Analysis of Techniques of Sentiment and Topic Detection

Supriya Paul
Mtech student
Dept of Comp. Sci. & IT,
Dr B. A. M. University, Aurangabad

SachinDeshmukh
Assistant Professor
Dept of Computer Science and IT,
Dr. B. A. M. University,
Aurangabad

ABSTRACT

User generated media like blogs, discussion forums is used as a tool by people to communicate their experiences with others. Presence of such huge data on Internet demands proper means to generate processed information from the unstructured data. What users need is more than mere sentiments. They need to know public opinion or experience of various aspects of a product like how is camera quality of the phone or energy efficiency of electronic products. For meeting the high demands of users, various techniques have been proposed till date. In this paper we are evaluating, all these techniques that discover topic along with sentiment associated with it. Many models were proposed to incorporate sentiment analysis with topic model to find aspects of a product and users sentiment about the aspect. Results of these models can be beneficial for various industries as well as users.

General Terms

Text classification, text mining, data mining

Keywords

Aspect detection, sentiment analysis, topic modeling, opinion mining, latent Dirichlet allocation (LDA).

1. INTRODUCTION

With evaluation of Internet, users can post their views or experiences regarding a product or service for global audience which was previously limited to a small circle. It provides users with a device to make their voice and choice hear by appropriate authorities in a faster and efficient way. Various types of social media like blogs, discussion forums offer tons of information regarding public opinion about various products and services. The high amount of user opinions or views in the form of unstructured text data demands for various techniques to unearth information vital for various applications. Applications developed on these opinions generally comprise of opinion search for users, keeping track of opinions for the purpose of business intelligence, and prediction of user behavior for target marketing. The sentiment of a word is dependent on the domain or topic so it is appropriate to consider the sentiment along with the topic. Furthermore, in addition to the overall sentiment polarity of the document, people may be interested in the sub topics expressed in the document. From the perspective of a user reading the reviews to get information about a product, the evaluations of the specific aspects are just as important as the overall rating of the product.

Most of methods employed on review dataset usually concentrate on sentiment analysis only and doesn’t light on driving force behind those sentiments. Awareness of positive, negative opinion of aspect/facet about the given product is crucial in order to make more accurate predictions and deductions. For example, energy efficiency as well as cost of an air conditioner may be considered after the cooling capacity. So it may happen that various aspects of a product or service may have different sentiments attached to them which in turn can be different from overall sentiment of the document. This demand for the means which can find out cause of sentiment along with the sentiment. In this paper our attempt is to study such approaches that comprehend topics and their related sentiments from text data.

In rest of the paper, section 2 describes various methodologies applied in individual papers for extracting sentiment and topics, section 3 involve result analysis of each technique.

2. METHODOLOGIES

2.1 Topic Sentiment Mixture (TSM) Model

This is one of the first attempts to obtain sentiment along with topic from available text. In this model, they were assuming that a blog article is obtained by sampling words from a mixture model which include a background language model, a set of topic models and two sentiment (positive and negative) language model. It considers sentiment and topic as two different language models. A word selected is assumed to be coming from either sentiment or topic model. TSM is basically constructed using the probabilistic latent semantic indexing (pLSI) [7] model with an extra background component and two additional sentiment subtopics.

In order to acquire labels for positive, negative sentiment, which in turn is used to learn model priors, this model uses an existing weblog sentiment retrieval system i.e. Opinmind[10]. They would submit various queries to obtain as much diversified results as possible, and then these results were blend together to obtain training collection for model building. They found out that results obtained with prior knowledge were more clear and distinctive than without prior. In this way after acquiring the sentiment models and topic models, results of sentences were ranked for topics and then categorized by sentiments. Based on these overall opinions for documents/topics were relieved.[4]

2.2 Multi-Aspect Sentiment (MAS) model

This model focuses on modeling topics to match a set of predefined aspects that are explicitly rated by users in reviews. Sentiment is modeled as a probability distribution over different sentiments for each of the aspects, and this distribution is derived from a weighted combination of discovered topics and words. MAS require training data that are rated by users for each aspect, so that discovered topics and sentiment can be mounted on the predefined aspects and their ratings. Sentiment text is aggregated to generate sentiment summary of each rating aspect extracted from MG-LDA.[5]

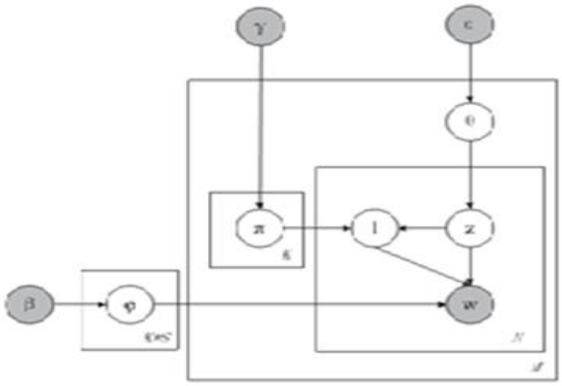

Figure 1a: Sentiment LDA

Multi-Grain Latent Dirichlet Allocation model (MG-LDA) [6] allows terms being generated from either a global topic or a local topic. Global topic is the topic chosen based on the document level context and local topic is chosen based on a sliding window context over the text. They assumed that ratable aspects will be captured by local topics and global topics will capture properties of reviewed items. In the experiments, the number of global topics K^{gl} must exceed the number of local topics K^{loc} by factor of 2 for better results.

The MAS model was designed for sentiment text extraction or aggregation. MAS works in a supervised setting as it requires that every aspect is rated at least in some documents, which is infeasible in real-world applications[5].

2.3 Leveraging Sentiment Analysis for Topic Detection (STD)

They have proposed end to end sentiment analysis framework which combines sentiment classification approaches with sentiment topic detection. The term snippet is a small part of text around a specified keyword in a given document.[9]

There are two key components in this model: The sentiment classification component and the sentiment topic recognition component. The sentiment classification component computes the sentiment polarity of each snippet and creates sentiment taxonomy. Based on the result of this component, the topic detection component further identifies the most significant information related to each sentiment category.

First the given snippet is portioned into positive/negative/neutral categories by first calculating numeric score based on relative sentiment expressed by words in the snippet. Two factors are used to detect sentiment topic words that are word PMI (Pointwise Mutual Information/specific mutual information) and word support. PMI value calculates the uniqueness of word for the given sentiment category. They have adopted the following expression to calculate the PMI value of word w against the category s.[9]

$$PMI(w,s)=\log((p(w,s))/((p(s)*(p(w)+0.05)))), \qquad (1)$$

where p(w, s) is the co-occurrence among w and s, p(s) gives the distribution of category s and p(w) evaluates the distribution of word win the whole snippet collection.

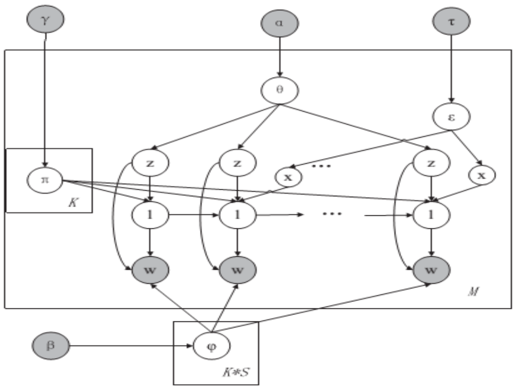

Figure 1b: Dependency Sentiment LDA

This formula was used to calculate the PMI value of the words in each sentiment category. Then the frequency of the words in each category was combined with its PMI value and selects the top frequent words with high PMI value as the final sentiment topic words. It had provided 41-58% precision for different types of user generated data and 55-61% recall. [9]

2.4 Sentiment Analysis with Global Topics and Local Dependency

Sentiment classification is referred as document level sentiment classification. It is unsupervised learning method. They considered that sentiment polarity of each word rely on local context. For that purpose they propose Sentiment-LDA, by appending a sentiment layer to the Latent Dirichlet Allocation (LDA) [2]. To acquire the dependency among the sentiments in the document, they had proposed a novel Dependency-Sentiment LDA model by considering the inter-dependency of sentiments through a Markov chain.

2.4.1 Sentiment LDA

Sentiment-LDA, as shown in Figure 1a, is a four-layer topic model. They had added a sentiment layer between the topic layer and the word layer. So that sentiment layer will be associated with topic layer, and words must associate with both sentiment labels and topics. As sentiment layer is added, sentiments and topics are considered together in this model. Sentiment-LDA can not only classify the overall sentiment polarity for the document, but also can calculate the polarity for each topic.

In order to perform Gibbs sampling with Sentiment LDA, we need to compute the conditional probability, P (where and are vectors of assignments of topics and sentiments for all the words in the collection except for the considered word at position i in document d. Probability P is calculated as follows

$$P(z_i = z, l_i = l | z_{-i}, l_{-i}, w) \propto \frac{\{n_m^{(z)}\}_{-i} + \alpha}{\{n_m\}_{-i} + K\alpha} * \frac{\{n_m^{(z,l)}\}_{-i} + \gamma_l}{\{n_m^{(z)}\}_{-i} + \sum_{l=1}^{S} \gamma_l} * \frac{\{n_{z,l}^{(t)}\}_{-i} + \beta}{\{n_{z,l}\}_{-i} + V\beta} \qquad (2)$$

Where $n_m^{(z)}$ is the number of times words assigned to topic z in document m.n_m is the total number of words in document m. $n_m^{(z,l)}$is the number of times words assigned to topic z and sentiment I in document m. $n_{z,l}^{(t)}$ is the number of times word appeared in topic z and sentiment l. $n_{z,l}$ is the number of times words assigned to topic z and sentiment l. $-i$ denotes a quantity except for the data in position. α, β and γ are prior variables. Sentiment-LDA puts sentiment layer corresponding to topic layer, thus analyzing whole document as well as document sub-topics becomes easy.

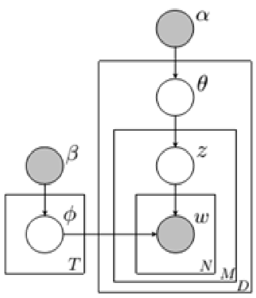

Figure 2a: SLDA

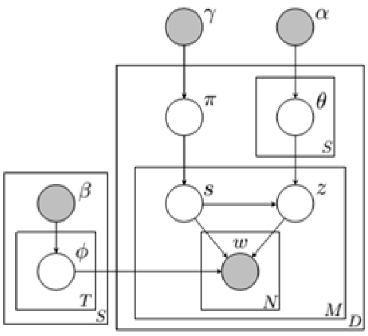

Figure 2b: ASUM

2.4.2 Dependency-Sentiment-LDA

Dependency-Sentiment-LDA let go of sentiment layer independence assumption in Sentiment-LDA. As shown in Figure 1b, Markov chain is formed based on the sentiments of the words in a document, in which the sentiment of a word rely on previous word. They had formulated the Dependency-Sentiment-LDA as a special type of HMM. They are considering the word layer as the observation, and the combination of sentiment layer and transition variable layer, in condition of topic layer, is considered as hidden variables. Sampling formulas can be seen as follows without detailed derivations. When $x_i \neq 0$and $x_{i+1} \neq 0,$ legal component i,

$$P(x_i = x, z_i = z, l_i = l | x_{-i}, z_{-i}, l_{-i}, w) \propto$$

$$\frac{\{n_m^{(z)}\}_{-i} + \alpha}{\{n_m\}_{-i} + K\alpha} * \frac{\{n_{z,l}^{(t)}\}_{-i} + \beta}{\{n_{z,l}\}_{-i} + V\beta} * \frac{\{n_m^{(x_i)}\}_{-i} + \tau_{x_i}}{(\{n_m\}_{-i} + \sum_{x=1}^{X} \tau_x)} *$$

$$\frac{(\{n_m^{(x_{i+1})}\}_{-i} + I(x_i = x_{i+1}) + \tau_{x_{i+1}})}{(\{n_m\}_{-i} + \sum_{x=1}^{X} \tau_x + 1)} \qquad (3)$$

Where $n_m^{(x_i)}$ is the number of transition variables assigned to x in document; $I(\cdot)$ is the indicator function.

Dependency-sentiment LDA not only analyzes the global topic and sentiment in a combined way, but also it adds the local dependency among sentiments. It also first time model the sentiment dependency in the joint sentiment and topic methods.[1]

2.5 Aspect and Sentiment Unification Model

This model employed the observation that one sentence tends to represent one aspect and one sentiment. They have proposed Sentence LDA (SLDA) which constrains that all words in a single sentence be drawn from one aspect. Results obtained from SLDA provide more specific and more coherent topics/aspects as compared to simple LDA[12]. Aspect and Sentiment Unification Model (ASUM) extends SLDA by uniting sentiments. For that purpose, in the initialization step, they assigned the sentiment seed words their seed sentiment. It demonstrates aspect and sentiment together to model sentiments toward various aspects. ASUM combines aspects and sentiment, in this way discovers pairs of {aspect, sentiment}, which they call senti-aspects. ASUM capture important aspects that are closely coupled with a sentiment. Fig 2a and 2b represents SLDA and ASUM respectively.[8]

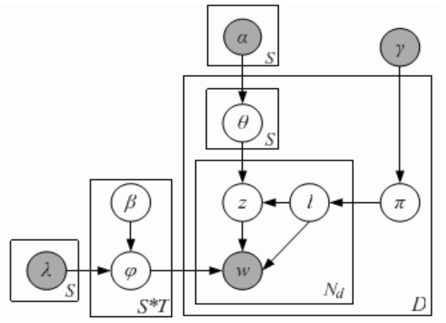

Figure 3: JST

2.6 Weakly Supervised Joint Sentiment-Topic Detection (JST)

JST is weakly supervised with only minimum prior information being assimilated, which is more flexible. It is found that JST is more useful in sentiment classification. JST model can be seen in figure 3. JST model is the extension of existing LDA [2] framework which has three layers, in which topics are coupled with documents, and topics are coupled with topics [11]. Sentiment and topic of sentiment are simultaneously detected from text at document level by Joint Sentiment-Topic (JST) which is weakly supervised. Prior information about the sentiment lexicons is combined into model learning by modifying the Dirichlet priors of the topic-word distributions. Sentiment layer is added in the topic model latent dirichlet allocation (LDA) [2]. It is different from other sentiment-topic model in following aspects: 1) it is weakly supervised. 2) It can detect topics and sentiment simultaneously.[11]

3. COMPARATIVE ANALYSIS

Models have used various types of datasets for result analysis like blogs, discussion forums, review websites data sets, online shopping websites data etc. Result of TSM model shows that prior knowledge of sentiments can provide more coherent and distinctive results. TSM system arranges results according to the hidden aspects of sentences which in turn provide user a deeper understanding of opinions. MAS considers aspects rated by users hence more relevant results were observed (as topics are linked to rated aspects [5]). But this approach comes with disadvantage that it becomes domain dependent. MAS is only supervised method which we have studied.

In STD technique, they acquired most relevant topics within top 10 identified words. Dependency-Sentiment-LDA is more powerful than Sentiment-LDA. Dependency-Sentiment-LDA can not only analyze the topic and sentiment simultaneously, but also consider the local dependency among sentiment labels thus increasing accuracy by margin of 3%~5%. They also achieved good results with union of various lexicon set used for sentiment analysis. Best results are achieved for both Sentiment-LDA and Dependency-Sentiment-LDA with MPQA as lexicons set. ASUM calculates senti-aspect pair at sentence level[8]. It provides aspect-specific sentiment words which can be used in applications such as reviewsummarization. But if ASUM is applied on comparatively small sentence then the results possibly may not be accurate. It can provide different sentiments for same aspect. When applied on same dataset accuracy of ASUM is more than JST [8]. It captures important evaluative details of the reviews and outperforms other models and come close to supervised classification. JST can perform better with subjective reviews and also with combination of unigram and bigrams.

Models	Sentiment analysis	Topic detection	Accuracy
TSM	Sentiment models Positive (θ_P) and negative (θ_N)	pLSI	50-55%
Dependency sentiment LDA	Lexicon based sentiment analysis	Markov chained LDA	69%
ASUM	Sentiment seed words	Sentence LDA	75-85%
JST	Lexicon based sentiment analysis	LDA	70%~75%

4. CONCLUSIONS

Topic and sentiment detection is necessary for many applications like summarizing the search results, surveying public opinion, user behavior prediction and making business decisions. Until now many models are proposed which are studied in this paper. Even though many models have been proposed there is still much scope for research as results of these models are not satisfactory. It is possible to get better performance by using Multi-gram in many of the models. Also with use of NLP like part of speech tagger can give better results. To provide better results, Subjective dataset along with dataset from which commonly used words can be removed.

5. ACKNOWLEDGMENTS

Our thanks to Department of Computer Science and Information Technology of DrBabasahebAmbedkarMarathwada University, Aurangabad for resources and their help.

6. REFERENCES

[1] Fangtao Li, Minlie Huang, Xiaoyan Zhu, J. 2010 Sentiment Analysis with Global Topics and Local Dependency. In Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10),1371-1376

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. J. 2007 Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022.

[3] C. Lin, and Y. He. 2009. Joint Sentiment/Topic Model for Sentiment Analysis, In 18th ACM Conference on Information and Knowledge Management (CIKM).

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

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

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

[7] T. Hofmann. 1999. Probabilistic Latent Semantic Indexing. Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval. pp. 50-57.

[8] Yohan Jo, Alice ho. Feb 9–12 2011. Aspect and Sentiment Unification Model for Online Review Analysis.WSDM'11.

[9] KekeCai, Scott Spangler, Ying Chen, Li Zhang. Leveraging Sentiment Analysis for Topic Detection. International Conference on Web Intelligence and Intelligent Agent Technology(IEEE/WIC/ACM), pp.265-271.

[10] Opinmind.http://www.opinmind.com.

[11] C. Lin, Yulan He, R. Everson. June 2012. Weakly Supervised Joint Sentiment-Topic Detection from Text. IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 6 .pp. 1134-1145.

[12] DP. D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.

[13] Bing Liu. 2010. Sentiment Analysis and Subjectivity. Handbook of natural Language Processing, Second Edition.

[14] L. Zhuang, F. Jing, and X.-Y. Zhu, 2006, Movie review mining and summarization In Proceedings of the 15th ACM international conference on Information and knowledge management (CIKM), pp. 43–50.