IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-10 E-ISSN: 2347-2693

Product Features Extraction for Feature Based Opinion Mining using Latent Dirichlet Allocation

Padmapani P. Tribhuvan[1*], Sunil G. Bhirud[2], Ratnadeep R.Deshmukh[3]

[1*]Dept. of Computer Science and Engineering, Deogiri Institute of Engineering and Management Studies, Aurangabad, India
[2] Dept. of Computer Engineering and IT, Veermata Jijabai Technological Institute, Mumbai, India
[3] Dept. of CS and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India

[*]*Corresponding Author: padmapanitribhuvan@dietms.org*

Available online at: www.ijcseonline.org

Received: 19/Sep/2017, Revised: 02/Oct/2017, Accepted: 20/Oct/2017, Published: 30/Oct/2017

***Abstract*—** Unstructured product reviews are difficult to analyse. By applying feature-based opinion mining on product reviews, we can analyse product reviews. In Feature Based Opinion Mining, method of extracting features plays very important role. Performance of feature based opinion mining is depends on how features are extracted from product reviews. In this paper, we discussed how Latent Dirichlet Allocation topic model can be used for product features extraction. We discussed a methodology to extract product features using Latent Dirichlet Allocation topic model. We applied basic Latent Dirichlet Allocation (LDA) topic model on 24259 product reviews of 7 product categories to extract product features. We inferred the model using Gibbs Sampler. The result shows that LDA model extracts product reviews efficiently.

***Keywords*—**Feature-Based Opinion Mining, Aspect-Based Sentiment Analysis, Topic Models, Latent Dirichlet Allocation

I. INTRODUCTION

Analysing unstructured product reviews is very critical task. Feature-Based Opinion Mining is an opinion mining task which analyses reviews at feature level and these reviews are unstructured in nature. Feature based opinion mining performs three different tasks: feature extraction, opinion polarity identification for each extracted feature and generating summary of features and opinions. As feature extraction is the task on which other two tasks are depends, it plays very important role in performance of feature based opinion mining.

There are different feature extraction approaches. In Frequency Based Approach, features are identified by applying a set of constraints on high frequency noun phrases. This approach is very simple and quite effective. The disadvantage of this type of approach is that low frequency features are missed. Feature-based summarization [1] and OPINE [2] are examples of this is type of approach.

Low frequency features can be found out by using another approach known as Syntax Based Approach. It uses feature-opinion relationships and find outs relation patterns. Opinion Observer [3], Multi-Facet Rating [4], Tree Kernel Approach [5] are examples of this type of approach.

In supervised learning approach, feature extraction considered as a special case of the general information extraction problem. Manual labelling of training data set is required to train a model.

Topic modelling approach assumes each review document consists of a mixture of topics i.e. features and each topic i.e. feature is a probability distribution over words. In this paper we discussed how topic models can be used for product feature extraction. In this paper, We discussed a methodology to extract product features using Latent Dirichlet Allocation topic model.

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

II. RELATED WORK

There are many topic models which are specifically implemented to accomplish Feature-Based Opinion Mining Task.

MG-LDA Model, proposed by Titov et al. [6], is model which based on LDA. It extracts features from reviews. This model considers two types of topics, global topic which corresponds to global features or attributes of the product in the review and another is local topics which corresponds to product features.

Brody et al. [7] proposed LDA based model which extracts local topic as product feature. It is applied at sentence level.

Jo and Oh [8] proposed ASUM discovers Senti-Aspect (Senti-Feature) pair. This model is also based on LDA.

Tan et al. [9] proposed FB-LDA (Foreground and Background LDA) model based on LDA. This model can filter out background topics and then extract foreground topics to reveal possible reasons.

© 2017, IJCSE All Rights Reserved

In this paper we discussed how basic LDA model can be used for product feature extraction for Feature Based Opinion Mining.

III. METHODOLOGY

The steps to exatract features are shown in Figure1. These steps are discussed below.

1. Preprocessing : The collected product reviews are preprocessed by three steps: Data cleaninig, Stop word removal and stemming.

2. Document Term Matrix Generation : DTM is mathematical matrix which gives the frequency of terms that occur in a collection of documents. Figure 2 shows Document Term Matrix. In this matrix each row represents Document, each term represents Term and each cell stores frequency of the Term in a Document. If we have M number of documents and V number of terms in vocabulary then DTM will be MXV matrix. This is one of the popular ways to convert unstructured data into structured data. Figure3 shows example of DTM generation.

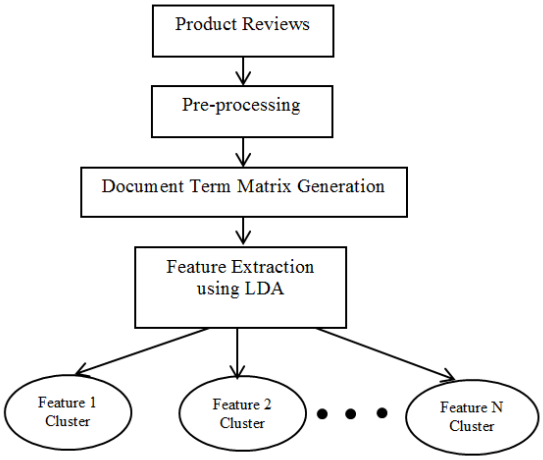


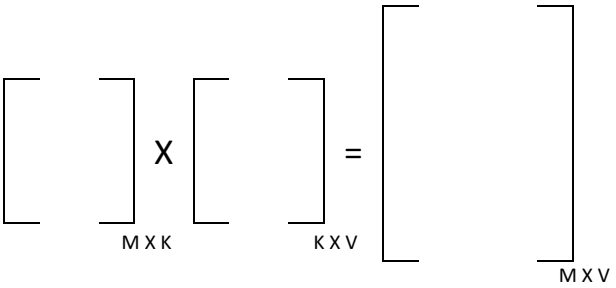
Figure 1. Feature Extraction using LDA

	$Term_1$	$Term_2$	.	.	.	$Terms_V$
Doc_1						
Doc_2						
.						
.						
.						
Doc_M						

Figure 2. Document Term Matrix

3. Feature Extraction using LDA [10]: Latent Dirichlet Allocation (LDA) is a topic model. It is generative probabilistic model. LDA generates K number of clusters each representing a topic. LDA uses a matrix factorization.

If DTM is M X V matrix then LDA considers



where M X K is per document topic distribution matrix, K X V is per topic word distribution matrix and K is number of topics.

Doc1. : Mobile is good.
Doc2. : Battery life is bad but color is good.
Doc3. : Color is bad.
Doc4. : Battery life is good.
Doc5. : Mobile color is good.

Pre-processing

Doc1. : mobil good.
Doc2. : batteri life bad color good.
Doc3. : color bad.
Doc4. : battery life good.
Doc5. : mobile color good.

DTM Generation

	mobil	**Good**	**batteri**	**life**	**bad**	**Color**
Doc1	**1**	**1**	**0**	**0**	**0**	**0**
Doc2	**0**	**1**	**1**	**1**	**1**	**1**
Doc3	**0**	**0**	**0**	**0**	**1**	**1**
Doc4	**0**	**1**	**1**	**1**	**0**	**0**
Doc5	**1**	**1**	**0**	**0**	**0**	**1**

Figure 3. Example of DTM Generation.

We assume a review document is generated as follows: a reviewer first decides a product on which review is to written. The reviewer decides the set of product features about which he/she wants to write and the decides which feature is liked and which one disliked. We assume that if we

© 2017, IJCSE All Rights Reserved

apply LDA on review documents of single product , it will gives us the clutsers representing product features.

We can apply different opinion polarity identification techniques on these features for feature based opinion mininig of the product review documents.

IV. 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. For stop word removal we use list of 571 words as Stop word list. Stemming is done using Porter Stemming algorithm. We consider K i.e. number of topics is 10.We applied LDA topic model on different product reviews. To infer we use Gibbs Sampler with number of iterations 500. We performed same experiment for all 7 product categories.

Table 1 shows dataset details and DTM details. It shows number of review documents, number of terms in all reviews after pre-processing. All DTM are sparse matrices. So Table 1 also details about sparsity. Table 1 also gives maximal term length for each product category product reviews. Table 2 shows sample output of LDA. In table2, only 10 rows are shown. Each column of Table 2 shows terms belonging to a topic i.e. product feature arranged probability wise.

V. CONCLUSION AND FUTURE DIRECTIONS

In this paper we discussed how basic LDA model can be used to extract product features. As basic LDA model is used for product feature extraction, the feature extracted are generic features and are not product specific. So we can implement a model based on basic LDA to extract product features from product reviews like the models in [6], [7], [8] and [9]. A method can be implemented to name is feature cluster extracted by basic LDA model.

REFERENCES

[1] Hu, Minqing, and Bing Liu. "*Mining and summarizing customer reviews."* In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168-177. ACM, 2004.

[2] Popescu, Ana-Maria, Bao Nguyen, and Oren Etzioni. "*OPINE: Extracting product features and opinions from reviews*." In Proceedings of HLT/EMNLP on interactive demonstrations, pp. 32-33. Association for Computational Linguistics, 2005.

[3] Liu, Bing, Minqing Hu, and Junsheng Cheng. "*Opinion observer: analyzing and comparing opinions on the web*." In Proceedings of the 14th international conference on World Wide Web, pp. 342-351. ACM, 2005.

[4] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. "*Multi-facet Rating of Product Reviews."* In ECIR, vol. 9, pp. 461-472. 2009.

[5] Jiang, Peng, Chunxia Zhang, Hongping Fu, Zhendong Niu, and Qing Yang. "*An approach based on tree kernels for opinion mining of online product reviews*." In Data Mining (ICDM), 2010 IEEE 10th International Conference on, pp. 256-265. IEEE, 2010

[6] 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.

[7] 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.

[8] 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.

[9] 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. 2014.

[10] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "*Latent dirichlet allocation*." Journal of machine Learning research 3, no. Jan (2003): 993-1022.2003

Authors Profile

Ms. Padmapani P.Tribhuvan is B.E.(IT), M.E. (CSE) and currently pursuing Ph.D. She is working as Assistant Professor in Department of Computer Science and Engineering, Deogiri Institute of Engineering and Management Studies, Aurangabad. Her main research intersts in Machine Learning, Sentiment Analysis.

Dr. Sunil G. Bhirud is BE (Electronics), ME (Electronics), Ph.D. He is working as Professor, Veermata Jijabai Technological Institute,Mumbai, India. His area of research is Computing in Mathematics, Natural Science, Engineering and Medicine, Artificial Intelligence, Computer Graphics.

Dr. Ratnadeep R. Deshmukh is M.Sc.,M.E.(CSE) Ph.D. FIETE. He is working as Professor, DST-FIST Program Coordinator, Department of CSIT, Dr. B.A.M. University, Aurangabad, (MS) India. He is Fellow & Chairman, IETE Aurangabad Center, Life Member ISCA, CSI, ISTE, IEEE, IAEng, CSTA, IDES & SMACEEE. Member of Management Council of University. Visited University of Santiago Compostela,Spain in PEIN Research Excellence Program, Honkong, Shainghai-Chaina,Bankok-Thailand, University of Washington Seattle, WA, USA.

© 2017, IJCSE All Rights Reserved

Table 1: Dataset Details and DTM Details

Sr. No.	Product Category	# Documents	# Terms	Non Sparse Entries	Sparse Entries	Sparsity	Maximal Term Length
1	Air Conditioners	572	9992	55493	5659931	99%	56
2	Canister Vacuums	3557	35319	316111	125313572	100%	74
3	Coffee Machines	4198	39164	360834	164049638	100%	132
4	Digital SLRs	4198	44953	482979	188229715	100%	96
5	Laptops	4202	49201	507687	206333317	100%	143
6	MP3 Players	3685	40255	346675	147993000	100%	128
7	Space Heaters	3845	34862	296865	133747525	100%	282

Table 2: LDA output for Laptops Product Reviews

	Topic.1	Topic.2	Topic.3	Topic.4	Topic.5	Topic.6	Topic.7	Topic.8	Topic.9	Topic.10
1	keyboard	comput	window	macbook	Whir	ram	Machin	son	cycl	Laptop
2	laptop	dell	mac	pro	Delet	graphic	Drive	649	begun	great
3	kei	year	run	appl	Loyal	model	Work	acer	instruct	price
4	button	bui	softwar	screen	Amp	price	Screen	charm	learn	fast
5	screen	toshiba	work	displai	Commend	processor	Bought	dislik	reflect	purchas
6	pad	bought	program	design	Differ	card	Inch	outstand	anoi	love
7	feel	problem	system	trackpad	Iso	drive	Thing	present	apach	recommend
8	touch	month	user	glossi	Remain	memori	Hard	revers	audiovideo	good
9	mous	work	easi	firewir	Respect	notebook	Back	sophist	batch	feature
10	type	replac	Time	aluminum	Stabil	core	Fast	brows	beach	bui

© 2017, IJCSE All Rights Reserved