.

Given the high accuracy rates for verb and adjective accuracies, the principal challenge we must tackle is propagating the lexical results up to the post level. In examining the adjectives in the test data, a small percentage of adjectives with manifest polarity are classified as such by our system: i.e., while adjective accuracy is high, the corresponding recall rates are low. Hopefully in improving recall rates at the lexical level we will improve accuracy and precision at the post level.

When compared with the tasks of other studies on sentiment analysis, our results are promising. For example, (Turney 2002) reviews movies and sees an average improvement of approximately 15% over baseline accuracy in a binary classification ("thumbs up" or "thumbs down"). Only our positive accuracy is over 15% higher than our baseline figure. Nevertheless, our results are topic- and genre-independent. It is perhaps more difficult to interpret sentiment expressed in letters to the editor, newspaper columnists' writings, and political blogs, all of which our classifier was trained on, than movie reviews.

Conclusions and Future Work

This paper presents initial experiments in classifying blog posts according to sentiment using verb classes and an online dictionary for determining adjective polarity. In so doing we obtain results that are well above the baseline accuracy rates for such tasks. Since we make use of new resources to determine polarity at the post level, we uncover new challenges, such as how to classify a post as having a certain orientation once we have determined specific lexical items in sentences to have a given orientation. Our approach is currently topic- and genre-independent, and as such we next plan to integrate our research into a topic-classifying search engine for blogs. In addition to finding predominantly objective vs. subjective blogs, this search engine will be able to selectively display posts that only express positive or negative opinions about a topic.

A remaining question for us, as well as for others working on sentiment analysis, is that of what features to use in the classification task to have the most accurate classification. In this work we present make use of verb classes, hitherto neglected, in studies on sentiment analysis. We also query an online dictionary, Wikipedia's Wiktionary, for orientation of adjectives in a post. Our Wiktionary method is quite accurate, and we predict the accuracy can be further improved, first as the HTML code of the dictionary becomes standardized, and secondly by taking into account other ways to express sentiment, such as "lacking X", where X is a polar noun. We also plan to extend our adjective lists of known polarity so as to improve recall. In subsequent research we plan to include word tokens as a textual feature for classification, since this information has proven useful to sentiment classification tasks (Wilson, Wiebe, & Hoffmann 2005). We will also investigate effective ways of incorporating negations of sentiment into sentiment classification. Our current strategy is to disregard the possible negation of sentiment, and there is no doubt a more effective way to make use of this additional sentiment expression.

We would also like to improve our methods of obtaining quality blog training data. While collecting our training sets, we discovered additional blog sources that are good candidates to provide larger amounts of topic-diverse objective and polarized-subjective posts meeting our strict criteria. We plan to incorporate them into future research and assess the impacts of blog-only training data to classifier performance, since high amounts of quality training and test data will lead us to infer significant improvements in the classification of sentiment due to a given linguistic or textual feature.

Additionally, we would like to examine in greater detail the rhetorical structure and linguistic features of posts expressing sentiment. Not only is this endeavor theoretically interesting, it will most likely yield increased precision. We stated in the introduction that blogs have a more heterogenous way of expressing opinions than, for example, an oped piece. That said, we should not assume that there is no regularity in the rhetorical structure bloggers employ when expressing sentiment. A principal question in this endeavor is that of *where* in the blog the most sentiment-rich text can be found. Would it be beneficial to examine only the first and/or last paragraphs for sentiment? We also aim to better tease out more linguistic clues with respect to sentiment, as (Hatzivassiloglou & McKeown 1997) show robust results in predicting the orientation of adjectives joined by conjunctions. In examining our data we observe that sentence-

initial clauses beginning with *despite* and *although* express a given polarity in the subordinating clause and its complement polarity in the main clause. We believe linguistic cues such as the examination of subordinating clauses and rhetorical structure analysis could marry well with information retrieval techniques, such as using strictly lexical information, for determining the sentiment of blogs.

Acknowledgments

We wish to thank Anmol Bhasin, Gaurav Chandalia, Mohit Devnani, and Shakthi Poornima for their valuable insights and help on this research.

References

Dave, K.; Lawrence, S.; and Pennock, D. M. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, 519–528. New York, NY, USA: ACM Press.

Glance, N.; Hurst, M.; Nigam, K.; Siegler, M.; Stockton, R.; and Tomokiyo, T. 2005. Deriving marketing intelligence from online discussion. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 419–428. New York, NY, USA: ACM Press.

Hatzivassiloglou, V., and McKeown, K. R. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, 174–181. Morristown, NJ, USA: Association for Computational Linguistics.

Hu, M., and Liu, B. 2004. Mining and Summarizing Customer Reviews. 168–177. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

Hurst, M., and Nigam, K. 2004. Retrieving Topical Sentiments from Online Document Collections. In *Document Recognition and Retrieval XI*, 27–34.

Kessler, B.; Numberg, G.; and Schuetze, H. 1997. Automatic detection of text genre. 32–38. Proceedings of the Thirty-fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics.

Klavans, J., and Kan, M. 1998. Role of Verbs in Document Analysis. 680–686. Proceedings of the Conference COLING-ACL, Montreal.

Levin, B. 1993. *English Verb Classes and Alternations*. U. Chicago Press.

Li, W.; Zhang, X.; Niu, C.; Jiang, Y.; and Srihari, R. 2003. An Expert Lexicon Approach to Identifying English Phrasal Verbs. 513–520. Proceedings of ACL 2003.

Mishne, G. 2005. Experiments with Mood Classification in Blog Posts. Stylistic Analysis Of Text For Information Access Workshop at SIGIR 2005.

Riloff, E., and Wiebe, J. 2003. Learning Extraction Patterns for Subjective Expressions. 105–112. Conference on Empirical Methods in Natural Language Processing (EMNLP-03). ACL SIGDAT.

Sawney, M. 2005. The HBR List: Breakthrough Ideas for 2005, Blog-trolling in the Bitstream. *Harvard Business Review* 83:2:25–31.

Srihari, R.; Li, W.; Niu, C.; and Cornell, T. 2003. InfoXtract: A Customizable Intermediate Level Information Extraction Engine. In *Proceedings of the HLT-NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology Systems*.

Turney, P. D. 2002. Thumbs up or thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics*, 417 – 424.

Wiebe, J.; Wilson, T.; and Cardie, C. 2004. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 1:2.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. HLT-EMNLP.

Wilson, T.; Wiebe, J.; and Hwa, R. 2004. Just how mad are you? Finding strong and weak opinion clauses. Proceedings of the 19th National Conference on Artificial Intelligence.