Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference

Aspect-Level Sentiment Analysis Based on a Generalized Probabilistic Topic and Syntax Model

Haochen Zhou[†] **and Fei Song**[‡]
[†]MI9 Retail, Toronto, ON, Canada M2J 4W9
[‡]School of Computer Science, University of Guelph
Guelph, ON, Canada N1G 2W1
hzhou@mi9retail.com and fsong@uoguelph.ca

Abstract

A number of topic models have been proposed for sentiment analysis in recent years, which rely on extensions of the basic LDA model. In this paper, we apply a generalized topic and syntax model called Part-of-Speech LDA (POSLDA) to sentiment analysis, and propose several feature selection methods that separate entities from the modifiers that describe the entities. Along with a Maximum Entropy classifier, we can use the selected features to conduct sentiment analysis at both document and aspect levels. The advantage of using POSLDA is that we can automatically separate semantic and syntactic classes, and easily extend it to aspect level sentiment analysis by mapping topics to aspects. However, words in the noun-related classes, which are also treated as semantic classes, should be removed as much as possible to reduce their impact on sentiment analysis. To evaluate the effectiveness of our solutions, we conducted experiments on two collections of review documents and obtained the accuracy results competitive to the previous work on sentiment analysis.

Introduction

With the fast growth and convenient access of the Internet, people can now easily share their opinions through blogs, discussion forums, and social networks. Sentiment analysis (SA) is a form of text classification which automatically determines the sentiment of a review document (usually positive or negative, but can be a scale of multiple levels such as 1 to 5). Sentiment analysis has gained its popularity due to many useful applications such as online customer review analysis (Pang, Lee, and Vaithyanathan 2002) and opinionated web search (Wang, Lu, and Zhai 2010).

Whereas general text classification is concerned with features that distinguish different topics, sentiment analysis deals with features about subjective feelings and opinions that describe or modify certain entities. Since a review document typically contains both kinds of features, any solutions for sentiment analysis ultimately face the challenge of separating the objective entities from the subjective expressions that modify these entities. In addition, there is a growing need to provide sentiment ratings for both an overall review document and the aspects described within it, called the Aspect-Level SA. For example, a reviewer may be positive about a product, but negative about some of its components or attributes. Consumers often need to know the sentiments about such aspects as well in order to make informed decisions about certain purchases.

In this paper, we apply a generalized topic and syntax model called POSLDA to SA and Aspect-Level SA. POSLDA can separate semantic classes (mostly made of content words such as nouns, verbs, adjectives, and adverbs) and syntactic classes (mostly made of functional words such as determiners, prepositions, and conjunctives), allowing us to identify semantic words that modify entities from those that specify the entities. In addition, POSLDA associates each topic with its own semantic classes, which helps us extract content words about related aspects by modeling topics as aspects. Based on these intuitions, we propose several new feature selection methods, which along with a Maximum Entropy classifier, allow us to conduct sentiment analysis at both document and aspect levels. To evaluate our solutions, we conducted experiments on two sets of review documents and obtained the accuracy results competitive to the previous work on sentiment analysis.

The remainder of the paper is organized as follows. First, we provide a review of the related work on topic modeling and sentiment analysis. Then, we present our proposed solutions for feature selection based on the POSLDA model. After that, we describe our experiments along with analyses. Finally, we conclude the paper and discuss some possible directions for future work.

Related Work

Topic Modeling

Topic modeling helps uncover the underlying topics for a collection of documents using probabilistic models. The basic LDA model proposed by Blei et al. (2003) is sufficiently modular and has been extended in various ways. One particular extension, Part-Of-Speech LDA (POSLDA) (Darling 2012), extends LDA with HMM (Hidden Markov Model) so that both the topic information about the long-range relationships of words and the syntax information about the local context of words can be captured in one model. As a result, it can not only identify topics such as sports and travel, but further separate them into specific POS distributions such as

"nouns about sports" and "verbs about travel". Compared with another similar topic and syntax model HMMLDA (Griffiths et al. 2005), POSLDA is more generalized in that it contains several other models as special cases, including the basic LDA, Bayesian HMM, and HMMLDA.

More specifically, the generative process for POSLDA can be described as follows:

1. For each row $\pi_r \in \boldsymbol{\pi}$:
   draw $\pi_r \sim Dir(\gamma)$
2. For each word distribution $\phi_\eta \in \boldsymbol{\phi}$:
   draw $\phi_\eta \sim Dir(\beta)$
3. For document $d \in D$:
   (a) draw a topic proportion $\theta_d \sim Dir(\alpha)$
   (b) For each word $w_i \in d$:
      i. draw a class $c_i \sim \pi_{c_i-1}$
      ii. If $c_i \in C_{sem}$:
         draw a topic $z_i \sim \theta_d$ and $w_i \sim \phi^{sem}_{c_i,z_i}$
      iii. Else:
         draw $w_i \sim \phi^{syn}_{c_i}$

Here, $\boldsymbol{\pi}$ is the HMM transition matrix and each row π_r corresponds to a POS class; the set of classes C includes both semantic classes C_{sem} and syntactic classes C_{syn}; for syntactic classes, word distributions are ϕ^{syn}, while for semantic classes, word distributions are ϕ^{sem}; $Dir(.)$ are Dirichletl distributions; and α, β, and γ are hyperparameters for POSLDA.

POSLDA has several advantages that are potentially helpful for Aspect-Level SA. First, it can automatically capture functional words (e.g., "the", "at", and "of") in syntactic classes so that the semantic classes are mostly populated with content words. Secondly, it further separates the content words into different semantic classes so that we can model certain Part-Of-Speech categories such as nouns, verbs, adjectives, and adverbs for SA. Finally, each topic is associated with its own semantic classes, making it possible to perform aspect-level SA if topics are modeled as aspects.

Sentiment Analysis

SA is considered as unconventional text classification in that it relies on subjective words, mostly made of adjectives and adverbs, to distinguish between different sentiments. Such words usually have low frequencies within a review document. For example, when describing a camera, the user is unlikely to use words like "great" and "excellent" repeatedly, but rather use different words such as "sharp" pictures, "sleek" design, and "high" resolution. In addition, the same words can carry different sentiments for different domains. For example, "unpredictable" can be negative for a car but positive for a thriller movie.

Earlier work for SA typically uses manually selected features for text representation. Further improvements start with seed words and extend lists of features through a thesaurus or training documents. Later work applies topic modeling to SA, especially aspect-level SA (Wang, Lu, and Zhai 2010; Jo and Oh 2011). In this paper, we are particularly interested in (Duric and Song 2012) where a topic and syntax model of HMMLDA is used to separate topics from syntactic classes so that features can be selected from words in the syntactic classes for SA. However, we propose to extend this work by replacing HMMLDA with POSLDA and further apply our solutions to aspect-level SA.

Proposed Solutions

The main idea behind our proposed solutions is to use POSLDA to select semantic words suitable for SA, and further apply these methods to the related aspects for aspect-level SA.

Optimizing the Modeling Process of POSLDA

Like many probabilistic models, modeling with POSLDA is a unsupervised process, which is usually optimized by measures like perplexity. However, as observed by Chang et al. (2009), such optimized results often do not match well with the human-labeled results. Also for SA, we are interested in selecting subjective words made mostly of adjectives and adverbs, which have to be matched with the semantic classes in POSLDA. To get around these problems, we follow the extension in Darling (2012) that uses a tagging dictionary to control the modeling process. A tagging dictionary is computed from the labeled POS data so that each word is associated with the POS categories it can participate. By explicitly labeling the syntax classes of POSLDA with known POS categories, we can create a semi-supervised modeling process. Any words that can be found in the tagging dictionary will be mapped to the corresponding syntax classes. For words not found in the tagging dictionary, we will simply map them to all syntax classes, as is the case in the unsupervised modeling process. We call this extension "POSLDA with Tagging", which not only help us to produce human-readable results but also avoid the need to map semantic classes to known POS categories.

Feature Selection for Sentiment Analysis

In this paper, we propose new feature selection methods for SA based on the results of POSLDA modeling, since it can not only separate functional words from semantic words, but further distinguish different kinds of semantic words. More specifically, we formulate three different methods for Feature Selection (FS):

FS Based on Semantic Classes Since POSLDA can separate the semantic classes from the syntactic classes, a simple solution is to choose the words with high probabilities from the semantic classes. Since the distribution of a class contains all words in a vocabulary, we can pick those highly ranked words from the class so that the accumulative probability for the selected words is greater than a pre-determined level (e.g., 75% or 90%). To select words for all semantic classes, we can merge those selected from the individual semantic classes into one set: W_{sem}. Similarly, we can get a set of highly ranked functional words for the syntactic classes: W_{syn}.

FS Based on Semantic Classes with Tagging With a tagging dictionary, we can explicitly label a semantic class with

a known POS category and also make the modeling results more human readable. However, generating a tagging dictionary requires human-labeled POS data, which is not easy to obtain for a new dataset. Fortunately, we can borrow such information from an existing dataset. For example, the WSJ dataset from the ACL_DCI release contains about 3 millions of words over 6,058 documents. As in (Darling 2012), we use a condensed set of 17 POS tags so that we can get a reasonable coverage for a new dataset in English. In addition, any words that are not covered by the tagging dictionary can participate in the semi-supervised modeling process for POSLDA so that they may still be placed into the relevant classes through co-occurrence patterns in the dataset. As an important benefit, since we now know which classes correspond to nouns and related categories, we can easily remove them from the set of features selected from the semantic classes.

FS with Automatic Stopword Removal Although POSLDA can separate semantic classes from syntactic classes, some functional words may still appear in certain semantic classes even though not as common as with other models such as LDA. Again, since we now know which are the syntactic classes, we can easily remove words in W_{syn} out of those selected from semantic classes. One clear advantage for POSLDA is that W_{syn} can be automatically generated for each new dataset, and thus well-customized for the dataset. If we use a manually constructed stopword list, we may either over- or under-remove certain functional words for a new dataset.

Feature Selection for Aspect-Level SA

In POSLDA, each topic is paired with the semantic classes so that we may have distributions like "nouns about sports", "verbs about travel", and so on. As a result, we can adapt the feature selection methods above to the semantic classes for each topic. The challenge here is to train the model so that topics can be more or less treated as aspects. For example, in the TripAdvisor dataset, each review can provide sentiment ratings for five aspects including value, room, location, cleanliness, and service. Due to the unsupervised process for topic modeling, both the number of topics and the contents of the topics may not match well with the given aspects for a particular dataset. Just like the way we use a tagging dictionary to control the modeling process for the syntax classes, we can also use pre-determined word lists for aspects to form a semi-supervised process for modeling aspects.

We follow the bootstrapping method in (Wang, Lu, and Zhai 2010) to generate pre-determined word lists for all aspects. Using these seed words for the correspoding topics, we can then identify different aspects and use the features selected from the related semantic classes to determine the sentiment ratings for these aspects. Take the TripAdvisor dataset as an example. If we are interested in the "Value" aspect, we can examine the semantic classes that are paired up with this aspect, and choose the top-ranked features with the FS methods discussed above. Then, by feeding these features to a text classifier such as Maximum Entropy classier, we can produce a sentiment rating for this particular aspect. Thus, we can see that POSLDA is well suited not only for document level SA, but also for aspect-level SA.

Experimental Results

Datasets and Evaluation Metrics

We conduct experiments on two datasets: the movie reviews[1] for document-level SA and the TripAdvisor data[2] for aspect-level SA. Each movie review has two outcomes: positive and negative, but each TripAdvisor review has a rating of 1 to 5. To simplify our analysis, we map these values into two outcomes: $\{1, 2\} \rightarrow$ negative and $\{4, 5\} \rightarrow$ positive. We omit the reviews with a rating of 3 since they do not show obvious polarity information (Wang, Lu, and Zhai 2010).

The movie data has 1,000 positive reviews and 1,000 negative reviews for a total of 1,583,820 words or an average of 791.91 words per document. The TripAdvisor data, extracted from TripAdvisor.com, is fairly large; so we choose a random sample of 15,242 reviews from the original set of 108,891 reviews. The sample has a total of 3,458,177 words or an average of 226.88 words per document. In addition to the overall ratings, the sample also has ratings on up to five aspects, including values (6,584 positives vs. 1,951 negatives), rooms (7,676 vs. 2,734), locations (6,969 vs. 2,172), cleanliness (7,071 vs. 1,242), and services (8,581 vs. 523).

Since SA is essentially a classification task, we use the accuracy measure to evaluate the performance of a SA system, which is calculated by the matched class labels between the computed and the annotated over all the reviews in a test dataset (Pang, Lee, and Vaithyanathan 2002).

Results on Sentiment Analysis

To establish a baseline for comparisons, we use document frequencies or df-cutoff method to select top-ranked features along with a MaxEnt classifier for SA. For the movie review data, we set aside 200 reviews (100 positives and 100 negatives) as a validation set to control the training process of a MaxEnt classifier. We use the remaining 1,800 reviews for a three-fold cross-validation to train and test our solutions. For all systems, we choose 2,500 features, as this is the number commonly used for text classification (Pang, Lee, and Vaithyanathan 2002).

For POSLDA modeling, we try different parameter settings and identify an optimal setting of K = 25 topics, S = 17 POS classes, and SS = 6 semantic classes for our model after 2,500 iterations for the modeling process. In addition, for the features selected with the methods proposed in this paper, we further cut them down to 2,500 using the df-cutoff method.

Table 1 shows the classification results in both accuracies and standard deviations, where SC stands for FS from semantic classes and SC/Tagging, for FS from semantic classes with Tagging. Compared with the baseline, both systems based on SC and SC/Tagging achieve better results. A t-test between SC and the baseline produces a p-value of

[1] https://www.cs.cornell.edu/people/pabo/movie-review-data/

[2] http://sifaka.cs.uiuc.edu/∼wang296/Codes/LARA.zip

Methods	Accuracies	Std-Dev's
Baseline	0.847	2.9×10^{-3}
SC	0.856	1.7×10^{-3}
SC/Tagging	0.866	5.7×10^{-3}

Table 1: SA for the Movie Review Data

Aspectss	TP	FP	FN	TN	Acc
Overall	2665	469	79	898	0.867
Value	2351	435	105	675	0.849
Room	2146	437	102	783	0.845
Location	1950	406	132	640	0.828
Cleanliness	2441	675	83	474	0.794
Service	2437	584	38	120	0.804

Table 2: Aspect-Level SA for the TripAdvisor Data

0.0196, indicating that the improvement is statistically significant at 95% confidence level. However, a t-test between SC and SC/Tagging produces a p-value of 0.087, showing no significant improvement at 95% level[3].

Since POSLDA allows us to generate a list of functional words automatically, we also try to remove stopwords from W_{syn} using different cut-off levels between 90% and 95%. The best results is obtained at 93% cut-off level, which generates a total of 351 stopwords and increases the accuracy to 0.881 with the SC/Tagging method.

Results on Aspect-Level Sentiment Analysis

We use the TripAdvisor data to test our solution for aspect-level SA because each review contains the information for up to five aspects, including value, room, location, cleanliness, and service. Since each aspect is paired with its own semantic classes in POSLDA, we simply apply the best feature selection method we identified above, which is SC/Tagging plus stopword removal, to these semantic classes in order to determine the sentiment rating for the corresponding aspect.

As shown in Table 2, the accuracy at the document level is 0.867 and the results for the individual aspects are comparable to this overall performance. The low accuracies for the "Cleanliness" and "Service" aspects may be caused by the ambiguity between them. For example, "a room is dirty" can be about "Cleanliness" or "Service" or both. On the whole, it is fairly straightforward to extend our feature selection methods to aspect-level SA due to the generalized probabilistic model of POSLDA.

Conclusions and Future Work

We approached feature selection for Sentiment Analysis (SA) by applying a recently developed probabilistic topic and syntax model, called POSLDA, to separate the semantic words from the functional words, and further extended it to aspect-level SA. First, by selecting top-ranked words from the semantic classes, we can extract features that perform better for SA than the baseline of simply choosing words with high document frequencies. Secondly, by training the modeling process for semantic classes with a POS tagging dictionary, we can further improve the performance for SA. Finally, by removing the stopwords automatically extracted from the functional classes, we get even better results for SA. Since POSLDA associates each topic with its own semantic classes, we can easily turn our solutions to solve the problem for aspect-level SA by controlling the modeling process for topics so that topics are modeled as aspects. Overall, the results for aspect-level SA are comparable to that for the document-level SA, illustrating that POSLDA is general and well-suited for both SA and aspect-level SA.

POSLDA is currently implemented to handle unigrams, but it could be extended to handle n-grams. N-grams are especially useful to extract adverb-adjective pairs that can enhance polarity rating with positive pairs like "very happy" and negation pairs like "too expensive".

We could also extend our FS schemes to classify documents with a scale of sentiment ratings (e.g., 1 to 5). Although a scale rating system looks more complex at first, it can be broken down into multiple binary classification tasks. For example, a review can be first classified into a positive or negative category, and then in each category, it can be further classified into a more specific category based on how positive or negative it is. However, in order to perform such a task, we need a well-annotated dataset, and the accuracy levels are likely to drop as more categories and a refined rating scale are added.

[3]This result is, however, significant at 90% level.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Chang, J.; Boyd-Graber, J. L.; Gerrish, S.; Wang, C.; and Blei, D. M. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22*, 288–296.

Darling, W. M. 2012. *Generalized Probabilistic Topic and Syntax Model for Natural Language Processing*. Ph.D. Dissertation, School of Computer Science, University of Guelph, Guelph, Ontario, Canada.

Duric, A., and Song, F. 2012. Feature selection for sentiment analysis based on content and syntax models. *Journal of Decision Support Systems* 53(4):704–711.

Griffiths, T. L.; Steyvers, M.; Blei, D. M.; and Tenenbaum, J. B. 2005. Integrating topics and syntax. In *Advances in Neural Information Processing Systems 17*, 537–544.

Jo, Y., and Oh, A. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, 815–824.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of Empirical Methods in Natural Language Processing*, 79–86.

Wang, H.; Lu, Y.; and Zhai, C. 2010. Latent aspect rating analysis on review text data: A rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 783–792.