Sentiment Classification Based on AS-LDA Model

Jiguang Liang¹, Ping Liu¹, Jianlong Tan¹, Shuo Bai¹,¹

¹National Engineering Laboratory for Information Security Technologies, Institute of Information Engineering Chinese Academy of Sciences, Beijing 100190, China
²Shanghai Stock Exchange, Shanghai 200120, China

Abstract

We address the task of sentiment classification - identification of the polarity of the subjective document in this paper. We introduce a sentiment classification method called AS-LDA. In this model, we assume that words in subjective documents consist of two parts: sentiment element words and auxiliary words which are sampled accordingly from sentiment topics and auxiliary topics. Sentiment element words include targets of the opinions, polarity words and modifiers of polarity words. Experimental results demonstrate that our approach outperforms Latent Dirichlet Allocation (LDA).

Keywords: sentiment analysis; sentiment classification; Latent Dirichlet Allocation; subjective document

1. Introduction

Sentiment analysis, also known as opinion mining, is the computational study of opinions, sentiments and emotions expressed in text (Liu, 2010). Sentiment analysis has been widely used across a wide range of domains in recent years, such as information retrieval (IR)(Pang and lee, 2008; Liu, 2010; Li et al., 2012), question answering systems (Oh et al., 2012; Kucuktunc et al., 2012) and social network (Diakopoulos et al., 2010; Tan et al., 2011).

Usually, sentiment analysis can be decomposed into three subtasks: subjective text detection, subjective information extraction and sentiment classification (polarity identification). This paper addresses the third subtask - sentiment classification. In short, Sentiment classification aims to automatically predict sentiment polarity (eg, positive or negative) of users publishing sentiment data (eg, reviews, blogs) (Pan et al., 2010).

In this paper, we target at describing an effective method, called AS-LDA, for the sentiment classification problem. Given a subjective document, we assume that there are two kinds of words in it, sentiment element words and auxiliary words. Sentiment element words are those words that can be used to express certain opinions, sentiments or emotions. Sentiment element words is an essential part of the subjective document. Comparing to sentiment element words, auxiliary words are not all that important. Auxiliary words are used to assist or enhance meaning expression. In AS-LDA model, two kinds of topics, corresponding to sentiment element words...
and auxiliary words, are produced. In other words, words in subjective documents are sampled either from the sentiment topics or from the auxiliary topics. Consider the following examples in which sentiment element words are shown with italics.

- **Review 1**: There are *good performances* by Hattie Jacques as the matron, however her *character* seems *a little more subdued* and quieter than her previous ‘matron’s’.

This is a review about film and Hattie Jacques is the leading actress of this film. From the italic items, we can clearly observe that targets of the opinion (*performances, character*), polarity words (*good, subdued and quieter*) and modifiers (a little more), are considered as sentiment element words. And the rest are auxiliary words which are sampled from auxiliary topics.

The rest of the paper is structured as follows. Section 2 begins with a discussion of the related work. Section 3 presents the AS-LDA model for sentiment classification. In Section 4 we provide the evaluation of the proposed method. Then we conclude in Section 5.

2. Related Work

Significant research effort has been invested into sentiment analysis, especially in the domain of movie reviews (Pang et al., 2002; Kennedy and Inkpen, 2006; Maas et al., 2011), product reviews (Cui et al., 2006; Wei and Gulla, 2010), Twitter (Go et al., 2009; Jiang et al., 2011) and microblogs (Bermingham and Smeaton, 2010; Stieglitz et al., 2013). Go et al. (2009) introduce a novel approach using distant supervision for automatically classifying the sentiment of Twitter messages. Jiang et al. (2011) address target-dependent Twitter sentiment classification; given a query, they classify the sentiments of the tweets as positive, negative or neutral. The research progress on Chinese sentiment analysis is limited by the lack of Chinese sentiment corpora. Glorot et al. (2011) propose a deep learning approach, which learns to extract a meaningful representation for each review in an unsupervised fashion, to tackle the problem of domain adaptation for sentiment classifications. Wan (2009) focuses on this problem and proposes a cross-lingual sentiment classification by making use of labeled English corpus and unlabeled Chinese data.

Particularly worth mentioning is that Latent Dirichlet Allocation has been successfully used to classify the sentiments in recent years. He and Lin (2009) propose a novel probabilistic modeling framework based on LDA, called joint sentiment/topic model (JSP) for sentiment analysis. This model is fully unsupervised. Li et al. (2010) introduce a Sentiment-LDA model for sentiment analysis with global topics and local dependency. Jo and Oh (2011) describe two models, Sentence-LDA (SLDA) and Aspect and Sentiment Unification Model (ASUM), to tackle the problem of automatically discovering what aspects are evaluated in reviews and how sentiments for different aspects are expressed. Sentiment classification are obtained more granular down to the sentences and aspects. However, these works don’t make use of sentiment information during modeling.

3. Method

As proposed in Section 1, we consider the task of sentiment classification. In this section, we describe a new sentiment model called Auxiliary-Sentiment Latent Dirichlet Allocation (AS-LDA) for sentiment classification.

3.1. Auxiliary-Sentiment Latent Dirichlet Allocation (AS-LDA)

In LDA for sentiment classification, all words in the document are sampled from the global topics which neglects the particularity of sentiment words. In AS-LDA, sentiment element words and auxiliary words are treated differently. The graphical representation of AS-LDA is shown in Figure 1(b) and the notations are explained in Table 1. And the generative process of SA-LDA is described in Table 2.
Fig. 1. (a) Latent Dirichlet Allocation (LDA) model. (b) An extension of LDA to obtain AS-LDA for sentiment classification.

Table 1. Meanings of the notations.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$</td>
<td>the number of documents.</td>
</tr>
<tr>
<td>$V_a$</td>
<td>the number of auxiliary words.</td>
</tr>
<tr>
<td>$w_a$</td>
<td>the auxiliary words.</td>
</tr>
<tr>
<td>$z_a$</td>
<td>the auxiliary topics.</td>
</tr>
<tr>
<td>$\theta_a$</td>
<td>multinomial distribution over auxiliary topics.</td>
</tr>
<tr>
<td>$\varphi$</td>
<td>multinomial distribution over auxiliary words.</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Bernoulli distribution over sentiment element words.</td>
</tr>
<tr>
<td>$\alpha_a$</td>
<td>Dirichlet prior vector for $\theta_a$.</td>
</tr>
<tr>
<td>$\beta_a$</td>
<td>Dirichlet prior vector for $\varphi$.</td>
</tr>
<tr>
<td>$N$</td>
<td>the number of words in a document.</td>
</tr>
<tr>
<td>$V_s$</td>
<td>the number of sentiment element words.</td>
</tr>
<tr>
<td>$w_s$</td>
<td>the sentiment element words.</td>
</tr>
<tr>
<td>$z_s$</td>
<td>the sentiment topics.</td>
</tr>
<tr>
<td>$\theta_s$</td>
<td>multinomial distribution over sentiment topics.</td>
</tr>
<tr>
<td>$\delta$</td>
<td>multinomial distribution over sentiment element words.</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>hyper-parameter for $\gamma$.</td>
</tr>
<tr>
<td>$\alpha_s$</td>
<td>Dirichlet prior vector for $\theta_s$.</td>
</tr>
<tr>
<td>$\beta_s$</td>
<td>Dirichlet prior vector for $\delta$.</td>
</tr>
</tbody>
</table>
Table 2. The generative process of reviews with AS-LDA model.

1: For each topic \( k \in \{1, 2, \ldots, K\} \) for auxiliary words, do
\[
\text{draw } \phi_k \sim \text{Dir}(\phi_\alpha) \\
2: \text{End for} \\
3: \text{For each topic } l \in \{1, 2, \ldots, L\} \text{ for sentiment words, do}
\[
\text{draw } \eta_l \sim \text{Dir}(\eta_\alpha) \\
4: \text{End for} \\
5: \text{For each document } m \in D \text{ do}
\begin{itemize}
  \item \text{Choose a distribution } \theta_m \sim \text{Dir}(\theta_\alpha) \text{ for auxiliary words.}
  \item \text{Choose a distribution } \theta_s \sim \text{Dir}(\theta_\alpha) \text{ for sentiment element words.}
\end{itemize}
6: \text{End for}
7: \text{For each word } w_{mn} \text{ in } m \text{ do}
\begin{itemize}
  \item If \( w_{mn} \) is a sentiment candidate word, then
    \[
    \text{draw } \gamma = \text{Ber}(\lambda) \\
    \text{If } \gamma = 1, \text{ then}
    \begin{itemize}
      \item Choose a topic \( z_{mn}^s \) \text{ for } w_{mn}, \text{ draw } \gamma_{mn} = \text{Multi}(\gamma_s) \\
      \item Choose } w_{mn} \text{ from the distribution } \delta_l \text{ for auxiliary words, draw } z_{mn}^\delta \sim \text{Multi}(\delta_m) \\
    \end{itemize}
    \text{else}
    \begin{itemize}
      \item Choose a topic \( z_{mn}^a \) \text{ for } w_{mn}, \text{ draw } \gamma_{mn} = \text{Multi}(\gamma_a) \\
      \item Choose } w_{mn} \text{ from the distribution } \phi_k \text{ for sentiment words, draw } z_{mn}^\phi \sim \text{Multi}(\phi_m) \\
    \end{itemize}
  \item \text{Otherwise, for each } v \text{ assigned to topic } j, \text{ do} \\
    \begin{itemize}
      \item \text{Choose a topic } z_{mn}^v \text{ for } w_{mn}, \text{ draw } \gamma_{mn} = \text{Multi}(\gamma_v) \\
      \item Choose } w_{mn} \text{ from the distribution } \phi_k \text{ for auxiliary words, draw } z_{mn}^\phi \sim \text{Multi}(\phi_m) \\
    \end{itemize}
\end{itemize}
8: \text{End for}

3.2. Inference in AS-LDA

The inference goal is to find the solution of AS-LDA. In this paper, we use Gibbs sampling to perform model inference. The joint probability of the topics and the words can be factored into the following:

\[
P(w, z \mid \alpha, \beta) = P( w \mid z, \beta) P( z \mid \alpha) \int P( z \mid \theta) P( \theta \mid \alpha) d\theta \int P( w \mid z, \varphi) P( \varphi \mid \beta) d\varphi
\]

Integrating out \( \theta \) and \( \varphi \), we can obtain:

\[
P(w, z \mid \beta) = \left( \frac{\Gamma(V \beta)}{\prod_v \Gamma(\beta)} \right)^K \prod_j \Gamma(n_{j,v} + \beta) \prod_j \frac{\Gamma(n_{j,v} + V \beta)}{\Gamma(n_{j,v} + V \beta)}
\]

\[
P(z \mid \alpha) = \left( \frac{\Gamma(T \alpha)}{\prod_j \Gamma(\alpha)} \right)^K \prod_m \Gamma(n_{m,j} + \alpha) \prod_m \frac{\Gamma(n_{m,j} + K \alpha)}{\Gamma(n_{m,j} + K \alpha)}
\]

where \( n_{j,v} \) is the number of times word \( v \) assigned to topic \( j \), \( n_{m,j} \) is the number of times word \( v \) occurs in document \( m \).

At each iteration, the topics of words are chosen according to the conditional probability:

\[
P(z_i = j \mid w, z_{\gamma_i}, \alpha, \beta) = \frac{P(z_i, w \mid \alpha, \beta)}{P(z_{\gamma_i}, w \mid \alpha, \beta)} \propto \frac{n_{\gamma_i,j,v} + \beta}{n_{\gamma_i,j,v} + V \beta} \cdot \frac{n_{\gamma_m,j} + \alpha}{n_{\gamma_m,j} + K \alpha}
\]

The approximate probability of auxiliary topic in document \( m \) is

\[
\theta_m^a = \frac{n_{m,j}^a + \alpha_a}{n_{m,j}^a + K \alpha_a}
\]
The approximate probability of auxiliary words in topic $z$ is

$$\varphi_z = \frac{n_{j,v}^a + \beta_a}{n_{j,v}^a + V_a \beta_a}$$

In a similar way, the solution formulas for sentiment element words are

$$\theta_s^m = \frac{n_{m,j}^s + \alpha_s}{n_{m,j}^s + L \alpha_s}$$

$$\delta_z = \frac{n_{j,v}^s + \beta_s}{n_{j,v}^s + V_s \beta_s}$$

The notations are similar to those for Equation (2)(3).

4. Experiments

In this section, we demonstrate the effectiveness of our model through comparison experiments on Chinese sentiment corpus ChnSentiCorp (Tan, 2008). More precisely, we use ChnSentiCorp-Htl-ba-4000 and ChnSentiCorp-NB-ba-4000 corresponding to the domains hotel and computer to test our model. The vectors of the documents are represented by LDA and AS-LDA after Chinese word segmentation, deleting stop word and other process. Then we choose Support Vector Machine (SVM) to classify the sentiment. The measurement we use to compare the classification performance is F-measure.

In our experiments we set that the number of topics in LDA is $T$ and $K, L$ are the numbers of sentiment topics and auxiliary topics in AS-LDA, where $T = K + L$. The results are shown in Figure 2. From Figure 2, we can observe that AS-LDA performs better than LDA on the two sentiment corpora. We also show how the precision changes with the increase of the number of topics. We can find that AS-LDA get a better result when $K + L = 12$ from Figure 2.

5. Conclusions

In this paper, we present a new probabilistic sentiment-topic model called AS-LDA for sentiment classification. With this model, we divide the words in the subjective documents into two categories: sentiment element words and auxiliary words. Sentiment element words are sampled from sentiment topics while auxiliary words are sampled from auxiliary topics. We evaluate our model on Chinese sentiment corpora; the results show that the AS-LDA model is more applicable to classify the sentiment than LDA.
Acknowledgements

This work was supported by Strategic Priority Research Program of Chinese Academy of Sciences (X-DA06030602), National Nature Science Foundation of China (No. 61202226), National 863 Program (No. 2011AA010703), IIE Program (No.Y3Z0062201).

References