Coarse- and Fine-Grained Sentiment Analysis of Social Media Text

Clayton R. Fink, Danielle S. Chou, Jonathon J. Kopecky, and Ashley J. Llorens

Sentiment analysis—the automated extraction of expressions of positive or negative attitudes from text—has received considerable attention from researchers during the past 10 years. During the same period, the widespread growth of social media has resulted in an explosion of publicly available, user-generated text on the World Wide Web. These data can potentially be utilized to provide real-time insights into the aggregated sentiments of people. The tools provided by statistical natural language processing and machine learning, along with exciting new scalable approaches to working with large volumes of text, make it possible to begin extracting sentiments from the web. We discuss some of the challenges of sentiment extraction and some of the approaches employed to address these challenges. In particular, we describe work we have done to annotate sentiment in blogs at the levels of sentences and subsentences (clauses); to classify subjectivity at the level of sentences; and to identify the targets, or topics, of sentiment at the level of clauses.

INTRODUCTION

People make judgments about the world around them. They harbor positive and negative attitudes about people, organizations, places, events, and ideas. We regard these types of attitudes as sentiments. Sentiments are private states,[1] cognitive phenomena that are not directly observable by others. However, expressions of sentiment can be manifested in actions, including written and spoken language. Sentiment analysis is the study of automated techniques for extracting sentiment from written language. This has been a very active area of research in the computational linguistics community over the past 10 years.

The past 10 years have also seen a rapid increase in the use of the World Wide Web as a forum where people share their opinions and the details of their lives. Web logs (known as blogs), online forums, comment sections on media sites, and social networking sites such as Facebook and Twitter all fall under the heading of social media and, via user-generated text, capture millions of people's points of view. A vast amount of these data are

public, and their availability is fueling a revolution in computational linguistics and social network analysis. These social media data provide a potential source of real-time opinion from people around the world. Timely, aggregated data on people's opinions can be of great value to policymakers, social scientists, and businesses.

In this article we discuss work we have done in identifying expressions of sentiment in text extracted from social media. In the first section, we give an operational definition of sentiment, discuss related work, and describe examples of coarse-grained and fine-grained sentiment. In the second section, we describe annotation studies we have performed on blog posts; these studies focus on annotating subjectivity and sentiment at the sentence level as well as identifying the targets of sentiment at the clausal level. In the third section, we discuss our development of pattern recognition algorithms to classify the subjectivity of sentences as well as the targets of sentiment within sentences. In the final section, we draw conclusions and discuss future directions for our work.

SENTIMENT ANALYSIS

We define sentiment as a positive or negative attitude, held by a person or a group of people, that is directed at some thing. Things in this case include entities (people, groups, organizations, or geographic locations), events, actions involving entities, and ideas or concepts. By this definition, a sentiment has a polarity or valence (it is positive or negative), a source (the person or group of people holding the sentiment), and a target (the thing toward which the sentiment is directed). Automated sentiment analysis tasks are concerned with detecting the presence of sentiment in a unit of text and identifying the valence, source, and target of that sentiment.

In text, sentiments can be captured at various levels of granularity: at the level of the document, paragraph, sentence, or clause. Regardless of the level at which sentiment is captured, multiple sentiments directed at the same or different targets can reside in a single sample. At each level of granularity, different components of sentiment (valence, source, target, etc.) hold, and different techniques can be used for identifying sentiment.

The primary focus of early research in the field was to classify entire documents as containing overall positive or negative sentiment. This research direction was driven by the availability of convenient samples, such as movie or product reviews, in which the writer indicates whether the text is positive or negative toward the target of the review. The following excerpts from reviews of the film *A Serious Man* provide an illustration:

> "Time and time again, the wily filmmakers sprinkle the overarching storyline of the fall and decline of Larry Gopnik's life (a masterful, wide-ranging and sensitive performance from Michael Stuhlbarg) with a fine combination of overt, discreet and subliminal set-ups whose payoffs give their film extra punch and an unstoppable pace."[2]

> "*A Serious Man* is a truly despicable film, and I ordinarily count myself among the Coen brothers' fans and/or defenders. So I was astonished that with this film, in one fell stroke, they had me believing that everything their detractors say might just be right."[3]

The first review[2] is clearly positive, with positive subjective words such as "masterful" and phrases such as "extra punch" and "unstoppable pace" supporting this conclusion. The second review[3] is clearly negative. Words such as "despicable" and the phrase "they had me believing that everything their detractors say might just be right" tip us off that the author did not like the film.

Sentiment does not just occur at the whole-document level, however; nor is it limited to a single valence or target. Contrary or complementary sentiments toward the same topic or multiple topics can be captured across sentences or within the same sentence. The following three sentences, for example, are from a single blog post[4]:

> "In the post I wrote yesterday after the Kagan announcement, I noted one genuinely encouraging aspect of her record: in 1995, she rightly excoriated the Supreme Court confirmation process as a 'vapid and hollow charade' because nominees refuse to answer any meaningful questions about what they think or believe.

> "But during a briefing with reporters in the White House, Ron Klain, a top legal adviser to Vice President Joe Biden who played a key role in helping President Obama choose Kagan, said that she no longer holds this opinion.

> "Does anyone, anywhere, believe that her 'reversal' is motivated by anything other than a desire to avoid adhering to the standards she tried to impose on others?"

These sentences are, respectively, positive, objective, and negative references to Elena Kagan. Because discourse tends to contain expressions of contrasting sentiment along with objective descriptions, sentiment analysis at the subdocument level is concerned with distinguishing sentiment-containing segments of text from nonsentimental segments. Kim and Hovy[5] have suggested some approaches to this problem of detecting subjective sentences on the basis of the presence of subjective words or phrases, and Pang and Lee[6] described a graph-based technique for segmenting sections of a document on the basis of their subjectivity.

A single sentence may contain contrasting sentiments toward the same target, as well as multiple sources and multiple targets of sentiment. The following are two readings of the same sentence. In these examples, the sources of sentiment are shown in italics, and the targets of sentiment are in boldface:

> "When *Hillary's fans* complained about **the incredible amount of sexism on the Left**, I took it with a grain of salt."

> "**When Hillary's fans complained about the incredible amount of sexism on the Left**, *I* took **it** with a grain of salt."

The first reading references "Hillary's fans" as the source of negative sentiment toward "the incredible amount of sexism on the Left." The second reading captures the author's dubious attitude—based on his use of the phrase "with a grain of salt"—toward the sentiment captured in the first reading, with "I" referring to the author as the source of sentiment and "When Hillary's fans complained about the incredible amount of sexism on the Left" and "it" referring to the attitude of Senator Clinton's supporters. This example illustrates the difficulty of identifying fine-grained sentiment. Sentences that contain multiple readings and nested sentiments are common, so sophisticated techniques to detect the components of sentiment are required. Wiebe et al.[7] introduced the notion of subjective expressive frames, which capture different readings of the sentiment content in a sentence, and they and others have investigated methods for identifying them in text. Choi et al.[8] have investigated identifying sources of sentiment in text by using sequential machine learning. Kim and Hovy[9] have investigated using semantic frames defined in FrameNet[10] for identifying the topics (or targets) of sentiment, and Wilson[11] has described methods for identifying the attributions (or sources) of sentiment. Circling back to online reviews, Kessler and Nicolov[12] investigated using a ranking support vector machine (SVM) for identifying the targets of sentiment that correspond to specific features of the product being reviewed.

ANNOTATING FINE-GRAINED AND COARSE-GRAINED SENTIMENT

Human annotations of text are an essential part of statistical natural language processing, both as ground truth data for measuring the accuracy of classification algorithms and as training data for supervised machine learning. A number of researchers have investigated the task of annotating text for sentiment,[7,11–13] including attempts to annotate the targets of sentiment. Our research focuses on identifying sentiment at a granularity below the document level. In particular, we are interested in identifying subjective, sentiment-bearing sentences and then identifying the sentiment targets in these sentences, along with the valence of the expressed sentiments. The annotation tasks we have carried out are intended to support this workflow and provide a source of training data for subjectivity and sentiment classifiers.

We have concentrated on annotating individual sentences that were randomly selected from blog posts. Prior research on blog text has focused on annotating entire posts that match queries. For example, annotations are applied to blog posts that contain the words "McCain" or "Obama,"[13] and sentiments toward those entities are identified. Our rationale for annotating sentences selected at random was to give us a wide sample of sentiment as it is expressed across large collections of blogs. Also, our approach was to base classification on intrasentential (i.e., within-sentence) features, rather than on intersentential or document-level information, in order to focus on a more bottom-up approach to classification. Hence, we only needed to annotate sentences in isolation.

We annotated blogs from three separate blogging communities: politically oriented blogs from the United States, blogs by knitting enthusiasts, and blogs by tango enthusiasts. Political blogs are common worldwide and provide a forum for opinion and citizen journalism. These blogs provide a source of sometimes extremely polar opinion, much of it negative. Knitters have taken to the blogosphere with enthusiasm[14] and are an example of a cohesive online community that is dedicated to a specific topic. This community, in contrast to the political bloggers, tends to be positive. Similarly, tango-related blogs represent another online community with a specific concentration and, as with all the online communities we have studied, unique vocabulary and modes of expressing sentiment.

Blogs from each community were identified manually, and posts from each blog were retrieved using the Google Reader application programming interface (API) and the really simple syndication (RSS) feed of the blog. Blog comments were ignored for our study. Sentences were randomly selected from posts, and collections of sentences were provided to annotators. Guidelines were developed with input from a team of annotators to provide us with a consistent set of rules for annotating sentences. For this type of annotation task, typically a common set of sentences is embedded in the sets distributed to each of the annotators, and this common set is used later for measuring agreement across annotators.

The annotation task for each sentence was broken up into three phases, and we tracked interannotator agreement for each phase. The phases correspond to a workflow for sentiment identification (Fig. 1) that relies on a cascade of classifiers, each trained to identify sentiment at finer and finer levels of granularity. Each annotation step, then, is intended to provide training and test data for a particular classifier.

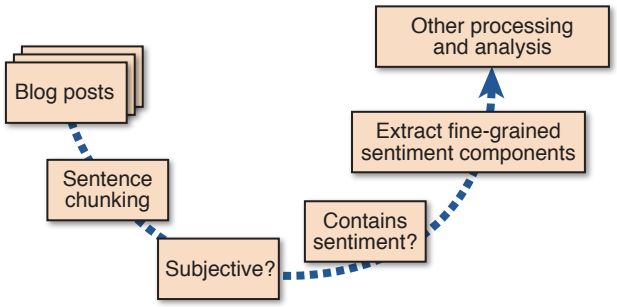

Figure 1. Notional sentiment workflow.

In the first phase, the annotator determined whether the sentence was subjective. For this determination, the annotator had to decide whether the sentence contained only verifiable statements of fact. For example, in the following sentence, the projection of McCain as the winner of the primary can be verified by looking directly at the source of the reported information:

1. "McCain is projected to be the winner of the Missouri Republican primary."

The annotator would not annotate this sentence at all and would move on to the next sentence in his or her annotation set. Another sentence, however, might contain speculations by the author that are not directly verifiable and represent a private state. For example, the following sentence contains the author's evaluation of some situation, but the truth of the assertion is unknown.

2. "At the end of the day, there'll always be disagreements, though."

For this sentence, the annotator would mark the sentence as subjective (Fig. 2) and move on to the second phase of annotating the sentence.

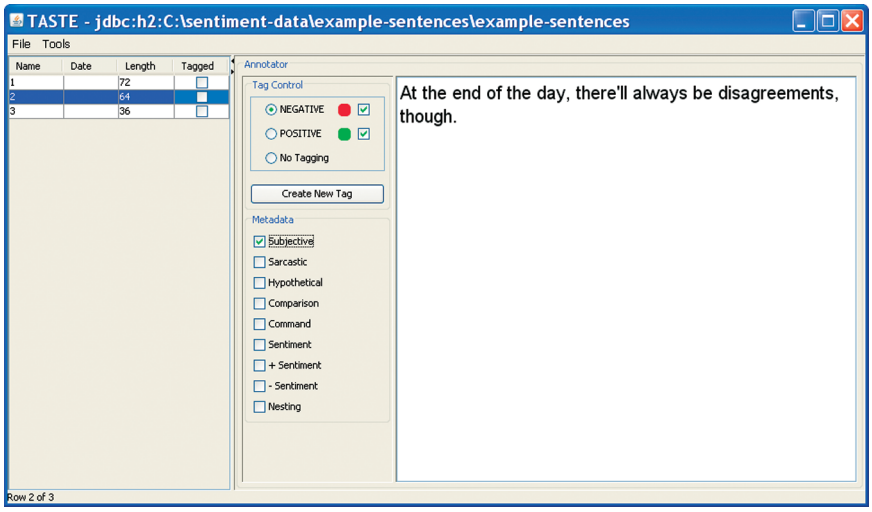

Figure 2. Sentence annotated as subjective.

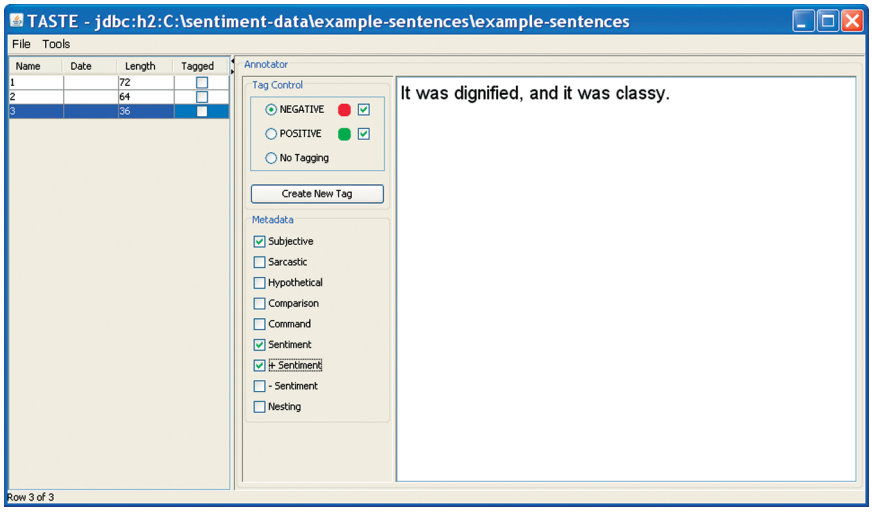

Figure 3. Sentence annotated as containing positive sentiment.

The second phase of annotation is concerned with deciding whether a sentence contains an expression of sentiment. Not all subjective sentences contain sentiment. In sentence 2, it is not clear whether the expression is negative or positive, nor is there a clear target of sentiment. However, in the following sentence, the author is clearly expressing a positive evaluation of "it":

3. "It was dignified, and it was classy."

In this case, the annotator would annotate this sentence as having sentiment and, in particular, positive sentiment (Fig. 3). In our study, the coarse-grained annotations for a sentence comprise subjectivity and sentiment.

The last phase of annotation involves identifying the targets of sentiment in a sentence. This identification process produces the fine-grained annotations for a sentence. In the case of sentence 3, the annotator would have annotated both instances of the word "it" in the sentence (Fig. 4).

We looked at agreement among three annotators for 300 sentences from political blogs, 300 from knitting blogs, and 200 from tango blogs. We examined interannotator agreement at both the coarse- and the fine-grained levels. Sentences were annotated at the fine-grained level only if the annotator believed the sentence had sentiment at the coarse-grained level.

To compute interannotator agreement for coarse-grained features, we used Krippendorff's α,[15] an agreement measure commonly used in the natural language processing community. Compared with other measures of agreement, such as pairwise percent agreement, Krippendorff's α gives a clearer definition of reliability and is informed by the base rate of each annotation category. Whereas 0% agreement implies annotators never agree, it also implies annotators went out of their way to disagree; hence, the equivalent α would be –1.0 (perfect negative correlation between annotators). This negative

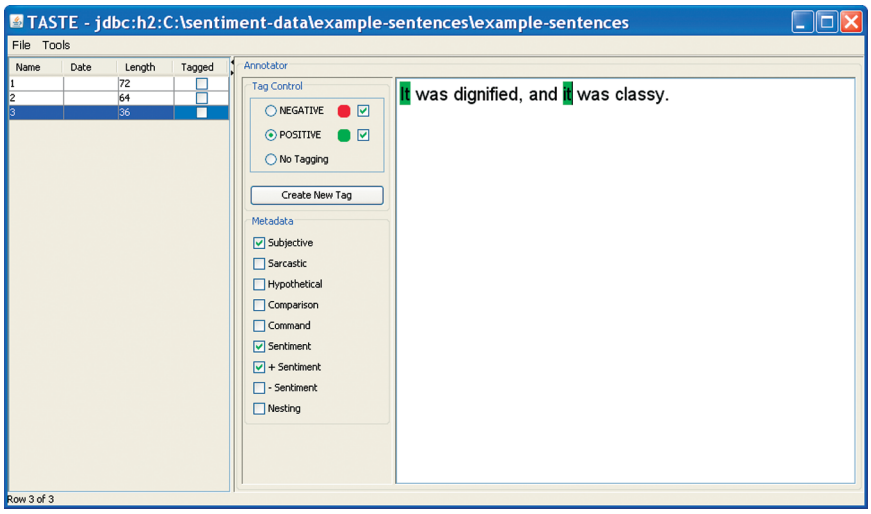

Figure 4. Sentence with annotated sentiment targets.

α hints that there is a problem with the instructions or that an annotator is deliberately performing the task incorrectly. The more positive α is, the more reliable the annotators are with respect to one another. Generally, α greater than 0.8 indicates superb agreement, whereas α between 0.667 and 0.8 indicates good agreement. Krippendorff's α also accounts for the base rate of each category, so if two annotators mark one particular category 90% of the time, a disagreement in annotating that category has a greater impact on the score than if the two annotators mark the most common category only 60% of the time.

For sentence-level annotations, the calculation of Krippendorff's α was based on an ordinal labeling: an unannotated sentence (i.e., the sentence was judged nonsubjective) was scored as a 0, a sentence annotated as subjective was scored as a 1, and a sentence annotated as sentimental was scored as a 2. For the results from the 800 sentences across all corpora, the overall three-way interannotator agreement at the coarse-grained level was good, with an overall Krippendorff's α of 0.768. However, there were large differences across domains. Sentences from the tango blogs had the highest agreement, with interannotator agreement of 0.852, followed by sentences from knitting blogs at 0.779, and then sentences from political blogs at 0.685. Politics is by far the toughest domain in which to achieve interannotator agreement, even when it comes to just agreeing about subjectivity!

We also calculated Krippendorff's α for our annotation of sentiment targets. This was done three times: for all sentences, for sentences where all annotators agreed at the sentence level on subjectivity, and then for sentences where all annotators agreed on both subjectivity and sentiment. To calculate the α for each set of sentences, we created a list of all words in the annotation set, sentence by sentence, and calculated the α on the basis of a nominal labeling of each word: a nontarget, a positive target, or a negative target.

Additionally, at the fine-grained level, we also looked at the percent agreement of annotations across annotators despite the problems with percent agreement mentioned above. There were a few reasons for this. For one, extant work in the literature about fine-grained sentiment analysis used this measure, so we wanted a comparable measure. Another reason is that, at the fine-grained level, annotators only mark the target of sentiment and not the targets of subjectivity. Thus, with fewer tags to compare, Krippendorff's α will likely underweigh actual interannotator agreement. Nonetheless, for completeness, we calculated Krippendorff's α at the fine-grained level as well.

The interannotator agreement results for sentiment targets are shown in Table 1. Controlling for sentiment (i.e., looking at only those sentences that annotators said were both subjective and had sentiment) resulted in the highest agreement for annotating targets: 84.17% for knitting (α = 0.672), 79.03% for tango (α = 0.706), and 76.97% for the political corpus (α = 0.701). These numbers can be construed as a measure of how complicated the sentences were in the various corpora because the lower the agreement, the harder it is to determine the target of sentiment. Overall agreement was also highest for knitting, at 74.67% (α = 0.618), compared with 66.56% for tango (α = 0.580) and 61.53% (α = 0.535) for politics.

SENTIMENT CLASSIFICATION STUDIES

Because the huge volume of text available in social media leads to extremely large datasets, we envision an

Table 1. Interannotator agreement results for blog sentiment annotations.

Statistics per Level of Agreement	Political Blogs	Knitting Blogs	Tango Blogs
Sentence-level agreement			
Krippendorff's α[15]	0.685	0.779	0.852
Target-level agreement (controlling for sentiment)			
Krippendorff's α	0.701	0.672	0.706
Percent agreement[a]	76.97	84.17	79.03
Overall target-level agreement			
Krippendorff's α	0.535	0.618	0.580
Percent agreement[a]	61.53	74.67	66.56

[a]Average of pairwise agreement.

automated classification process in which computationally expensive fine-grained sentiment analysis is preceded by inexpensive automated screeners to help ensure scalability. Specifically, we envision a three-stage cascade of binary classifiers that mimics the manual annotation process previously described (Fig. 1). The first-stage classifier separates the subjective sentences from the objective sentences in the dataset. The sentences identified as subjective are then passed to the second-stage classifier, which identifies the subset of those sentences that contain an expression of sentiment. In the final stage, sentences are broken into clauses, and in each clause, the target and valence of the sentiment are identified. The first two classification stages would essentially be conducting automated coarse-grained analysis preceding the automated fine-grained analysis conducted in the final stage. Although the three-stage cascaded system has not yet been implemented, our preliminary results explore the effectiveness of each stage individually. Once sentiment targets are extracted, entity and topic coreference processing, both within a single document and across different documents, will need to be applied to aggregate sentiment. For this study we focused exclusively on sentiment extraction and will focus on the aggregation of sentiment in later work.

Two different classifier types, Naive Bayes (NB)[16] and Least-Squares SVM (LS-SVM)[17] with a linear kernel, were trained and evaluated against a set of objective sentences from the Internet Movie Database and a set of subjective sentences from the movie review site Rotten Tomatoes (see Pang et al.[18]) as well as sentences from the political blog corpus that were annotated as being subjective or not annotated and, hence, objective. Versions of both classifier types are used widely in the statistical natural language processing literature. In our implementation, the NB classifier models each feature dimension with a multinomial distribution. This is well suited to text processing because the feature sets are typically large and general across various corpora and, hence, sparsely supported in most individual datasets (for the movie review dataset, we used approximately 7500 features; for the political blog dataset, we used approximately 1400 features). The NB classifier inherently assumes each feature dimension is independent, whereas the LS-SVM classifier, in contrast, does not rely on the assumption of independence. However, because features in the text classification domain appear to exhibit relatively weak correlations, the NB classifier has been shown to perform well despite the assumption of independence. The LS-SVM has the capacity to create a highly complex decision boundary and hence can almost always perform well on the training data, but it runs a greater risk of overfitting if the model parameters are not chosen carefully.

To obtain the reported results, we developed classifiers to separate subjective from objective sentences. As in any complex classification problem, the science of training the classifiers was preceded by the development of an effective feature set, which can be more of an art. The feature-development process involved defining a numerical representation of the input sentences on the basis of characteristics that are, one hopes, germane to one data class to the exclusion of the other, providing separability that can be exploited by the classifier. Our feature set included the presence of weakly or strongly subjective words based on a standard subjective lexicon[19]: unigrams (single words) in the sentence, bigrams (ordered pairs of words) in the sentence, number of adjectives in the sentence, number of words in the sentence, whether certain types of punctuation occurred in the sentence (i.e., exclamation mark, quotation mark), and whether the sentence has any word written in all capital letters (which is often an indication of sarcasm). All features used were binary, meaning that a given feature was used for an instance if and only if it held true for that instance. To pare down the feature set, we considered only unigrams and bigrams that occurred across the dataset more than five times. Additionally, we tried both stemmed and nonstemmed versions of the unigrams and bigrams, where a stemmed version of a word refers to retrieving the root of the word

Table 2. Sentence subjectivity classification results.

		Movies			U.S. Politics		
Classifier	Feature Set	Accuracy (mean ± SD)	Precision (mean ± SD)	Recall (mean ± SD)	Accuracy (mean ± SD)	Precision (mean ± SD)	Recall (mean ± SD)
Baseline	Two or more weak subjective words or one or more strong subjective words	56.2	54.2	78.5	63.7	78.9	52.4
NB	No stemming	89.6 ± 0.3	89.1 ± 0.4	90.3 ± 0.5	65.4 ± 1.8	68.5 ± 1.2	76.4 ± 2.1
	Stemmed	88.7 ± 0.2	88.5 ± 0.3	89.0 ± 0.4	64.3 ± 1.4	67.0 ± 1.4	77.7 ± 2.3
LS-SVM	No stemming	83.4 ± 0.3	83.8 ± 0.6	82.7 ± 0.7	66.0 ± 1.7	74.6 ± 2.2	64.3 ± 2.1
	Stemmed	81.2 ± 0.4	82.0 ± 0.6	81.2 ± 0.4	67.4 ± 1.4	74.0 ± 1.4	69.0 ± 2.3

SD, Standard deviation.

minus any suffixes that can affect having exact matches.

The results of this study are shown in Table 2. The mean and standard deviation of the various performance metrics were generated by evaluating the classifiers on 10 random training and test splits of the data. In the case of the movie review data, each classifier type consistently beat the results of a baseline classifier that marked sentences as subjective if they contained one or more strongly subjective words or two or more weakly subjective words. However, for the political blogs, the classifiers did not fare as well when compared with the same baseline. Both classifiers produced an accuracy in line with the baseline but did consistently more poorly on precision. Recall for the political blog data, however, was consistently better for both classifiers over the baseline. The political blogs may be more difficult because of the use of sarcasm (e.g., "A novel with a shelf life of yogurt") and subjective words that are rare and suggest sentiment only within the political domain (e.g., "flip-flopping" and "whopper"). The subjective word lexicon we used was general and not specific to the political domain, which may partially account for the poor performance. In future studies, we will look at developing a specialized feature set, including the use of domain-specific subjective words, that may be more appropriate for the data from political blogs.

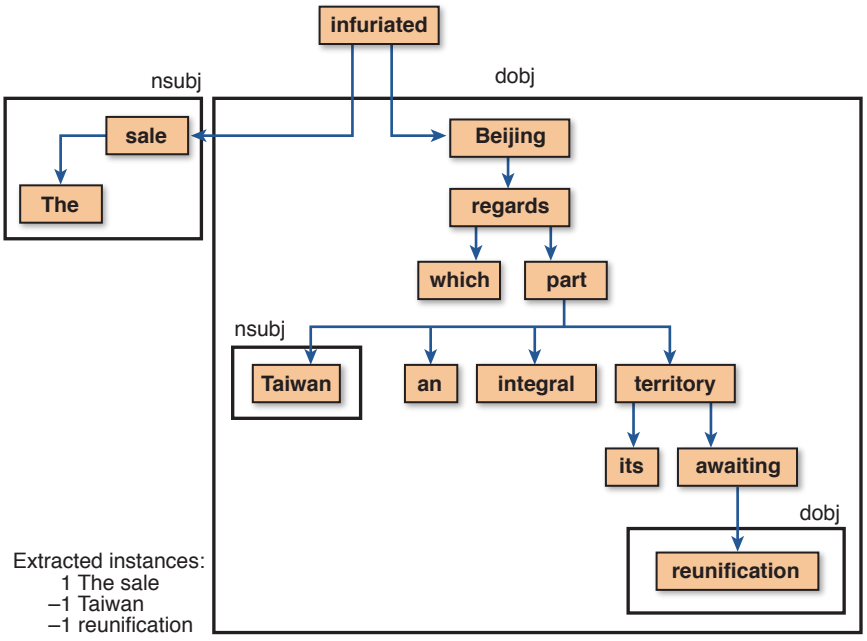

Figure 5. Example of a dependency parse of the sentence "The sale infuriated Beijing which regards Taiwan an integral part of its territory awaiting reunification, by force if necessary" with chunking results. dobj, direct object; nsubj, nominal subject.

We next explored identifying the targets of sentiment in sentences. For this study, we used the attitude annotations from version 2 of the Multi-Perspective Question Answering (MPQA 2) corpus.[19] These annotations of news stories and editorials from print media are all concerned with a number of predetermined topics. Expressions of attitudes—including sentiments—contained in a document's text were annotated, along with any targets of the attitude. The attitudes were also annotated for valence and valence strength. For our work, we looked only at attitudes of positive and negative sentiment with strengths of medium or higher.

Sentiment targets include nouns, verbs, prepositional phrases, or any number of nested grammatical components of a sentence. For this work, we focused in particular on phrases that serve as verb arguments (i.e., the objects, subjects, or complements of verbs). The classifiers then decided whether a particular verb argument was or was not a target of sentiment. For the current work, the valence of the target was not considered.

We used a dependency parse[20] of each sentence to extract the verb arguments from each sentence. A dependency parse is a directed graph containing the words of a sentence as vertices, with the edges capturing the grammatical dependencies between the words. Each dependency relationship has a governor and a dependent. For example, a nominal subject dependency has the verb as the governor and the verb subject as the dependent. We used the Stanford dependency parser[21] to generate dependency parses for each sentence from the MPQA 2 corpus that was annotated with sentiment attitudes and had targets. The Stanford parser optionally generates a tree structure that is a compressed representation of the dependencies; this facilitates the identification of contiguous sentence segments that fall under a given dependency relationship. To determine verb arguments, we looked for verb-specific grammatical dependencies such as nominal subject or direct object and then extracted the complete subtree with the dependent (target) as its root. For example, the sentence "The sale infuriated Beijing which regards Taiwan an integral part of its territory awaiting reunification, by force if necessary" has three verbs: "infuriated," "regards," and "awaiting." The dependency parse of the sentence is shown in Fig. 5. The verb arguments are "The sale," "Beijing regards . . .," "Taiwan," and "reunification." Given this segmentation of the sentence, we drop verb arguments that contain any nesting (in this case, "Beijing regards . . .") and use the remaining arguments as training instances. The extracted verb arguments are then aligned with the annotations. An extracted verb argument that over-

Table 3. Sentiment target classification results.

Classifier	Accuracy (mean ± SD)	Precision (mean ± SD)	Recall (mean ± SD)
Baseline (phrase contains or is preceded or followed by a subjective word)	0.63	0.37	0.44
NB[16]	0.5873 ± 0.014159	0.60582 ± 0.016567	0.50102 ± 0.044913
LS-SVM[17]	0.61159 ± 0.014934	0.62189 ± 0.012633	0.5687 ± 0.037732

SD, Standard deviation.

laps with text annotated as a sentiment target is considered a positive instance of a sentiment target. Those arguments that overlap no annotations are considered negative instances. Figure 5 shows the class assignments for the verb arguments extracted from the example sentence. In this case, only "The sale" is annotated as a target.

For the results reported here, we used a combination of lexical and semantic features to classify verb argument phrases as targets of sentiment or not. The features used were all binary and included whether the phrase contained a subjective word, whether the root word of the phrase was a subjective word, whether the governor of the root word was subjective, the dependency relation associated with the root word, and a feature that captured the verb class of the controlling verb and the argument type of the phrase. For this last feature, we relied on the set of verb classes developed by Levin.[22] This assignment of verbs and their arguments into semantic classes is especially important for sentiment analysis because in English, the direction of sentiment from subject to object is dependent on the semantic characteristics of the verb. For example, in the sentence above, the verb "infuriated" identifies a sentiment on the part of the object toward the subject.

For this experiment, we extracted sentences from the MPQA 2 corpus that were annotated for sentiment, had targets identified, and had a valence strength of medium or higher. This gave us a set of 1049 sentences. The verb argument phrases were identified as described above, resulting in 983 verb argument phrases that were targets of sentiment and 2412 that were not. We then created a balanced training set of 784 positive and 784 negative instances and a balanced test set of 199 positive and 199 negative instances. There were a total of 379 unique features. We trained using NB and LS-SVM, and the results are shown in Table 3.

The performance for both classifiers was low, but given the difficulty that human annotators have agreeing on sentiment targets, the results were not out of line with our expectations. The results did improve, however, over the precision and recall of a baseline classifier that classifies a verb argument as a sentiment target if it contains a subjective word or is immediately preceded or followed by a subjective word. This suggests that the presence of subjective words alone is not sufficient for determining whether a clause is a target of sentiment.

The feature set used for this study has considerable room for expansion, which suggests that improved results are attainable. Potential new features include unigrams and bigrams local to the verb argument phrase, FrameNet roles, and dependency parse path information.

CONCLUSION

We have framed the problem of sentiment analysis and presented an approach for identifying sentiment at the sentence and subsentence levels. Future directions for this work include developing classifiers to identify targets of sentiment and their valences, using entity coreference resolution to aggregate targets of sentiment within and across documents, performing tests to demonstrate the scalability of sentiment analysis on large volumes of social media text, and annotating sentences from various domains for training classifiers specific to those domains.

REFERENCES

[1] Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J., *A Comprehensive Grammar of the English Language*, Longman, New York (1985).

[2] Wegg, S. J., "Answer Me This," Review of *A Serious Man*, *JWR*, http://www.jamesweggreview.org/Articles.aspx?ID=1078 (7 Jan 2010).

[3] Phillips, M. W. Jr., "*A Serious Man* (2009)," *Goatdog's Movies*, http://goatdog.com/moviePage.php?movieID=985 (17 Dec 2009).

[4] Greenwald, G., "'Washington Intellectual Dishonesty' Defined," *Salon*, http://www.salon.com/news/opinion/glenn_greenwald/2010/05/11/kagan (11 May 2010).

[5] Kim, S., and Hovy, E., "Automatic Detection of Opinion Bearing Words and Sentences," in *Companion Volume to Proc. 2nd International Joint Conf. on Natural Language Processing (IJCNLP-05)*, Jeju Island, South Korea, pp. 61–66 (2005).

[6] Pang, B., and Lee, L., "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," in *Proc. 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, Barcelona, Spain, pp. 271–278 (2004).

[7] Wiebe, J., Wilson, T., and Claire C., "Annotating Expressions of Opinions and Emotions in Language," *Lang. Resour. Eval.* **39**(2/3), 164–210 (2005).

[8]Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S., "Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns," in *Proc. Human Language Technology Conf./Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, Canada, pp. 355–362 (2005).

[9]Kim, S.-M., and Hovy, E. "Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text," in *Proc. 21st International Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006), Workshop on Sentiment and Subjectivity in Text*, Sydney, Australia, pp. 1–8 (2006).

[10]Baker, C. F., Fillmore, C. J., and Lowe, J. B., "The Berkeley FrameNet Project," in *Proc. 17th International Conf. on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 1998)*, Montreal, Quebec, Canada, vol. 1, pp. 86–90 (1998).

[11]Wilson, T. A., *Fine-Grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*, Doctoral dissertation, University of Pittsburgh, Pittsburgh, PA, http://www.cs.pitt.edu/~wiebe/pubs/papers/twilsonDissertation2008.pdf (2008).

[12]Kessler, J., and Nicolov, N., "Targeting Sentiment Expressions Through Supervised Ranking of Linguistic Configurations," in *Proc. 3rd International AAAI Conf. on Weblogs and Social Media (ICWSM 2009)*, San Jose, CA, pp. 90–97 (2009).

[13]Hsueh, P., Melville, P., and Sindhwani, V., "Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria," in *Proc. North American Chapter of the Association for Computational Linguistics—Human Language Technologies (NAACL HLT 2009), Workshop on Active Learning for Natural Language Processing*, Boulder, CO, pp. 27–35 (2009).

[14]Wei, C., "Formation of Norms in a Blog Community," in *Into the Blogosphere: Rhetoric, Community and Culture of Weblogs*, L. Gurak, S. Antonijevic, L. A. Johnson, C. Ratliff, and J. Reyman (eds.), http://blog.lib.umn.edu/blogosphere/formation_of_norms.html (2004).

[15]Krippendorff, K., *Content Analysis: An Introduction to Its Methodology* (2nd ed.), Sage, Thousand Oaks, CA (2004).

[16]Metsis, V., Androutsopoulos, I., and Paliouras, G., "Spam Filtering with Naive Bayes—Which Naive Bayes?" in *Proc. 3rd Conf. on Email and Anti-Spam (CEAS 2006)*, Mountain View, CA, pp. 1–9 (2006).

[17]Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., and Vandewalle, J., *Least Squares Support Vector Machines*, World Scientific, Singapore (2002).

[18]Pang, B., Lee, L., and Vaithyanathan, S., "Thumbs Up? Sentiment Classification Using Machine Learning Techniques," in *Proc. 2002 Conf. on Empirical Methods Natural Language Processing (EMNLP 2002)*, Philadelphia, PA, pp. 79–86 (2002).

[19]Wilson, T., Ruppenhofer, J., and Wiebe, J., "README," *MPQA Opinion Corpus Release Page, Version 2.0*, http://www.cs.pitt.edu/mpqa/databaserelease/ (10 Dec 2008).

[20]Marneffe, M., MacCartney, B., and Manning, C., "Generating Typed Dependency Parses from Phrase Structure Parses," in *Proc. 5th International Conf. on Language Resources Evaluation (LREC 2006)*, Genoa, Italy, pp. 449–454 (2006).

[21]Klein, D., and Manning, C., "Fast Exact Inference with a Factored Model for Natural Language Parsing," in *Proc. Neural Information Processing Systems 15 (NIPS 2002): Advances in Neural Information Processing Systems (NIPS)*, Vancouver, British Columbia, Canada, pp. 3–10 (2003).

[22]Levin, B., *English Verb Classes and Alternations: A Preliminary Investigation*, University of Chicago Press, Chicago (1993).

The Authors

Clayton R. Fink

Danielle S. Chou

Jonathon J. Kopecky

Ashley J. Llorens

Clayton R. Fink is a Senior Software Engineer in the System and Information Sciences Group of the Milton S. Eisenhower Research Center at APL and was the principal investigator for the Dynamics of Sentiment in Social Media Independent Research and Development project under which this work was funded. His main contribution to this work was in developing approaches for fine-grained sentiment analysis. **Danielle S. Chou** works in the Guidance, Navigation, and Control Group of the Air and Missile Defense Department. For this project, she was responsible for developing annotation guidelines and implementing interannotator agreement measures, and she served as a member of the annotation team. She also investigated and evaluated machine learning approaches used in this study. **Jonathon J. Kopecky** is a postdoctoral fellow in the Milton S. Eisenhower Research Center and was responsible for developing annotation guidelines, implementing interannotator agreement measures, and annotation. He also developed approaches and feature sets for sentence-level classification of subjectivity. **Ashley J. Llorens** is a Senior Algorithm Developer and Project Manager in the Systems Group of the National Security Technology Department. He was responsible for implementing and testing the machine language algorithms used in this study for both coarse-grained and fine-grained sentiment analysis. For further information on the work reported here, contact Clayton Fink. His e-mail address is clayton.fink@jhuapl.edu.

The *Johns Hopkins APL Technical Digest* can be accessed electronically at **www.jhuapl.edu/techdigest**.