Do We Really Need Lexical Information? Towards a Top-down Approach to Sentiment Analysis of Product Reviews

Yulia Otmakhova
Computational Linguistics Lab
Department of Linguistics
Seoul National University
Gwanakro 1, Gwanak-gu
Seoul, 151-742, South Korea
julia.nixie@gmai1.com

Hyopil Shin
Computational Linguistics Lab
Department of Linguistics
Seoul National University
Gwanakro 1, Gwanak-gu
Seoul, 151-742, South Korea
hpshin@snu.ac.kr

Abstract

Most of the current approaches to sentiment analysis of product reviews are dependent on lexical sentiment information and proceed in a bottom-up way, adding new layers of features to lexical data. In this paper, we maintain that a typical product review is not a bag of sentiments, but a narrative with an underlying structure and reoccurring patterns, which allows us to predict its sentiments knowing only its general polarity and discourse cues that occur in it. We hypothesize that knowing only the review's score and its discourse patterns would allow us to accurately predict the sentiments of its individual sentences. The experiments we conducted prove this hypothesis and show a substantial improvement over the lexical baseline.

1 Introduction

For years, sentiment analysis has heavily relied on lexical resources, whether compiled by hand (Wilson et al., 2005) or automatically extracted from a large corpus (Hu and Lui, 2004). In addition to an overwhelming task of trying to capture *all* words and expressions that can convey a sentiment there are many other problems to solve: resolving the scope of negation to determine the shift of polarity (Lapponi et al., 2012), determining if an opinion is present in interrogative or conditional sentences (Narayanan et al., 2009), dealing with irony (Tsur, 2010), etc. But even if we manage to solve all aforementioned problems and create an efficient classifier, there will always be cases where reliance on lexical cues for subjectivity will betray us. Consider, for instance, the following examples from reviews of online universities[1]:

(1) The lectures are interactive and recorded. So, if you can't attend you can listen in later.
(2) I assure you, online learning at Capella was the most difficult form of education I have undergone!
(3) UMUC provided really good quality education until about 5 years ago.

In the first example, the author expresses a positive opinion of the university, but it will fail to be detected because it does not include any explicit sentiment cues (such opinions are referred to as "implicit" by Liu (2012) or as "polar facts" by Toprak et al. (2010)). Because the sentiment (and its presence) of such sentences is highly domain-dependent, they cannot be covered by any lexicons or learned in a supervised or a non-supervised way.

The second example does have a sentiment cue *difficult*, and judging by it the sentiment should be

[1] The examples in this section are taken from Darmstadt Service Review Corpus, available from https://www.ukp.tu-darmstadt.de/data/sentiment-analysis/darmstadt-service-review-corpus (Toprak et al., 2010). The corpus was also used as a development set for extracting features for this study.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1559–1568,
Denver, Colorado, May 31 – June 5, 2015.

negative. However, in this case the author actually expresses a positive view of an online university, defending it from people who claim that online education is "too easy". In the third example, the correct sentiment (negative) would again be impossible to determine because of a complicated structure.

These are just a few examples of what is currently impossible to classify correctly relying on lexical resources. To improve the classification results, there have been attempts to use local discourse information, such as discourse cues and polarity of adjacent sentences, in order to correct some of the misclassified sentences (Somasundaran, 2010). However, though such attempts resulted in some improvements, they also required quite complicated frameworks.

While such bottom-up approach (starting from lexical polarity and adding supplementary information to improve classification on a phrase and text level) is commonly used in sentiment analysis, we are wondering if it is the only valid one. Provided that we have a reliable external measure of a text's general polarity (such as a product rating for a product review) and the narrative has a predictable discourse structure, would not it be possible to classify its sentences in a top-down manner, without using any sentiment lexicons? In this paper, we experiment with this approach and compare its results with those of the traditional bottom-up method.

This paper is organized as follows. Section 2 presents a brief overview of previous studies related to sentiment analysis of product reviews, while section 3 explains the motivation behind taking an alternative approach. In section 4 we give the details of the experiments, and then in section 5 present their results. Lastly, section 6 summarizes our findings.

2 Previous Studies

Sentiment analysis so far has largely relied on explicit lexical information, either in the form of sentiment dictionaries and lexicons, such as SentiWordNet[2] or Subjectivity lexicon[3], opinion phrases extracted from a manually-annotated corpus or a dataset compiled in real time using machine learning with such lexical features as bag-of-words features, n-grams, collocations, or more sophisticated lexical patterns (Tang, 2009). As researchers realized the limitations of a purely lexical approach, they tried to augment it by using negation resolution, word meaning disambiguation or hand-crafted rules (Ding, 2008). However, though such efforts improved classification on the sentence level, they were not able to deal with the sentences where an opinion was implicit (i.e. there were no appraisal words or other lexical cues, see example 1 above) or the polarity of the sentiment word was different from the usual one (see example 2 above). To correct such misclassified instances, another level of complexity was added by using discourse features. Somasundaran (2010) defines opinion frames to enforce discourse constraints on the polarity of segments with the same or alternative target relations. Using a similar approach, Zhou et al. (2011) employ simplified Rhetorical Structure Theory (RST) relation cues (contrast, condition, continuation, cause and purpose) to eliminate polarity ambiguities. Yang (2014) concentrates on discovering opinionated sentences which do not have strong sentiment signals (implicit opinions), using discourse knowledge to improve the results of a Conditional Random Fields classifier. While such approaches are a definite improvement over the lexical baseline, they are computationally complex and still overly dependent on the lexical cues.

While machine learning algorithms such as Naïve Bayes or SVM are still the primary tools used for sentiment analysis, lately such texts as product reviews have been recognized as having an internal structure and inter-sentential relations, and thus structural conditional frameworks have been used for their classification. One popular tool is Conditional Random Fields (CRF), which was used, among others, by Zhao (2008) to classify sentiments on a sentence level, by Breck (2007) to identify subjective expressions, and by Li (2010) to summarize product reviews taking their structure into account.

3 Motivation behind Top-down Approach

Though most of the previous studies treat product reviews as a bag of sentences or even words, in fact they are narratives that have a specific structure. While their structure is less rigid and

[2] http://sentiwordnet.isti.cnr.it/ (Baccianella et al., 2010)

[3] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/ (Wilson et al., 2005)

predicable than, say, that of research papers, it nevertheless has some recurring patterns which lend themselves to generalization.

The same principle applies to sentiments appearing in product reviews. The authors of reviews do not simply pile up some random facts about the product or their evaluations of it. To the best of their abilities, they try to convince the reader to buy or not to buy a particular product, and, according to Grice's Maxims (Grice, 1975), they do it in the clearest and most effective way possible. Thus, if an author has in general a positive opinion of a product, the probability of a negative sentence appearing in a review is lower than that of a positive sentence, and even if a negative sentence is introduced, it is likely to appear together with a concession or a contrast marker, such as *although* or *but*, or be modified by a hedging expression, such as *might, only, could* be, which mitigate the negative effect on the reader. Thus the author makes us understand that his primary opinion of the product is still positive, and uses the discourse relation of *contrast* to present an opposite opinion.

Likewise, if an objective sentence appears in a review, it is not a random event, but a tool serving some purpose, such as interacting with a reader by asking questions which do not require an answer (*Where do I start?*) or supporting one's view by showing that you have some expertise necessary to provide a valid opinion. While in this paper we cover only objective sentences that are used to provide background information (the discourse relation of *background*), it is clear that other reasons for usage of objective sentences are present and capable of being formalized.

The facts mentioned above make us consider a product review as a text which has a primary polarity and optionally includes some segments which have an opposite polarity or no polarity at all (objective sentences). Instead of relying on lexical sentiment information, which makes it difficult to distinguish between objective and subjective sentences on one hand (implicit opinions) and between positive and negative opinions on the other (sarcasm, context-dependent polarity), we suggest using a top-down approach: determining the primary polarity of a review based on an external source of information, such as product rating, and then locating segments which do not conform with this polarity (have no polarity or an opposite polarity) by finding cues that mark a change in a discourse flow.

In the next section we describe an experiment which we conducted to confirm that this approach is viable.

4 Experiment

4.1 Data and Task

For the experiments in this study we use Filatova's Amazon product reviews corpus (so called Sarcasm Corpus[4]), consisting of 817 ironic and regular reviews. We chose to use this corpus because we believed that segments in ironic reviews would be difficult to classify by purely lexical means. Out of these 817 reviews we randomly selected 100 reviews for training and 20 reviews for test data. We did not use the whole dataset because the number of reviews with a particular review score differs greatly (60% of reviews are 5-star, while only 5% are 2-star). To prevent a skew towards positive labels we used equal-size random samples of reviews with all possible scores. Reviews were annotated by one of the authors and an external annotator on a clause level if a sentence contained opinions with opposite polarities, and on a sentence level otherwise. The inter-annotator agreement was measured by using Fleiss' kappa and Krippendorff's alpha, and the results showed that the annotation was highly reliable ($\kappa = 0.912$, $\alpha = 0.913$). Overall the training corpus consisted of 843 segments (438 negative, 268 positive and 137 objective), while the test set contained 145 segments (78 negative, 41 positive and 26 objective).

While the studies in sentiment analysis usually make distinction between subjective and objective sentences on one hand and between negative, positive and neutral sentences on the other, in this paper we make a twofold distinction, first classifying a segment as objective or subjective, and then, in case of subjective (polar) sentences, further subdividing them into positive and negative. To our mind the classification into positive, negative and neutral sentences, commonly adopted for product reviews, is incorrect, as neutral sentiments rarely, if ever, appear in reviews. What is com-

[4] http://storm.cis.fordham.edu/~filatova/SarcasmCorpus.html (Filatova, 2012).

monly referred to as neutral sentences should be classified as objective segments, as they do not carry any sentiment related to the subject matter.

When annotating the corpus we considered the intended semantic orientation of a segment, not its literal meaning and the presence and polarity of lexical cues. Thus, segments without any lexical cues could be annotated both as subjective and objective:

(4) I bought this mobo from Amazon, after buying the same month the DG31PR Classic for my wife. (*objective*)

(5) After I install my new PC, the 2do. day of use, the LAN failed. (*subjective, negative*)

Segments with a lexical cue of a certain polarity could be annotated both as positive and negative:

(6) The ring is **nice** and heavy. (*positive*)

(7) It's going to be a **nice** paperweight. (*negative*, from a review of a camera)

Finally, segments where an alternative product was praised or preferred were understood to be a criticism towards the reviewed product:

(8) I will never buy another Panasonic product. There are plenty of other brands that are loyal to their customers. (both segments are *negative)*

We view each of the reviews as a separate discourse with its own sentiment flow, and thus treat the sentiment analysis problem as a sequence classification task. We employ the CRF method, which outperforms other methods of sequence labeling (Lafferty, 2001). In CRFs the probability of a sequence is defined as

$$p_\lambda(Y \mid X) = \frac{\exp \lambda \cdot F(Y, X)}{Z_\lambda(X)}$$

where

$$Z_\lambda(x) = \sum_y \exp \lambda \cdot F(y, x)$$

where X is a set of input random variables, Y is a set of labels, and λ is a weight for the feature function $F(Y,X)$. (Sha, 2003).

All experiments in this paper were conducted using a C++ implementation of a linear Conditional Random Fields classifier (CRF++)[5]. Though more complex or constrained types of CRF classifiers and models based on them proved to be more suitable for sentiment analysis (Mao, 2006; Yang, 2014), we use the simplest model as a proof of concept in this study.

Each review in the training and test data is converted into a sequence of polarity segments assigned to it. For example, the following short review:

(9) The ring is nice and heavy. Have been wearing if for almost a month and still not a scratch!

is presented as a sequence of tokens POSITIVE POSITIVE, based on the sentiment labels from the annotation. The tokens are assigned features, as defined in the following sections, which are then fed into the classifier.

4.2 Features for Experiments

4.2.1 Lexical Features

To set a baseline, we use a state-of-art lexical classifier – Stanford Sentiment Analysis Classifier from Stanford CoreNLP toolkit[6] – to determine the lexical polarity of each individual sentence. Thus the *lexical* classifier considers only lexical features available in a particular sentence without looking at neighboring sentences or discourse cues. For the *local context classifier* we also determine the lexical polarity of the previous and next sentences and use the sequence of {prev_polarity, current_polarity, next_polarity} as a feature (a similar approach is taken by Somasundaran (2010)). This is done to disambiguate and, if necessary, to correct the polarity of misclassified instances that are sandwiched between the correctly classified ones. For example, if the lexical classifier fails to detect an implicit opinion

[5] Available from http://taku910.github.io/crfpp/
[6] Available from http://nlp.stanford.edu/sentiment/code.html

in a sentence that appears between two explicit opinions, it might correct it as follows:

POSITIVE OBJECTIVE POSITIVE ->
POSITIVE POSITIVE POSITIVE

4.2.2 Contrast Features

The main drawback of the *local context* classifier is that it can misclassify sentences with the opposite polarity, lumping them together with sentences of the primary polarity. To prevent this, for the *contrast* classifier we add another set of features – discourse cues with a Rhetoric Structure Theory (RST) (Mann and Thompson, 1988) relation of *contrast*. We consider both explicit and implicit discourse markers of contrast for this set of features:

4.2.2.1 Explicit Contrast Markers

Contrast relations are primarily realized by using explicit *discourse markers*, which, depending on their type, mark the sentence they appear in (in case of *although* type) or the previous sentence (in case of *but* type) as contrasting:

(10) The Phillips screwdriver on the end of one of the tines is helpful for things like tightening eyeglasses, POSITIVE CONTR
but it is slightly offset from the opposing blade and I've nicked or jabbed myself with it more than once while it's in my pocket. NEGATIVE NCONTR

(11) **Although** it has 10 workable buttons which come in handy for some games, POSITIVE CONTR
it has some major flaws. NEGATIVE NCONTR

The segment with the NCONTR marker usually has the primary polarity of the review, while the segment with a CONTR marker presents a contrasting opinion.

4.2.2.2 Implicit Contrast Relations

Contrast relations can also be manifested implicitly through the use of *hedges*. Hedging is often used when the review's author wants to mention some negative side of a product they like (or a positive aspect of a product they hate), but does not want to put an unnecessary emphasis on it. Such hedging expressions as *the only good/bad point, the only drawback, I would only recommend it*… etc are used for this purpose:

(12) With all the upgrades that Apple has done with their macbooks, I think I finally feel that it's worth the spending to buy my first mac. NHEDGE
My *only complain* is that it's still a lot more expensive than PC laptops with similar specs. HEDGE

4.2.3 Background Classifier

The *background* classifier allows us to capture some of the objective sentences that are related to the polar ones using a *background* RST relation. We identify three types of patterns where background relations are used:

1. *Acquirement patterns*: people often start reviews with an explanation of how they got the product.
2. *Personal background patterns*: people often support their evaluation of a product by stating who there are, what they do for a living, what kind of lifestyle they lead etc.
3. *Personal experience patterns*: again, to support their views the writers prime their readers by describing their experiences or achievements.

Unfortunately, background relations, unlike contrast relations, are almost never explicit. They are paratactic and lack discourse cues, so we need to rely on lexical and grammatical features for classification. However, we believe that because background information is usually presented in easy-to-predict patterns, it is more feasible and computationally inexpensive to use lexical cues to single out objective sentences than to try to capture infinitely large number of ways sentiments can be expressed. In the following subsections, we describe these patterns in more detail and explain which lexical and grammatical cues can be used to detect them.

4.2.3.1 Acquirement Patterns

At the beginning of a review people often explain how they acquired the product:

(13) I bought this camera for my deployment to Iraq. (*objective*)
It was in my cargo pants pocket one day I took it out and the lens was cracked and the silver trim ring had fallen off. (*negative*)

We formalize this feature as follows:

[I | we] [verb synonymous to "acquire"|verb of decision + verb synonymous to "acquire"],

or, more specifically:

[I | we] [ordered | bought | got .* as a gift | purchased | decided to buy…]

All verbs are in past simple tense, as only in this tense they are unlikely to bear any sentiment (compare, for example, sentences with the same verbs in present perfect tense:

(14) However, I am glad that I **have bought** a mac.(*positive*)

(15) This is probably the worst book **I've bought.** (*negative*)

Moreover, this pattern is likely to be used at the beginning of the review, so we add ACQUIRE feature only to those segments which appear in the first 25% of a review.

4.2.3.2 Personal Background Patterns

In these patterns, the authors offer their personal information that is relevant to the subject matter of the review and can support their opinion. For instance, in the following review the author refers to his pets as the major reason for buying a particular vacuum cleaner:

(16) I **have a cat and a dog**, and there is lots of shedding hair, all the time. (*objective, personal background*)
When I saw the DC25 Animal, I **decided to spend the money** hoping that this vacuum would do the job. (*objective, acquirement*)
It has lived up to my wildest dreams, it is wonderfully easy to handle, so easy to maneuver, the 16 lbs make such a difference compared to those very heavy machines I had before, I had no problem carrying it upstairs. (*positive*)

We formalize this feature as follows:

[I|we] [am (a|an)|have (a|an)|'m (a|an)|am not (a|an)]

The indefinite article is used to prevent matching such polar expressions, as *I'm very pleased with the quality of this product.* Again, such patterns are searched for only at the beginning of a review.

4.2.3.3 Personal Experience Patterns

These patterns also serve to provide some background information about user's experiences to back up his opinion on a product:

(17) Usually **I am a huge fan** of hats that look like food. (*objective, personal background*).
My meatloaf hat **has been** a hit for years. (*objective, personal experience*)
When I received my turkey hat I carefully unwrapped the bubble wrap and gazed upon its tan beauty. (*positive*).

To capture this pattern we search for verbs in perfect forms (except for the verbs of possession). We exclude verbs in perfect continuous forms, as they are more often used to describe positive or negative results of using a product. Compare, for example:

(18) I **have been using** it for almost a month and my lashes are so long, they touch my eyebrows... (*positive*)

We also exclude phrases that have "should/could" before "have", as they often express negative sentiments (Liu, 2014):

(19) Would have been nice if the stilts could accommodate multiple/varying heights. (*negative*)

4.2.4 Primary Polarity Features

Lastly, we use reviews' scores to predict their global semantic orientation (primary polarity). The intuition behind this is that the reviews with a higher score will contain more positive sentences than reviews with a lower score, and thus global polarity information might help us to amend incorrect predictions of a lexical classifier (a similar approach was taken, among others, by (Yang, 2014)). This is supported by the statistics of our corpus: the polarity of sentences in a review in general correlates with its score. Highly positive (5-star) and highly negative (1-star) reviews contain few segments of the opposite polarity, and even reviews with a less extreme score demonstrate a clear preference of one of the polarities (see Table 1). Thus it can be predicted that the classifier using this feature will tend to assign the primary polarity (positive for 4- and 5-star reviews, negative for 1-, 2-, and 3-star reviews) unless there is some strong evidence preventing it.

Review score	Positive	Negative	Objective
1	0.01	0.85	0.13
2	0.10	0.77	0.12
3	0.22	0.65	0.13
4	0.62	0.23	0.15
5	0.68	0.04	0.27

Table 1. Percentage of positive, negative and objective sentences in reviews with different product ratings

4.3 Bottom-up vs Top-down Approach: Experiment Design

4.3.1 Bottom-up Approach

This is a widely-used approach which relies on a lexical polarity classifier to determine the semantic orientation of each segment and then corrects the misclassified segments by employing more general features: discourse features (in our study – *contrast* and *background*) and global semantic orientation features (called *primary polarity* features in this paper).

The bottom-up approach has become a standard in sentiment analysis, so we believe there is no need to explain it in detail. The main focus of this study is on the top-down approach, which we describe below.

4.3.2. Top-down Approach

In this set of experiments, we do not use any lexical information about the presence of sentiments in segments and their types. Instead, we rely on rating scores assigned to the reviews to determine their primary polarity, and then correct the misclassified instances using discourse features. In general, the feature set used for this classifier is the same as for the bottom-up approach. The only important exception is that lexical features are completely omitted.

We adopt the following process for sentiment classification:

1. All sentences in a review are assigned a polarity label determined by the corresponding review rating.
2. We look for discourse patterns that are associated with a change of the primary polarity (POSITIVE -> NEGATIVE, NEGATIVE -> POSITIVE). These are usually manifested through **contrast** relation and enable us to correct some of misclassified polarity labels.
3. We look for discourse patterns where a polar statement is accompanied by an objective statement. A common example of such discourse relations in product reviews is **background**. At this stage, unnecessary POSITIVE and NEGATIVE labels are changed into OBJECTIVE.

Schematically this can be shown as follows using an arbitrary example of a 4-star review, where light-gray blocks stand for positive segments, dark-gray for negative segments and white for objective ones (here we assume that all segments will be initialized as positive, as it is the primary class for 4-star reviews):

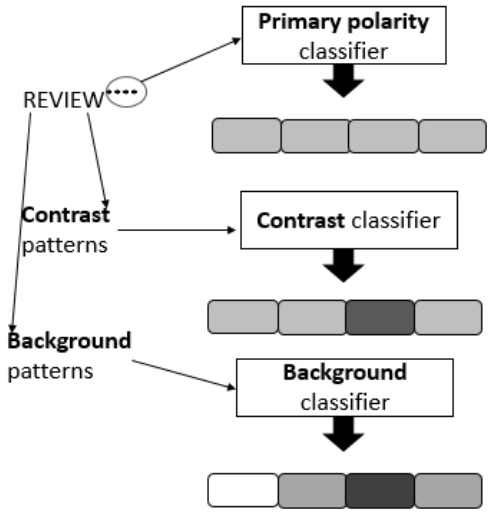

Figure 1. Top-down approach classification flow

In the next section, we discuss the results of experiments conducted to show that such features can improve sentiment classification.

5 Experiment Results

In this section, we compare the results achieved by using the top-down approach with those of the traditional bottom-up method.

5.1 Bottom-up Approach

The *lexical* classifier results, used as a baseline in this study, are listed in Table 2 below. As can be seen from the results, the recall and precision of positive and especially objective segments is low, which shows that purely lexical classifier cannot reliably distinguish between objective and subjective sentences and between positive and negative ones. The accuracy of the classifier is also low (0.6138).

When we add local discourse context, the recall of positive and negative segments improves, as does the overall accuracy (to 0.6758). However, the *local discourse* classifier completely ignores the objective sentences, assigning polarity to them. The precision of the negative class and overall precision also gets lower, as some positive segments sandwiched between negative ones are assigned a negative label.

Adding *contrast* discourse cues does not help to improve this situation, because it leads to overestimation of positive segments and lower accuracy (0.6620). In fact, the *contrast* classifier performs even worse than the local discourse one. It seems that lexical information introduces too much noise, and building up on such an unreliable basis does not produce expected results.

The *background* classifier improves the performance, especially for objective sentences, classifying them with a high precision and at least some recall. It also improves the overall accuracy (to 0.6896).

However, the most significant improvement is seen after adding the review scores (*primary polarity*) as features. It helps improve almost all scores, including accuracy, which reaches 0.7241.

5.2 Top-down Approach

The *primary polarity* classifier, which uses the review's rating to predict its overall polarity, has a high recall and an accuracy of 0.7379 (see Table 3). However, it again ignores the objective class, because it is distributed more or less evenly between reviews with different ratings and thus cannot be correlated with a particular review score.

	Subjective				**Objective**		**Total**			
	Negative		**Positive**							
	Prec	**Rec**	**Prec**	**Rec**	**Prec**	**Rec**	**Prec**	**Rec**	**F1**	**Acc**
Lexical	0.71	0.77	0.61	0.54	0.29	**0.27**	0.60	0.61	0.61	0.6138
Local discourse	0.69	**0.87**	**0.64**	0.73	0.00	0.00	0.55	0.68	0.61	0.6758
+ Contrast	0.69	**0.87**	0.62	0.68	0.00	0.00	0.55	0.66	0.60	0.6620
+ Background	0.73	0.85	0.60	0.75	**1.00**	0.12	0.74	0.69	0.65	0.6896
+ Primary pol.	**0.78**	**0.87**	0.62	**0.82**	**1.00**	0.12	**0.77**	**0.72**	**0.68**	**0.7241**

Table 2. Precision, recall, F1 and accuracy scores for the bottom-up approach

	Subjective				Objective		Total			
	Negative		Positive							
	Prec	**Rec**	**Prec**	**Rec**	**Prec**	**Rec**	**Prec**	**Rec**	**F1**	**Acc**
Primary polarity	0.75	0.94	0.71	0.83	0.00	0.00	0.61	0.74	0.66	0.7379
+ Contrast	0.81	0.90	0.68	**0.95**	0.00	0.00	0.63	0.75	0.68	0.7517
+ Background	**0.82**	**0.91**	**0.73**	**0.95**	**1.00**	**0.19**	**0.83**	**0.79**	**0.76**	**0.7931**

Table 3. Precision, recall, F1 and accuracy scores for the top-down approach

Also, because 3-star reviews contain more negative sentences than positive ones, all of them are lumped into the negative class. Thus the recall for the positive class is substantially low than for the negative class.

To single out the segments whose polarity is different from the primary one, we add explicit and implicit *contrast* features and train the contrast classifier. *Contrast* features help to raise recall for positive and precision for negative sentences, though, as might be expected, they do not affect the classification of objective segments. However, the overall precision and recall are improved, as well as the overall accuracy (to 0.7517)

The *background* classifier allows us to find some of objective sentences. *Acquirement, personal background* and *personal experience* patterns turn out to be precise features that also guarantee us at least some recall for objective sentences. Overall precision, recall and F1 scores are improved accordingly, as well as accuracy (to 0.7931).

As can be seen from comparing these results, even the most primitive rating-based classifier (*primary polarity*) achieves better recall and accuracy than any of lexical classifiers (even the one with primary polarity features). Moreover, adding discourse features to it consistently improves the results, allowing us to build a high-precision, high-recall sentiment classifier. On the other hand, building up on the lexical classifier does not show such consistent improvements.

6 Conclusion

Until now the sentiment analysis has been primarily done in a bottom-up way, starting with the classification of lexical items, then resolving the polarity of the sentence, then using discourse information to improve the lexical classification. However, lexical classifiers so far produce results that are too unreliable to become a basis of a discourse-level classification. We assert that starting from the top by roughly defining a text's polarity and assigning it to all its segments, and then fine-tuning the classification by "chiseling out" incorrect bits based on reliable discourse relations can be a more productive and effective approach. Our experiments show that such approach can lead to a substantial improvement over lexical baseline at least in texts with a predictable structure and recurring patterns, such as product reviews. Because each of the discourse features we tested led to improvement, we believe that the top-down classifier can be made even more accurate by employing other discourse relations in the form of carefully selected linguistic features.

References

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *LREC, 10*, 2200-2204.

Breck E., Choi Y., & Cardie C. (2007). Identifying Expressions of Opinion in Context. *IJCAI, 7*, 2683-2688.

Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. *WSDM*, 231-240.

Filatova, E. (2012). Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing. *LREC*, 392-398.

Grice, P. (1975). Logic and conversation. *Syntax and semantics, 3*, 41-58.

Hu, M., & Liu, B. (2004). Mining opinion features in customer reviews. *AAAI, 4*(4), 755-760.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning*, 282-289.

Lapponi, E., Read, J., & Ovrelid, L. (2012). Representing and resolving negation for sentiment analysis. *Proceedings of the 2012 ICDM Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction*, 687-692.

Li F., Han C., Huang M., Zhu X., Xia Y. J., Zhang S., & Yu H. (2010). Structure-aware review mining and summarization. *Proceedings of the 23rd International Conference on Computational Linguistics*, 653-661.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Liu, Y., Yu, X., Liu, B., & Chen, Z. (2014). Sentence-Level Sentiment Analysis in the Presence of Modalities. *Computational Linguistics and Intelligent Text Processing*, 1-16.

Mann, W. & Thompson, S. (1988). Rhetorical structure theory: Towards a functional theory of text organization. *Text*, *8*(3), 243-281.

Mao, Y., & Lebanon, G. (2006). Isotonic conditional random fields and local sentiment flow. *Advances in neural information processing systems*, 961-968.

Narayanan, R., Liu, B., & Choudhary, A. (2009). Sentiment analysis of conditional sentences. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 1*, 180-189.

Polanyi, L., & Zaenen, A. (2006). Contextual valence shifters. *Computing attitude and affect in text: Theory and applications*, 1-10.

Sha, F., & Pereira, F. (2003). Shallow parsing with conditional random fields. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, *1*, 134-141.

Somasundaran S. (2010). *Discourse-Level Relations for Opinion Analysis* (Doctoral dissertation). University of Pittsburgh.

Tang, H. F., Tan, S. B., & Cheng, X. Q. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications*, *36*(7), 10760-10773.

Toprak, C., Jakob, N., & Gurevych, I. (2010). Sentence and expression level annotation of opinions in user-generated discourse. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 575-584.

Tsur, O., Davidov, D., & Rappoport, A. (2010). Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. *ICWSM*, 162-169.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the conference on human language technology and empirical methods in natural language processing*, 347-354.

Yang, B., & Cardie, C. (2014). Context-aware learning for sentence-level sentiment analysis with posterior regularization. *Proceedings of ACL*.

Zhao J., Liu K. & Wang G. (2008). Adding Redundant Features for CRFs-based Sentence Sentiment Classification. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 117-126.

Zhou L., Li B., Gao W., Wei Z. & Wong K. F. (2011). Unsupervised Discovery of Discourse Relations for Eliminating Intra-sentence Polarity Ambiguities. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 162-171.