2015 IEEE 15th International Conference on Data Mining Workshops

Sentiment Polarity Classification using Structural Features

Daniel Ansari
SenticNet
e-mail: daniel@sentic.net

***Abstract*—This work investigates the role of contrasting discourse relations signaled by cue phrases, together with phrase positional information, in predicting sentiment at the phrase level. Two domains of online reviews were chosen. The first domain is of nutritional supplement reviews, which are often poorly structured yet also allow certain simplifying assumptions to be made. The second domain is of hotel reviews, which have somewhat different characteristics. A corpus is built from these reviews, and manually tagged for polarity. We propose and evaluate a few new features that are realized through a lightweight method of discourse analysis, and use these features in a hybrid lexicon and machine learning based classifier. Our results show that these features may be used to obtain an improvement in classification accuracy compared to other traditional machine learning approaches.**

Keywords-sentiment polarity classification; contrast discourse relation; sentence position

I. INTRODUCTION AND BACKGROUND

Discourse relations describe how different discourse segments, or non-overlapping spans of text, interact [1].

A number of researchers in sentiment analysis have investigated the role that discourse relations can play in reversing the polarity of opinions in segments of text [2, 3, 4], and [5] investigated the use of discourse relations in improving opinion polarity classification.

Studies [2, 3, 4, 6, 7, 8] have also examined how connectives such as *but, however, despite, although,* etc. are involved in contrasting discourse relations.

In the realm of Rhetorical Structure Theory [9], [7] found that the CONCESSION rhetorical relation was signaled by a connective 90% of the time in the newspaper article domain.

Researchers have recognized that word order and syntactic relations between words are important and useful for sentiment classification. Reference [10] obtained sub-pattern features by mining frequent sub-patterns from word sequences and dependency trees, and used these for document-level sentiment classification.

Reference [11] employed a boosting algorithm with subtrees of word dependency trees as features for sentence-level polarity classification.

Reference [2] used word dependencies and dependency trees to analyze how individual phrases combined, in the presence of conjuncts, to decide the overall sentiment of a sentence in a rule-based system. They compiled rules for the determination of polarity from dependency tree structures, for over 80 conjunctions.

More recent work [12] applies dependency-based rules for analyzing the flow of concepts in a *sentic computing* framework, which is concerned with concept-level sentiment analysis using affective and common-sense knowledge [13].

Recognizing that discourse and dependency parsers are resource-intensive and unreliable for unstructured text such as Twitter *tweets*, [4] also examined linguistic constructs such as *connectives, modals, conditionals,* and *negation,* and identified ones that affected sentiment. However, they incorporated features in their ML model for (strong) modals and conditionals, but not for conjunctions that signal discourse relations.

The system of [14] includes extraction of relevant aspects of a service from restaurant and hotel reviews, and detection of the sentiment of these aspects. One of the novel aspects of their work is in the use of user-provided labels—i.e., overall star ratings—as an additional signal in determining the polarity of individual phrases in the review, which may be different than the overall sentiment the user has provided. Similar to [15], they construct a sentiment lexicon from WordNet [16], which consists of words and their sentiment scores.

Reference [17] postulated that the position of a word in the text might make a difference to the sentiment, where movie reviews in particular might begin with an overall sentiment statement. Similar information was also formulated as a feature in [8].

The rest of the paper is organized as follows: Section II examines the relevant discourse relation. In Section III we present the features of our baseline, together with the new features. In Section IV we describe our corpus and annotation scheme, the experiments we performed and their results. Finally, conclusions and future work are discussed in Section V.

II. CONTRAST DISCOURSE RELATION

Out of the essential discourse relations for sentiment analysis put forth by [4], the focus of this work is on VIOLATED EXPECTATIONS and CONTRAST [18], which we unify as CONTRAST.

This relation also subsumes the CONCESSION rhetorical relation from RST, and is signaled by the cue phrases *but, while, yet, despite, although, though, while, other than,* and *apart from,* amongst others [7].

In the following example, the segment before the cue word *although* is slightly positive, and reverses the sentiment of the segment after:

978-1-4673-8493-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDMW.2015.57

```
I gave it a 2 because it did give
me a minor boost in energy,
although it didn't last very long
at all.
```

In this example, the segment after the word *but* is positive, in contrast to the polarity before which is negative:

```
Taste of the Fruit Punch is a bit
off but the energy, focus and drive
all the way through and beyond my
workouts makes up for the taste.
```

III. Feature Selection

We first describe the set of features comprising our baseline.

For negation handling, we adopted a similar approach to [17] and [19] by using a variable to store the negation state. We prepend "*not_*" to every word between a negation word (*not, isn't, didn't*, etc.) and either the first punctuation mark following the negation word, or another negation word. A maximum window of 10 words was found to provide the best increase in performance, for our limited corpus. This feature thus comprises a bag of words model, together with negation.

We use SentiWordNet [20], a lexical resource that assigns positivity and negativity scores to each word in WordNet, to calculate a sentiment for each phrase, using the following function from [14]:

$$\text{raw-score}(x) := \sum_{i=1}^{n} s_i ,$$

where x is a tokenized string ($w_1, w_2, \ldots, w_n$) of words, and s_i is the sentiment score of the word. However, diverging from [14], we further apply the principle of negation handling discussed in the previous paragraph, using a negation state variable to reverse the score s_i from SentiWordNet, if the word is deemed to be negated. We also do not use the floating point value of raw-score, but instead map this to a discrete value of 0 (where *abs*(raw-score) < 0.1), -1, or +1. We furthermore examined features for the *purity* of current and prior segments, as well as raw-score for prior segments [14], but found that for our domains, these hurt rather than improved performance.

Finally for our baseline, we utilize the overall sentiment of the review as provided by the user as a feature. We found that instead of using the actual integer value (1–10), a tri-valued feature for the ranges 1–4, 5–6, and 7–10, resulted in better classification performance.

The main features we introduce are for the handling of contrasting discourse relations. We do not use a discourse or a dependency parser, but alternatively, we employ a lightweight method of regular expression matching to segment each sentence.

Our first new feature is the outcome of the classifier for the prior segment, which has one of the values *positive, negative,* or *neutral*.

TABLE I. Statistics for segments labeled for each polarity in the corpus

Domain	Positive	Negative	Neutral	Total
supplements	1098	970	425	2493
hotels	1107	1014	403	2524

An assumption we made is that the cue phrase signaling a contrasting discourse relation is present at the beginning of segment. Thus, our second new feature is the cue phrase itself.

The two aforementioned features are closely related, yet distinct. They account for both the presence of the contrast relation, together with the cue phrase signaling the contrast.

Instead of using the position of words in the text as a feature, we exploited the position of the segment within the review as our last new feature. This is a structural feature that is related to the sentential structure of the review, rather than any coherence relations. The feature has a value from *1…n*, where *n* is the number of fragments in the review.

IV. Experiments

A. Corpus

Our supplements data set consisted of 545 reviews of nutritional supplements, obtained by crawling a popular fitness website, Bodybuilding.com. These reviews consisted of some metadata such as author and date, overall rating out of 10, and the review text.

We used UAM CorpusTool [21]. This tool provided an initial, automatic segmentation at the sentence level for each review, and we further divided these sentences into segments according to the contrasting discourse relations described in Section 2. We then annotated each segment for aspects and polarity. Two simplifications were made at this step: that the segment had a single polarity, and that this polarity applied to each relevant aspect contained in the segment.

The supplement reviews domain allowed us to make some assumptions that might not easily extend to other domains: that the holder of the opinion is the author of the review (as opposed to another party the reviewer may introduce), and that an aspect is regarded as universally either positive or negative from all reviewers' perspectives. For example, the aspects *strength* and *fat loss* are desirable and thus positive for everybody, whereas aspects for undesirable side effects such as *insomnia* or *high blood pressure* are always negative. Moreover, we selected a domain where there is broad consensus amongst reviewers of the aspects that are most important, and thus we did not need to concern ourselves with automatic aspect extraction, which was not a focus of the current work.

Although the assumptions above are normally reliable, other factors are sometimes present that confound our ideal that the user comments only on their experiences with the supplement in question. Consider the following phrase, which talks about the user's goals rather than the supplement itself: "*As my goals right now are mainly fat loss whilst keeping mass rather than building it…*" In the next phrase,

TABLE II. SENTIMENT CLASSIFICATION PRECISION, RECALL, F1, AND ACCURACY FOR SUPPLEMENTS (BOLDED NUMBERS INDICATE THE BEST RESULT)

Feature	Precision	Recall	F1	Accuracy
	Positive			
Base	75.5	79.8	77.6	79.7
base+conj	**76.4**	81.1	78.7	**80.7**
base+conj+idx	76.1	**81.6**	**78.8**	80.6
	Negative			
Base	70.8	77.8	74.2	78.9
base+conj	71.5	**79.1**	75.1	79.6
base+conj+idx	**72.0**	79.0	**75.3**	**79.9**
	Neutral			
Base	44.2	**27.8**	**34.1**	81.7
base+conj	43.3	25.9	32.4	81.6
base+conj+idx	**44.4**	26.4	33.1	**81.8**

TABLE III. SENTIMENT CLASSIFICATION PRECISION, RECALL, F1, AND ACCURACY FOR HOTELS

Feature	Precision	Recall	F1	Accuracy
	Positive			
Base	**77.9**	78.5	**78.2**	**80.8**
base+conj	76.8	**78.6**	77.7	80.2
base+conj+idx	73.8	74.7	74.2	77.3
	Negative			
Base	70.8	**79.8**	**75.0**	**78.7**
base+conj	**70.9**	79.2	74.8	78.6
base+conj+idx	70.3	79.7	74.7	78.3
	Neutral			
Base	44.2	**29.3**	**35.2**	82.8
base+conj	43.8	28.0	34.2	82.8
base+conj+idx	**45.3**	28.5	35.0	**83.1**

the author makes a comparison with a different product: "*I can't say exactly how goos* [sic] *this stuff is, but like the H-cut, the old one worked wonders last time around, so this time I'm sticking ith* [sic] *what I know.*" Here, the author discusses their intention to update their review: "*I'll update this review later on when i'm stacking muscle again.*" The following two fragments involve holders of opinions other than the author: "*It has been said that it does'nt* [sic] *get to your stomach at all and is very concentrated,*" and "*but i heard good things about it so we will see.*" Despite problematic segments like these having strong potential to skew our results, we do not attempt to account for them, since these are research problems in themselves.

Another problem with this domain is that sentences are sometimes poorly formed and structured into the reviews. Spaces might be missing between the period and the following sentence, thus causing automatic segmentation to fail. Grammatical and spelling errors are also rife, perhaps more so when compared with reviews in certain other domains.

Our corpus of hotel reviews was built from the LARA TripAdvisor data set from [22]. It consisted of 166 reviews of hotels in the Seattle, Washington area. There was an assumption that by limiting the corpus in such a fashion, an ML model might be able to perform better than if the reviews were of hotels spread out geographically.

Again, we used UAM CorpusTool to perform automatic, followed by further manual, segmentation of the review texts. We annotated each segment for polarity.

There were some notable differences from the supplements domain. Reviewers seem to be more likely to make comparisons with other hotels, thus a phrase could have mixed sentiment: negative for one hotel, and positive for another. The hotel reviews also tended to contain more neutral sentences near the beginning, as the authors gave a general background of themselves or their reason for traveling.

One problem that we encountered with the data set was in the review texts; a review on the TripAdvisor website consists of an optional title, the review text, and various metadata. In the data set, the review text was appended to the review title; thus, the review began with the title and the first sentence of the actual review, often with no punctuation separating them. This potentially affected our results.

The overall sentiments were given on a 5 point rating scale. We doubled these values so that our tri-valued feature would function correctly.

Data statistics for the corpus are given in Table I.

B. Results

We experimented with n-grams of varying lengths, Porter stemming [23], spelling correction, and purity [14], but do not include these results here.

We chose the Java-based *OpenNLP* [24] toolkit for its maximum entropy classifier. The MaxEnt models were trained on approximately half of the reviews and tested on the other half; the training and testing sets were then flipped, and the results aggregated.

We evaluated systems with the following features to train the models:

- **base:** our baseline feature set, comprising the following features:
  - bag-of-words with negation handling, converting all words to lower case (which we found performed better)
  - raw-score of the segment
  - overall sentiment of the review, provided by the reviewer
- **conj:** the discourse relation conjunction words beginning each segment, and the outcome of the prior segment
- **idx:** the position of the segment within the review, if it falls within the first 3 segments (for the supplements domain), or the first 6 segments (for the hotels domain)

We compared these systems by measuring precision, recall, F1 and accuracy, for the positive, negative, and neutral classes in each domain.

The results are shown in Table II and Table III.

V. CONCLUSIONS

We cannot make a direct comparison between our results and those of other work, simply because of the differences in the corpora and how we annotated ours. Our baseline feature set was, nonetheless, somewhat similar to that used by [14].

This work shows that modest improvements in precision, recall, and accuracy for the positive and negative classes in the supplements domain can be obtained by using a lightweight method of discourse analysis to segment sentences, and using features for contrasting discourse relations against these segments.

The hotels domain presented much more of a mixed bag of results, with the baseline features performing overall better. It is possible that the flawed data set had a small part to play in this; another possible explanation may be that polarity shifts between segments in this domain include neutral sentiment more than the supplements domain, and this label is much more difficult to predict by the MaxEnt classifier.

We also examined the position of segments within the review as a feature. Interestingly, this feature didn't really improve our results for positive reviews, yet caused a marginal improvement for negative reviews, again in the supplements domain. This indicates that, for our supplements dataset at least, the sentiments expressed in the first few sentences are a better predictor of the sentiments of the remainder of the review when the review is mostly negative.

We did not obtain any consistent improvement by using our features for the neutral class. We found that although positive and negative segments may at times be surrounded by neutral segments, we could not form any intuition ourselves in determining the sentiment of a neutral segment on the basis of the polarities of the surrounding segments.

The approach we have taken is particularly beneficial in our supplement reviews domain, where heavy linguistic resources such as parsing do not perform well and may be prone to failure. Our features are simple, yet we demonstrated that maximum entropy classifiers based on them fare quite well.

REFERENCES

[1] F. Wolf and E. Gibson. 2005. Representing discourse coherence: a corpus-based study. In *Computational Linguistics, 31(2), pp. 249–287.*

[2] A. Meena and T. V. Prabhakar. 2007. Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. In *ECIR, LNCS 4425, pp. 573–580*.

[3] S. Li and C. Huang. 2009. Sentiment Classification Considering Negation and Contrast Transition. In *23rd Pacific Asia Conference on Language, Information and Computation, pp. 297–306.*

[4] S. Mukherjee and P. Bhattacharyya. 2012. Sentiment analysis in twitter with lightweight discourse analysis. In *Proc. COLING 2012: Technical Papers, pp. 1847–1864.*

[5] S. Somasundaran, G. Namata, J. Wiebe, and L. Getoor. 2009. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proc. EMNLP, pp. 170–179.*

[6] L. Polanyi and A. Zaenen. 2006. Contextual Valence Shifters. In *Computing attitude and affect in text: Theory and application.* Springer Verlag.

[7] M. Taboada. 2006. Discourse markers as signals (or not) of rhetorical relations. In *Journal of Pragmatics 38, pp. 567–592.*

[8] K. Chawla, A. Ramteke, and P. Bhattacharyya. 2013. IITB-Sentiment-Analysts: Participation in sentiment analysis in Twitter SemEval 2013 task. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 495–500.*

[9] W. C. Mann and S. A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. In *Text, 8 (3), 243–281.*

[10] S. Matsumoto, H. Takamura, and M. Okumura. 2005. Sentiment classification using word sub-sequences and dependency sub-trees. In *Proc. 9th Pacific-Asia International Conference on Knowledge Discovery and Data Mining (PAKDD 2005), pp.301–311.*

[11] T. Kudo and Y. Matsumoto. A boosting algorithm for classification of semi-structured text. 2004. In *Proc. 9th EMNLP, pp.301–308.*

[12] S. Poria, E. Cambria, G. Winterstein, and G.B. Huang. 2014. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. In *Elsevier Knowledge-Based Systems, pp. 45–63.*

[13] E. Cambria and A. Hussain. 2015. Sentic computing: A common-sense-based framework for concept-level sentiment analysis. *Springer, Cham, Switzerland.*

[14] S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. A. Reis, J. Reynar. 2008. Building a sentiment summarizer for local service reviews. In *Proc. NLPIX.*

[15] M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD).*

[16] G. A. Miller. 1995. WordNet: A lexical database for English. In *Communications of the ACM, (11):39–41.* http://wordnet.princeton.edu/.

[17] B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proc. EMNLP. pp. 79–86.*

[18] J. R. Hobbs. 1985. On the coherence and structure of discourse. *Technical Report 85-37, Center for the Study of Language and Information (CSLI), Stanford, CA.*

[19] V. Narayanan, I. Arora, and A. Bhatia. 2013. Fast and accurate sentiment classification using an enhanced Naive Bayes model. *In LNCS 8206, pp 194–201.*

[20] A. Esuli and F. Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06. pp. 417–422.*

[21] M. O'Donnell. 2007. UAM CorpusTool. Version 2.8.14. http://www.wagsoft.com/CorpusTool/.

[22] H. Wang, Y. Lu, and C. Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2010), pp. 783–792.*

[23] M. Porter. 1980. An algorithm for suffix stripping. In *Program, Vol. 14, no. 3, pp. 130–137.*

[24] Apache OpenNLP. https://opennlp.apache.org