Classification of Online Reviews by Computational Semantic Lexicons

Boris Kraychev[1] and Ivan Koychev[1,2]

[1]Faculty of Mathematics and Informatics,
University of Sofia "St. Kliment Ohridski", Sofia, Bulgaria
[2]Institute of Mathematics and Informatics at Bulgarian Academy of Sciences
{kraychev, koychev}@fmi.uni-sofia.bg

Abstract. The paper presents a method for opinion polarity classification of online reviews, which includes a web crawler, part of speech tagger, constructor of lexicons of sentiment aware words, sentiment scoring algorithm and training of opinion classifier. The method is tested on 500 000 online reviews of restaurants and hotels, which are relatively simple and short texts that are also tagged by their authors and limited to a set of topics like service, food quality, ambience, etc. The results from conducted experiment shows that the presented method achieves an accuracy of up to 88%, which is comparable with the best results reported from similar approaches.

Keywords. Opinion Polarity Classification, Sentiment Analysis, Natural Language Processing.

1 Introduction

The task of large scale sentiment analysis draws increasing research interest in recent years. With the rise of the social networks and different types of web media like forums, blogs, video sharing, it became very important to develop methods and tools that are able to process the information flow and automatically analyse opinions and sentiment from online texts and reviews. Such analysis has various applications in the business and government intelligence and the online public relationships.

The paper presents a method that builds semantic lexicons for online review polarity classification. It includes building a sentiment aware dictionary, morphological approaches for feature extraction, label sequential rules, opinion orientation identification by scoring and linear regression algorithms. The method was implemented and tested with 500 000 recent online user reviews about restaurants and hotels.

The domain of restaurant and hotel reviews suggests the usage of feature oriented analysis because customers are discussing few aspects like food, service, location, price, and general ambiance. Our goal is to estimate the sentiment polarity using multiple approaches. We built two independent lexicons: the first consisting of sentiment aware parts of speech and the second one representing evaluation pairs of adjectives and nouns extracted from the reviews. The second lexicon actually represents a set of extracted features from the online reviews.

We implemented the above method and conducted experiments with 500 000 recent online user reviews about restaurants and hotels. As a result our sentiment classifier achieves an 88% of accuracy, which can be considered as very good result, given that the raw online data contains spam reviews and human errors in the self assessment (the number of 'stars' assigned by the author to the review).

2 Related Work

Recently, the area of automated sentiment analysis has been very actively studied. Two major streams of research can be distinguished: The first relates to the building of sentiment aware lexicons and the second group consists of the work on complete sentiment analysis systems for documents and texts.

The early works in this field has been initiated by psychological researches in the second half of twentieth century (Deese, 1964; Berlin and Kay, 1969; Levinson, 1983) which postulated that words can be classified along semantic axes like "big-small", "hot-cold", "nice-unpleasant", etc. This enabled the building of sentiment aware lexicons with explicitly labelled affect values.

The recent work on this subject involves the usage of statistical corpus analysis (Hatzivassiloglou and McKeown, 1997) which expands manually built lexicons by determining the sentiment orientation sentiment orientation of adjectives by analyzing their appearance in combination with adjectives from the existing lexicon. Usually adjectives related with "*and*" like the clause *"The place is awesome and clean"* suppose that both adjectives have the same orientation, while the conjunction with "*but*" supposes that the adjectives have opposite orientation.

Other recent research is made by Grefenstette, Shanahan, Evans and Qu [4] [7] with exploration of the number of findings by search engines where an adjective, supposed to enter the lexicon is being examined towards a set of other well determined adjectives over several semantic axes. The authors consider that adjectives would appear more frequently closer to their synonyms and their sentiment orientation can be determined statistically by the number of search engine hits where the examined word appears close to any of the seed words.

The movie reviews have been a subject of research for Pang, Lee and Vathyanathan [8] and Yang Liu [2]. The first system achieves an accuracy of roughly 83% and shows that machine learning techniques perform better than simple counting techniques. The second system implements linear regression approaches, (an interesting introduction in that area is presented by C. Bishop[1]) and combines the box office revenues from previous days, together with the people's sentiments about the movie to predict the sales performance of the current day. The best results of the algorithm achieve an accuracy of 88%.

Some of the authors as Pang [8] try to separate the text on factual and opinion propositions, while other as Godbole [6] considers that both mentioned facts and opinions contribute to the sentiment polarity of a text.

Other approach for product reviews is the feature-based sentiment analysis explored by B. Liu, Hu and Cheng [9] which extracts sentiment on different features of the subject. The techniques used are Label Sequential Rules (LSR) and Pointwise

Mutual Information (PMI) score, introduced by Tourney [10]. General review of the sentiment analysis methods is made by Pang and Lee [3] in 2008.

A recent approach is proposed by Hassan and Radev [13] in 2010 which determines the sentiment polarity of words by applying Markov random walk model to a large word relatedness graph where some of the words are used as seeds and labelled with their sentiment polarity. To determine the polarity of a word the authors generate Markov random chains, supposing that walks started from negative words would hit first a word labelled as negative. The algorithm has excellent performance and does not require large corpus.

Our approach for the current experiment is to use scoring algorithms, enhanced by sequential rules in order to improve the sentiment extraction for the different estimation axes for restaurants and perform the polarity classification by standard machine learning algorithms, based on numerical attributes, issued from the scoring process.

3 Sentiment Lexicon Generation and Sentiment Analysis

We apply two algorithms which, to our knowledge, have not been explored until now. The first one is the expansion of the dictionary through WordNet by keeping the sentiment awareness and positivity value by applying a histogram filter from the learning set of text. The second is the discovery of propositional patterns, determined as label sequential rules using relatively large test set of online reviews (250 000).

The major processing steps of our sentiment analysis system are:

1. *Construction of lexicons of sentiment aware words.* Actually all major sentiment analysis systems rely on a list of sentiment aware words to build initial sentiment interpretation data. We developed the following dictionaries of sentiment aware words and pairs of words.
   (a) *Lexicon of sentiment aware adjectives and verbs* - a manually built list of seed words, expanded with databases of synonyms and antonyms to a final list of sentiment aware words.
   (b) *Lexicon of sentiment aware adjective-noun pairs.* It is obtained with feature extraction techniques using propositional models and Label Sequential Rules (LSR) introduced by [9]. LSR discover sequential patterns of parts of speech. They are very effective extracting the sentiment for specific features, mentioned in the review.
2. *Sentiment scoring algorithms.* We are using scoring techniques to calculate a list of attributes per review. The aim is to build numerical depiction of the sentiment attributes of the text, taking care of negation, conditionality and basic pronoun resolution. The reviews represented in this attribute space are passed to the machine learning module.
3. *Opinion polarity classification.* We trained Machine learning algorithms based on attributes provided by the scoring algorithm then we evaluated the performance of the learned classifiers on new reviews.

3.1 Determining Lexicon Seeds and Lexicon Expansion through WordNet

We sorted the parts of speech from the training set to find out the most frequently used ones. Then we manually classified adjectives and verbs as seeds for future classification expansion. This forms our seeds for future lexicon development.

We used WordNet to expand the dictionary with synonyms and antonyms. It is well known that WordNet offers a very large set of synonyms and there are paths that connect even good and bad as synonyms, so we limited the expansion to two levels and applied a percentage to decrease the confidence weight of words found by that method.

Significance weight for lexicon expansion through WordNet is calculated with a method proposed by Godbole [6]. The significance weight of a word is equal to $w = 1/c^d$, where c is a constant > 1 and d is the distance from the considered to the original word. The expansion is planned in two stages – the first stage is to simple enlarge the dictionary by the 1^{-st} and 2^{-nd} level synonyms of words, then as a second stage – apply a filter on the resulting words to eliminate words ending in contradictory positivity assessment. This can happen by building a histogram for each word over the sentiment tagged reviews from the learning set. We exclude the words having different histogram than their corresponding seeds. The final polarity weight is calculated as follows: for a given term we can mark with p the appearances in positive texts, with n the appearances in negative texts and with P, N and U the total number of positive, negative and neutral texts, respectively. The polarity weight is then calculated by the equation $polarity\_weigth = \frac{p - n}{P + N + U} w$.

Unknown words which are not mentioned in the learning set are kept with the weight of their first ancestor with calculated weight, multiplied by a coefficient between 0 and 1 following the formula above. In our case the value chosen was 0.8 e.g. $c = 1.25$ and words without clear evidence in the learning set were kept with decreased weight by 20%.

3.2 Lexicon Generation with Label Sequential Rules

The label sequential rules [9] provide a method for feature extraction and discovery of common expression patterns. Our targeted area of short online reviews suggests that people would follow similar expression models. The label sequential rules are mapping sequences of parts of speech and are generated in the following form:

```
{$feature,noun}{(be),verb}{$quality,adjective}
[{and,conjunction}{$quality,adjective}] => 90%
{$actor,pronoun}{*,verb}
[{*,determiner}]{$feature,noun} => 90%
```

where the square brackets indicate that the part is non mandatory and each rule has a confidence weight to be considered further. The conjunctions '*and'* and '*but'* in the phrases were used to enlarge the lexicon with adjectives having similar or opposite sentiment orientation. It is important to note that the LSR method allows splitting the analysis to features and further summarize and group the reviews by features.

The construction of LSR patterns is important part of the learning algorithm. By sorting all N-term part-of-speech sequences, the ones which frequency is over a predefined threshold are kept and added to LSR knowledge base, declaring the nouns as features and the adjectives and verbs as sentiment positivity evaluators.

3.3 Methods for Sentiment Analysis

Our sentiment analysis algorithm is based on sentiment aware term scoring which is then evaluated by machine learning algorithms.

The scoring algorithm determines sentiment aware terms in text and assigns their sentiment weight in the dictionary of sentiment aware words. The weight values are real numbers, positive or negative according to the determined sentiment orientation. The algorithm takes into account negation like "not, don't, can't" and inverses the relative weight value. It also takes care of simple conditional propositions like '*if the staff was polite, I would…*' and applies a simple technique for pronoun resolution. For our results we rely on the fact that short online reviews are kept simple and the lack of profound conditionality and pronoun resolution analysis would not impact our final results. We have to admit that these modules could be improved further.

The final result of the scoring algorithm is a set of weight sums, counts and expression of previously estimated values that would facilitate further machine learning classification.

With this set of attributes, we obtained a regular problem for machine learning which we explored in our experiments.

4 The Sentiment Analysis Experiment

4.1 Design

Our experiment involves the following steps:

1. Web crawling to collect online reviews and their self assessment by their authors.
2. Part of speech analysis to all acquired texts using MorphAdorner [11].
3. Sorting the data from the test set to determine the seed words and LSR patterns for the generation of the lexicons.
4. Generation of the lexicons by expansion through WordNet [5] and LSR extraction [12].
5. Numerical representation of the texts by scoring sentiment aware words.
6. Experiments with machine learning algorithms over the attributes' space.

The goal of the experiment is first to extract live data from the web, then analyze the contents and extract seed words and patterns for lexicon generation.

The final sentiment analysis consists of calculating numerical attributes like sum of weighted positive/negative items, count of contradiction related words and mathematical expressions using previously calculated parameters. The expressions are actually forming the scores that can be assessed. The sentiment polarity classification is then performed in the environment for machine learning benchmarking WEKA.

4.2 Determining the Positive and Negative Weights of the Text

The sum of the weights of positive and negative items in the text forms the first two classification attributes: `PosW` and `NegW` respectively. We obtain these sums by the scoring algorithm which identifies the sentiment aware words and phrases from both lexicons. It also counts the negations, conditionality and pronoun resolution, and procedure the `Contr` attribute. For example if the word is preceded by negation like 'not', 'don't', 'can't' the polarity of the item is exchanged. For example 'not good' goes to the sum of negative words instead of the one for positive, with its default weight. The Table 1 describes the final list of attributes.

Table 1: The list of attributes passed to the machine learning algorithm.

Attribute	Description	Implementation
`PosW`	Σ *of the weights of positive items*	Scoring algorithm
`NegW`	Σ *of the weights of negative items*	Scoring algorithm
`Contr`	*Count of contradiction elements*	Scoring algorithm
`score1`	$f(posw, negw)$	`{posw}+{negw}`
`score2`	$f(posw, negw)$	`{posw}+2*{negw}`
`score3`	$f(posw, negw)$	`2*{posw}+{negw}`
`score4`	$f(posw, negw, contr)$	`{posw}+{negw}-{contr}`

4.3 Results of Sentiment Polarity Classification with WEKA

In order to be able to experiment with more machine learning algorithms we added supplementary attributes, formed by the original three ones. The most evident one is a simple addition of the positive weight and the negative weight (they have indeed opposite signs) which forms a simple score of positive minus negative items in the text. We also experimented with doubling the value of negative or positive items to handle the fact that reviewers might tend to give more strength on one of these groups.

The classification through three machine learning algorithms gives the results shown in. The accuracy of 87-88% is satisfying our expectation because our raw review data contains classification errors. The estimation of the classification errors should be explored further and requires voluminous manual data revision.

Table 2: Results by different machine learning algorithms

Algorithm	Accuracy	Precision
NaiveBayes	87%	87%
VotedPerceptron	83%	69%
ADTree	88%	87%

5 Discussion: Thumbs Up or Thumbs Down for Restaurants

The sentiment classification tasks vary for different domains. In the current experiment we showed that sentiment analysis algorithms can perform better when it is restricted to particular domain, where it is easier to perform feature extraction algorithm. Interesting results can be obtained by examining the expressed sentiment over all scanned reviews of UK restaurants by features as food, staff, ambiance, etc.

We should note that restaurants are a very competitive domain and reviewers are attentive to all details. The feature that annoys most of the clients is the non-politeness of the staff. Next to it stands the quality of the food and the price comes as the third most bothering feature.

If we count the general customer sentiment about all evaluated restaurants we should conclude ‘Thumbs up’ because the bigger part of expressed reviews and features are positive.

6 Conclusion

In the present work we built method for online review classification, which was tested on a large data set of UK restaurant reviews. The approach constructs a lexicon of sentiment aware words and phrases over the application domain. Then it estimates the sentiment polarity by applying scoring techniques over the reviews and providing the results to machine learning algorithms. The final classification is made using machine learning algorithms from the WEKA environment.

The results are showing a clear path to follow – topic related sentiment analysis is a prominent area where automatic sentiment classification can be considered as effective and robust monitoring tool. Future researches could include demographic and geographic data to show peoples’ preferences and provide deeper analysis.

Future work might include improvement of the scoring algorithm – better pronoun resolution, improvement in the detection of conditional propositions. The generation of the lexicon of sentiment aware words could be improved in the area of feature extraction by implementing more sequential rules and detecting more part-of-speech patterns. Last but not least the lexicon building algorithm could be applied on different topic areas like sentiment analysis of reviews of movies, books, news stories, and certainty identification in text.

Acknowledgements. This research is supported by the SmartBook project, subsidized by the Bulgarian National Science Fund, under Grant D002-111/15.12.2008.

References

1. Bishop, C.M. Pattern Recognition and Machine Learning, Springer (2006).
2. Liu, Y. Review Mining from Online Media: Opinion Analysis and Review Helpfulness Prediction for Business Intelligence. VDM Verlag (2010).
3. Pang B., Lee, L. Opinion Mining and Sentiment Analysis, now Publishers Inc. (2008).

4. Shanahan, J. G., Qu Y., Wiebe J. (Eds.). Computing Attitude and Affect in Text: Theory and Applications, Springer (2006).
5. Witten, I. H., Frank, E. Data Mining. Practical Machine Learning Tools and Techniques, Elsevier (2005).
6. Godbole, N., Srinivasaiah, M., Skiena, S. Large-scale Sentiment Analysis for News and Blogs, Int. Conf. on Weblogs and Social Media ICWSM (2007).
7. Grefenslette, G. Qu, Y., Evans, D.A., Shanahan, J. G. Validating the Coverage of Lexical Resources for Affect Analysis and Automatically Classifying New Words along Semantic Axes, Springer (2006).
8. Pang, B., Lee, L., Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques, Proceedings of the 2002 Conference on Empirical Methods of Natural Language Processing (EMNLP) (2002).
9. Liu, B., Hu, M., Cheng, J. Opinion Observer: Analyzing and comparing opinions on the web, Proceedings of WWW (2005).
10. Tourney, P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, Proceedings of the Association for Computational Linguistics (ACL) (2002).
11. MorphAdorner Part of Speech Tagger, http://morphadorner.northwestern.edu/morphadorner/postagger/
12. Miller, G.A. WordNet: A lexical database. Communications of the ACM 38(11),(1995)
13. Hassan, A., Radev, D. Identifying Text Polarity Using Random Walks, Proceedings of the Association for Computational Linguistics (2010).