INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 11(3):414-427, June 2016.

Efficient Opinion Summarization on Comments with Online-LDA

J. Ma, S. Luo, J. Yao, S. Cheng, X. Chen

Jun Ma, Senlin Luo
School of Information and Electronics
Beijing Institute of Technology
Beijing, China
{junma,luosenlin}@bit.edu.cn

Jianguo Yao*, Shuxin Cheng
School of Software
Shanghai Jiao Tong University
800 Dongchuan Road
Minhang, Shanghai 200240, China
{jianguo.yao,reallytrue1262}@sjtu.edu.cn
*Corresponding author: jianguo.yao@sjtu.edu.cn

Xi Chen
School of Computer Science
McGill Universiy
Montreal QC Canada
xi.chen7@mail.mcgill.ca

Abstract: Customer reviews and comments on web pages are important information in our daily life. For example, we prefer to choose a hotel with positive comments from previous customers. As the huge amounts of such information demonstrate the characteristics of big data, it places heavy burdens on the assimilation of the customer-contributed opinions. To overcoming this problem, we study an efficient opinion summarization approach for a set of massive user reviews and comments associated with an online resource, to summarize the opinions into two categories, i.e., positive and negative. In this paper, we proposed a framework including: (1) overcoming the big data problem of online comments using the efficient online-LDA approach; (2) selecting meaningful topics from the imbalanced data; (3) summarizing the opinion of comments with high precision and recall. This framework is different from much of the previous work in that the topics are pre-defined and selected the topics for better opinion summarization. To evaluate the proposed framework, we perform the experiments on a dataset of hotel reviews for the variety of topics contained. The results show that our framework can gain a significant performance improvement on opinion summarization.
Keywords: Opinion summarization, Latent Dirichlet Allocation (LDA), online - LDA, imbalanced data, big data.

1 Introduction

The rapid development of the Web 2.0 application makes tremendous and diverse information flood the web. We have to admit that the information shows a wide variety of the meanings which may hardly grasp without summarization. Even worse, the data contained this information shows the characteristics of big data and brings the challenge to the efficiency of the data processing. With more and more user-contributed reviews and comments on the Web, the corresponding websites can become more popular resources that reflect the attitudes and interests of the users in a way that depart from the advertisement and the content of the underlying information resource itself.

Many techniques have been developed to extract concise information from these contents, such as sentiment classification, text summarization and topic modeling [3] [4] [7]. Nevertheless, the comments on the web are updated unceasingly, it is hard to perform online opinion summarizing with these current techniques. Even though these comments are meant to be useful, the vast opinions summarized are still not easily digested and exploited by the users.

When we want to make a comparison of electronic products such as cell phones and laptops, common attributes of the products under consideration include ease of use, battery life, sound quality, Add on s etc. Actually, on most of eCommerce websites, these attributes are pre-defined topics/features and mainly describe hardware performance. Let we say laptops, because of the system's original configuration, the user's experiences can be completely different even with the same hardware. And the after-sales services are also a major concern of the user which only can be reflected in the comments. Thus the pre-defined topics do not demonstrate much diversity on different products. The user's comments are valuable information resource that needs to be summarized.

On *tripadvisor*[1], in order to make an easy comparison of hotels, the scalar rating mechanism is built on the websites for users. But the scalar ratings, e.g. scores between 1 and 5, are not very helpful for hotel managers or tourists because the numeric value does not provide the subjectivities or opinions that come from customer experiences. Also, these scalar ratings are not comparable: for example, when a 3-star hotel receives a high score from 10 tourists while a 4-star hotel receives a medium score from 1 tourist, that does not imply that the former one is better than the latter. In this situation, how to obtain valuable information from users' comments is more important. Furthermore, personal experiences about each hotel cannot totally the same. Consider two typical hotel comments shown in figure 1. These two comments discuss several different topics of the hotel, such as price, room, food etc. The same topics are also in the comments, such as room, breakfast. Apparently, the topics in hotel comments show more diverse topics than electronics products comments. It is impossible to list all the topics tourists may share. Extracting meaningful topics from the comments is not an easy task.

Hotel comments show a very interesting phenomenon of imbalance. The hotels with more comments imply that this hotel is popular and the tourists posting the comments are more likely to share their good experience with others. So the positive comments are far more than the negative comments. The situation of the less popular hotel is quite different, in that less comments will be posted if the tourists had a bad experience. The imbalanced data is the big problem for summarization in form of binary classification.

In our framework, we use scheme of online topic extraction in coping with big data problem. Online inference is employed to handily analyzes the huge number of comments in stream form. There is a superior advantage that makes online LDA process massive collection without heavy computational cost and memory necessity.

Due to the imbalance of the hotel comments, the meaningful topic selection is another challenging problem to opinion summarization. In our framework, topic selection is carried on with multi-facets of consideration. In comparing with three *ROC* based topic selection methods, FAST [19] is the best one in handling the extracted topics with the problem of the imbalance, and relative low computation is needed. Furthermore, better opinion summarization is obtained with any redundancy topics filtered out and accuracy in classification. In our evaluation, we observe that our framework avoids several problems faced by supervised classification approach.

The aim of the present work is to study the manner in which hotel comments can be summarized into positive and negative opinions with meaningful selected topics, so that the obtained summary can be used in real life. Our main contributions are summarized as follows:

[1] www.tripadvisor.com

Figure 1: Different Topics on Hotel Comments

- We present a framework of comments summarization and the online variational methods are used to handle huge amounts of comments from the web in coping with the big data.
- We address the problem of data imbalance of hotel comments. Different from existing works on pre-defined topics, topic selection is performed with the consideration of the more positive comments and less negative comments.
- The ratable topics can be a form of summary and the opinion summarization is performed with these topics for easy digest and exploitation. The experiments are conducted on comments crawled from *tripadvisor*. Several metrics are used for the evaluation, and experimental results show that our proposed framework can summarize the comments in a good manner.

The rest of the paper is organized as follows. Section II surveys existing studies on comments summarization in topic models. In Section III, we propose a framework of the opinion summarization and discuss the topics/features involved in this task and the challenges it implies, in comparison to other LDA based text summarization. In Section IV, we propose a different approach to analyze data imbalance. The evaluation results using several metrics are reported in Section V. In Section VI, we offer insights on the challenges of opinion summarization and point out clear directions in which further improvements can be made.

2 Related Work

We first review the research works related to topic modeling. We then give a brief overview of opinion summarization using other techniques, and we discuss the difference between our framework and sentiment classification lastly.

Topic modeling. Topic Sentiment Model (TSM) [17] is based on the pLSI model [7] which is used to extract the topics. While they set the topics into three sub-topics: neutral, positive and negative topics, the generation progress of documents is considered as first choosing the sub-topics, then choosing the topics in these sub-topics. Y. Lu et al. [13] used a two-step strategy to integrate the opinions with a pulse. The first step is to divide the opinion documents into expert opinions and ordinary opinions. They called it semi-supervised pLSA because the topics are found from expert opinions on the second step, then use as the defined aspect to cluster ordinary opinions.

Latent Dirichlet Allocation (LDA) [3] is another representative topic model which provides a basis for textual-level summarization in an unsupervised way. Supervised latent Dirichlet allocation (sLDA) model [4] accommodates a response variable to make the LDA model work under a supervised condition with the facility of the classification. Multi-grain LDA (MG-LDA) model [10] manipulates the LDA model to induce multi-grain topics. The main idea of the model is to find a ratable aspect within texts on a given topic and use this rating information to identify more coherent aspects. Labeled LDA [5] model is a supervised model with the ability of the k-classification. Joint sentiment/topic (JST) model [11] is a four-layer probabilistic model with the extension of three hierarchical layers LDA which can perform sentiment classification under fully unsupervised way. Z. Ma et al. [21] proposed two topic models, MSTM and EXTM to extract the topics from the documents and its comments respectively, then select representative comments from comment clusters.

Opinion summarization with the LDA related model is multi-faceted and very involved. These approaches can have scalability issues.

Comments summarization. Comments summarization involves two major steps, topic identification and classification. Generally existing research is to classify the comments according to their polarity, which is positive or negative [1] [2] [**?**]. This kind of summarization on comments can give a very general notion of what the users feel about the product. The accuracy of the classification heavily depends on the identified topics and the distance measure. LDA models are one way of topic identification. NLP-based techniques are the other ways to identify topics in the text [13] [14] [15] [16] employed pointwise mutual information and cosine distance as distance measures to perform the binary classification and found that the latter one leads to better accuracy.

Our proposed framework is different from the LDA with the employment of online inference in handling big data. And we also address the imbalance problem of hotel comments. Our work aims at improving the accuracy and scalability of opinion summarization model and inferring meaningful topics for better summarization of the comments.

3 A Better Way to Extract the Topics

In this section, we demonstrate the framework of our model. A brief analysis is given to hotel comments. Then we compare different approaches of topic extraction and highlight the advantage of online inference LDA model for comments topic extraction.

3.1 The Framework

Imagine booking a hotel on the web. We may not review every comment on each hotel and furthermore we could not find the sentiment changes within a long term. So how can we manage and digest the large information other tourists provided? LDA model is well-known algorithm for discovering the main themes of large and unstructured documents. Our framework is combining the LDA model and a *ROC*-based topic selection.

Opinion summarization includes three steps. The first step is topic extraction, the online inference for the LDA model is used for improving its scalability. The potential topics of the comments may not be evaluated properly because the topic number k of LDA is pre-defined by the user. This means that not all the topics extracted are meaningful for opinion summarization, or good for the classification of positive and negative. So we perform the topics selection (extracted topics are the features for classification) in a second step. The third step is opinion summarization or binary classification. As we can see in figure 2, the collected comments show

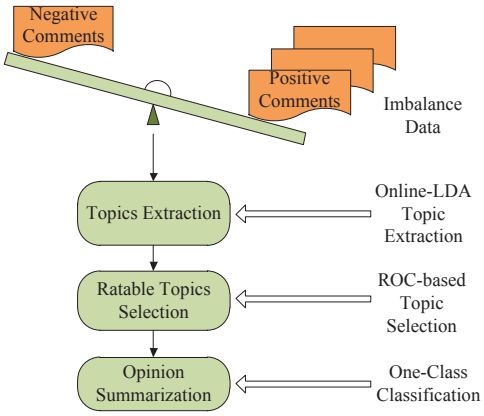

Figure 2: Framework of our model

the characteristic of imbalance (We detail the reason in the next section) which pose a server problem on classification. The *ROC*-based topic selection is used in our framework for better classification. The relevant algorithm is described in section 3.

3.2 Topic Extraction on Comments

There are several probabilistic models to extract the topics - unigram model, multi-gram model, policy model and LDA model. The fundamental idea of these models is that the comments analyzed are considered to have one or some pre-defined topics. The difference is that each one is based on different statistical assumptions.

Probabilistic latent semantic indexing (pLSI) model introduces two probability layers to reduce the constrains on the number of topics and mixture weights of each topic. The probability of a comment is:

$$p(d, w_n) = p(d) \sum_z p(w_n|z)p(z|d). \tag{1}$$

But these topics mixtures are only for those training comments and cannot be used for previously unseen comments. Furthermore, pLSI is a model prone to overfitting in training. So this model is not a well-defined generative model either.

Latent Dirichlet Allocation (LDA) is an extension of the pLSI which introduce a Dirichlet prior on topics. Here we denote as θ. The generative process includes two steps: one is to choose θ for the Dirichlet prior on topics, then choose a word from $p(w_n|\theta, \beta)$.

This process is a continuous mixture distribution:

$$p(c_i|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} p(w_n|\theta, \beta) \right) d\theta, \tag{2}$$

$p(\theta|\alpha)$ are the mixture weights on topics.

In this work, LDA is used to extract topics from hotels comments collection. A unified topic model is trained on the integrated content by combining the multiple text fields within each comment together. Given a comments collection $C = c_1, c_2, , c_N$ where N denotes the comments number, each comment c_i is assigned a distribution over K topics learned from the comments collection where K denotes the pre-defined topic number.

Hotels comments are updated frequently on their pages. A model using a supervised learning algorithm cannot well generalize profiles of new comments. LDA model can be used for new comments and can characterize the comments under the unsupervised form in term of the estimated posterior distribution. Usually this posterior cannot be computed directly [3], and is mostly approximated using Markov Chain Monte Carlo (MCMC) methods or variational inference. The realization of the particular MCMC method, the Gibbs sampling algorithm [18], is widely used to LDA based comments collection modeling.

The applicability of Gibbs sampling depends on the ease with which the sampling process creates separate variables for each piece of observed data and fix the variables in question to their observed values, rather than sampling from those variables. Gibbs sampling generates a Markov chain of variables, each of which is correlated with nearby variables. Each step of the Gibbs sampling procedure involves replacing one of the variables with a value drawn from the distribution of that variable conditioned on the values of the remaining variables. Thus the algorithm converges much slowly when handling high-dimensional data.

As we know, the variational method is a deterministic alternative to sampling-based algorithms. The only assumption made for variational method is the factorization between hidden variables and visible variables. Thus, the inference problem is transformed into an optimization problem as the equation shows:

$$\mathcal{L}(\omega, \phi, \gamma, \lambda) \triangleq \mathbb{E}_q[\log p(\omega, z, \theta, \beta|\alpha, \eta] - \mathbb{E}_q[\log q(z, \theta, \beta)] \tag{3}$$

ϕ, γ, are the parameters of z and θ, λ is the parameter of the topics β. The variational inference may converge faster than Gibbs sampling. However, it still requires a full pass through the entire collection each iteration. It can therefore be time and memory consuming in the application to large and stream coming comments collection.

Hoffman et al. [8] proposed a much faster online algorithm for the variational inference of LDA. This time a fully factorized variables is used, then the lower bound is defined as

$$\mathcal{L} \triangleq \Sigma_d \ell(n_d, \phi_d, \gamma_d, \lambda), \tag{4}$$

The online variational inference comes from the best setting of the topics λ. After estimating the $\gamma(n_d, \lambda)$ and $\phi(n_d, \lambda)$ on seen comments, then set λ to maximize

$$\mathcal{L}(n, \lambda) \triangleq \Sigma_d \ell(n_d, \gamma(n_d, \lambda), \phi(n_d, \lambda), \phi_d), \tag{5}$$

The convergence of the online inference had been analyzed and proved much faster than other variational methods.

The hotel comments on *tripadvisor* keep increasing dramatically as we have seen. The scalability is an unavoidable challenge for processing the data set in real time. Online variational inference for LDA can be much more useful in dealing with a high volume of data and it can handily analyze massive collections of comments. Moreover, online LDA need not locally store or collect the comments- each can arrive in a stream and be discarded after one look. Refer to [8] for a detail analysis of online variational inference for LDA.

4 Opinion Summarization from Topic Selection

In this section we briefly illustrate the data imbalance problem and two topics selection methods, then we describe the algorithm, used in our framework for opinion summarization from imbalanced data topic selection.

Further to the topic extraction described in the previous section, we explore the impact of the topics on the summarization performance. Intuitively, opinion summarization can be different from the summarization of factual data, as comments regarded as informative from the factual point of view may contain little or no sentiment. So, eventually, they are useless from the sentimental point of view. The main question we address at this point is: how can we determine the informative extracted topics for opinion summarization.

The comments collection we clawed from *tripadvisor* demonstrates the imbalance problem of more positive and less negative. The data imbalance presents a unique challenge to classify the comments from the extracted topics. Precision and recall are widely used measurements for classification performance. The precision for a class is the number of true positives divided by the total number of elements labeled as belonging to the positive class. Recall is the number of true positives divided by the total number of elements that actually belong to the positive class. Consider the two topics sets on text classification, the first topic set may yield higher precision, but lower recall, than the second topics set. By varying the decision threshold, the second topic set may produce higher precision and lower recall than the first topic set. Thus, one single threshold cannot tell us which extracted topic set is better. The topic selection needs serious consideration.

Commonly, there are two methods to select topics: the first is rank topics in descending order with the related criteria, such as *ROC*, then choose the top, say, l topics. The second is more complicated with computation of cross-correlation coefficient between topics. Scatter matrices are belong to the first method.

$$J_3 = \mathrm{trace}\{S_w^{-1} S_b\} \tag{6}$$

where $S_w = \sum_{i=1}^{c} P_i S_i$, P_i is the a priori probability of class w, S_i is the mean vector of class w, S_w is the within-class scatter matrix, and S_b is the between-class scatter matrix.

The cross-correlation coefficient is the second method to topic selection. Let i_1 be the best topic selected using the first method.

$$i_2 = \max_j \{a_1 R_j - a_2 |\rho_{i_1,j}|\}, \quad j \neq i_1 \tag{7}$$

This equation considers the cross-correlation ($\rho_{i_1,j}$)between the best topic and topic $j \neq i_1$. The rest of the topics are ranked according to

$$i_k = \max_j \{a_1 R_j - \frac{a_2}{k-1} \sum_{r=1}^{k-1} |\rho_{i_r,j}|\}, \tag{8}$$

$$j \neq i_r, \quad r = 1, 2, \ldots, k-1$$

These two methods are designed for well-balanced data and if the data dimension is high, the effectiveness of the topic selection is a severe problem for classification. Most comments we crawled from *tripadvisor* are about the length of 150 words. In generalizing the comments with the LDA model, the pre-defined topic number k is set to 20, 30 rather than 100 because of the relative short length of each comment. Even this moderate number may produce high dimension problems for the topic selection. The detailed analysis is presented in section V. Meanwhile, the computational cost is another bottleneck, even we employ the topic extraction with online-LDA,

the topic selection is a time-consuming task with these two methods because of the computing of matrix inverse. So in the proposed framework, we use FAST [19] to perform the topic selection for opinion summarization. The topic selection metric is based on an *ROC* curve generated on optimal simple linear discriminants. Then those topics with the highest *AUC* (Area Under Curve) are selected as the most relevant. This method is designed for the topics selection of imbalanced data classification.

A *ROC* curve is a criterion for ranking the topics, FAST employs a new threshold determination method which fixes the number of points to fall in each bin to obtain the threshold for *ROC*. Bin means the width of data separation. We use more bins in high density data areas and fewer bins in sparse data area, each bin containing the same number of data. Thus more thresholds computed from each bin are placed into the density area for the calculation of the *ROC*. On the opposite, fewer thresholds placed into the sparse area. This effective procedure can be described as following pseudo-code in Algorithm 1:

Algorithm 1 Pseudo-code of effective procedure.

```
K: number of bins
N: number of comments
T: number of topics
InBin=0 to N with a step size T/20
for i = 1 to T do
  Sort T_i(T_i is the ith value of vector T)
  for j = 1 to K do
    Bottom=round(InBin(j))+1
    Top=round(InBin(j+1))
    Threshold=mean(T_i(Bottom to Top))
    Classify T_i
  end for
  Calculate the AUC (Area Under Curve)
end for
```

The detailed analysis of the algorithm is in [19]. The benefit applied is not merely about selected topics for classification, the computational cost of the algorithm is relatively low because no matrix inverse is calculated. Because the area under the *ROC* curve is a strong predictor of performance, especially for imbalanced data classification problems, we can use this score as our topic selection: we choose those topics with the highest areas under the curve because they have the best predictive power for comment collection.

5 Experiments

The experiments is to evaluate the model that produces opinion summaries of comments, in the context of which we assess the best manner to use summarization opinion for the users to quickly digest. In this section, we will present and discuss the experimental results of topic selection and opinion summarization on the hotel comments dataset.

5.1 Dataset

We crawled 250,004 hotel comments from *tripadvisor* in one month period (from Nov, 2012 to Dec, 2012). The comments in the dataset are labeled according to 5 scales of 'star' expressing the polarity of the opinion of the reviewers (1, 2 corresponding to negative comments and 4,

5 corresponding to positive comments). Since we summarize the opinions into two classes of sentiment (positive and negative), the neural comments (scale 3) are excluded from the comments collection. Figure 4 shows the statistical information about the comments collection. Most comments are within a length of 150 words.

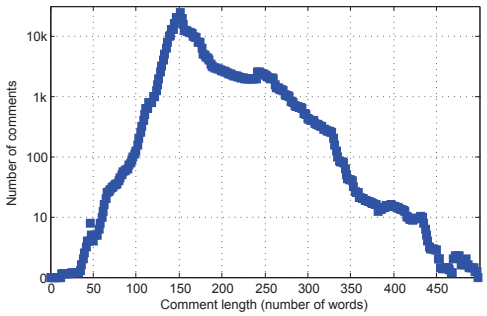

Figure 3: The statistical information of the comments collection

5.2 Topic Selection

The first experiment is performed to evaluate the generalization performance of the online-LDA model. As we pointed out in Section III that the LDA with online inference can handle massive datasets much faster than other methods such as variational inference, Gibbs sampling. We need to verify that there is no generalization performance degeneration using online inference for LDA. We compared online-LDA model with pLSI models and LDA model described in Section 3. In this experiment, we used all the comments crawled from *tripadivor* containing 250,004 comments. We held out 10% of the collection for test purposes and trained the models on the remaining 90%. We have found $\alpha = 50/T$ and $\beta = 0.01$ to work well with hotel comments collection for LDA model and online-LDA model.

The perplexity [3] is used as the measurement for the evaluation of the models. As it is the standard metric and it measures the model's ability of generalizing unseen data; lower perplexity indicates the higher likelihood and better model performance.

We trained these three models using EM with the stopping criteria, that the average change in expected log likelihood is less than 0.001%.

Figure 4 presents the perplexity for each model in terms of the comments analyzed. Three models were trained from the crawled comments without looking at the same comment twice. It can be seen that the online-LDA model have a lower perplexity than pLSI and LDA model after analyzsis of the same number of comments. This superior advantage comes from the fact that the online variational inference converged much faster than variational bayes used in LDA [8]. After analyzing the total comments collection, both LDA models reached the same level of perplexity about 1700. The generalization performance of online-LDA is as good as the LDA with an extra advantage of much faster fitting to the comments. The perplexity of pLSI is 1900. So the results show that online-LDA is more adapted to coming comments under online environment.

Besides fast topics extraction, our summarization framework employed a lower cost topics selection method in coping with the imbalanced nature of hotel comments. The experiment performed next is to evaluate the performance of topic selection. The balanced error rate (BER) is the main judging criterion for the topic selection [20].BER is the average of the error rates of positive comments and negative comments. If these two classes are balanced, the BER is equal to the error rate as the rate of inverse recall [6]. We evaluated the performance of selected topics

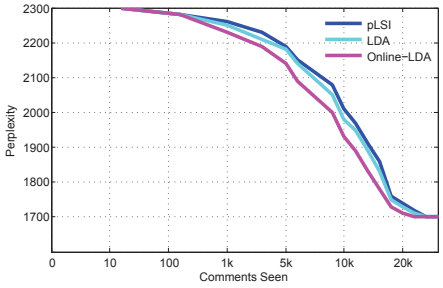

Figure 4: The perplexity results for pLSI, LDA and Online-LDA.

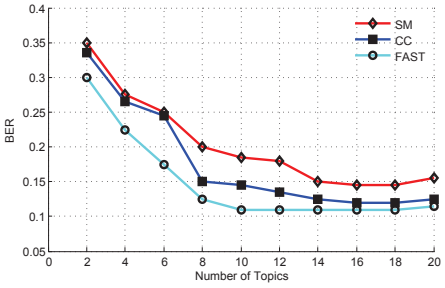

Figure 5: BER with $k = 20$ using SVM.

in our framework (FAST) with the comparison of the topics selected by scatter matrices (SM) and cross-correlation coefficient (CCC). The main concern in our framework is the performance of the topic selection metric, so we simply choose the popular SVM classifier to evaluate the performance without the detail analysis of the difference with other classifiers. Table 2 presents the description of the comments used in BER evaluation.

Table 1: Comments used in evaluation of BER

	Number of Topics	Ratio
Positive Comments	180,023	95.5%
Negative Comments	8,053	4.5%

These comments collection demonstrates the strong imbalanced nature that the negative comments are less than 5% of the total comments. From the previous analysis of data set, we know that most of the comment length is less than 260. So we set $k = 20$, 30 respectively for the online-LDA model, then the extracted topics are used for BER evaluation with the methods described in our work.

Figure 5 and 6 show the result of the performance in terms of BER. We can see that the BER changes dramatically with the different numbers of the topics. We observe that BER decreases as the topics increase when the topics number is less than 9. And then the BER reaches to the relatively stable values of 0.15, 0.1 and 0.08 respectively to SM, CC and FAST. The explanation for this behavior is that the redundant topics have little impact on the performance of the classifier. This robustness might be useful to redundant topics for classification. But our goal is comment summarization rather than classification. Redundant topics can bury informative topics and make the user hard to exploit.

The topics selected using FAST significantly outperformed SM and CCC topics with lower BER when using the SVM classifier. Several experimental results reveal that the lowest BER comes from the 9 selected topics. I.Tiov et al. [10] show that 9 topics out of 45 LDA topics correspond to ratable aspects. This is a quite interesting discovery that the topics we selected are same as the ratable aspects defined by a manual analysis of the documents, as the more topics used in classification, the more freedom we can have to distinguish the polarity of the comments in a finer granularity. But the performance of the classification remains stable to a certain level. The optimal topics come from selected topics with the emerging of this level. This suggests that topic extracted using LDA is not a sufficiently representative topic of the importance of comments for summarization purposes. Thus, using *ROC*-based topic selection that has proven useful for opinion summarization can yield better results.

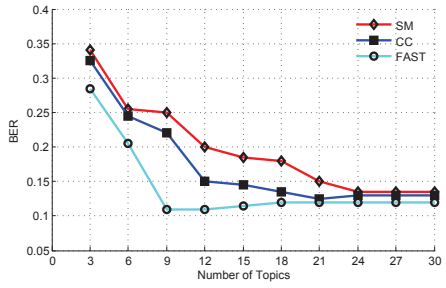

Figure 6: BER with $k = 30$ using SVM.

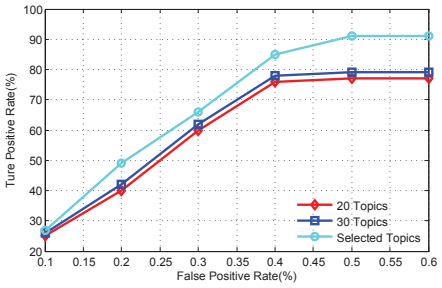

Figure 7: Performance of Summarization in term of ROC.

5.3 Opinion Summarization

As discussed before, a significant advantage of our framework over existing models in topic selection and classification is the lower computational cost in topic extraction and topic selection in imbalanced comments. We only consider the positive and negative comments given data set, with the neutral comments being ignored. There are two main reasons. Firstly, hotel comments opinion summarization in our case is effectively a binary classification problem, i.e. comments are being classified as either positive or negative, without the alternative of neutral. Secondly, the selected topics merely contribute to the positive and negative words, and consequently there will be much more influence on the summary results of positive and negative comments given data set. Furthermore, the classification with less negative comments shows the unique similarity of the outlier detection. As a result, we choose to evaluate the overall performance of the opinion summarization indirectly through outlier detection based on the selected topics. More specifically, we apply one-class SVM classifier on the comments with selected topics.

$$f(x) = \text{sign}((\omega \cdot \Phi(x)) - \rho). \tag{9}$$

The regularization parameters ω and ρ solve the quadratic programming problem of the ν-SVM. The classification using these methods is computationally simple and does not require significant memory.

The main problem we encountered is that the lexicon is needed first in our proposed summarization framework. In our case, however, comments are often composed of ungrammatical sentences and, additionally, a high number of unusual combinations of escape characters (corresponding to the vivid sentiment expressed), which make the comments much noisier and harder to process than the standard data sets traditionally used for summarization evaluation. Nevertheless, the online-LDA, being a generative model, proved to be quite robust to variations in the input data and, most importantly, to the change of the domain (Micro-blog etc.) [8]. There are $10,314$ words in our lexicon for the experiments.

Table 2: Performance of Opinion Summarization

Topics	Positive Comments Detection Rates	Negative Comments Detection Rates
Selected Topics	89%	91%
20 topics	80%	77%
30 topics	80%	79%

The results of the opinion summarization are shown in Table 2. The first thing to note in Table 2 is that the opinion summarization model is doing a much better job at classifying the

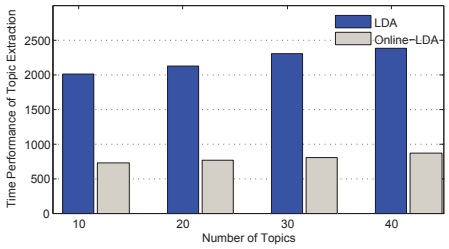

Figure 8: Time Performance of the Topic Extraction.

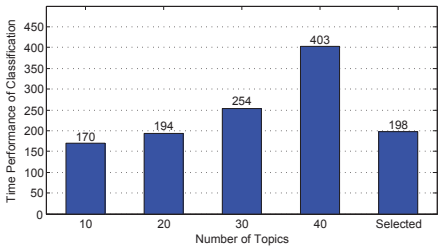

Figure 9: Time Performance of The Classification.

comments according to its polarity than the solo LDA model, the main problem with the latter being a relatively low precision. The main reason for this is an insufficient number of annotated negative examples when performing the topic selection.

The results show that the model is capable of reliably identifying negative comments (Figure 7). It can be observed that there is a considerable improvement in classification accuracy after performing the topics selection with FAST, with 5.3% improvement for our framework.

We evaluate the topic extraction time. We extract the 10, 20, 30 and 40 topics with LDA and online-LDA respectively. Results show that the online-LDA model outperforms the LDA model (Figure 8).

We evaluate the framework's time performance. We extract the 10, 20, 30 and 40 topics with online-LDA, and then perform the topic selection with two topics sets, 20 and 30 topics. We classify the comments in two ways: the first is to classify the comments with original topics, the second is to classify the comments with selected topics. The time performance of the second classification is averaged over two selected topics.

The topic selection used more time when classifying but the classification time does not increase dramatically due to the lower dimension of the data. Results show an extra 28 seconds in comparison with 10 topics (Figure 9) due to the time-consumption of the topic selection. We believe, however, that for the best performance of summarization a 28 second period is considered low enough in handling over $200,000$ comments so that our results indicate an acceptable time performance penalty.

6 Conclusion

In this paper, we have presented a new framework for the summarization of hotel comments. The most useful usage of opinion summarization is a web application. While most of the existing approaches to opinion summarization have not put into much consideration of the scalability of the models. Scalability is the most important task in our proposed framework. The online-LDA model is used for extracting the topics from the huge and increasing comments collection. The generalization performance remains the same but the computational cost is lower in comparison with LDA model. We address the imbalance problem of the comments. And the topics selection method, FAST is used for better classification performance. The selected topics are informative and easy for the user to digest the comments.

There are several directions we plan to investigate in the future. One is the best comments selection when the aim is to brief the comments collection. Another one is the ratable aspects regression analysis for certain kinds of reviews.

Bibliography

[1] A. Divtt and K. Ahmad (2007); Sentiment polarity identification in financial news: A cohesion-based approach, *In ACL'07*, Prague, Czech Republic, June 2007, 1-8.

[2] B. Pang, L. Lee and S. Vaithyanathan (2002); Thumbs up?: sentiment classification using machine learning techniques, *EMNLP'02: Proc of the ACL'02 conference on Empirical Methods in Natural Language Processing, Morristown, NJ, USA*, 10: 79-86.

[3] D.M. Blei, A. Ng and M. Jordan (2003); Latent Dirichlet Allocation, *Journal of Machine Learning Research*, January 2003, 3:993-1022.

[4] D.M. Blei and J.D. McAuliffe (2007); Supervised topic models, *In NIPS'07*, Vancouver, B.C., Canada, 1-8.

[5] D. Ramage, D. Hall, R. Nallapati and C.D. Manning (2009); Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora, *In EMNLP'02: Proc. of the ACL'02 conference on Empirical Methods in Natural Language Processing*,Stroudsburg, PA, USA, 2009.

[6] D.M.W. Powers (2001); Evaluation: Precision, Recall and F-measure to ROC, Informedness, Markedness & Correlation, *Journal of Machine Learning Technologies*, 2(1):37-63.

[7] T. Hofmann (1999); Probabilistic latent semantic indexing, *In SIGIR'99: Proc. of the 22nd Annual Intl. ACMSIGIR Conf. on Research and Development in Information Retrieval*, New York, NY, USA.

[8] M.D. Hoffman, D.M. Blei and F. Bach (2010); On-line learning for Latent Dirichlet Allocation, *NIPS2010, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, Lake Tahoe, Nevada, USA, 50-57.

[9] H. Wang, Y. Lu and CX. Zhai (2011); Latent Aspect Rating Analysis without Aspect Keyword Supervision, *KDD'11, Proc. of the 17th ACM SIGKDD intl. conf. on Knowledge discovery and data mining*, San Diego, California, USA, 618-626.

[10] I. Titov and R. McDonald (2008); A Joint Model of Text and Aspect Ratings for Sentiment Summarization, *Proc. of ACL'08*, Columbus, Ohio, USA, 308-316.

[11] C. Lin and Y. He (2009); Joint Sentiment/Topic Model for Sentiment Analysis, *CIKM'09, Proceedings of the 18th ACM conference on Information and knowledge management*, Hong Kong, China, 375-384.

[12] L.-W. Ku, Y.-T. Liang and H.-H. Chen (2006); Opinion extraction, summarization and tracking in news and blog corpora, *AAAI-CAAW'06, Proceedings ofAAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs*, Stanford, California, USA, 1-8.

[13] Y. Lu, C. Zhai and N. Sundaresan (2009); Rated aspect summarization of short comments, *WWW'09, Proceedings of the 18th international conference on World wide web*, ACM, NY, USA, 131-140.

[14] P.D. Turney (2002); Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, *ACL'02, Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Morristown, NJ, USA, 417-424.

[15] P.D. Turney and D.L. Littman (2003); Measuring praise and criticism: Inference of semantic orientation from assocation, *ACM Trans. Inf. Syst.*, 21(4):315-346.

[16] P. Stenetorp, S. Pyysalo, G. Topic, S. Ananiadou and J. Tsujii (2012); BRAT: a web-based tool for NLP-Assisted text annotation, *EACL '12 Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, 102-107.

[17] Q. Mei, X. Ling, M. Wondra, H. Su and C. Zhai (2007); Topic sentiment mixture: modeling facets and opinions in weblogs, *WWW '07 Proceedings of the 16th international conference on World Wide Web*, Banff, Alberta, Canada, 171-180

[18] B. Walsh (2002); Markov chain Monte Carlo and Gibbs sampling, *Lecture notes for EEB 596z*, 2002.

[19] X. Chen and M. Wasikowski (2008); Fast: A roc-based feature selection metric for small samples and imbalanced data classification problems, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 124-132.

[20] YW. Chen and CJ. Lin (2015); Combining SVMS with various feature selection strategies, *Available: www.csie.ntu.edu.tw/ cjlin/papers/fearutes.pdf.*

[21] Z. Ma, A. Sun, Q. Yuan and G. Cong (2012); Topic-Driven reader comments summarization, *CIKM'12*, Maui, HI, USA, 265-274.