www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 2 Issue 7 (July 2013), Page No. 2189-2194

OPINION AND TOPIC DETECTION USING SENTIMENT CLASSIFIER

Rajendran Rahul [(1)], K.Kishore kumar[(2)] , S.Selvakumar

Department of computer science & Engineering,
PB College of engineering,
Sriperumbudur, Chennai - 602117.
rrl5050@gmail.com [(1)], **mail2kumar90@gmail.com** [(2)]

ABSTRACT

Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks. Sentiment analysis systems are being applied in almost every business and social domain because opinions are central to almost all human activities. This paper proposes a novel probabilistic modeling framework called joint sentiment-topic (JST).JST model based on latent Dirichlet allocation (LDA), supervised approaches to sentiment classification which often fail to produce satisfactory performance when shifting to other domains. The weakly-supervised nature of JST makes it highly portable to other domains. This is verified by the experimental results on datasets from five different domains. We hypothesize that the JST model can readily meet the demand of large-scale sentiment analysis from the web.

Key Terms

Semantics, Topics, Sentiment analysis, opinion mining, latent Dirichlet allocation (LDA), Joint sentiment-topic(JST).

1. INTRODUCTION

With the explosion of Web 2.0, various types of social media such as blogs, discussion forums, and peer-to-peer networks present a wealth of information that can be very helpful in assessing the general public's sentiment and opinions toward products and services. Recent surveys have revealed that opinion-rich resources like online reviews are having greater economic impact on both consumers and companies compared to the traditional media.Among various sentiment analysis tasks, one of them is sentiment classification, i.e.,identifying whether the semantic orientation of the given text is positive, negative, or neutral. Although much work has been done in this line, most of the existing approaches rely on supervised learning models trained from labeled corpora where each document has been labeled prior to training.

1.1 LATENT DIRICHLET ALLOCATION

Latent Dirichlet allocation (LDA**)** is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics.

In LDA, each document may be viewed as a mixture of various topics. This is similar to probabilistic latent semantic analysis (pLSA), except that in LDA the topic distribution is assumed to have a Dirichlet prior. In practice, this results in more reasonable mixtures of topics in a document. It has been noted, however, that the pLSA model is equivalent to the LDA model under a uniform Dirichlet prior distribution. A document is given the topics. This is a standard bag of words model assumption, and makes the individual words exchangeable.

1.1.1 Problems With Latent Dirichlet Allocation

Latent Dirichlet allocation is purely topic-based without considering the associations between topics and sentiments. This can be overcome with Joint Sentiment Topic (JST) which reads information both topics and sentiments.It does not make thorough analysis of text instead gocus is upon the overall sentiment of a document.

1.2 OPINION AND TOPIC DETECTION

Among various sentiment analysis tasks, one of them is sentiment classification, i.e., identifying whether the semantic orientation of the given text is positive, negative or neutral. Although much work has been done in this line most of the existing approaches rely on supervised learning models trained from labelled corpora where each document has been labelled as positive or negative prior to training. However, such labelled corpora are not always easily obtained in practical applications. Also, it is well-known that sentiment classifiers trained on one domain often fail to produce satisfactory results when shifted to another domain, since sentiment expressions can be quite different in different domains These observations have thus motivated the problem of using unsupervised or weakly-supervised approaches for domain-independent sentiment classification. Another common deficiency of the aforementioned work is that it only focuses on detecting the overall sentiment of a document, without performing an indepth analysis to discover the latent topics and the associated topic sentiment.

In general, a review can be represented by a mixture of topics. For instance, a standard restaurant review will probably discuss topics or aspects such as food, service, location, price, and etc. Although detecting topics is a useful step for retrieving more detailed information, the lack of sentiment analysis on the extracted topics often limits the effectiveness of the mining results, as users are not only interested in the overall sentiment of a review and its topical information, but also the sentiment or opinions towards the topics discovered.

2.RELATED WORKS

2.1 SENTIMENT ANALYSIS

With the explosion of Web 2.0, various types of social media such as blogs, discussion forums, and peer-to-peer networks present a wealth of information that can be very helpful in assessing the general public’s sentiment and opinions toward products and services. Recent surveys have revealed that opinion-rich resources like online reviews are having greater economic impact on both consumers and companies compared to the traditional media . Driven by the demand of gleaning insights into such great amounts of user-generated data, work on new methodologies for automated sentiment analysis and discovering hidden knowledge from unstructured text data has bloomed splendidly. Among various sentiment analysis tasks, one of them is sentiment classification, i.e., identifying whether the semantic orientation of the given text is positive, negative, or neutral. Although much work has been done in this line , most of the

existing approaches rely on supervised learning models trained from labeled corpora where each document has been labeled as positive or negative prior to training.

2.2 SENTIMENT CLASSIFICATION

Machine learning techniques have been widely deployed for sentiment classification at various levels, e.g., from the document level, to the sentence and word/phrase level. On the document level, one tries to classify documents as positive, negative, or neutral, based on the overall sentiments expressed by opinion holders. There are several lines of representative work at the early stage . Turney used weakly supervised learning with mutual information to predict the overall document sentiment by averaging out the sentiment orientation of phrases within a document. Pang et al. classified the polarity of movie reviews with the traditional supervised machine learning approaches and achieved the best results using SVMs. In their subsequent work, the sentiment classification accuracy was further improved by employing a subjectivity detector and performing classification only on the subjective portions of reviews. The annotated movie review data set (also known as polarity data set) used in and has later become a benchmark for many studies. Whitelaw et al. used SVMs to train on combinations of different types of appraisal group features and bag-of-words features, whereas Kennedy and Inkpen leveraged two main sources, i.e., General Inquirer and Choose the Right Word, and trained two different classifiers for the sentiment classification task.

As opposed to the work that only focused on sentiment classification in one particular domain, some researchers have addressed the problem of sentiment classification across domains. Aue and Gamon explored various strategies for customizing sentiment classifiers to new domains, where training is based on a small number of labeled examples and large amounts of unlabeled in-domain data. It was found that directly applying a classifier trained on a particular domain barely outperforms the baseline for another domain. In the same vein, more recent work focused on domain adaptation for sentiment classifiers. However, their approach relies on labeled data from all domains to train an integrated classifier and thus may lack flexibility to adapt the trained classifier to other domains where no label information is available.

2.3 SENTIMENT-TOPIC MODELS

JST models sentiment and mixture of topics simultaneously. Although work in this line is still relatively sparse, some studies have preserved a similar vision. Most closely related to our work is the Topic-Sentiment Model (TSM) [1], which models mixture of topics and sentiment predictions for the entire document. However, there are several intrinsic differences between JST and TSM. First, TSM is essentially based on the probabilistic latent semantic indexing (PLSI) model with an extra background component and two additional sentiment subtopics, whereas JST is based on LDA. Second, regarding topic extraction, TSM samples a word from the background component model if the word is a common English word. Other models by Titov and McDonald [2], [3] are also closely related to ours, since they are all based on LDA. The Multi-Grain Latent Dirichlet Allocation model (MG-LDA) [3] is argued to be more appropriate to build topics that are representative of ratable aspects of customer reviews, by allowing terms being generated from either a global topic or a local topic. Being aware of the limitation that MG-LDA is still purely topic-based without considering the associations between topics and sentiments, Titov and McDonald further proposed the Multi-Aspect Sentiment model (MAS) [2] by extending the MG-LDA framework.

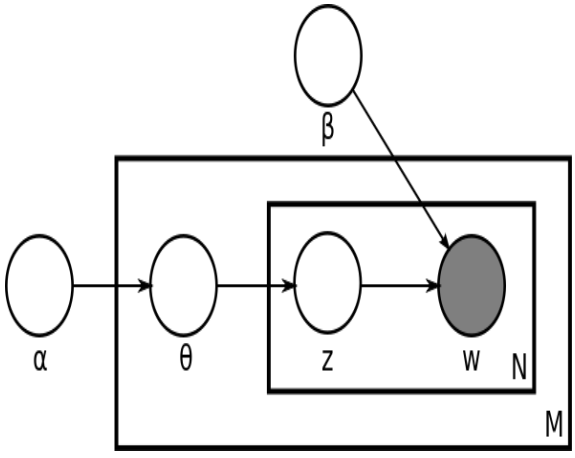

Fig. 1. LDA Model

α is the parameter of the Dirichlet prior on the per-document topic distributions.

β is the parameter of the Dirichlet prior on the per-topic word distribution.

θ_i is the topic distribution for document *i*,

ϕ_k is the word distribution for topic *k*,

z_{ij} is the topic for the *j*th word in document *i*,

w_{ij} is the specific word.

3. PROPOSED FRAMEWORK

3.1 JST (JOINT SENTIMENT TOPIC)

The LDA model, as shown in Fig. 1, is based upon the assumption that documents are mixture of topics, where a topic is a probability distribution over words [9], [10]. Generally, the procedure for generating a word in a document under LDA can be broken down into two stages. One first chooses a distribution over a mixture of T topics for the document. Following that, one picks a topic randomly from the topic distribution, and draws a word from that topic according to the corresponding topic-word distribution.

The existing framework of LDA has three hierarchical layers, where topics are associated with documents, and words are associated with topics. In order to model
document sentiments, we propose a joint sentiment-topic model [4] by adding an additional sentiment layer between the document and the topic layers. Hence, JST is effectively a four-layer model, where sentiment labels are associated with documents, under which topics are associated with sentiment labels and words are associated with both sentiment labels and topics.

Assume that we have a corpus with a collection of D
documents denoted by C ¼ fd1; d2; . . . ; dDg; each document in the corpus is a sequence of N_d words denoted by
d ¼ ðw1; w2; . . . ; w_{N_d} Þ, and each word in the document is an
item from a vocabulary index with V distinct terms denoted
by f1; 2; . . . ; V g. Also, let S be the number of distinct sentiment labels, and T be the total number of topics. The procedure for generating a word w_i in document d under
JST boils down to three stages. First, one chooses a sentiment label l from the per-document sentiment distribution 6_d. Following that, one chooses a topic from the topic distribution $/_{d;l}$, where $/_{d;l}$ is conditioned on the sampled sentiment label l. It is worth noting that the topic distribution of JST is different from that of LDA. In LDA, there is only one topic distribution / for each individual document. In contrast, in JST each document is associated with S (the number of sentiment labels) topic distributions, each of which corresponds to a sentiment label l with the same number of topics. This feature essentially provides means for the JST model to predict the associated sentiment with the topics extracted.

3.2 JST MODEL

This model extends the state-of-the-art topic model latent Dirichlet allocation (LDA), by constructing an additional sentiment layer**.** It is highly portable when shifted to other domains .

JST is effectively a four-layer model, where sentiment labels are associated with documents, under which topics are associated with sentiment labels and words are associated with both sentiment labels and topics.

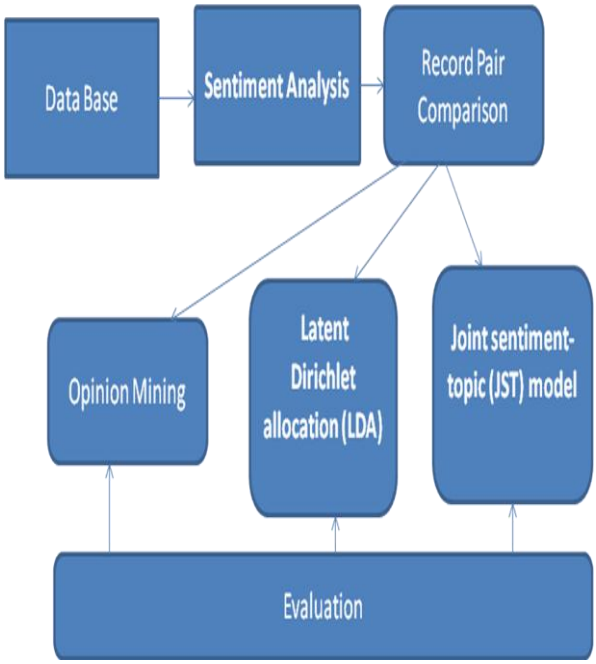

Fig.2. JST Model

4.EXPERIMENTS

4.1 DATA SETS

Two publicly available data sets, the MR and MDS data sets, were used in our experiments. The MR data set has become a benchmark for many studies since the work of Pang et al. [5]. The version 2.0 used in our experiment consists of 1,000 positive and 1,000 negative movie reviews crawled from the IMDB movie archive, with an average of 30 sentences in each document. We also experimented with another data set, namely subjective MR, by removing the sentences that do not bear opinion information from the MR data set, following the approach of Pang and Lee [6]. The resulting data set still contains 2,000 documents with a total of 334,336 words and 18,013 distinct terms, about half the size of the original MR data set without performing subjectivity detection.

.

4.2 CLASSIFYING DOCUMENT

The document sentiment is classified based on P ðljdÞ, the probability of a sentiment label given document. In our experiments, we only consider the probability of positive and negative labels for a given document, with the neutral label probability being ignored. There are two reasons for this. First, sentiment classification for both the MR and MDS data sets is effectively a binary classification problem, i.e., documents are being classified either as positive or negative, without the alternative of neutral. Second, the prior information we incorporated merely contributes to the positive and negative words, and consequently there will be much more influence on the probability distribution of positive and negative labels for a given document, rather than the distribution of neutral labels in the given document. Therefore, we define that a document d is classified as a positive-sentiment document if the probability of a positive sentiment label P $ðl_{pos}jdÞ$ is greater than its probability of negative sentiment label P $ðl_{neg}jdÞ$, and vice versa.

ship	good	recip	children	action	funni
titan	realli	food	learn	good	comedi
crew	plai	cook	school	fight	make
cameron	great	cookbook	child	right	humor
alien	just	beauti	ag	scene	laugh
jack	perform	simpl	parent	chase	charact
water	nice	eat	student	hit	joke
stori	fun	famili	teach	art	peter
rise	lot	ic	colleg	martial	allen
rose	act	kitchen	think	stunt	entertain
boat	direct	varieti	young	chan	funniest
deep	best	good	cours	brilliant	sweet
ocean	get	pictur	educ	hero	constantli
dicaprio	entertain	tast	kid	style	accent
sink	better	cream	english	chines	happi
prison	bad	polit	war	horror	murder
evil	worst	east	militari	scari	killer
guard	plot	middl	armi	bad	cop
green	stupid	islam	soldier	evil	crime
hank	act	unit	govern	dead	case

Fig.3. Classifying Labels By JST

4.3 TOPIC EXTRACTION

The second goal of JST is to extract topics from the data sets, and evaluate the effectiveness of topic sentiment captured by the model. Unlike the LDA model where a word is drawn from the topic-word distribution, in JST one draws a word from the per-corpus word distribution conditioned on both topics and sentiment labels.

HelloSir,

I am glad to inform you that my marriage has been fixed for 1st May 2013. I will be grateful if you could grant me 15 days leave from May

1st2013.

The wedding cards are still not printed. I will send you a copy of the same when its printed. I would be glad if you could make your presence felt for my marriage. hoping that you will accept my humble request

Thanks for your consideration.

Yours Sincerely,

Kumar

Fig4.Opinions And* Topics *Extracted By JST

5.CONCLUSION

Joint sentiment-topic (JST) model targets sentiment and topic detection simultaneously in a weakly-supervised fashion, while most of the existing approaches to sentiment classification favor in supervised learning.,JST consistently outperformed existing models. For general domain sentiment classification, by incorporating a small amount of domain-independent prior knowledge,the JST model achieved either better or comparable performance compared to existing semi-supervised approaches despite using no labelled documents, which demonstrates the flexibility of JST in the sentiment classification task. The future work is extended to opinion mining by analyzing the people's attitudes and feelings towards products and services by making an indepth analysis.

REFERENCES

[1] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs," Proc. 16th Int'l Conf. World Wide Web (WWW), pp. 171-180, 2007.

[2] I. Titov and R. McDonald, "A Joint Model of Text and Aspect Ratings for Sentiment Summarization," Proc. Assoc. Computational Linguistics—Human Language Technology (ACL-HLT), pp. 308-316, 2008.

[3] I. Titov and R. McDonald, "Modeling Online Reviews with Multi-Grain Topic Models," Proc. 17th Int'l Conf. World Wide Web,pp. 111-120, 2008.

[4] C. Lin and Y. He, "Joint Sentiment/Topic Model for Sentiment Analysis," Proc. 18th ACM Conf. Information and Knowledge Management (CIKM), pp. 375-384, 200911-120, 2008.

[5] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up?: Sentiment Classification Using Machine Learning Techniques," Proc. ACLConf. Empirical Methods in Natural Language Processing (EMNLP),pp. 79-86, 2002

[6] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts,"Proc. 42th Ann. Meeting on Assoc. for Computational Linguistics
(ACL), pp. 271-278, 2004.

[7] S. Lacoste-Julien, F. Sha, and M. Jordan, "DiscLDA: DiscriminativenLearning for Dimensionality Reduction and Classification," Proc.Neural Information Processing Systems (NIPS),2008.

[8] D. Ramage, D. Hall, R. Nallapati, and C. Manning, "Labeled LDA:A Supervised Topic Model for Credit Attribution in Multi-Labeled Corpora," Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), pp. 248-256, 2009.

[9] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," J. Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[10] M. Steyvers and T. Griffiths, "Probabilistic Topic Models," Handbook of Latent Semantic Analysis, vol. 427, no. 7, pp. 424-440, 2007.