Contextual Phrase-Level Polarity Analysis using Lexical Affect Scoring and Syntactic N-grams

Apoorv Agarwal
Department of Computer Science
Columbia University
New York, USA
aa2644@columbia.edu

Fadi Biadsy
Department of Computer Science
Columbia University
New York, USA
fadi@cs.columbia.edu

Kathleen R. Mckeown
Department of Computer Science
Columbia University
New York, USA
kathy@cs.columbia.edu

Abstract

We present a classifier to predict contextual polarity of subjective phrases in a sentence. Our approach features lexical scoring derived from the Dictionary of Affect in Language (DAL) and extended through WordNet, allowing us to automatically score the vast majority of words in our input avoiding the need for manual labeling. We augment lexical scoring with n-gram analysis to capture the effect of context. We combine DAL scores with syntactic constituents and then extract n-grams of constituents from all sentences. We also use the polarity of all syntactic constituents within the sentence as features. Our results show significant improvement over a majority class baseline as well as a more difficult baseline consisting of lexical n-grams.

1 Introduction

Sentiment analysis is a much-researched area that deals with identification of positive, negative and neutral opinions in text. The task has evolved from document level analysis to sentence and phrasal level analysis. Whereas the former is suitable for classifying news (e.g., editorials vs. reports) into positive and negative, the latter is essential for question-answering and recommendation systems. A recommendation system, for example, must be able to recommend restaurants (or movies, books, etc.) based on a variety of features such as food, service or ambience. Any single review sentence may contain both positive and negative opinions, evaluating different features of a restaurant. Consider the following sentence (1) where the writer expresses opposing sentiments towards food and service of a restaurant. In tasks such as this, therefore, it is important that sentiment analysis be done at the phrase level.

(1) The Taj has great food but I found their service to be lacking.

Subjective phrases in a sentence are carriers of sentiments in which an experiencer expresses an attitude, often towards a target. These subjective phrases may express neutral or polar attitudes depending on the context of the sentence in which they appear. Context is mainly determined by content and structure of the sentence. For example, in the following sentence (2), the underlined subjective phrase seems to be negative, but in the larger context of the sentence, it is positive.[1]

(2) The robber entered the store but his efforts were crushed when the police arrived on time.

Our task is to predict contextual polarity of subjective phrases in a sentence. A traditional approach to this problem is to use a prior polarity lexicon of words to first set priors on target phrases and then make use of the syntactic and semantic information in and around the sentence to make the final prediction. As in earlier approaches, we also use a lexicon to set priors, but we explore new uses of a Dictionary of Affect in Language (DAL) (Whissel, 1989) extended using WordNet (Fellbaum, 1998). We augment this approach with n-gram analysis to capture the effect of context. We present a system for classification of neutral versus positive versus negative and positive versus negative polarity (as is also done by (Wilson et al., 2005)). Our approach is novel in the use of following features:

- **Lexical scores derived from DAL and extended through WordNet**: The Dictionary of Affect has been widely used to aid in interpretation of emotion in speech (Hirschberg

[1] We assign polarity to phrases based on Wiebe (Wiebe et al., 2005); the polarity of all examples shown here is drawn from annnotations in the MPQA corpus. Clearly the assignment of polarity chosen in this corpus depends on general cultural norms.

et al., 2005). It contains numeric scores assigned along axes of pleasantness, activeness and concreteness. We introduce a method for setting numerical priors on words using these three axes, which we refer to as a "scoring scheme" throughout the paper. This scheme has high coverage of the phrases for classification and requires no manual intervention when tagging words with prior polarities.

- **N-gram Analysis: exploiting automatically derived polarity of syntactic constituents** We compute polarity for each syntactic constituent in the input phrase using lexical affect scores for its words and extract n-grams over these constituents. N-grams of syntactic constituents tagged with polarity provide patterns that improve prediction of polarity for the subjective phrase.

- **Polarity of Surrounding Constituents:** We use the computed polarity of syntactic constituents surrounding the phrase we want to classify. These features help to capture the effect of context on the polarity of the subjective phrase.

We show that classification of subjective phrases using our approach yields better accuracy than two baselines, a majority class baseline and a more difficult baseline of lexical n-gram features.

We also provide an analysis of how the different component DAL scores contribute to our results through the introduction of a "norm" that combines the component scores, separating polar words that are less subjective (e.g., *Christmas* , *murder*) from neutral words that are more subjective (e.g., *most*, *lack*).

Section 2 presents an overview of previous work, focusing on phrasal level sentiment analysis. Section 3 describes the corpus and the gold standard we used for our experiments. In section 4, we give a brief description of DAL, discussing its utility and previous uses for emotion and for sentiment analysis. Section 5 presents, in detail, our polarity classification framework. Here we describe our scoring scheme and the features we extract from sentences for classification tasks. Experimental set-up and results are presented in Section 6. We conclude with Section 7 where we also look at future directions for this research.

2 Literature Survey

The task of sentiment analysis has evolved from document level analysis (e.g., (Turney., 2002); (Pang and Lee, 2004)) to sentence level analysis (e.g., (Hu and Liu., 2004); (Kim and Hovy., 2004); (Yu and Hatzivassiloglou, 2003)). These researchers first set priors on words using a prior polarity lexicon. When classifying sentiment at the sentence level, other types of clues are also used, including averaging of word polarities or models for learning sentence sentiment.

Research on contextual phrasal level sentiment analysis was pioneered by Nasukawa and Yi (2003), who used manually developed patterns to identify sentiment. Their approach had high precision, but low recall. Wilson et al., (2005) also explore contextual phrasal level sentiment analysis, using a machine learning approach that is closer to the one we present. Both of these researchers also follow the traditional approach and first set priors on words using a prior polarity lexicon. Wilson et al. (2005) use a lexicon of over 8000 subjectivity clues, gathered from three sources ((Riloff and Wiebe, 2003); (Hatzivassiloglou and McKeown, 1997) and The General Inquirer[2]). Words that were not tagged as positive or negative were manually labeled. Yi et al. (2003) acquired words from GI, DAL and WordNet. From DAL, only words whose pleasantness score is one standard deviation away from the mean were used. Nasukawa as well as other researchers (Kamps and Marx, 2002)) also manually tag words with prior polarities. All of these researchers use categorical tags for prior lexical polarity; in contrast, we use quantitative scores, making it possible to use them in computation of scores for the full phrase.

While Wilson et al. (2005) aim at phrasal level analysis, their system actually only gives "each clue instance its own label" [p. 350]. Their gold standard is also at the clue level and assigns a value based on the clue's appearance in different expressions (e.g., if a clue appears in a mixture of negative and neutral expressions, its class is negative). They note that they do not determine subjective expression boundaries and for this reason, they classify at the word level. This approach is quite different from ours, as we compute the polarity of the full phrase. The average length of the subjective phrases in the corpus was 2.7 words, with a standard deviation of 2.3. Like Wilson et al.

[2]http://www.wjh.harvard.edu/ inquirer

(2005) we do not attempt to determine the boundary of subjective expressions; we use the labeled boundaries in the corpus.

3 Corpus

We used the Multi-Perspective Question-Answering (MPQA version 1.2) Opinion corpus (Wiebe et al., 2005) for our experiments. We extracted a total of 17,243 subjective phrases annotated for contextual polarity from the corpus of 535 documents (11,114 sentences). These subjective phrases are either "direct subjective" or "expressive subjective". "Direct subjective" expressions are explicit mentions of a private state (Quirk et al., 1985) and are much easier to classify. "Expressive subjective" phrases are indirect or implicit mentions of private states and therefore are harder to classify. Approximately one third of the phrases we extracted were direct subjective with non-neutral expressive intensity whereas the rest of the phrases were expressive subjective. In terms of polarity, there were 2779 positive, 6471 negative and 7993 neutral expressions. Our Gold Standard is the manual annotation tag given to phrases in the corpus.

4 DAL

DAL is an English language dictionary built to measure emotional meaning of texts. The samples employed to build the dictionary were gathered from different sources such as interviews, adolescents' descriptions of their emotions and university students' essays. Thus, the 8742 word dictionary is broad and avoids bias from any one particular source. Each word is given three kinds of scores (pleasantness – also called evaluation, *ee*, activeness, *aa* and imagery, *ii*) on a scale of 1 (low) to 3 (high). Pleasantness is a measure of polarity. For example, in Table 1, *affection* is given a pleasantness score of 2.77 which is closer to 3.0 and is thus a highly positive word. Likewise, activeness is a measure of the activation or arousal level of a word, which is apparent from the activeness scores of *slug* and *energetic* in the table. The third score, imagery, is a measure of the ease with which a word forms a mental picture. For example, *affect* cannot be imagined easily and therefore has a score closer to 1, as opposed to *flower* which is a very concrete and therefore has an imagery score of 3.

A notable feature of the dictionary is that it has different scores for various inflectional forms of a word (*affect* and *affection*) and thus, morphological parsing, and the possibility of resulting errors, is avoided. Moreover, Cowie et al., (2001) showed that the three scores are uncorrelated; this implies that each of the three scores provide complementary information.

Word	*ee*	*aa*	*ii*
Affect	1.75	1.85	1.60
Affection	2.77	2.25	2.00
Slug	1.00	1.18	2.40
Energetic	2.25	3.00	3.00
Flower	2.75	1.07	3.00

Table 1: DAL scores for words

The dictionary has previously been used for detecting deceptive speech (Hirschberg et al., 2005) and recognizing emotion in speech (Athanaselis et al., 2006).

5 The Polarity Classification Framework

In this section, we present our polarity classification framework. The system takes a sentence marked with a subjective phrase and identifies the most likely contextual polarity of this phrase. We use a logistic regression classifier, implemented in Weka, to perform two types of classification: Three way (positive, negative, vs. neutral) and binary (positive vs. negative). The features we use for classification can be broadly divided into three categories: I. Prior polarity features computed from DAL and augmented using WordNet (Section 5.1). II. lexical features including POS and word n-gram features (Section 5.3), and III. the combination of DAL scores and syntactic features to allow both n-gram analysis and polarity features of neighbors (Section 5.4).

5.1 Scoring based on DAL and WordNet

DAL is used to assign three prior polarity scores to each word in a sentence. If a word is found in DAL, scores of pleasantness *(ee)*, activeness *(aa)*, and imagery *(ii)* are assigned to it. Otherwise, a list of the word's synonyms and antonyms is created using WordNet. This list is sequentially traversed until a match is found in DAL or the list ends, in which case no scores are assigned. For example, *astounded*, a word absent in DAL, was scored by using its synonym *amazed*. Similarly, *in-humane* was scored using the reverse polarity of

its antonym *humane*, present in DAL. These scores are Z-Normalized using the mean and standard deviation measures given in the dictionary's manual (Whissel, 1989). It should be noted that in our current implementation all function words are given zero scores since they typically do not demonstrate any polarity. The next step is to boost these normalized scores depending on how far they lie from the mean. The reason for doing this is to be able to differentiate between phrases like "fairly decent advice" and "excellent advice". Without boosting, the pleasantness scores of both phrases are almost the same. To boost the score, we multiply it by the number of standard deviations it lies from the mean.

After the assignment of scores to individual words, we handle local negations in a sentence by using a simple finite state machine with two states: RETAIN and INVERT. In the INVERT state, the sign of the pleasantness score of the current word is inverted, while in the RETAIN state the sign of the score stays the same. Initially, the first word in a given sentence is fed to the RETAIN state. When a negation (e.g., not, no, never, cannot, didn't) is encountered, the state changes to the INVERT state. While in the INVERT state, if 'but' is encountered, it switches back to the RETAIN state. In this machine we also take care of "not only" which serves as an intensifier rather than negation (Wilson et al., 2005). To handle phrases like "no better than evil" and "could not be clearer", we also switch states from INVERT to RETAIN when a comparative degree adjective is found after 'not'. For example, the words in phrase in Table (2) are given positive pleasantness scores labeled with positive prior polarity.

Phrase	has	no	greater	desire
POS	VBZ	DT	JJR	NN
(ee)	0	0	3.37	0.68
State	RETAIN	INVERT	RETAIN	RETAIN

Table 2: Example of scoring scheme using DAL

We observed that roughly 74% of the content words in the corpus were directly found in DAL. Synonyms of around 22% of the words in the corpus were found to exist in DAL. Antonyms of only 1% of the words in the corpus were found in DAL. Our system failed to find prior semantic orientations of roughly 3% of the total words in the corpus. These were rarely occurring words like *apartheid*, *apocalyptic* and *ulterior*. We assigned zero scores for these words.

In our system, we assign three DAL scores, using the above scheme, for the subjective phrase in a given sentence. The features are (1) μ_{ee}, the mean of the pleasantness scores of the words in the phrase, (2) μ_{aa}, the mean of the activeness scores of the words in the phrase, and similarly (3) μ_{ii}, the mean of the imagery scores.

5.2 Norm

We gave each phrase another score, which we call the *norm*, that is a combination of the three scores from DAL. Cowie et al. (2001) suggest a mechanism of mapping emotional states to a 2-D continuous space using an Activation-Evaluation space (AE) representation. This representation makes use of the pleasantness and activeness scores from DAL and divides the space into four quadrants: "delightful", "angry", "serene", and "depressed". Whissel (2008), observes that tragedies, which are easily imaginable in general, have higher imagery scores than comedies. Drawing on these approaches and our intuition that neutral expressions tend to be more subjective, we define the norm in the following equation (1).

$$norm = \frac{\sqrt{ee^2 + aa^2}}{ii} \quad (1)$$

Words of interest to us may fall into the following four broad categories:

1. **High AE score and high imagery:** These are words that are highly polar and less subjective (e.g., *angel* and *lively*).
2. **Low AE score and low imagery:** These are highly subjective neutral words (e.g., *generally* and *ordinary*).
3. **High AE score and low imagery:** These are words that are both highly polar and subjective (e.g., *succeed* and *good*).
4. **Low AE score and high imagery:** These are words that are neutral and easily imaginable (e.g., *car* and *door*).

It is important to differentiate between these categories of words, because highly subjective words may change orientation depending on context; less subjective words tend to retain their prior orientation. For instance, in the example sentence from Wilson et al.(2005)., the underlined phrase

seems negative, but in the context it is positive. Since a subjective word like *succeed* depends on "what" one succeeds in, it may change its polarity accordingly. In contrast, less subjective words, like *angel*, do not depend on the context in which they are used; they evoke the same connotation as their prior polarity.

(3) They haven't succeeded and will never succeed in breaking the will of this valiant people.

As another example, AE space scores of *goodies* and *good* turn out to be the same. What differentiates one from the another is the imagery score, which is higher for the former. Therefore, value of the norm is lower for *goodies* than for *good*. Unsurprisingly, this feature always appears in the top 10 features when the classification task contains neutral expressions as one of the classes.

5.3 Lexical Features

We extract two types of lexical features, part of speech (POS) tags and n-gram word features. We count the number of occurrences of each POS in the subjective phrase and represent each POS as an integer in our feature vector.[3] For each subjective phrase, we also extract a subset of unigram, bigrams, and trigrams of words (selected automatically, see Section 6). We represent each n-gram feature as a binary feature. These types of features were used to approximate standard n-gram language modeling (LM). In fact, we did experiment with a standard trigram LM, but found that it did not improve performance. In particular, we trained two LMs, one on the polar subjective phrases and another on the neutral subjective phrases. Given a sentence, we computed two perplexities of the two LMs on the subjective phrase in the sentence and added them as features in our feature vectors. This procedure provided us with significant improvement over a chance baseline but did not outperform our current system. We speculate that this was caused by the split of training data into two parts, one for training the LMs and another for training the classifier. The resulting small quantity of training data may be the reason for bad performance. Therefore, we decided to back off to only binary n-gram features as part of our feature vector.

[3]We use the Stanford Tagger to assign parts of speech tags to sentences. (Toutanova and Manning, 2000)

5.4 Syntactic Features

In this section, we show how we can combine the DAL scores with syntactic constituents. This process involves two steps. First, we chunk each sentence to its syntactic constituents (NP, VP, PP, JJP, and Other) using a CRF Chunker.[4] If the marked-up subjective phrase does not contain complete chunks (i.e., it partially overlaps with other chunks), we expand the subjective phrase to include the chunks that it overlaps with. We term this expanded phrase as the *target phrase*, see Figure 1.

Second, each chunk in a sentence is then assigned a 2-D AE space score as defined by Cowie et al., (2001) by adding the individual AE space scores of all the words in the chunk and then normalizing it by the number of words. At this point, we are only concerned with the polarity of the chunk (i.e., whether it is positive or negative or neutral) and imagery will not help in this task; the AE space score is determined from pleasantness and activeness alone. A threshold, determined empirically by analyzing the distributions of positive (pos), negative (neg) and neutral (neu) expressions, is used to define ranges for these classes of expressions. This enables us to assign each chunk a prior semantic polarity. Having the semantic orientation (positive, negative, neutral) and phrasal tags, the sentence is then converted to a sequence of encodings $[Phrasal - Tag]_{polarity}$. We mark each phrase that we want to classify as a "target" to differentiate it from the other chunks and attach its encoding. As mentioned, if the target phrase partially overlaps with chunks, it is simply expanded to subsume the chunks. This encoding is illustrated in Figure 1.

After these two steps, we extract a set of features that are used in classifying the target phrase. These include n-grams of chunks from the all sentences, minimum and maximum pleasantness scores from the chunks in the target phrase itself, and the syntactic categories that occur in the context of the target phrase. In the remainder of this section, we describe how these features are extracted.

We extract unigrams, bigrams and trigrams of chunks from all the sentences. For example, we may extract a bigram from Figure 1 of $[VP]_{neu}$ followed by $[PP]_{neg}^{target}$. Similar to the lexical

[4]Xuan-Hieu Phan, "CRFChunker: CRF English Phrase Chunker", http://crfchunker.sourceforge.net/, 2006.

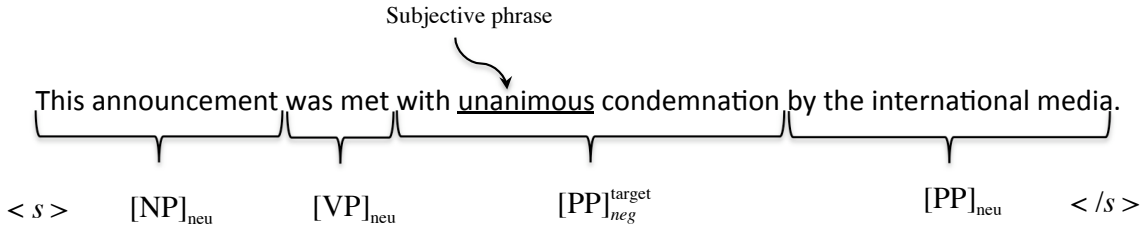

Figure 1: Converting a sentence with a subjective phrase to a sequence of chunks with their types and polarities

n-grams, for the sentence containing the target phrase, we add binary values in our feature vector such that the value is 1 if the sentence contains that chunk n-gram.

We also include two features related to the target phrase. The target phrase often consists of many chunks. To detect if a chunk of the target phrase is highly polar, minimum and maximum pleasantness scores over all the chunks in the target phrase are noted.

In addition, we add features which attempt to capture contextual information using the prior semantic polarity assigned to each chunk both within the target phrase itself and within the context of the target phrase. In cases where the target phrase is in the beginning of the sentence or at the end, we simply assign zero scores. Then we compute the frequency of each syntactic type (i.e., NP, VP, PP, JJP) and polarity (i.e., positive, negative, neutral) to the left of the target, to the right of the target and for the target. This additional set of contextual features yields 36 features in total: three polarities: {positive, negative, neutral} * three contexts: {left, target, right} * four chunk syntactic types: {NP, VP, PP, JJP}.

The full set of features captures different types of information. N-grams look for certain patterns that may be specific to either polar or neutral sentiments. Minimum and maximum scores capture information about the target phrase standalone. The last set of features incorporate information about the neighbors of the target phrase. We performed feature selection on this full set of n-gram related features and thus, a small subset of these n-gram related features, selected automatically (see section 6) were used in the experiments.

6 Experiments and Results

Subjective phrases from the MPQA corpus were used in 10-fold cross-validation experiments. The MPQA corpus includes gold standard tags for each phrase. A logistic classifier was used for two polarity classification tasks, positive versus negative versus neutral and positive versus negative. We report accuracy, and F-measure for both balanced and unbalanced data.

Feature Types	Accuracy	Pos.*	Neg.*	Neu.*
Chance baseline	33.33%	-	-	-
N-gram baseline	59.05%	0.602	0.578	0.592
DAL scores only	59.66%	0.635	0.635	0.539
+ POS	60.55%	0.621	0.542	0.655
+ Chunks	64.72%	0.681	0.665	0.596
+ N-gram (all)	67.51%	0.703	0.688	0.632
All (unbalanced)	70.76%	0.582	0.716	0.739

Table 3: Results of 3 way classification (Positive, Negative, and Neutral). In the unbalanced case, majority class baseline is 46.3% (*F-Measure).

Feature Types	Accuracy	Pos.*	Neg.*
Chance baseline	50%	-	-
N-gram baseline	73.21%	0.736	0.728
DAL scores only	77.02%	0.763	0.728
+ POS	79.02%	0.788	0.792
+ Chunks	80.72%	0.807	0.807
+ N-gram (all)	82.32%	0.802	0.823
All (unbalanced)	84.08%	0.716	0.889

Table 4: Positive vs. Negative classification results. Baseline is the majority class. In the unbalanced case, majority class baseline is 69.74%. (* F-Measure)

6.1 Positive versus Negative versus Neutral

Table 3 shows results for a 3-way classifier. For the balanced data-set, each class has 2799 instances and hence the chance baseline is 33%. For the unbalanced data-set, there are 2799 instances of positive, 6471 instances of negative and 7993 instances of neutral phrases and thus the baseline is about 46%. Results show that the accuracy increases as more features are added. It may be seen from the table that prior polarity scores do not do well alone, but when used in conjunction with other features they play an important role in achieving an accuracy much higher than both baselines (chance and lexical n-grams). To re-

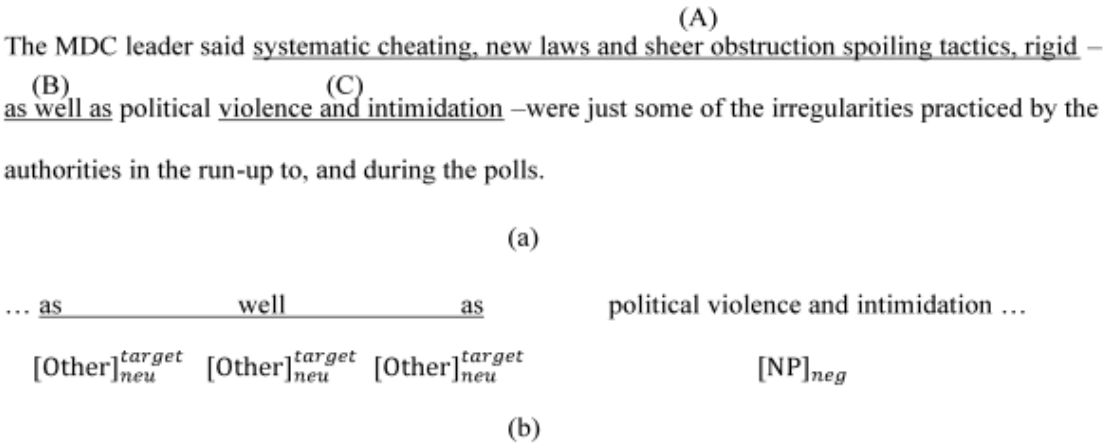

Figure 2: (a) An example sentence with three annotated subjective phrases in the same sentence. (b) Part of the sentence with the target phrase (B) and their chunks with prior polarities.

confirm if prior polarity scores add value, we experimented by using all features except the prior polarity scores and noticed a drop in accuracy by about 4%. This was found to be true for the other classification task as well. The table shows that parts of speech and lexical n-grams are good features. A significant improvement in accuracy (over 4%, p-value = $4.2e$-15) is observed when chunk features (i.e., n-grams of constituents and polarity of neighboring constituents) are used in conjunction with prior polarity scores and part of speech features.[5] This improvement may be explained by the following observation. The bigram "$[Other]^{target}_{neu}$ $[NP]_{neu}$" was selected as a top feature by the Chi-square feature selector. So were unigrams, $[Other]^{target}_{neu}$ and $[Other]^{target}_{neg}$. We thus learned n-gram patterns that are characteristic of neutral expressions (the just mentioned bigram and the first of the unigrams) as well as a pattern found mostly in negative expressions (the latter unigram). It was surprising to find another top chunk feature, the bigram "$[Other]^{target}_{neu}$ $[NP]_{neg}$" (i.e., a neutral chunk of syntactic type "Other" preceding a negative noun phrase), present in neutral expressions six times more than in polar expressions. An instance where these chunk features could have been responsible for the correct prediction of a target phrase is shown in Figure 2. Figure 2(a) shows an example sentence from the MPQA corpus, which has three annotated subjective phrases. The manually labeled polarity of phrases (A) and (C) is negative and that of (B) is neutral. Figure 2(b) shows the relevant chunk bigram which is used to predict the contextual polarity of the target phrase (B).

It was interesting to see that the top 10 features consisted of all categories (i.e., prior DAL scores, lexical n-grams and POS, and syntactic) of features. In this and the other experiment, pleasantness, activation and the norm were among the top 5 features. We ran a significance test to show the importance of the norm feature in our classification task and observed that it exerted a significant increase in accuracy (2.26%, p-value = $1.45e$-5).

6.2 Positive versus Negative

Table 4 shows results for positive versus negative classification. We show results for both balanced and unbalanced data-sets. For balanced, there are 2779 instances of each class. For the unbalanced data-set, there are 2779 instances of positive and 6471 instances of neutral, thus our chance baseline is around 70%. As in the earlier classification, accuracy and F-measure increase as we add features. While the increase of adding the chunk features, for example, is not as great as in the previous classification, it is nonetheless significant (p-value = 0.0018) in this classification task. The smaller increase lends support to our hypothesis that polar expressions tend to be less subjective and thus are less likely to be affected by contextual polarity. Another thing that supports our hypothesis that neutral expressions are more subjective is the fact that the rank of imagery (*ii*), dropped significantly in this classification task as compared to the previous classification task. This implies that imagery has a much lesser role to play when we are dealing with non-neutral expressions.

[5] We use the binomial test procedure to test statistical significance throughout the paper.

7 Conclusion and Future Work

We present new features (DAL scores, norm scores computed using DAL, n-gram over chunks with polarity) for phrasal level sentiment analysis. They work well and help in achieving high accuracy in a three-way classification of positive, negative and neutral expressions. We do not require any manual intervention during feature selection, and thus our system is fully automated. We also introduced a 3-D representation that maps different classes to spatial coordinates.

It may seem to be a limitation of our system that it requires accurate expression boundaries. However, this is not true for the following two reasons: first, Wiebe et al., (2005) declare that while marking the span of subjective expressions and hand annotating the MPQA corpus, the annotators were not trained to mark accurate expression boundaries. The only constraint was that the subjective expression should be within the mark-ups for all annotators. Second, we expanded the marked subjective phrase to subsume neighboring phrases at the time of chunking.

A limitation of our scoring scheme is that it does not handle polysemy, since words in DAL are not provided with their parts of speech. Statistics show, however, that most words occurred with primarily one part of speech only. For example, "will" occurred as modal 1272 times in the corpus, whereas it appeared 34 times as a noun. The case is similar for "like" and "just", which mostly occur as a preposition and an adverb, respectively. Also, in our state machine, we haven't accounted for the impact of connectives such as "but" or "although"; we propose drawing on work in argumentative orientation to do so ((Anscombre and Ducrot, 1983); (Elhadad and McKeown, 1990)).

For future work, it would be interesting to do subjectivity and intensity classification using the same scheme and features. Particularly, for the task of subjectivity analysis, we speculate that the imagery score might be useful for tagging chunks with "subjective" and "objective" instead of positive, negative, and neutral.

Acknowledgments

This work was supported by the National Science Foundation under the KDD program. Any opinions, ndings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reect the views of the National Science Foundation. score.

We would like to thank Julia Hirschberg for useful discussion. We would also like to acknowledge Narayanan Venkiteswaran for implementing parts of the system and Amal El Masri, Ashleigh White and Oliver Elliot for their useful comments.

References

J.C. Anscombre and O. Ducrot. 1983. Philosophie et langage. l'argumentation clans la langue. Bruxelles: Pierre Mardaga.

T. Athanaselis, S. Bakamidis, , and L. Dologlou. 2006. Automatic recognition of emotionally coloured speech. In *Proceedings of World Academy of Science, Engineering and Technology, volume 12, ISSN 1307-6884*.

R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, and W. Fellenz et al. 2001. Emotion recognition in human-computer interaction. In *IEEE Signal Processing Magazine, 1, 32-80.*

M. Elhadad and K. R. McKeown. 1990. Generating connectives. In *Proceedings of the 13th conference on Computational linguistics*, pages 97–101, Morristown, NJ, USA. Association for Computational Linguistics.

C. Fellbaum. 1998. Wordnet, an electronic lexical database. In *MIT press*.

V. Hatzivassiloglou and K. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of ACL.*

J. Hirschberg, S. Benus, J.M. Brenier, F. Enos, and S. Friedman. 2005. Distinguishing deceptive from non-deceptive speech. In *Proceedings of Interspeech, 1833-1836*.

M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of KDD*.

J. Kamps and M. Marx. 2002. Words with attitude. In *1st International WordNet Conference.*

S. M. Kim and E. Hovy. 2004. Determining the sentiment of opinions. In *In Coling*.

T. Nasukawa and J. Yi. 2003. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of K-CAP*.

B. Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL.*

R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. 1985. A comprehensive grammar of the english language. Longman, New York.

E. Riloff and J. Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of EMNLP*.

K. Toutanova and C. D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.*

P. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL.*

C. M. Whissel. 1989. The dictionary of affect in language. In *R. Plutchik and H. Kellerman, editors, Emotion: theory research and experience, volume 4, Acad. Press.*, London.

C. M. Whissell. 2008. A psychological investigation of the use of shakespeare=s emotional language: The case of his roman tragedies. In *Edwin Mellen Press.*, Lewiston, NY.

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. In *Language Resources and Evaluation, volume 39, issue 2-3, pp. 165-210.*

T. Wilson, J. Wiebe, and P. Hoffman. 2005. Recognizing contextual polarity in phrase level sentiment analysis. In *Proceedings of ACL.*

J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of IEEE ICDM*.

H. Yu and V. Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of EMNLP*.