High-level Features for Learning Subjective Language across Domains

Gaël Dias[1], Dinko Lambov[1], Veska Noncheva[2]
[1] Universidade da Beira Interior, Covilhã, Portugal
ddg@di.ubi.pt, d_lambov@mail.bg
[2] University of Plovdiv, Plovdiv, Bulgaria
wesnon@pu.acad.bg

Abstract

In this paper, we propose to study the characteristics for analyzing subjective content in documents. For that purpose, we present and evaluate a novel method based on level of abstraction of nouns. By comparing state-of-the-art features and the level of abstraction of nouns between three annotated corpora and texts downloaded from Wikipedia and Web Blogs, we show that, building data sets for the classification of opinionated texts can be done automatically from the web, at the document level. Moreover, we present accuracy levels within domains of 96.5% and across domains of 74.5%.

Introduction

Over the past years, there have been an increasing number of publications focused on the detection and classification of sentiment and subjectivity in texts. Most research have focused on the construction of models within particular domains and have shown difficulties in crossing domains. As a consequence, our aim in constructing a classifier is to maximize accuracy both on a single topic and across topics. For that purpose, we propose to use high-level features (e.g. level of affective words, level of abstraction) rather than low-level features (e.g. unigrams, bigrams) to learn a model of subjectivity which may apply to different domains: movie reviews, newspaper articles, manually and automatically annotated texts downloaded from Wikipedia and Web Blogs.

Since sentiment in different domains can be expressed in different ways (Boiy et al., 2007; Aue and Gammon, 2005), supervised classification techniques require large amounts of labeled training data. However, the acquisition of these labeled data can be time-consuming and expensive. From that assumption, we propose to automatically produce learning data from web resources. To do so, we propose to compare Wikipedia and Web Blogs texts to reference objective and subjective corpora. Our methodology uses state-of-the-art high-level characteristics that have been used to classify opinionated texts and proposes a new feature to classify sentiment texts, based on the level of abstraction of nouns. Finally, an exhaustive evaluation shows that (1) the level of abstraction of nouns is a strong clue to identify subjective texts which crosses domains, (2) high-level features allow cross-domain learning models and (3) automatically labeled dataset extracted from Wikipedia and Web Blogs give rise, on average, to the best cross-domains classifiers reaching accuracy levels of 74.5%.

Related Work

At document level, (Wiebe et al., 2004) derive a variety of subjectivity from corpora and demonstrate their effectiveness on classification tasks. They determine a relationship between low frequency terms and subjectivity and find that their method for extracting subjective n-grams is enhanced by examining those that occur with unique terms.

(Chesley et al., 2006) present a method using verb class information, and an online resources, the Wikipedia dictionary, for determining the polarity of adjectives. They use verb-class information in the sentiment classification task, since exploiting lexical information contained in verbs has shown to be a successful technique for classifying documents.

Other research in the sentiment classification field regards cross-domain classification. Tests have been done by (Finn and Kushmerick, 2003), (Aue and Gammon, 2005) and (Boiy et al., 2007). Overall, they show that sentiment analysis is a domain-specific problem, and it is hard to create a domain independent classifier. One possible approach is to train the classifier on a domain-mixed set of data instead of training it on one specific domain (Finn and Kushmerick, 2003), (Aue and Gammon, 2005) and (Boiy et al., 2007). Another possibility is to propose high-level features which do not depend so much on topics such as Part-of-Speech statistics as in (Finn and Kushmerick, 2003). Just by looking at part-of-speech statistics, improved results can be obtained comparatively to unigram models (low-level models) when trying to cross domains.

Characterizing Subjectivity

Subjectivity can be expressed in different ways as summarized in (Boiy et al., 2007) who identify the following dimensions: evaluation (positive or negative), potency (powerful or unpowerful), proximity (near or far), specificity (clear or vague), certainty (confident or doubtful) and identifiers (more or less), direct expressions, elements of actions and remarks. Based on these assumptions, our methodology aims at classifying texts at the subjectivity level (i.e. subjective vs. objective) based on high-level features which can apply to different domains. For that purpose, we use state-of-the-art features proposed in related works and propose a new feature based on the level of abstraction of nouns.

State-of-the-art Features

Intensity of Affective Words: In most of previous works, sentiment expressions mainly depend on some words which can express subjective sentiment orientation. (Strapparava and Mihalcea, 2008) have used a set of words extracted from WordNet Affect (Strapparava and Valittuti, 2004) to annotate the emotions in a text simply based on the presence of words from the WordNet Affect lexicon.

Dynamic Adjectives and Semantically Oriented Adjectives: (Hatzivassiloglou and Wiebe, 2000) consider two features for the identification of opinionated sentences: (1) semantically oriented adjectives and (2) dynamic adjectives. They noted that all sets involving dynamic adjectives and adjectives with positive or negative polarity are better predictors of subjective sentences than the class of adjectives as a whole.

Classes of Verbs: (Chesley et al., 2006) present a method using verb class information. To obtain relevant verb classes, they use an automatic text analyzer which groups verbs according to classes that often correspond to their polarity. We reproduce their methodology by using the classification of verbs available in Levin's English Verb Classes and Alternations (Levin, 1993).

Level of Abstraction of Nouns

There is linguistic evidence that level of generality is a characteristic of opinionated texts, i.e. subjectivity is usually expressed in more abstract terms than objectivity (Osgood et al., 1971). Indeed, descriptive texts tend to be more precise and more objective and as a consequence more specific. In other words, a word is abstract when it has few distinctive features and few attributes that can be pictured in the mind. One way of measuring the abstractness of a word is by the hypernym relation in WordNet (Miller, 1995). In particular, a hypernym metric can be the number of levels in a conceptual taxonomic hierarchy above a word. So, a word having more hypernym levels is more concrete than one with fewer levels.

Corpora

To perform these experiments, we used three manually annotated standard corpora and built one corpus based on Web resources and automatically annotated.

Mpqa: The Multi-Perspective Question Answering (MPQA) Opinion Corpus[1] contains 10,657 sentences in 535 documents from the world press on a variety of topics. All documents in the collection are marked with expression-level opinion annotations.

Rotten/Imdb: The second corpus is the subjectivity dataset v1.0[2] which contains 5000 subjective and 5000 objective sentences collected from movie reviews data (Pang and Lee, 2004).

Chesley: (Chesley et al., 2006) manually annotated a dataset of objective and subjective documents[3]. It contains 496 subjective and 580 objective documents.

Wiki/Blog: For our data set, we downloaded part of the static Wikipedia dump archive[4] and automatically spidered Web Blogs from different domains. We propose to compare Wikipedia and Web Blogs texts to reference objective and subjective corpora and show that Wikipedia texts are representative of objectivity and Web Blogs are representative of subjectivity.

Wilcoxon Rank-Sum Test

Before performing any classification task, it is useful to evaluate to what extent the given features are discriminative and allow representing distinctively the datasets in the given space of characteristics. For that purpose, we propose to do feature selection by applying the Wilcoxon rank-sum test.

The two sample Wilcoxon test with one-sided alternative is carried out for all experiments. The samples contain 200 values for each of the sets (100 objective texts and 100 subjective) and the exact p-value is computed. The exact 95% confidence interval for the difference of the location parameters of each of the sets is obtained by the algorithm described in (Bauer, 1972) for which the Hodges-Lehmann estimator is employed. As a consequence, for each of the sets, we are 95% confident that the interval contains the actual difference between the features values of subjective and objective texts.

[1] http://www.cs.pitt.edu/mpqa/

[2] www.cs.cornell.edu/People/pabo/movie-review-data/

[3] http://www.tc.umn.edu/~ches0045/data/

[4] http://download.wikimedia.org/enwiki/

Corpus: Feature:	Mpqa	Rotten/ Imdb	Chesley	Wiki/ Blog
Affective words	< 0,0001	< 0,0001	< 0,0001	< 0,0001
dynamic adj.	< 0,0001	< 0,0001	0,014	< 0,0001
semantical adj.	< 0,0001	< 0,0001	0,045	< 0,0001
conjecture verbs	0,00024	< 0,0001	0,021	< 0,0001
marvel verbs	< 0,0001	< 0,0001	0,44	< 0,0001
see verbs	< 0,0001	< 0,0001	0,006	< 0,0001
positive verbs	< 0,0001	0,00011	0,075	0,00061
level of abstraction	0,003	< 0,0001	< 0,0001	< 0,0001

Table 1: Computed p-values using Wilcoxon test

As illustrated in Table 1, we can see that only the level of positive verbs does not significantly separate the objective sample from the subjective one over training corpora. As a consequence, we discarded this feature from our classification task.

Experiments

In this section, we report the results of machine learning experiments for which we used two classifiers (Support Vector Machines and Linear Discriminant Analysis) to learn models of subjectivity over seven different features for four different domains. In this paper, we propose to use Linear Discriminant Analysis (LDA) as an alternative to SVM. Indeed, SVM has proved to work better when the size of features is high. All experiments have been performed on a leave-one-out 5 cross validation basis. In particular, we used Joachim's (2002) SVMlight package[5] for training and testing with SVM and the free software for statistical computing R[6] for LDA. As part-of-speech tagger we used the MontyTagger module[7] (Liu, 2004).

In-Domain Data and Level of Abstraction

In order to evaluate the importance of the level of abstraction of nouns as a clue for subjectivity identification, we first propose to study the six state-of-the-art features without the level of abstraction of nouns and then compare with the full set of seven features. Then, we present the importance of each class of features individually to assess how discriminative each class of features is. For that purpose, we defined four classes of features: affective words, adjectives (semantically oriented and dynamic), verbs (conjecture, marvel and see) and level of abstraction of nouns. The results are illustrated in Table 2 and 3 for leave-one-out 5 cross validation for in-domain data i.e. each model is tested with documents from the same domain of the training texts. The evaluation assesses that LDA reaches higher levels of accuracy than the SVM for all datasets, with a maximum of 96.5% for Rotten/Imdb and seven features.

	Mpqa	Rotten/ Imdb	Chesley	Wiki/ Blogs
7 features	60.5%	**87.5%**	64.5%	74%
6 features	**88%**	84.5%	**66%**	**82%**
Affective words only	**90.5%**	77.5%	**66.5%**	74.5%
Adjectives only	73.5%	79.5%	62%	75%
Verbs only	76%	77.5%	58.5%	**84%**
Level of Abstraction only	58%	**85.5%**	63%	74%

Table 2: Results of SVM for In-Domain tests

	Mpqa	Rotten/ Imdb	Chesley	Wiki/ Blogs
7 features	**93.5%**	**96.5%**	**71%**	**94%**
6 features	93%	92%	68%	89.5%
Affective words only	**90%**	76.5%	**66.5%**	76%
Adjectives only	72.5%	81.5%	60%	76%
Verbs only	76.5%	80%	58%	**84.5%**
Level of Abstraction only	68.7%	**86%**	63.5%	74%

Table 3: Results of LDA for In-Domain tests

Results for Cross-Domain Data

In order to test models across domains we propose to train different models based on one domain only at each time and test the classifiers over the other domains. In Table 4 and 5, we present the results for the classification experiment for domain transfer. Each percentage can be expressed as the average results over all datasets. Best results overall are obtained with LDA for the Wiki/Blog dataset with accuracy of 74.5%.

		Mpqa	Rotten/ Imdb	Chesley	Wiki/ Blogs
All	Accuracy	52.6%	69.5%	**73.9%**	71%
Subjective	Precision	51.5%	**74.2%**	70.3%	74.2%
	Recall	**100%**	59%	82%	63.5%
Objective	Precision	25%	67.1%	**81.7%**	68.7%
	Recall	5.3%	**79.8%**	65.8%	78.5%

Table 4: Results of SVM for Cross-Domain tests

		Mpqa	Rotten/ Imdb	Chesley	Wiki/ Blogs
All	Accuracy	67.6%	70.9%	73.6%	**74.5%**
Subjective	Precision	64.7%	**80.2%**	69.3%	78.8%
	Recall	**96%**	48.8%	89.5%	65.5%
Objective	Precision	67.7%	67.4%	**89.2%**	75.6%
	Recall	39.3%	**93%**	57.8%	83.5%

Table 5: Results of LDA for Cross-Domain tests

It is important to notice that the best model on average is obtained with automatically labeled data i.e. texts extracted from Wikipedia and Web Blogs in an uncontrolled way. As such we are capable to create models automatically without manually annotated corpora. Indeed, if one is more interested in precision for subjectivity, Rotten/Imdb should be used as training set.

It is also interesting to notice that both learning algorithms, SVM and LDA, present similar results on average although with a small advantage for LDA. Tables 4 and 5 also show that precision, recall and accuracy levels are equally distributed by both algorithms.

[5] http://svmlight.joachims.org/

[6] http://www.r-project.org/

[7] http://web.media.mit.edu/~hugo/montylingua/

Cross-Domain and Level of Abstraction

It is also important to understand how different features manage to cross domains. Table 6 and 7 show the classification results for all features alone, for all features together with and without the level of abstraction. The results show average accuracy levels over all datasets in a leave-on-out 5 cross validation basis. Similarly to the results presented in the section above, SVM and LDA show similar results on average for almost all experiments. Indeed, the level of abstraction of nouns is the best feature to cross domains except for the case of the Mpqa dataset for the SVM algorithm.

	Mpqa	Rotten/ Imdb	Chesley	Wiki/ Blogs
7 features	52.6%	**69.5%**	**73.9%**	**71%**
6 features	**65.3%**	66.6%	69.8%	69.5%
Affective words only	**67%**	60.5%	67.5%	65.8%
Adjectives only	60.1%	65.3%	67.9%	68.6%
Verbs only	65.9%	69.5%	68%	70.6%
Level of Abstraction only	52%	**71.9%**	**71.9%**	**72%**

Table 6: Results of SVM for Cross-Domain tests

	Mpqa	Rotten/ Imdb	Chesley	Wiki/ Blogs
7 features	**67.6%**	**70.9%**	**73.6%**	**74.5%**
6 features	67.1%	66.8%	69.3%	71.9%
Affective words only	67.3%	60.1%	67.6%	65.8%
Adjectives only	64.4%	66.5%	68.4%	69.1%
Verbs only	67.1%	69.1%	68.1%	68.8%
Level of Abstraction only	**72.7%**	**72.5%**	**72.1%**	**73.5%**

Table 7: Results of LDA for Cross-Domain tests

This is due to the over-evaluation strong features by the SVM. Indeed, in Table 6, as Affective words are the best feature for the Mpqa dataset, best results are obtained without the level of abstraction. In the other cases, the level of abstraction of nouns is the best feature thus implying best results with seven features. As a consequence, we can trust the value of predictability of level of abstraction of nouns for learning subjective language.

Conclusions

Sentiment classification is a domain specific problem i.e. classifiers trained in one domain do not perform so well in others. At the same time, sentiment classifiers need to be customizable to new domains in order to be useful in practice. In this paper, we proposed new experiments based on high-level features to learn subjective language across domains. Best results showed accuracy of 96.5% within domain experiments and 74.5% across domains. Unfortunately we were not able to achieve accuracy on the subjectivity classification problem comparable to those reported for standard topic-based categorization. But the results produced via automatically constructed data are better than or at least comparable to the predictability produced via manually annotated corpora. A direct application of this study is to automatically produce data sets for other languages than English and allow classification of multilingual opinionated texts.

References

Aue, A. and Gamon, M. 2005. *Customizing sentiment classifiers to new domains: a case study*. In Proceedings of International Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria.

Bauer, D.F. 1972. *Constructing Confidence Sets Using Rank Statistics*. In Journal of the American Statistical Association 67, 687–690.

Boiy, E., Hens, P., Deschacht, K., Moens, M. 2007. *Automatic Sentiment Analysis in On-line Text.* In Proceedings of the International Conference on Electronic Publishing held in Vienna, Austria.

Chesley, P., Vincent, B., Xu, L. and Srihari,R. 2006. *Using Verbs and Adjectives to Automatically Classify Blog Sentiment*. In Proceedings of AAAI Spring Symposium.

Finn, A. and Kushmerick, N. 2003. *Learning to Classify Documents According to Genre*. J. American Society for Information Science and Technology, Special issue on Computational Analysis of Style, 57(9).

Hatzivassiloglou, V. and Wiebe, J. 2000. *Effects of Adjective Orientation and Gradability on Sentence Subjectivity*. In Proceedings of International Conference on Computational Linguistics.

Joachims, T. 2002. *Learning to Classify Text Using Support Vector Machines*. Dissertation, Kluwer.

Levin, B. 1993. *English Verb Classes and Alternations*. University of Chicago Press.

Liu., H. 2004. Montylingua: An End-to-End Natural Language Processor with Common Sense. Available at: web.media.mit.edu/~hugo/montylingua.

Miller, G.A. 1995. *Wordnet: A Lexical Database*. In Communications of the ACM 38.

Osgood, C.E., Suci, G.J. and Tannebaum, P.H. 1971. *The Measurement of Meaning*. University of Illinois Press.

Pang, B. and Lee, L. 2004. *A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts*. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics.

Strapparava, C. and Mihalcea, R. 2008. *Learning to Identify Emotions in Text*. In proceedings of the Symposium on Applied Computing. 1556-1560.

Strapparava, C. and Valitutti, A. 2004. *WordNet-Affect: An Affective Extension of WordNet*. In Proceedings of the Language Resources and Evaluation International Conference. Lisbon, Portugal.

Wiebe, J., Wilson, T., Bruce, R., Bell, M. and Martin, M. 2004. *Learning Subjective Language*. In Computational Linguistics, 30(3), 277–308.