DOI: http://dx.doi.org/10.26483/ijarcs.v8i9.4939

ISSN No. 0976-5697

Volume 8, No. 9, November-December 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

ANNOTATING FEATURES EXTRACTED THROUGH LATENT DIRICHLET ALLOCATION FOR FEATURE BASED OPINION MINING

Padmapani P. Tribhuvan
Department of Computer Science and Engineering,
Deogiri Institute of Engineering and Management Studies,
Aurangabad, India.

Sunil G. Bhirud
Department of Computer Engineering and IT,
Veermata Jijabai Technological Institute,
Mumbai, India.

Ratnadeep R.Deshmukh
Department of CS and IT,
Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad, India.

Abstract: Online product reviews contains opinions about products and their features. These product reviews are plain text and therefore analysis of these reviews requires more efforts. In this paper, we tackle the problem of features based opinion mining of product reviews using LDA topic model and proposed annotation algorithm. We proposed an architecture for feature based opinion mining based on topic models and an algorithm that automatically annotates features extracted through LDA topic model. The experimental result shows that the algorithm gives average feature annotation accuracy 77.14%, average positive polarity annotation accuracy 86.02% and average negative polarity annotation accuracy 88.57%. The algorithm can be used with different topics models as well.

Keywords: Feature-Based Opinion Mining, Review Summarization, Topic Annotation, Aspect-Based Sentiment Analysis, Topic Models, Latent Dirichlet Allocation

1. INTRODUCTION

There are many online shopping websites which ask their customers to review products. The numbers of customer reviews that product receives grows rapidly day by day. These reviews are very useful for product manufactures as well as people planning to purchase that product. As the numbers of reviews are very large in number it is difficult to analyze people opinions, sentiments, evaluations, appraisals, attitudes and emotions towards product and product features. So different techniques are used in area of Feature Based Opinion Mining or Aspect Based Sentiment Analysis to analyze and summarize product reviews. These techniques are expected to extract product features about which reviewer has commented on along with the opinion or sentiment expressed, find out expressed opinion is positive or negative and then summarize how many positive and negative opinions are expressed on particular product and product features.

Different approaches are proposed to solve problem of feature based opinion mining. Liu [1] classified these approaches into four categories: 1.Finding frequent nouns and noun phases, 2. Using opinion and target relations, 3. Using supervised learning, 4. Using topic models and mapping implicit aspects. Schouten and Frasincar [2] discussed taxonomy for aspect-level sentiment analysis approaches. They classified different feature based opinion mining tasks into different approaches. For feature based opinion mining they discussed four approaches: 1. Syntax based approach, 2.Supervised machine learning approach, 3. Unsupervised machine learning approach, and 4. Hybrid machine learning approach.

One of the most popular unsupervised machine learning approaches for feature based opinion mining is topic modeling. In machine learning and natural language processing, topic modeling is type of statistical modeling for discovering the abstract topics that occurs in a collection of documents [3]. Topic modeling is unsupervised learning method and it assumes each document is consists of a mixture of topics and each word is probability distribution over words [4].

Mainly there are two basic topic models Probabilistic Latent Semantic Indexing (PLSI) and Latent Dirichlet Allocation (LDA). There are many topic models which are specifically proposed to solve problem of feature based opinion mining, which are either based on PSLI or LDA. These models are very good for feature extraction from reviews. But we need to annotate these features manually and they do not summarize the reviews. So in this paper, we proposed system architecture and the annotation algorithm which automatically annotate of features extracted through LDA topics model, finds the polarity orientation for each extracted feature and summarize the product reviews feature-wise.

Contributions of this paper are -

1. System architecture for feature based opinion mining and summarization
2. The annotation algorithm which automatically annotates features extracted through LDA topics model for feature based opinion mining and summarize the product reviews.

This paper is organized as follows. Section-2 discusses related work, Section-3 discusses proposed system, Section-4 focuses on experiments and results and Section-5 concludes the paper with future directions.

2. RELATED WORK

Titov and McDonald [5] proposed Multi-grain topic models for extracting ratable features from reviews. This model is extended in [6] for summarizing the extracted features. It is a joint model of text and aspect ratings for extracting text to be displayed in sentiment summaries. Our proposed algorithm also generate summary for feature wise rating of particular product.

Lu et al. [7] proposed a topic model based on PLSI to generate a rated aspect/ feature summary. The model decomposes view of overall ratings for major features and gives feature-wise rating. This model is useful for short comments and it needs overall rating of short comments as input. Our proposed system work on complete product reviews which are not rated.

Brody and Elhadad [8] proposed LDA based model for feature based opinion mining and summarization. Model extracts topics as features at sentence level. Polarity identification is done for each feature using seed adjectives with known polarity. This model is flexible with regard to domain and language of review. This model is closely related to our proposed system. Proposed model also uses positive and negative seed words. In [8], conjunction graph is built over adjectives for each feature for polarity identification and we proposed a simple algorithm for this purpose.

Wang et al. [9] proposed model which does not need pre-specified features of products. It uses review ratings. This model mines latent topical features, ratings on each identified feature aspect and weights placed on different features by a reviewer. This model can be applied to various domain data.

In each above work topic model is proposed for feature extraction and then summary is generated. We used Latent Dirichlet Allocation topic model to extract features and then proposed algorithms is used for annotation and summary generation.

3. PROPOSED SYSTEM

Figure 1 shows proposed system architecture. The system performs summarization in four steps: (1) preprocess the dataset, (2) apply topic model, (3) annotate the topics, (4) generate summary. These steps are performed in multiple sub-steps:

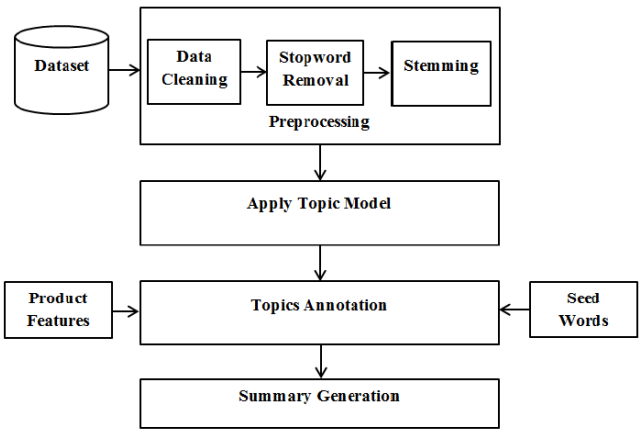

Figure 1. System Architecture for Feature Based Opinion Mining and Summarization.

A. *Preproceesing*

Preprocessing is done in three steps. Data cleaning is done to remove unwanted part of reviews in dataset. For example, we do not need ReviewerID, ReviewerName in reviews so we removed this information from reviews in dataset. After data cleaning, stopwords are removed from reviews. Stopwords are the common English words with high frequency of occurrence which are not useful for further processing. Stemming is done after removing stopwords from reviews. In this step, we reduce derived words to their stem word.

B. *Apply Topic Model [3]*

LDA topic model is a applied on preprocessed review datset to extract product features. Figure 2 shows graphical representation of Latent Dirichlet Allocation topic model.

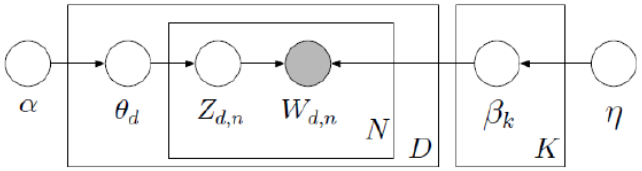

Figure 2. A Graphical Representation of Latent Dirichlet Allocation

C. *Annotation*

Figure 3 shows proposed annotation algorithm which automatically annotates product features extracted through LDA topic model. This algorithm assumes that each topic extracted through LDA topic model represents a product feature. Algorithm also assumes that all product features are nouns. The algorithm accepts four inputs. *Terms X Topics* matrix which is output of LDA topic model. Positive Seed words $PS_i = \{ps_1, ps_2, \ldots, ps_m\}$ .Negative Seed words $NS_i = \{ns_1, ns_2, \ldots, ns_m\}$. PS_i and NS_i are the seed word list containing positive and negative word respectively. Features of product $F_i = \{f_1, f_2, \ldots, f_n\}$. Output of the algorithm is Topics with feature and polarity annotation.

We use Turney's [11] paradigm words as positive and negative seed words. The list of positive and negative seed words is shown in Figure 4 and Figure 5 respectively. We use WordNet [12] to grow this seed list automatically. Synonyms and antonyms of words in seed list are searched using WordNet-3.0.

D. *Generate Summary*

Finally, we generate summary based on the annotation done by the proposed annotation algorithm.

4. EXPERIMENTS AND RESULTS

We used amazon dataset which is available on http://uilab.kaist.ac.kr/research/WSDM11. This dataset contains 24259 reviews of 7 product categories. Table 1 shows details of dataset. We performed all the steps on all 7 product categories separately. For stopword removal we use list of 571 words as Stopword list. Stemming is done using Porter Stemming algorithm. We consider K i.e. number of topics 20.

To infer we use Gibbs Sampler with number of iterations 600. Table 2 shows sample output of LDA topic model and Table 3 shows sample output of annotation

algorithm a Column heading the table is annotation. All these table shows 10 topics and first 10 rows only.

Table 4 and Figure 6 shows result of the algorithm. We calculate the feature annotation and polarity annotation accuracy separately. The generated sample summary is shown in Table 5. In the Table 5 'null' value in feature indicates that the topic is not annotated by the algorithms and therefore it is annotated as 'null'. For 'null' feature also algorithm gives polarity classification.

Algorithm : Annotation Algorithm.

Input:

1. Matrix $Terms \times Topics$ created through LDA topic model.
2. Positive Seed words $PS_i = \{ps_1, ps_2, \ldots, ps_m\}$.
3. Negative Seed words $NS_i = \{ns_1, ns_2, \ldots, ns_m\}$.
4. Features of product $F_i = \{f_1, f_2, \ldots, f_n\}$.

Output: Topics with feature and polarity annotation

1. **Initialization:**
   $rowcount \leftarrow number\_of\_terms\_in\_vocabulary/number\_of\_topics$,
   $positive \leftarrow 0, negative \leftarrow 0$.
2. **for** each row i of $Terms \times Topics$ matrix
3. **for** each column j of $Terms \times Topics$ matrix
4. part-of-speech tag $Term_{i,j}$
5. **end**
6. **end**
7. $i \leftarrow 0$
8. **do**
9. **for** each column j of $Terms \times Topics$ matrix
10. **if** $Topic_j$ annotated **then**
11. **break**
12. **else**
13. **if** $Term_{i,j}$ is noun and $Term_{i,j} \in F$ **then**
14. **for** $k \leftarrow 1$ to $rowcount$
15. **if** $Term_{i,j} \in PS$ **then**
16. $positive \leftarrow positive + 1$
17. **else if** synonym of $Terms_{i,j} \in PS$
18. or antonym of $Terms_{i,j} \in NS$ **then**
19. Add $Terms_{i,j}$ to PS
20. **end if**
21. **if** $Term_{i,j} \in NS$ **then**
22. $negetive \leftarrow negative + 1$
23. **else if** synonym of $Terms_{i,j} \in NS$
24. or antonym of $Terms_{i,j} \in PS$ **then**
25. Add $Terms_{i,j}$ to PS
26. **end if**
27. **end for**
28. **continue**
29. **end if**
30. annotate cluster $Topic_j$ as F_k_positive_negative
31. **end if**
32. **end for**
33. $i \leftarrow i + 1$
34. **while** all $Topics$ are not annotated

Figure 3: Annotation Algorithms

good, nice, excellent, positive, fortunate, correct, superior, amazing, attractive, awesome, best, comfortable, enjoy, fantastic, favorite, fun, glad, great, happy, impressive, love, perfect, recommend, satisfied, thank, worth

Figure 4: List of Positive Seed Words

bad, nasty, poor, negative, unfortunate, wrong, inferior, annoying, complain, disappointed, hate, junk, mess, not good, not like, not recommend, not worth, problem, regret, sorry, terrible, trouble, unacceptable, upset, waste, worst, worthless

Figure 5: List of Negative Seed Words

Table 1: Dataset Details

Amazon Dataset available at : http://uilab.kaist.ac.kr/research/WSDM11		
No. of Reviews : 24259		
Sr. No.	**Product Category**	**#Reviews**
1	Air Conditioners	572
2	Canister Vacuums	3557
3	Coffee Machines	4198
4	Digital SLRs	4198
5	Laptops	4204
6	MP3 Players	3685
7	Space Heaters	3845

5. CONCLUSION

We proposed architecture for feature based opinion mining and an algorithm that automatically annotates features extracted through LDA topic model. LDA topic model is applies on product reviews to extract product features and proposed annotation algorithm is applied on these extracted features. The proposed algorithm is the algorithm gives average feature annotation accuracy 77.14%, average positive polarity annotation accuracy 86.02% and average negative polarity annotation accuracy 88.57% Accuracy of feature annotation is completely depends on the list of product features. For better performance, the list should contain frequent product features. Accuracy of polarity annotation increases with the seed list which grows automatically.

Instead of LDA, if we use topic models specifically proposed for feature extraction from product reviews i.e. models proposed in [5], [7], [8], [9], [10], [13], [14], [15], [16], [17], [18], [19] and [20], then result of proposed algorithm will be increased.

For future work, the proposed model can be applies to the topic model specifically proposed for feature extraction from product reviews.

Table 2: Sample Output of LDA

air	product	water	dai	call	cooler	easi	window	unit	work
condition	return	drain	run	part	humid	review	hose	problem	great
minut	back	empti	summer	servic	live	good	exhaust	instal	make
blow	box	floor	hour	month	fan	quiet	vent	offic	well
size	week	hose	hous	custom	dry	set	hot	heat	machin
long	item	full	bedroom	ship	expect	happi	heat	plug	area
bother	ship	bucket	night	replac	low	move	portabl	amcor	place
perfect	amazon	requir	temp	start	make	pretty	insul	space	out
cold	arriv	tank	sleep	delonghi	evapor	big	fit	issu	hope
reason	order	plug	hot	warranti	put	noisi	back	point	recommend

Table 3: Sample Output of Annotation Algorithm

Size	**Product**	**Capacity**	**Capacity**	**Warantee**	**Energy**	**Null**	**Energy**	**Quality**	**Price**
air	product	water	dai	call	cooler	easi	window	unit	work
condition	return	drain	run	part	humid	review	hose	problem	great
minut	back	empti	summer	servic	live	good	exhaust	instal	make
blow	box	floor	hour	month	fan	quiet	vent	offic	well
size	week	hose	hous	custom	dry	set	hot	heat	machin
long	item	full	bedroom	ship	expect	happi	heat	plug	area
bother	ship	bucket	night	replac	low	move	portabl	amcor	place
perfect	amazon	requir	temp	start	make	pretty	insul	space	out
cold	arriv	Tank	sleep	delonghi	evapor	big	fit	issu	hope
reason	order	Plug	hot	warranti	put	noisi	back	point	recommend

Table 4: Result of Annotation Algorithm

Product Category	**Feature Annotation Accuracy**	**Positive Polarity Annotation Accuracy**	**Negative Polarity Annotation Accuracy**
Air Conditioners	95	85.1488	81.7261
Canister Vacuums	75	84.8055	91.1666
Coffee Machines	70	80.8787	91.119
Digital SLRs	90	84.2519	86.994
Laptops	60	91.7708	93.6111
MP3 Players	70	87.0575	91.0887
Space Heaters	80	88.2205	84.25

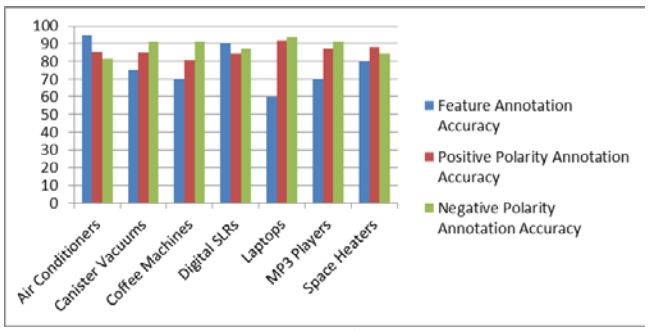

Figure 6: Result of Annotation Algorithm

Table 5 : Summary Generated for Air Conditioner

Feature	**% of Positive Opinions**	**% of Negative Opinions**
Size	77.78	22.22
Product	33.33	66.67
Capacity	66.67	33.33
Warantee	0	100
Energy	75	25
Null	100	0
Quality	50	50
Price	42.86	57.14
Cooling	40	60
Remote	50	50

REFERENCES

[1] Liu, Bing. ”Sentiment analysis and opinion mining.” Synthesis lectures on human language technologies 5.1 (2012): 1-167.

[2] K. Schouten and F. Frasincar, ”Survey on Aspect-Level Sentiment Analysis,” in IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 3, pp. 813-830, March 1 2016.

[3] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. ”Latent dirichlet allocation.” the Journal of machine Learning research 3 (2003): 993-1022.

[4] Abbasi Moghaddam, Samaneh. ”Aspect-based opinion mining in online reviews.” PhD diss., Applied Sciences: School of Computing Science, 2013.

[5] Titov, Ivan, and Ryan McDonald. ”Modeling online reviews with multi-grain topic models.” In Proceedings of the 17th international conference on World Wide Web, pp. 111-120. ACM, 2008.

[6] Titov and R. McDonald, A Joint Model of Text and Aspect Ratings for Sentiment Summarization, in Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT 2008). ACL, 2008, pp. 308316.

[7] Lu, Yue, Cheng Xiang Zhai, and Neel Sundaresan. ”Rated aspect summarization of short comments.” In Proceedings of the 18th international conference on World wide web, pp. 131-140. ACM, 2009.

[8] Brody, Samuel, and Noemie Elhadad. ”An unsupervised aspect-sentiment model for online reviews.” In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 804-812. Association for Computational Linguistics, 2010.

[9] Wang, Hongning, Yue Lu, and ChengXiang Zhai. ”Latent aspect rating analysis without aspect keyword supervision.” In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 618-626. ACM, 2011.

[10] Jo, Yohan, and Alice H. Oh. ”Aspect and sentiment unification model for online review analysis.” In Proceedings of the fourth ACM international conference on Web search and data mining, pp. 815-824. ACM, 2011.

[11] Turney, Peter D., and Michael L. Littman. ”Measuring praise and criticism: Inference of semantic orientation from association.” ACM Transactions on Information Systems (TOIS) 21.4 (2003): 315-346.

[12] Miller, George A. ”WordNet: a lexical database for English.” Communications of the ACM 38.11 (1995): 39-41.

[13] Mei, Qiaozhu, Xu Ling, MatthewWondra, Hang Su, and ChengXiang Zhai. ”Topic sentiment mixture: modeling facets and opinions in weblogs.” In Proceedings of the 16th international conference onWorld Wide Web, pp. 171-180. ACM, 2007.

[14] Lin, Chenghua, Yulan He, Richard Everson, and Stefan Rger. ”Weakly supervised joint sentimenttopic detection from text.” Knowledge and Data Engineering, IEEE Transactions on 24, no. 6 , pp. : 1134-1145 2012.

[15] Kim, Suin, Jianwen Zhang, Zheng Chen, Alice H. Oh, and Shixia Liu. ”A Hierarchical Aspect- Sentiment Model for Online Reviews.” In AAAI. 2013.

[16] Moghaddam, Samaneh, and Martin Ester. ”The flda model for aspect-based opinion mining: addressing the cold start problem.” In Proceedings of the 22nd international conference on World Wide Web, pp. 909-918. International World Wide Web Conferences Steering Committee, 2013.

[17] Xueke, Xu, Cheng Xueqi, Tan Songbo, Liu Yue, and Shen Huawei. ”Aspect-level opinion mining of online customer reviews.” Communications, China 10, no. 3 (2013): 25-41.

[18] Liang, Jiguang, Ping Liu, Jianlong Tan, and Shuo Bai. ”Sentiment Classification Based on AS-LDA Model.” Procedia Computer Science 31 (2014): 511-516.

[19] Wu, Yao, and Martin Ester. ”FLAME: A Probabilistic Model Combining Aspect Based Opinion Mining and Collaborative Filtering.” In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 199-208. ACM, 2015.

[20] Tan, Shulong, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Jiajun Bu, Chun Chen, and Xiaofei He. ”Interpreting the public sentiment variations on twitter.” IEEE Transactions on Knowledge and Data Engineering 26, no. 5 (2014): 1158-1170.