International Journal of Applied Science and Engineering

Development of a two-step LDA based aspect extraction technique for review summarization

Subha Jyoti Das[1], Riki Murakami[1], Basabi Chakraborty[2*]

[1] *Graduate School of Software and Information Science, Iwate Prefectural University, Iwate, Japan*
[2] *Faculty of Software and Information Science, Iwate Prefectural University, Iwate, Japan*

ABSTRACT

Summarization of online reviews by customers is a popular practice for evaluation of products or services. As the reviews accumulate, the large size and the unstructured nature of the reviews hinder manual summarization. Automatic categorization of the reviews as a whole into only positive and negative group cannot represent a clear picture. An aspect based automatic summarization technique can provide better visualization. However, automatic extraction of proper aspects from the huge reviews of any product is not very easy. There are some research works in this direction, but any definite method is yet to come. In this work, a two-step Latent Dirichlet Allocation (LDA) technique, which is popularly used for topic modelling has been developed for efficient aspect extraction. The method has been evaluated by simulation experiments on Amazon product reviews and Yelp restaurant and hotel reviews. The results have been found quite matching with human annotated results.

Keywords: Opinion analysis, Aspect extraction, Review summarization, Two-step LDA.

OPEN ACCESS

Received: June 2, 2020
Revised: August 27, 2020
Accepted: December 16, 2020

Corresponding Author:
Basabi Chakraborty
basabi@iwate-pu.ac.jp

Publisher:
Chaoyang University of Technology
ISSN: 1727-2394 (Print)
ISSN: 1727-7841 (Online)

1. INTRODUCTION

E-commerce has seen an exponential growth with the internet and other technological advancement. People have found a convenient place for expressing their voice in online platforms. Opinion summarization have garnered a lot of attention due to the potential research opportunities (Rao and Shah, 2018). The online reviews contain a huge amount of information which is necessary for the business owners for getting user feedback to improve their services and also important for the future potential buyers for making an informed decision. The main hurdle for achieving the goal is the enormous amount of reviews available for each product and summarizing the reviews by human element is a tiresome work. As a result, developing techniques for category independent summarization of the unstructured data has become an important research topic (Wawer, 2015).

Opinion mining has three implementation levels, these are document level, sentence level and aspect level (Hajmohammadi et al., 2012). Document or sentence based summarizations usually present an opinion overview of the whole document or sentence, whereas the aspect based opinion summarization gives a detailed view based on the aspects of the product or service in question. Usually the researches on opinion analysis produce the result in a bipolar manner, positive or negative. This analysis can produce some overall idea about user sentiment but a detailed analysis is possible only through aspect based summarization (Bagheri et al., 2013).

In an aspect based summarization method, the opinion summarization part of the work can be done in few ways, such as bipolar system or numerical scale based system. One of the most important part of the process is to detect aspects. In this work, effort has been put on aspect extraction in an efficient manner. Over the years, various approaches have been evolved such as, rule based, supervised and unsupervised. Previously a rule based Word2Vec model (Das and Chakraborty, 2020) based method was developed by author. The proposed system was based on frequency of the words in a corpus, as a result many low frequency word but contextually important word and implicit aspects have been overlooked while summarizing the reviews. Also the rules for the processing had to be decided manually beforehand. To reduce human interference, traditional Latent Dirichlet Allocation (LDA) method, a topic modelling technique has been used for aspect extraction in Das and Chakraborty (2019). LDA (Blei et al., 2003) is an unsupervised process, proposed by David and his co-researchers in 2003. Though LDA exploits latent relationship to find topics, often contextually unrelated words are put together, whereas other contextually important words for the topic go unnoticed, which makes extracting topics harder. As a remedy to this problem, a two-step LDA based approach for aspect extraction has been proposed in this work. This method helps to find out the latent, implicit aspects as well as the explicit aspects. The proposed algorithm has been used to extract aspects from various reviews of benchmark data sets and the results have been evaluated by checking with manual annotation of the reviews.

The rest of the paper is organized as follows: section 2 describes a few related works in the area of aspect extraction followed by our proposed method in section 3. In section 4, datasets used in this experiment and proposed method have been described. Section 5 represents results and discussion while section 6, the last section, contains conclusion.

2. ASPECT EXTRACTION METHODS

Aspect extraction is important for summarizing the reviews in a detailed manner and easy comprehension. Reviews normally are written in an unstructured manner and aspects can be mentioned in any form. Detecting the aspects thus becomes a challenge. Usually aspects can be grouped into two classes, explicit and implicit. Explicit aspects are easier to detect, but the problem lies with detecting the implicit aspects, especially when human intervention is not used.

Aspect extraction has been one of the most challenging area of research in the field of opinion mining or sentiment analysis. Here is a very brief review of the important existing methods of aspect extraction related to our approach.

2.1 Unsupervised Approaches:

There are several unsupervised approaches. In Popescu and Etzioni (2005), the researchers have tried to extract aspects by using OPINE and web PMI, this method is dependent on web services for measuring the Pointwise Mutual Information (PMI). In Hu and Liu (2004), researchers also tried to extract aspects by an unsupervised method based on association mining. Here the researchers POS tagged data to find the noun words and extract features based on the co-occurrence of those terms. In Yi et al. (2003), researchers also used an unsupervised approach by developing methods based on mixture language model and likelihood ratio model.

2.2 Supervised Approaches:

Few researchers have taken supervised approaches. In Jakob and Gurevych (2010), researchers developed a conditional random field (CRF) based method to find out the aspects, in this method they provide some information, POS (Parts of Speech) tags, short dependency path, distance between words and opinion sentence. In Kessler and Nicolov (2009), researchers trained a Support Vector Machine (SVM) classifier to find out related opinion words and target aspects. In Jin et al. (2009), researchers developed a Hidden Markov Model based approach, which employs several techniques such as POS tags, internal information of pattern of phrases and other contextual information. In Fang and Huang (2012), researchers proposed a method which implements SVM and latent discriminate method to find the aspects and cluster them. It was implemented on Chinese restaurant reviews.

2.3 LDA Based Approaches:

There are several topic modelling based approaches for aspect extraction. LDA is very helpful in discovering topic with the help of semantic information in unstructured text data. Topic modelling assists in finding the hidden pattern in enormous data. Several researchers have tried to find optimum methods based on topic modelling to identify clusters of similar aspects.

1. Employing online tools:

   In Ekinci and Omurca (2017), researchers proposed a method to extract implicit aspects, based on topic modelling, with the help of an application named 'Babelfy', which is a multilingual, graph based semantic network.

2. Combined with other models:

   In Debortoli et al. (2016), researchers discussed the challenges usually faced by everyone, and proposed a topic modelling, infused with LASSO (Least Absolute Shrinkage and Selection Operator) multinomial logistic regression and implemented on online review data holding more than 12000 reviews. In Zhao et al. (2010), the researchers proposed a topic modelling based hybrid model, MaxEnt-LDA. This is a semi-supervised model, which uses maximum entropy with topic modelling for identifying aspects and opinions together. This model, along with adjective words also take into account the non-adjective words for opinion analysis.

   In Jo and Oh (2011), researchers proposed a method of employing two models, and accumulating the end results simultaneously. The first model is Sentence-LDA which detects aspects in sentence level. The second model is an Aspect and Sentiment Unification Model (ASUM) which identifies the aspects and their sentiment words together.

3. Knowledge based approaches:

In Moghaddam and Ester (2011), researchers proposed three probabilistic graphical models to generate aspect summary. First one is an extension of PLSI (Probabilistic Latent Semantic Indexing) model, second one is an extension of LDA model and the third one is ILDA (Interdependent LDA). To make the system more efficient the dependency on preexisting knowledge should be minimized. In Allahyari et al. (2017), the researchers proposed a model named KB-LDA (Knowledge Based LDA). In this method they integrated an ontology based knowledge with the LDA to label the topics in a more better and meaningful way.

In Wang et al. (2014), researchers proposed two models, one is semi-supervised method Fine Grained Labeled LDA (FL-LDA), where preconceived knowledge from the e-commerce website can be used as seed to train the model for extracting aspects and cluster them properly. Another model is Unified Fine Grained Labeled LDA (UFL-LDA) where the aspects overlooked in the previous model, can be extracted. In Chen et al. (2013), researchers developed a method called MDK-LDA which employs Multi Domain knowledge for the same word, which can mean different things in different domain, even sometimes in the same domain. The same researchers proposed another method called GK-LDA (Chen et al., 2013), where the wrong knowledge learned by MDK-LDA can be handled with the general knowledge learned by the model. To remove the problems of both MDK-LDA and GK-LDA, researchers proposed another method called MC-LDA (LDA with *Must-link* and *Cannot-link* constraints) in Chen et al. (2013).

4. Other methods:

In Xueke et al. (2013), researchers proposed a Joint Aspect Sentiment (JAS) model. This model tries to identify implicit aspects by the explicit aspects extracted by the LDA model. In Xu et al. (2012), the researchers also followed the same suit. In Brody and Elhadad (2010), researchers proposed an unsupervised technique based on local LDA. This method relied on keeping the number of topics small and operating on sentence level. In Bagheri et al. (2014), researchers developed an unsupervised method to extract aspects called ADM-LDA, where they considered every word in a sentence as a state of Markov chain, and the subsequent words in the chain are more probable to be in the same topic.

In Teh et al. (2006) researchers proposed a method called Hierarchical Dirichlet Process. This method is considered as an extension of the LDA method. This method is a non-parametric Bayesian method, which clusters data involving multiple groups. In Srivastava and Sutton (2017) researchers propose a method named Prod LDA based on Autoencoded Variational Inference for Topic Model (AVITM). AVITM is an inference method based on Auto-Encoding Variational Bayes (AEVB).

The methods mentioned above usually employ preconceived human knowledge or domain knowledge to extract aspects. In our proposed approach based on two-step LDA, no preconceived knowledge is needed and both the explicit and implicit aspects can be extracted efficiently. It is well known that LDA cannot perform well with small corpora for coherent topics extraction. In the proposed two-step method, smaller corpora can also be processed efficiently and coherent clusters can be formed. In the next section, the proposed method is presented.

3. PROPOSED METHOD

In the proposed method, LDA is used to detect aspect related words having latent relationship with one another. LDA identifies the latent relationships between words and collect the related words into a single topic. The word distribution of any topic is associated with a probability value and the words are ranked in order of their probability values. Often contextually unimportant word becomes higher ranked whereas contextually important word falls in lower position. As the number of words in a topic is very large, a part of the high ranked words is considered to represent a topic. Thus in the selection process, many contextually important words can be missed. In the second step, a guided LDA proposed in Singh (2017) in supervised mode, is used to alleviate this problem by extracting seed words from the first step to guide the second phase of LDA. The final result contains a lot of common words which are then clustered to get the proper aspect words.

The proposed method for aspect extraction can be divided into five steps as explained below. The whole process is shown in Fig. 1 and the algorithm is presented in Algorithm 1.

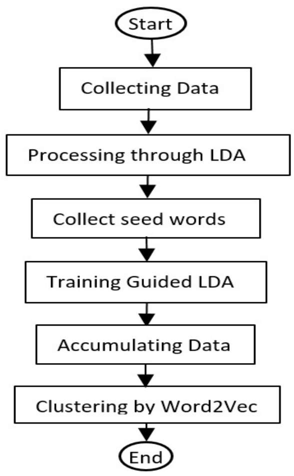

Fig. 1. Two-step LDA for aspect extraction

Das et al., International Journal of Applied Science and Engineering, 18(1), 2020120

Algorithm 1 Two step LDA method algorithm

```
Input: Raw text file of reviews
Output: Aspects initialization
    First step:
for Every review in corpora do
    Convert all the words in lower case
end for

function REMOVE_STOPWORD(Texts)
    Remove stop words
    return Remaining words
end function

function MAKE_BIGRAMS(Texts)
    Find probable bigrams
    return bigrams and rest of the words
end function

function LEMMATIZATION(bigrams, other words, POS tags to be allowed)
    lemmatize the target words
    return lemmatized form of allowed words
end function

call remove_stopwords
call make_bigrams
call lemmatize

create dictionary(unique id assigned to words)
create corpus(term document frequency)

for topic numbers 10 to 110 at interval of 10 do
    Train LDA and find coherence score for every topic number
end for
Acquire the results with highest coherence number
    Second step:
for every topic up to N (predefined number) do
    Choose the first 2 words as seeds.
end for
create voabulary with lemmatized words
assign unique ids to word
create term document frequency matrix
train Guided LDA for N topics with collected seeds
Collect the resultant topics in a single list to remove duplicate words
Cluster the words with pre-trained Word2Vec model, and label the clusters
```

3.1 Collecting the Desired Product Reviews

The proposed algorithm is developed in a manner so that it can work for any kind of reviews irrespective of the target product or service. The reviews range from electronics, office products to service based businesses like restaurants and hotels. All the different review sets have to be collected and stored in different text files.

3.2 Extracting the Topics from Traditional LDA

The text files containing the reviews are to be processed through following steps:

1. ***Preprocessing:*** First the corpus is to be processed through the preprocessing step in which the punctuation, special characters and stop words are to be removed.
2. ***Creating bigrams:*** Bigrams are to be formed of those words which occur frequently together.
3. ***Lemmatizing the words:*** In this step the words are to be lemmatized. Only **noun, adjective, verb** and **adverb** are chosen.
4. ***Creating dictionary and corpus:*** To construct the topic model two important inputs are dictionary and corpus. At first the dictionary has to be created with the corpora, from the lemmatized data, then with help of the dictionary the corpus has to be created. Dictionary contains all the lemmatized data, with a unique Id assigned to them. Corpus is a mapping of the Id to the word frequency in the documents.
5. ***Creating the topic model with optimum topic number:*** To create the topic model, the LDA module of Gensim is used here with proper setting of parameters like number of topics etc., the parameters 'alpha' and 'beta' are set as default. Fig. 2. represents the process of aspect extraction by traditional LDA.

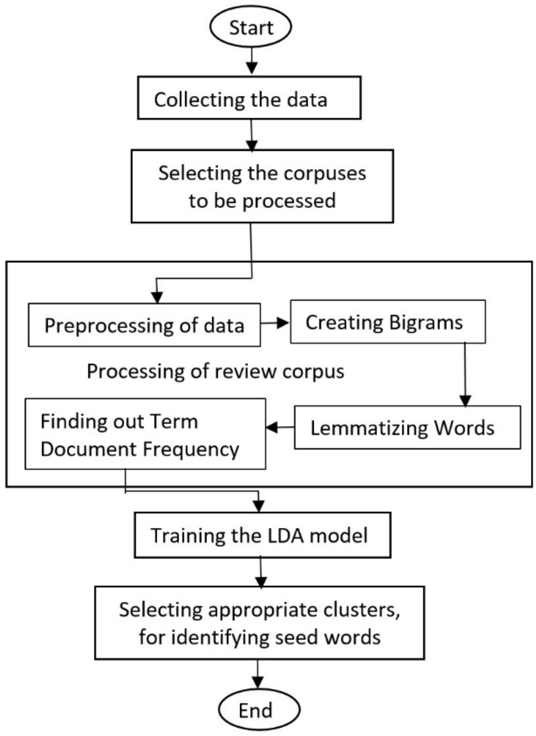

Fig. 2. Traditional LDA for aspect extraction

3.3 Aspect Extraction by Guided LDA

In traditional LDA, the extracted topics do not always reflect desired result. Often context wise unimportant words are included in topics and the important words are missed out as the words in the topic are selected according to their probability values. To alleviate this problem, a guided LDA has been used in the second step. In this step the higher probability words from the previous step are used as seed words to guide the training of LDA.

The process is as follows:

1. ***Collecting the seed words***: The words in the topics after first LDA step are inspected and high probability

words are to be chosen as seed words. Those words are to be set as training words for the second step guided LDA.

2. ***Creating vocabulary***: A vocabulary is to be created with all the lemmatized words from the previous step.
3. ***Creating the dictionary***: A dictionary is to be created with the lemmatized words with an assigned unique id.
4. ***Creating term-document matrix***: A term document matrix is to be created with the distribution of words for all the documents.
5. ***Training the LDA***: With all the necessary information available at hand, the guided LDA is to be trained and used to find the final topics.

3.4 Making a List of Aspect Words

The topics generally contain a lot of common words with higher probability, which are high frequency words, but along with those words, the important but lower probability words also appear in those topics. A unique list of aspect words has to be prepared by taking the words from all the topics. After this step all the duplicate words are to be removed.

3.5 Clustering the Aspect Words

The words in the list are to be clustered with the help of Word2Vec word embedding module to get the final aspects as the labelled clusters. The process is shown in Fig. 3.

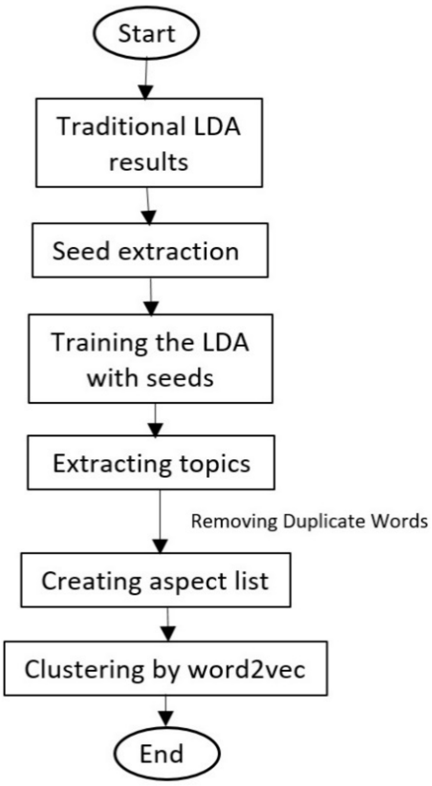

Fig. 3. Process of guided LDA and final clustering

4. DATASETS AND EXPERIMENT

4.1 Data Set Used

The review corpora used for current research purpose were collected from Amazon product data, put together by Julian McAuley, UCSD (University of California, Santa Davis) (He and McAuley, 2016) and the YELP hotel and restaurant reviews put together by Rayana and Akoglu. Reviews (unlabelled) of 6 product review corpora and 12 hotel and restaurant corpora are chosen for implementing our method. All the reviews are in English language.

1. ***Amazon corpus:*** The product corpora are Camera lens protector (2547 reviews), Headphone (2074 reviews), Paper shredder (2531 reviews), Television mount (1050 reviews), Phone (4397 reviews), Printer (3017 reviews).
2. ***Yelp corpus:*** For hotel and restaurant, there are 12 corpora, among them 6 corpora are big and 6 are small corpora.
   - 2.1 ***Big corpora*** are Maialino (892 reviews), ABC kitchen (1780 reviews), Casa mono (896 reviews), Pylos (847 reviews), Cook shop (1107 reviews), Sakagura (1165 reviews).
   - 2.2 ***Small corpora*** are Greek restaurant (210 reviews), Peppino's pizza (253 reviews), Dekalb restaurant (59 reviews), Blue spoon coffee (86 reviews), Hunter's (178 reviews), Alameda (52 reviews).

4.2 Simulation Experiment

Tools used for the experiment are Google colab, python 3.6, gensim library, guidedlda library, nltk library and a pre trained Word2Vec model.

The reviews of only one product or hotel/restaurant stored in a text file are processed at a time in the proposed method. Every review in the corpus is considered a document in this method.

1. At first, the corpus is pre-processed for removing stopwords and bigrams are formed to find the words which occur frequently and a function was created to create bigrams and the corpus was processed through it to find the probable bigrams.
2. Words are then lemmatized and POS tagged to prioritize noun, adjective, verb and adverb words as aspect words.
3. A dictionary is created according to the description in section 3.2. The Corpora module was used to implement this part. This module has some helpful features, such as, i) creating a corpus, ii) appending documents to a corpus, iii) easily accessing a document by the unique document id, iv) accessing the documents sequentially.
4. As topic numbers are supposed to be decided before the training for the model starts, it is necessary to find the optimum topic number for training the model. There are several methods to decide the optimum

number of topics, among them coherence value method was chosen for being closest to the human judgment. Coherence value measure the quality of the topics by measuring the semantic similarity of the words constituting the topics. There are several standards for measuring coherence value, as explained in Kumar (2018). Here in this work ***c_v*** method have been chosen. For every review corpus, coherence score has been calculated for a range of topic numbers (10-110) and the topic number with highest coherence value has been chosen. After training the model with optimum topic number, the topics are extracted.

The change of coherence value with the increase of number of topics for amazon product reviews, yelp hotel and restaurant reviews (larger and smaller corpora) have been shown in Fig. 4-6 respectively. The different colors of the graph signify different product corpora. In the box at lower right corner, the number of reviews in a corpus and corresponding optimum topic number have been mentioned in the format (number of reviews – number of topics)

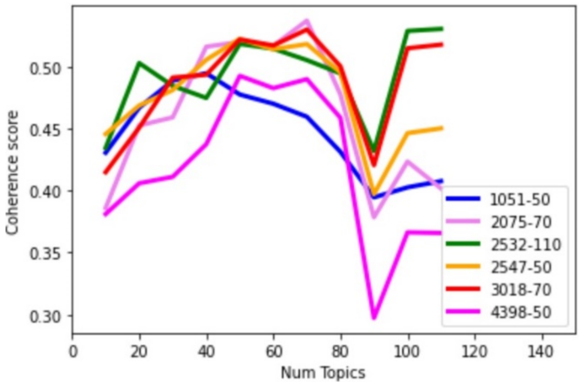

Fig. 4. Coherence values for different number of topics for amazon reviews

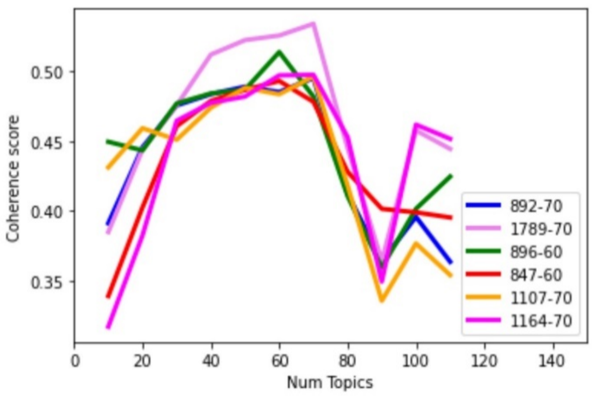

Fig. 5. Coherence values for different number of topics for Yelp reviews (large corpora)

In Fig. 6, the graphs representing the trends of coherence number for different product corpora have been shown. At the lower right corner, the numbers of reviews have been shown with the color of the graph representing them.

From the Fig. 4 and 5, the optimum topic numbers for the reviews are found as follows:

Product reviews:
Camera lens protector- 50
Headphone- 70
Paper shredder- 110
Television mount- 50
Phone- 50
Printer- 70

Yelp reviews (Big):
Maialino- 70
ABC kitchen- 70
Casa mono- 60
Pylos- 60
Cook shop- 70
Sakagura- 70

Though traditional LDA works well with sizeable corpora, it cannot produce good result when the corpus size is small as is reflected in Fig. 6. It is found that the coherence value keeps growing with increase of topic numbers. From this result we can understand that LDA does not perform well with smaller corpora. As coherence number cannot give conclusive optimum topic number, the optimum topic number was assumed to be 30 for each smaller corpus.

Yelp reviews (small corpora):
Greek restaurant- 30
Peppino's pizza- 30
Dekalb restaurant- 30
Blue spoon coffee- 30
Hunter's- 30
Alamada - 30

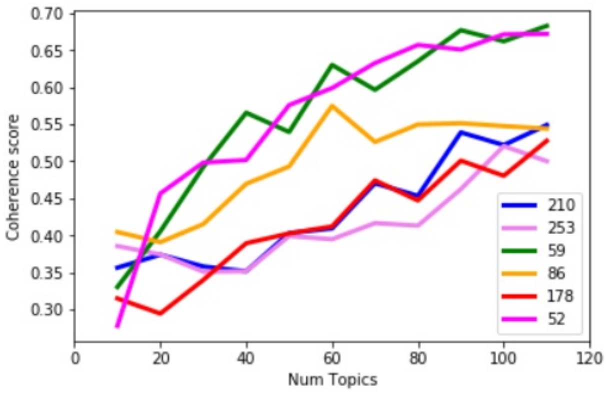

Fig. 6. Coherence values for different number of topics for Yelp reviews (smaller corpora)

5. In the second step, first 10 topics are chosen and from each topic first 2 words (or bigram) are chosen to train the guided LDA. To train the guided LDA, a dictionary has been created and a term-document matrix is created with the list of vocabulary and documents

(reviews). From the topics found from guided LDA, the aspect list is formed and clustered to have the final list of extracted aspects.

To show the efficiency of the proposed method, two other LDA based methods are also implemented and the results are compared.

5. RESULTS AND DISCUSSION

5.1 Results

Table 1 to Table 6 show the results for amazon corpora while Table 7 to Table 12 show the results for yelp data for larger corpora. The results for the smaller corpora of yelp data are listed in Table 13 to Table 18. There are six rows in every table. The first row represents the results of the Word2Vec based method, the second row represents the results of traditional LDA, the third row represents the results of Hierarchical Dirichlet Process, the fourth row represents the results of ProdLDA, the fifth row represents the results of proposed two step LDA. The explicit aspects as well as implicit aspects have been shown in the results. The implicit aspects have been shown in bold and italic format. The final row represents the ground truth annotated by human. The human annotations were done by the members of our research laboratory. All the aspects were extracted by them and rated by them.

5.2 Discussion: A Case Study

The results represented in Table 1 is described in detail here. The first row is the result from Word2Vec and rule based method (Das and Chakraborty, 2020) which worked on frequency based aspect extraction, i.e. high frequency noun words are most likely to be the aspects. In the next step, traditional LDA is used with fixed number of topics as 50, the two words with highest probability from the first ten most important topics are then chosen to input to the second step. The second row represents the 10 pairs of aspects which are to be fed to the guided LDA in the second step. It is found that the topics are not comprised of contextually important words, also the words with similar meaning are not being stored in a single topic.

To improve the quality of the aspects, two-step LDA, with the seed words from the topics of the traditional LDA, is proposed. For this case, twenty seed words for ten topics are chosen. Guided LDA is implemented for ten output topics. The first three topics are as follows:

Topic lists-

Topic 0: *filter, lens, protect, glass, buy, good, lense, use, get, well, protection, put, need, clean, camera, tiffen, come, quality, cheap, great.*

Topic 1: *filter, take, get, go, reflection, lens, shoot, use, image, light, picture, shot, photo, clean, try, cheap, see, remove, cause, know.*

Topic 2: *filter, buy, good, lens, protect, lense, cheap, price, camera, protection, product, put, make, much, think, take, time, quality, money, job.*

After closer inspection of the topics, it can be seen that the topics reflect the contextually important words with respect to the corpus with lots of duplication. The reason behind this is, as every review is being considered as a document, and the training is being forced by seed words, so the words matching with the seed words in most reviews along with the scarce words are being selected. All the duplicated words are removed after making a list of all the possible aspect words and then the words are again clustered to get more meaningful topics to be used as final aspects. The final aspect words are shown in the final row.

Table 1. Results for camera lens protector

Word2Vec	filter, quality, reflection, camera, lens, glass, price, image, protection, product
Traditional LDA	['filter'(0.444),'lens'(0.229)], ['take'(0.085),'go'(0.075)], [buy'(0.331),'good'(0.211)], ['really'(0.134),'lot'(0.082)], ['lense'(0.454),'also'(0.152)], ['camera'(0.446),'item'(0.107)], ['say'(0.228),'much'(0.201)], ['get'(0.420),'order'(0.127)], ['protection'(0.363),'scratch'(0.226)], ['photo'(0.152),'even'(0.139)]
Hierarchical Dirichlet Process	[‘filter’,‘lens’], [‘filter’,‘lens’], [‘filter’,‘auto_focus’], [‘lens’,‘blurry’], [‘filter’,‘basis’], [‘filter’,‘lens’], [‘cmo’,‘skier’], [‘hardly’,‘lee’], [‘favor’,‘recomendado’], [‘lens’,‘april’]
ProdLDA	['cool', 'equally'], ['higher', 'monitor'], ['perceive', 'leather'], ['scrathce', 'qualitymade'], ['disappointment', 'rainy'], ['rapido', 'method'], ['accept', 'everyone'], ['family', 'possibly'], ['candid', 'exceed'], ['density', 'filthy']
Two-step LDA	Protection-[‘Protection’,’protect’], Price-[‘cost’,’price’], Image-[‘picture’,’image’,’photo’], Lens-[‘lense’,’lens’,’camera’], Purchase experience-[‘purchase’,’buy’], Shooting picture-[‘***shoot***’,’***shot***’], Image distortion-[‘reflection’], [‘flare’], Glass quality-[‘glass’], Ease of scratch-[‘scratch’]
Human extracted	protection, glare/ reflection, distortion, mount, images, coating, price, cleaning, durability, light source

Table 2. Results for headphone

Word2Vec	sound, bass, headphones, quality, music, koss, pouch, set, head, headband
Traditional LDA	['headphone'(0.147),'sound'(0.096)], ['ear'(0.091),'wear'(0.065)], ['look' (0.105), 'sound_quality'(0.088)], ['portapro'(0.205),'lot'(0.090)], ['year'(0.141),'break'(0.110)], ['review'(0.134),'recommend'(0.079)], ['phone'(0.249),'other'(0.061)], ['comfort'(0.149),'long'(0.103)], ['product'(0.225),'day'(0.108)], ['portable'(0.174),'easily'(0.130)]
Hierarchical Dirichlet Process	[‘headphone’, ‘sound’], [‘headphone’, ‘great’], [‘headphone sound’], [‘tiny’, ‘slightly’], [‘durability’, ‘google’], [‘funk’, ‘frick’], [‘ohm’, ‘headphone’], [‘dismantle’, ‘modification’], [‘controlling’, ‘distributor’], [‘bug’, ‘ofthe’]
ProdLDA	[‘penchant’, ‘davismetal’], [‘deny’, ‘unparalleled’], [‘delivering’, ‘muddie’], [‘archos’, ‘initial’], [‘sizing’, ‘wireless’], [‘scare’, ‘pocketpod’], [‘siii’, ‘belt’],[‘iucky’, ‘passable’], [‘bithead’, ‘multiply’], [‘sannheiser’, ‘aged’]
Two-step LDA	Price-[‘price’], Purchase experience-[‘buy’,’purchase’], Quality-[‘Amazing’,’great’], Sound-[‘sound’], Design-[‘design’] Cable-[‘cord’,’wire’] Speaker-[‘speaker’] Bass-[‘bass’] Case-[‘case’]
Human extracted	ear cushion, speaker, case/pouch, cable, headband, durability, sound, isolation, bass, price

Table 3. Results for paper shredder

Word2Vec	shreds, shredder, cutters, sheets, card, paper, price, machine, problems, piece
Traditional LDA	['shredder'(0.412),'shred'(0.156)], ['paper'(0.542),'get'(0.182)], ['buy'(0.307),'far'(0.292)], ['price'(0.336),'little'(0.155)], ['use'(0.514),'lot'(0.239)], ['put'(0.302),'say'(0.234)], ['good'(0.480),'year'(0.187)], ['return'(0.169),'may'(0.161)], ['also'(0.329),'cd'(0.247)], ['empty'(0.273),'top'(0.184)]
Hierarchical Dirichlet Process	[‘shredder’,‘paper’],[‘professional’,‘cop e’], [‘shredder’,‘highly’], [‘nonfunctional’,‘lie’], [‘shredder’,‘difficulty’], [‘proceed’, ‘shredder’], [‘cancel’, ‘porch’], [‘okay’, ‘utilize’], [‘shredder’, ‘hardly’], [‘surprised’, ‘training’]
ProdLDA	[‘max’, ‘control’], [‘unintentionally’, ‘minutesfor’], [‘greasy’, ‘one’], [‘lift’, ‘series’], [‘privet’, ‘exspectation’], [‘quota’, ‘cannon’], [‘jet’, ‘visible’], [‘around’, ‘scam’], [‘trigger’, ‘discard’], [‘unplugged’, ‘set’]
Two-step LDA	Shredding-['shredding', 'shred', 'shredder'], Things to shred-['cardboard', 'plastic', 'paper'], ['card'], [‘cd’] Jamming-['jam', 'jammed'], Heating-['overheat'], Buying-['price', 'buy', 'purchase'], Problems-['issue', 'problem'], Basket-[‘Basket’], Noise-[‘***noisy***’,’***loud***’], Blade-[‘blade’] Weight-[‘heavy’]
Human extracted	shredding, cutters, performance, product quality, price, purchase, credit card, cutting, paper jamming, built, heating

Table 4. Results for television mount

Word2Vec	tv, mount, quality, drill, bolts, screws, angle, price, purchase, anchors
Traditional LDA	['tv'(0.217),'mount'(0.092)], ['wall'(0.143),'screw'(0.124)], ['work'(0.106),'tilt'(0.100)], ['purchase'(0.118),'want'(0.110)], ['easy'(0.389),'install'(0.290)], ['also'(0.106),'instruction'(0.099)], ['look'(0.147),'find'(0.140)], ['use'(0.378),'sturdy'(0.166)], ['unit'(0.125),'perfect'(0.112)], ['monitor'(0.089),'may'(0.079)]
Hierarchical Dirichlet Process	[‘mount’, ‘crooked’], [‘fastener’, ‘different’], [‘tv’, ‘grateful’], [‘traffic’, ‘like’], [‘platform’, ‘love’], [‘patient’, ‘heart’], [‘weld’, ‘sanus’], [‘smallish’, ‘address’], [‘realy’, ‘spray’], [‘vesa’, ‘position’]
ProdLDA	[‘store’, ‘wonderful’], [‘world’, ‘plastic’], [‘together’, ‘finder’], [‘week’, ‘hiccup’], [‘earth’, ‘cruise’], [‘finder’, ‘resistance’], [‘buen’, ‘tightening’], [‘racket’, ‘extra’], [‘operation’, ‘bottem’], [‘hell’, ‘satrs’]
Two-step LDA	Durabiltiy-['***strong***', '***sturdy***', 'good', '***stable***'], Install-['instal', 'installation', 'install'], Manual-['describe', 'say', 'articulate'], Hardware-['screw', 'bolt', 'swivel'], Purchase experience-[‘Sell’,’purchase’,’price’,’buy’], Tv hardware-[‘cable’,’television’,’tv’] Tool-['bracket'] Usage setup-['corner', 'side', 'angle'] Mount-[‘mount’]
Human extracted	Installation, mount, fit, hardware, price, built, quality, joints, bolts, delivery

Table 5. Results for phone

Word2Vec	ooma, phone, device, calls, voice, internet, services, telo, cell, features
Traditional LDA	['phone'(0.058), 'call'(0.053)], ['service'(0.341),'great'(0.120)], ['month'(0.318),'pay'(0.224)], ['minute'(0.078),'router'(0.078)], ['unit'(0.153),'keep'(0.102)], ['set'(0.295),'easy'(0.255)], ['internet'(0.185),'charge'(0.145)], ['let'(0.038),'build'(0.017)], ['number'(0.385),'port'(0.257)], ['still'(0.214),'much'(0.102)]
Hierarchical Dirichlet Process	[‘ooma’, ‘phone’], [‘ooma’, ‘phone’], [‘ooma’, ‘phone’], [‘ooma’, ‘seam’], [‘ooma’, ‘service’], [‘ooma’, ‘call’], [‘ooma’, ‘permit’], [‘ooma’, ‘worse’], [‘mgs’, ‘ooma’], [‘indeed’, ‘everyone_know’]
ProdLDA	['calrity', 'vpn'], ['boiling', 'infographic'], ['esp', 'unpack'], ['dollar', 'adverse'], ['numbercall', 'slice'], ['led', 'attract'], ['changeover', 'positiv'], ['kind', 'registered'], ['alternif', 'powering'], ['gizmo', 'supportemma']
Two-step LDA	Problem- ['issue', 'problem'] Price- ['cost', 'fee', 'price'] Type of connection- ['phone', 'landline'] Quality- ['great', 'excellent', 'good'] Number porting-[‘port’] Brand-[‘ooma’] Internet connection-[‘router’] Quality- ['great', 'excellent', 'good'] Voice quality- ['voice']
Human extracted	call, voice, phone bill, installation, instruction, internet, support services, performance, price

Table 6. Results for printer

Word2Vec	brother, prints, printers, paper, instruction, cartridge, cable, driver, quality, laser
Traditional LDA	['printer'(0.201),'print'(0.084)], ['go'(0.146),'page'(0.142)], ['setup'(0.105),'computer'(0.103)], ['easy'(0.446),'fast'(0.287)], ['hour'(0.048),'directly'(0.027)], ['set'(0.634),'speed'(0.076)], ['put'(0.089),'add'(0.076)], ['printing'(0.419),'home'(0.179)], ['install'(0.176),'software'(0.120)], ['problem'(0.368),'quality'(0.224)]
Hierarchical Dirichlet Process	[‘printer’, ‘print’], [‘printer’, ‘print’], [‘printer’, ‘sequence’], [‘printer’, ‘use’], [‘printer’, ‘statuspage’], [‘printer’, ‘print’], [‘angle’, ‘fastness’], [‘bubble’, ‘contrast’], [‘avid’, ‘task’], [‘xerox’, ‘worsen’]
ProdLDA	['lol', 'receiver'], ['minimalist', 'brief'], ['envy', 'sidestep'], ['breezy', 'deliberately'], ['caveat', 'art'], ['includedwir', 'profile'], ['blotch', 'wizardy'], ['ultrafine', 'touch'], ['tank', 'registration'], ['caveat', 'there']
Two-step LDA	Setup-[‘set’,’***setting***’], Brand-[‘brother’], Cartridge- [‘inkjet’, ’toner’, ’print’, ’cartridge’, ’printer’, ’printing’, ’ink’] Installation- [‘Installation’,’instal’,’install’] Driver- [‘software’,’download’,’computer’], [‘driver’] Problems-[‘problem’,’issue’,’thing’] Manual-[‘instruction’] Wireless connection- [‘wireless’,’cable’,’router’, ’network’] Paper tray-[tray] Price-[‘price’,’cost’]
Human extracted	printing, instruction, installation, paper jam, toner cartridge, inkjet, laser, wireless, paper tray, price

Table 7. Results for maialino

Word2Vec	restaurant, meals, service, table, appetizers, waitress, waiter, pork, pasta, dish, menu.
Traditional LDA	['good'(0.060),'food'(0.049)], ['go'(0.071),'eat'(0.049)], ['dish'(0.112),'great'(0.097)], ['friend'(0.067), 'amazing'(0.060)], ['dessert'(0.089),'bread'(0.078)], ['really'(0.222), 'brunch'(0.126)], ['nice'(0.181),'tasty'(0.113)], ['olive'(0.036),'notice'(0.029)], ['vegetable'(0.026),'mushroom'(0.018)], ['pig'(0.155),'like'(0.075)]
Hierarchical Dirichlet process	[‘good’, ‘great’], [‘good’, ‘selezione’], [‘really_enjoyed’, ‘mint’], [‘wobbly’, ‘obligate’], [‘cuttlefish’, ‘absorbency’], [‘dare’, ‘logo’], [‘number’, ‘presentation’], [‘entry’, ‘pesto’], [‘agnolotti’, ‘sth’], [‘peroni’, ‘mmmm’]
ProdLDA	['one', 'lazy'], ['edible', 'mostly'], ['soul', 'broth'], ['irv', 'cooling'], ['sucked', 'williamsburg'], ['vecchio', 'chatting'], ['mil', 'escort'], ['borderline', 'mirror'], ['depth', 'doll'], ['deserve', 'completely']
Two-step LDA	Type of meals- ['meal', 'brunch', 'eat', 'dish', 'dessert', 'dinner', 'lunch', 'breakfast'], Bar/restaurant- ['restaurant', 'bar'], Pasta- ['cheese', 'pasta'], Taste- ['taste', 'flavor'], ['***tasty***', '***delicious***'] Reservation-['reservation'], Pork delicacy- ['pig', 'pork'], Drink-[‘wine’], Service-[‘service’], Staff-[‘waiter’], Experience- ['amazing', 'nice', 'great']
Human extracted	staff , service, reservation, food, entrée, appetizer, drink, dessert, pork, spaghetti, pasta, price.

Das et al., International Journal of Applied Science and Engineering, 18(1), 2020120

Table 8. Results for ABC kitchen

Word2Vec	place, salad, pasta, farm, order, price, kitchen, menu, dish, fish, food
Traditional LDA	['food'(0.033),'good'(0.032)], ['salad'(0.120),'back'(0.092)], ['start'(0.148),'sundae'(0.134)], ['toast'(0.156), 'perfect'(0.146)], ['taste'(0.225),'interesting'(0.066)], ['tasty'(0.146), 'portion'(0.091)], ['sweet'(0.131),'brunch'(0.126)], ['meat'(0.064),'soft'(0.050)], ['always'(0.135),'actually'(0.126)], ['bread'(0.146),'ricotta'(0.061)]
Hierarchical Dirichlet Process	[‘good’, ‘food’], [‘good’, ‘food’], [‘good’, ‘food’], [‘resi’, ‘smallish’], [‘hiccup’, ‘smidge’], [‘bar_area’, ‘mapuche’], [‘newspaper’, ‘throw’], [‘chervil’, ‘playful’], [‘fashion’, ‘wear’], [‘sad’, ‘teriffic’]
ProdLDA	['decadence', 'california'], ['barley', 'trend'], ['duplicate', 'shortcoming'], ['longing', 'scrumptious'], ['shouting', 'lowlight'], ['scrawny', 'fight'], ['unparalleled', 'effectively'], ['chard', 'fizz'], ['discount', 'antic'], ['scrumptious', 'coconut']
Two-step LDA	Food elements- ['vegetable', 'beet'], Meals- ['entree', 'meal', 'menu', 'salad'], Taste- ['taste', 'flavor'], Dessert- ['sundae', '***delicious***', 'dessert'], Price-[‘price’], Types of meal- ['dinner', 'lunch'] Service-[‘service’] Reservation-[‘reservation’], Staff-[‘server’], Non veg items-[‘chicken’], [‘lobster’] Pizza-[‘pizza’]
Human extracted	Food, entrée, salad, pasta, pizza, staff, service, fish, ice cream, cheeseburger, atmosphere, drink, menu, price

Table 9. Results for casa mono

Word2Vec	Service, mono, experience, plate, restaurant, tapas, bar,
Traditional LDA	['food'(0.075),'good'(0.058)], ['place'(0.052),'order'(0.049)], ['dish'(0.103),'think'(0.038)], ['find'(0.035),'top'(0.032)], ['wait'(0.108),'reservation'(0.100)], ['sweet'(0.055), 'fry'(0.039)], ['duck_egg'(0.109),'bread'(0.087)], ['excellent'(0.099),'overall'(0.088)], ['enough'(0.110),'portion'(0.101)], ['barely'(0.016),'highlight'(0.015)]
Hierarchical Dirichlet Process	[‘dish’, ‘good’], [‘place’, ‘good’], [‘food’, ‘go’], [‘mon’, ‘adorably’], [‘detract’, ‘surprise’], [‘fortunately’, ‘pattern’], [‘evoo’, ‘dog’], [‘cloy’, ‘unsolicited’], [‘cranberry’, ‘better’], [‘hickory’, ‘write’]
ProdLDA	['suanne', 'mmmmmmm'], ['upsetting', 'professional'], ['trevor', 'cloud'], ['refer', 'cunchy'], ['shame', 'artfully'], ['denoting', 'bastianich'], ['piquillo', 'unfussy'], ['want', 'bullshit'], ['compete', 'western'], ['hunk', 'succeed']
Two-step LDA	Dessert- ['***tasty***', 'dessert', 'sweet'], Restaurant-['restaurant', 'bar', 'menu', 'waiter'], Staff-['server'], ['cook', 'chef', 'eat'], ['hostess'] Food-['clam', 'mussel'], ['bread'], ['food'], ['pork_belly'], ['sauce', 'dish'], ['tapa'], Meal-[‘dinner', 'lunch', 'meal'], Service-['service'], Reservation-['reservation'], Drink-['wine'], Taste-['taste', 'flavor'], ['***salty***'], Experience-['excellent', 'amazing', 'great', 'perfect'], ['nice', 'good'], ['experience'] Service-['serve'], Price-['price'],
Human extracted	Staff, food, chorizo, service, meatball, drink, salad, dish, octopus, menu, wine, price, space

Table 10. Results for pylos

Word2Vec	reservations, dinner, place, restaurant, drink, service , octopus, meal, server, experience
Traditional LDA	['good'(0.070),'order'(0.044)], ['food'(0.249),'place'(0.164)], ['love'(0.103),'pylo'(0.094)], ['dish'(0.144),'also'(0.101)], ['appetizer'(0.134),'wait'(0.078)], ['ceiling'(0.107),'fresh'(0.082)], ['friend'(0.151),'waiter'(0.102)], ['entree'(0.156),'much'(0.113)], ['even'(0.104),'may'(0.092)], ['check(0.071)','yet'(0.048)]
Hierarchical Dirichlet Process	[‘good’, ‘talk’], [‘front’, ‘thrill’], [‘zeus’, ‘wander’], [‘support’, ‘exchellend’], [‘magic’, ‘anyhow’], [‘imaginative’, ‘spiciness’], [‘really’, ‘die’], [‘charge’, ‘perfectly_seasoned’], [‘ass’, ‘pet’], [‘made_reservation’, ‘musaka’]
ProdLDA	['enought', 'custard'], ['crowd', 'glaze'], ['opt', 'accomodat',], ['pressure', 'limit'], ['cyprus', 'deborah'], ['napolean', 'melt'], ['cling', 'watery'], ['neatly', 'confused'], ['haunt', 'punchy'], ['payment', 'hypothetical']
Two-step LDA	Service-[‘service’], Drink-[‘ wine’]. Price-[‘price’], Meal- [‘meal’, ‘dinner’], [‘entree’, ’appetizer`] Staff-[‘server’], [‘waiter’] Dessert- ['delicious', 'dessert'], Reservation-[‘reservation’], Side- ['sauce', 'salad'], Taste- ['flavor', 'taste'],
Human extracted	Staff, food, chorizo, service, meatball, drink, salad, dish, octopus, menu, wine, Price, space, music

Table 11. Results for cook shop

Word2Vec	service, food, reservation, restaurants, brunch, lunch, salad, staff, menu, experience
Traditional LDA	['good'(0.062),'food'(0.052)], ['restaurant'(0.153),'try'(0.142)], ['would'(0.114),'egg'(0.106)], ['reservation'(0.112),'bar'(0.098)], ['know'(0.145),'price'(0.099)], ['high'(0.056),'pizza'(0.052)], ['way'(0.152),'experience'(0.121)], ['forget'(0.046),'include'(0.038)], ['want'(0.163),'find'(0.114)], ['always'(0.138),'end'(0.118)]
Hierarchical Dirichlet Allocation	[‘good’, ‘brunch’], [‘luckily’, ‘food’], [‘detail’, ‘whiny’], [‘hold’, ‘unable’], [‘marryland’, ‘food’], [‘brunchwise’, ‘replish’], [‘wonderfully’, ‘attractive’], [‘sudden’, ‘serve’], [‘griddled’, ‘amazingly’], [‘beginning’, ‘surprised’]
ProdLDA	['travel', 'universe'], ['hectic', 'traffic'], ['kudo', 'pulse'], ['brisk', 'disappoint'], ['smear', 'probably'], ['teapot', 'porridge'], ['lisa', 'chi'], ['goodness', 'verde'], ['tranquility', 'sighting'], ['knowledgeable', 'kevin']
Two-step LDA	Menu-['menu', 'salad'] Order- ['order'] Restaurant- ['bar', 'restaurant'] Service- ['service'] Taste- ['flavor', 'taste'] Foods- ['dish', '***tasty***', 'dessert', 'entree'] Staff- ['server'] Types of meal- ['dinner', 'breakfast', 'lunch', 'brunch', 'meal'] Drink- ['cocktail'], [‘wine’] Experience-[‘experience’]
Human extracted	Service, staff, décor, food, reservation, chicken, turkey sausage, steak, salmon, salad, drink

Table 12. Results for sakagura

Word2Vec	Service, dishes, taste, price, beef, sake, sashimi, sushi, food, chicken.
Traditional LDA	['food'(0.086),'sake'(0.081)], ['restaurant'(0.081),'get'(0.074)], ['dessert'(0.084),'time'(0.080)], ['really'(0.131),'find'(0.130)], ['would'(0.095),'amazing'(0.091)], ['large'(0.037),'next'(0.032)], ['top'(0.124),'nice'(0.099)], ['back'(0.117),'light'(0.055)], ['table'(0.143),'remember'(0.079)], ['flavor'(0.168),'piece'(0.090)]
Hierarchical Dirichlet Process	['good', 'sake'], ['sake', 'example'], ['deli', 'sens'], ['drunken', 'patron'], ['feat', 'gryll'], ['good', 'worry'], ['purposely', 'bartender'], ['element', 'luggage'], ['jumai', 'art'], ['factor', 'prettiest']
ProdLDA	['superb', 'belly'], ['upfront', 'delight'], ['steamed', 'sleek'], ['neo', 'culinarily'], ['erst', 'comforting'], ['prime', 'work'], ['minor', 'ini'], ['gaijin', 'lone'], ['pleased', 'amount'], ['priceless', 'chrysanthemum']
Two-step LDA	Service-['serve'], ['service'] Restaurant staff- ['restaurant', 'menu', 'bar', 'waiter'], Drink- ['drink', 'bottle', 'eat'], ['sake'] Meal type- ['dinner', 'lunch'], Food- ['noodle'], ['pork_belly', 'eggplant'], ['fish','salmon_roe'], ['meat', 'pork', 'food'], ['dish', 'meal', '***tasty***', 'dessert', 'sauce', '***delicious***'], Experience- ['beautiful', 'nice', 'love'], ['excellent', 'good', 'great', 'amazing'], ['experience'] Food ethnicity- [***'japanese***'], Reservation- ['reservation'], Taste- ['flavor', 'taste'], ['sweet'], Price- ['price'], ['***expensive***'],
Human extracted	Food, service, staff, atmosphere, presentation, price, reservation, sake, pork, beef, décor, sashimi

Table 13. Results for Greek restaurant

Word2Vec	food, salad, lamb, chicken, service, place, restaurant, bread, pita, meal
Traditional LDA	['place'(0.128),'food'(0.069)], ['go'(0.063),'lunch'(0.063)], ['restaurant'(0.044),'love'(0.037)], ['sandwich'(0.057),'bread'(0.054)], ['eat'(0.095),'snack'(0.074)], ['eggplant'(0.059),'meze'(0.051)], ['baklava'(0.074),'dish'(0.046)], ['healthy'(0.096),'definitely'(0.072)], ['meat'(0.051),'choice'(0.044)], ['favorite'(0.043),'dinner'(0.041)]
Hiearchical Dirichlet Process	['fave', 'abstractly'], ['imagine', 'dressed'], ['topping', 'actually'], ['kefi', 'bargain'], ['lucky', 'stick'], ['smile', 'start'], ['important', 'penni'], ['enough', 'white'], ['storefront', 'nearby'], ['dining', 'comment']
ProdLDA	['weary', 'consistent'], ['fast', 'simplicity'], ['roasted', 'natty'], ['especially', 'considerable'], ['honestly', 'boureki'], ['latter', 'low'], ['depend', 'bed'], ['lady', 'amount'], ['difficult', 'pay'], ['quietly', 'fell']
Two-step LDA	Side course-['soup', 'meal', 'bread', 'dish', 'sandwich', 'sauce', 'salad'] Overall experience-['nice', 'really', 'awesome', 'perfect', 'wonderful', 'definitely', 'good', 'amazing', 'alright', 'great'] Taste-['***tasty***', 'spinach_pie', ***'delicious'***], ['taste', 'flavor'] Non-vegetarian foord-['meat', 'lamb', 'chicken', 'food'] Price-['price'] Types of meal-['lunch', 'dinner', 'snack'] Vegetables used-['cucumber', 'tomato', 'onion'] Food ethnicity-['***greek***'] Staff-['staff'], ['restaurant','waitress'] Drink-['beer']
Human extracted	Food, snack, greek dish, appetizer, salad, drink, lamb sandwich, chicken sandwich, staff, service, atmosphere

Table 14. Results for peppino's pizza

Word2Vec	staff, place, order, pizza, pie, peppino, crust, sauce, price, mozzarella.
Traditional LDA	['pizza'(0.126),'good'(0.070)], ['crust'(0.50),'make'(0.047)], ['slice'(0.042),'real'(0.030)], ['pizza'(0.042),'really'(0.039)], ['live'(0.045),'friendly'(0.042)], ['expensive'(0.033),'attentive'(0.031)], ['dough'(0.041),'great'(0.035)], ['may'(0.049),'go'(0.044)], ['would'(0.041),'bread'(0.040)], ['day'(0.032),'grimaldi'(0.019)"]
HIearchical Dirichlet Process	['tomato', 'operate'], ['show', 'indifferent'], ['gentleman', 'amazed'], ['sew', 'gem'], ['average', 'diavolas'], ['clove', 'ordering'], ['tolerant', 'delightfully'], ['alla_vodka', 'cold'], ['homemade', 'brooklyn'], ['trip', 'forage']
ProdLDA	['favor', 'gobble'], ['bite', 'san'], ['music', 'incomparable'], ['fluffy', 'biaca'], ['sunset', 'blind'], ['reminiscent', 'thumb'], ['margharita', 'marguerita'], ['pepper', 'say'], ['mastery', 'fancy'], ['canned', 'legit']
Two-step LDA	Staff-['***friendly***'], ['waitress', 'restaurant'] Food-['sausage', 'cheese', 'pizza'] Sides-['salad', 'sauce', '***delicious***'] Pizza-['crust', 'oven'] Meal-['lunch', 'dinner'] Food ethnicity-['***italian***'] Drink-['soda'] Service-['service'] Price-['price'] Pizza delivery-['delivery']
Human extracted	Staff, service, atmosphere, pizza, bread, pie, pasta, salad, sauce, price, delivery

Table 15. Results for dekalb restaurant

Word2Vec	food, restaurant, menu, staff
Traditional LDA	['order'(0.028),'food'(0.028)], ['food'(0.032),'service'(0.015)], ['would'(0.022),'taste'(0.022)], ['good'(0.024),'say'(0.016)], ['great'(0.023),'go'(0.020)], ['get'(0.028),'find'(0.024)], ['place'(0.033),'love'(0.025)], ['food'(0.031),'price'(0.017)], ['get'(0.026),'really'(0.026)], ['find'(0.019),'space'(0.019)]
Hierarchical Dirichlet Process	['line', 'need'], ['intrigue', 'last'], ['brownstone', 'mussel'], ['mean', 'fresh'], ['hang', 'general'], ['executive', 'pay'], ['tend', 'total'], ['fare', 'uneven'], ['yesterday', 'rave'], ['license', 'lack']
ProdLDA	['believe', 'sloppy'], ['breaker', 'lovely'], ['beautiful', 'scramble'], ['diet', 'friend'], ['natural', 'fresh'], ['longstanding', 'owner'], ['corner', 'dad'], ['culture', 'outstanding'], ['hot', 'run'], ['extensive', 'consume']
Two-step LDA	Food-['omelette'], Sauce-['puree', 'leek'], Place-['place'], Review-['rave'], Taste-['***flavorless***'], Staff-['helpful']
Human extracted	Food, service, recommendation, atmosphere, exterior, staff, menu, drink, seat, burger

Table 16. Results for blue spoon coffee

Word2Vec	Coffee, place, latte.
Traditional LDA	['drink'(0.021),'go'(0.018)], ['get'(0.026),'place'(0.026)], ['coffee'(0.056),'serve'(0.024)], ['place'(0.033),'coffee'(0.026)], [‘place’(0.034), ‘coffee’(0.023)] ['coffee'(0.023),'get'(0.016)], ['coffee'(0.041),'drink'(0.027)], ['place'(0.031),'coffee'(0.030)], ['good'(0.030),'serve'(0.020)], ['great'(0.049),'place'(0.032)]
Hierarchical Dirichlet Process	[‘world’, ‘local’], [‘tasty’, ‘decaf’], [‘plum’, ‘cloud’], [‘size’, ‘cool’], [‘something’, ‘shut’], [‘name’, ‘bfast’], [‘max’, ‘baked’], [‘close’, ‘grab’], [‘hooked’, ‘next’], [‘dump’, ‘positive’]
ProdLDA	['frou', 'soggy'], ['inclined', 'lug'], ['tell', 'enough'], ['crawl', 'block'], ['conversation', 'suppose'], ['soho', 'utmost'], ['vote', 'fair'], ['ham', 'soho'], ['certainty', 'downpour'], ['next', 'condiment']
Two-step LDA	Menu-['choice', 'selection'], Types of coffee-['espresso', 'coffee'], Service-['fast', 'quickly'], Flavors-['creamy', 'goat_cheese', 'flavor'] Price-[‘overprice’], Taste enhancer-[‘honey’,’syrup’], Overall experience-['good', 'awesome', 'great'], [‘***charming***’] Experience of coffee-['enticing', 'entice'] Coffee beans-[‘roasted’, ‘roast’] Café-['cafe', 'bakery', 'restaurant']
Human extracted	Staff, drinks, coffee, honey lavender latte, cookies, bagel, soup, salad, pastries, scone, recommend

Table 17. Results for hunter’s

Word2Vec	Food, drinks, staff, places, service, menu, salad, dinner, brunch, hunters
Traditional LDA	['great'(0.030),'get'(0.026)], ['want'(0.016),'go'(0.015)], ['good'(0.064),'restaurant'(0.038)], ['hunter'(0.029),'menu'(0.028)], ['cocktail'(0.036),'delicious'(0.029)], ['hunter'(0.029),'main'(0.029)], ['meal'(0.029),'fresh'(0.024)], ['lot'(0.036),'place'(0.032)], ['dinner'(0.030),'place'(0.027)],
Hierarchical Dirichlet Process	[‘day’, ‘look’], [‘literally’, ‘enjoyable’], [‘tingle’, ‘surprised’], [‘airy’, ‘creme’], [‘super’, ‘find’], [‘rigeur’, ‘mason’], [‘katness’, ‘perfection’], [‘gloopy’, ‘square’], [‘brother’, ‘available’], [‘group’, ‘possible’]
ProdLDA	['crave', 'cheap'], ['inspire', 'disappoint'], ['runny', 'vegetarian'], ['sundays', 'pindar'], ['hop', 'round'], ['bowl', 'mussel'], ['awesome', 'lucked'], ['freeze', 'liking'], ['potpie', 'elvis'], ['temperature', 'review']
Two-step LDA	Meal type-['brunch', 'dinner', 'dessert', 'cocktail', 'meal'] Service-[‘service’] Menu-[‘menu’, ‘entree’] Drink-[‘coffee’], ['bottle', 'drink'] Experience-['amazing', 'nice', 'great'], ['experience'] Food items-['cheese', 'sauce', 'bread', 'wine', 'burger'] Reservation-['reservation'] Staff-['***friendly***'], ['waitress'] Taste-['sweet', 'delicious'] Price-['price']
Human extracted	Food, service, ambience, staff, price, space, coffee, cocktail, appetizer, fish, salad

Table 18. Results for alameda

Word2Vec	Place, food, menu
Traditional LDA	['little'(0.024),'good'(0.024)], ['bar'(0.025),'time'(0.022)], ['really'(0.020),'alameda'(0.016)], ['menu'(0.022),'place'(0.015)], ['place'(0.023),'old'(0.016)], ['great'(0.022),'place'(0.019)], ['really'(0.029),'also'(0.029)], ['great'(0.036),'would'(0.027)], ['drink'(0.024),'food'(0.024)], ['serve'(0.025),'old'(0.017)]
Hierarchical Dirichlet Process	['cheddar', 'sign'], ['specific', 'outdoor'], ['expectation', 'Handful'], ['work', 'housemade'], ['key', 'dock'], ['expect', 'enhancement'], ['option', 'affair'], ['buck', 'beautiful'], ['octopus', 'song'], ['enter', 'brine']
ProdLDA	['combo', 'man'], ['offer', 'pair'], ['worker', 'reach'], ['fun', 'grab'], ['brussel', 'material'], ['friendly', 'mackerel'], ['tough', 'velvet'], ['min', 'blare'], ['street', 'remind'], ['else', 'hamburger']
Two step LDA	Price-['dollar'], ['***pricier***', '***expensive***'], ['price'], ['overprice'] Taste-['sweet', '***tasty***', '***delicious***', 'cute'], ['***salty***'], Atmosphere- ['atmosphere', 'vibe'], ['neighborhood', 'area'], Food- ['food', 'meat'], ['oyster'], ['sauce', 'salad', 'flavor','dish'], ['burger', 'menu', 'cheeseburger', 'sausage'], ['bread', 'cheese'] Experience-['decent', 'good', 'excellent', 'solid'], Service-['serve'], Drink-['cocktail'], ['bar'], ['wine'], ['drink', 'beer'], Staff-['waitress'], ['***friendly***'], ['staff'] Restaurant-['dining', 'restaurant'], Name-['alameda']
Human extracted	Food, staff, drinks, bar, ambiance, décor, price, service, menu

5.3 Analysis:

From the results of two-step LDA, it can be inferred that the two-step LDA gives better results than the Word2Vec based method and LDA only method.

1. The two-step LDA method extracts implicit aspects very well, which was lacking in Word2Vec method and only traditional LDA based method.
2. Two-step LDA method performs better in case of small corpora also. As Word2Vec extracts only explicit aspects but two-step LDA process detects implicit aspects also.
3. Two-step LDA results are much closer to the ground truth annotated by human.

6. CONCLUSIONS

Aspect extraction is an important part of the review summarization process in order to have a full picture for evaluation of the products or services from the reviews. Lot of researches have proposed various techniques in this field, but still there is no general accepted method to achieve the best result. As automatic comprehension of natural language is itself difficult for proper understanding of inherent semantics, review summarization is more difficult as those are mostly comprised of unstructured texts. The proposed two-step LDA based method addresses the shortcoming of the previously developed rule based method based on Word2Vec.

Two-step LDA seems to be better at extracting aspects as the clusters formed at the end of proposed two-step LDA are more coherent than the traditional LDA only method. Even in some cases where the traditional LDA produces duplicate aspects, two-step LDA produces diverse coherent clusters. Also it has been found, at least qualitatively, that the extracted aspects match the ground truth. In the case of smaller corpora, Word2Vec based method could not perform well and traditional LDA cannot be trained properly with smaller corpora, the two step LDA has shown promising result. Thus the proposed approach seems to be a candidate for automatic review summarization. It is also has been shown that the proposed approach performs well compared to two other LDA based approaches for aspect extraction.

ACKNOWLEDGMENT

We are grateful to Julian McAuley, UCSD for providing us with the valuable Amazon dataset and Shebuti Rayana, Leman Akoglu for providing us with Yelp dataset. We are also grateful to Pattern Recognition and Machine Learning laboratory of Iwate Prefectural University for providing us with tools to accomplish the research activities.

REFERENCES

Allahyari, M., Pouriyeh, S., Kochut, K., Arabnia, H.R. 2017. A knowledge-based topic modeling approach for automatic topic labeling. International Journal of Advanced Computer Science and Applications(ijacsa), 8. http://dx.doi.org/10.14569/IJACSA.2017.080947

Bagheri, A., Saraee, M., De Jong, F. 2014. ADM-LDA: An aspect detection model based on topic modelling using the structure of review sentences. Journal of Information Science, 40, 621–636.

Bagheri, A., Saraee, M., Jong, F.D. 2013. Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. Knowledge-Based Systems 52 201213, DOI:10.1016/j.knosys.2013.08.011

Blei, D.M., Ng, A.Y., Jordan, M.I. 2003. Latent dirichlet allocation, J Mach Learn Res 3 (Jan.). 993–1022.

Brody, S., Elhadad, N. 2010. An unsupervised aspect-sentiment model for online reviews. Human Language Technologies: in Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT-10), Los Angeles, USA, 804-812.

Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R. 2013. Leveraging multi-domain prior knowledge in topic models. Proceedings of the Twenty-Third international joint conference on Artificial Intelligence (IJCAI-13), Beijing, China, AAAI Press, 2071–2077.

Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R. 2013. Discovering coherent topics using general knowledge. Proceedings of the 22nd ACM international conference on Conference on information & knowledge management (CIKM-13), San Francisco, USA, 209–218.

Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R. 2013. Exploiting domain knowledge in aspect extraction. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP-13), Seattle, USA, 1655–1667.

Das, S.J., Chakraborty, B. 2019. An approach for automatic aspect extraction by latent dirichlet allocation. IEEE 10th International Conference on Awareness Science and Technology (iCAST), Morioka, Japan, 1-6. doi: 10.1109/ICAwST.2019.8923417

Das, S.J., Chakraborty, B. 2020. Design of a category independent, aspect based automated opinion analysis technique for online product reviews. International Journal of Applied Science and Engineering, 17, 175–189. https://doi.org/10.6703/IJASE.202005_17(2).175

Debortoli, S., Müller, O., Junglas, I., vom Brocke, J. 2016. Text mining for information systems researchers: An annotated topic modeling tutorial. Communications of the Association for Information Systems, 39. https://doi.org/10.17705/1CAIS.03907. 110–135.

Ekinci, E., Omurca, S.I. 2017. Extracting implicit aspects based on latent dirichlet allocation. Doctoral Consortium - DCAART, (ICAART 2017) ISBN, 17–23.

Fang, L., Huang, M. 2012. Fine granular aspect analysis using latent structural models. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju, South Korea: Short Papers- 2, 333–337.

Hajmohammadi, M.S., Ibrahim, R., Othman, Z.A. 2012. Opinion mining and sentiment analysis: A survey. International Journal of Computers & Technology, 2. ISSN:2277-3061(online)

He, R., Jullian, McAuley, 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. World Wide Web conference. DOI : http://dx.doi.org/10.1145/2872427.2883037. 507–517.

Hu, M., Liu, B. 2004. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, WA, USA. https://doi.org/10.1145/1014052.1014073. 168–177

Jakob, N., Gurevych, I. 2010. Extracting opinion targets in a single- and cross-domain setting with conditional random fields. 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, Massachusetts.

Jin, W., Ho, H.H., Srihari, R.K., 2009. OpinionMiner: a novel machine learning system for web opinion mining and extraction. 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Paris, France.

Jo, Y., Oh, A.H. 2011. Aspect and sentiment unification model for online review analysis. Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM-11), Hong Kong, 815–824.

Kessler, J.S., Nicolov, N. 2009. Targeting sentiment expressions through supervised ranking of linguistic configurations. Third International AAAI Conference on Weblogs and Social Media, San Jose, California, USA, 90–97.

Kumar, K. 2018. Evaluation of topic modeling: Topic coherence. https://datascienceplus.com/evaluation-of-topic-modeling-topic-coherence

Moghaddam, S., Ester, M., 2011. ILDA interdependent LDA model for learning latent aspects and their ratings from online product reviews. SIGIR'11, July 24–28, Beijing, China. Copyright 2011 ACM 978-1-4503-0757-4/11/07. https://doi.org/10.1145/2009916.2010006. 665–674.

Popescu, A.M., Etzioni, O. 2005. Extracting product features and opinions from reviews. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada. https://doi.org/10.3115/1220575.1220618. 339–346.

Rao, A., Shah, K. 2018. A domain independent technique to generate feature opinion pairs for opinion mining. WSEAS transactions on information science and applications, ISSN / E-ISSN: 1790-0832 / 2224-3402, 15, 61–69.

Rayana, S., Akoglu. L. Stony Brook University. http://odds.cs.stonybrook.edu/yelpzip-dataset/.

Singh, V. 2017. Guided LDA: Guided topic modeling with latent Dirichlet allocation. https://guidedlda.readthedocs.io/en/latest/

Srivastava, A., Sutton, C. 2017. Autoencoding variational inference for topic models, Proc. Int. Conf. Learn. Representations. arXiv:1703.01488

Teh, Y., Jordan, M., Beal, M., Blei, D. 2006. Hierarchical dirichlet processes. Journal of the American Statistical Association. 101. 10.2307/27639773. 1566–1581.

Wang, T., Cai, Y., Leung, H.-f., Lau, R.Y., Li, Q., Min, H. 2014. Product aspect extraction supervised with online domain knowledge. Knowledge-Based Systems, 71, 86–100.

Wawer, A. 2015. Towards domain independent opinion target extraction. IEEE 15th International Conference on Data Mining Workshops(ICDMW), DOI: 10.1109/ICDMW.2015.255, 1326 –1331.

Xu, X., Tan, S., Liu, Y., Cheng, X., Lin, Z. 2012. Towards jointly extracting aspects and aspect-specific sentiment knowledge. Proceedings of the 21st ACM International Conference on Information and Knowledge management (CIKM-12). Maui Hawaii, USA, 1895–1899.

Xueke, X., Xueqi, C., Songbo, T., Yue, L., Huawei, S. 2013. Aspect level opinion mining of online customer reviews. China Communications, 10, 25–41.

Yi, J., Nasukawa, T., Bunescu, R., Niblack, W. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. Proceedings of the Third IEEE International Conference on Data Mining. doi: 10.1109/ICDM.2003.1250949. 427–434.

Zhao, W.X., Jiang, J., Yan, H., Li, X. 2010. Jointly modeling aspects and opinions with a Maxent-LDA hybrid. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP-10), Massachusetts, USA, 56–65.