DOMAIN-INDEPENDENT BAYESIAN MODEL FOR ASPECT CATEGORY DETECTION AND DISTRIBUTED VECTOR FOR IMPLICIT ASPECT EXTRACTION

AL JANABI OMAR MUSTAFA ABBAS

UNIVERSITI SAINS MALAYSIA

2021

DOMAIN-INDEPENDENT BAYESIAN MODEL FOR ASPECT CATEGORY DETECTION AND DISTRIBUTED VECTOR FOR IMPLICIT ASPECT EXTRACTION

by

AL JANABI OMAR MUSTAFA ABBAS

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

October 2021

DEDICATION

To the honour of our first teacher:

Prophet Mohammed Ibn Abdullah, Peace Be Upon Him

To My Beloved Parents: Mustafa Abbas AL-Janabi

Siham Hasan AL-Janabi

To My Beloved Brothers and Sisters

ACKNOWLEDGEMENT

All praises are due to Allah, the merciful the compassionate for all the knowledge, and for His permission for this work to be completed at this time.

My sincere gratitude and appreciation to my supervisor **Associate Professor Nurul Hashimah Ahamed Hassain Malim** who provided the necessary supervision and support, which culminated into the successful completion of this study.

I owe a deep sense of gratitude to my co-supervisor **Associate Professor Cheah Yu-N** for his keen interest on me at every stage of my research.

It is my privilege to thank my friend **Dr. Abdulkarim Kanaan Jebna** for his technical help.

I truly would thank my friend **Dr. Mohammad Alzayed** for the financial support.

It is important to always remember that, this work “PH.D.” won’t be possible without my father **Mustafa Abbas AL-Janabi**.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ALGORITHMS

Page

LIST OF SYMBOLS

	LDA model
α	the parameter for Per-document topic distribution
β	the parameter for Per-topic word distribution
φ_k	is the word distribution for each topic k.
$z_{i,j}$	is the word distribution for the $j-th$ word in a document i,
$w_{i,j}$	is the particular word (only observed component).
	TSLDA model
α	the parameter for Per-document topic distribution
β	the parameter for Per-topic word distribution
ρ	the parameter for Per-category aspect distribution
θ_d	is the topic distribution for document i.
$\emptyset_t$	is the word distribution for topic T.
μ_s	is the topic-seeds distribution
$z_{d,s}$	is the topic for the $s-th$ seed word in document d, and
w	is the aspect-term in TSLDA
	HDP_CGS
k	The initial number of topics
α	Parameter of topics prior
γ	Parameter of tables prior
V	Number of unique words in the vocabulary
Doc	List of documents
$\vec{\tau}$	Prior proportion
$\vec{\vartheta}_k$	Multinomial parameter of document to topic distribution

LIST OF ABBREVIATIONS

ABSA	Aspect-based sentiment analysis
ACD	Aspect Category Detection
ADM	Aspect detection model
ASUM	Aspect and Unification Model
BNP	Bayesian Nonparametric
BOWs	Bag of Words
CBOW	Continuous Bag of Words
CGS	Collapsed Gibbs Sampling
coAR	co-occurrence association rule
CRP	Chinese Restaurant Process
CV	Count Vectorizer
DD	Dirichlet Distribution
DP	Dirichlet Process
DV	Distributed Vector
GS	Gibbs sampling
HDP	Hierarchical Dirichlet Process
HDP_CGS	Hierarchical Dirichlet Process_Collapsed Gibbs Sampling
IOB-tagger	Inside-Outside-Beginning Tagger
JST	Joint Sentiment Topic
KL	Kullback–Leibler
LDA	Latent Dirichlet Allocation
MaxEnt-LDA	Maximum Entropy Latent Dirichlet Allocation
MC	Markov Chain
MCMC	Monte Carlo Markov Chain
NCE	Noise Contrastive Estimation
NLP	Natural Language Processing
PM	Phrase Model
PMI	Pointwise Mutual Information
POS	Part of Speech
SemEval	Semantic Evaluation
SML	Supervised Machine Learning
SVM	Support Vector Machine
TF-IDF	Term-frequency Inverse document frequency
TSLDA	Topic-seeds Latent Dirichlet Allocation
VI	Variational Inference
VS	Vector Space
VSM	Vector Space Model

MODEL BAYESIAN DOMAIN-INDEPENDEN BAGI PEPERIKSAAN KATEGORI DAN PENGANGKUTAN VEKTOR YANG DISEDIAKAN UNTUK PEMBEKAL ASPEK IMPLIK

ABSTRAK

Perkembangan web 2.0 telah meningkatkan kemampuan masyarakat untuk berkongsi sentimen atau pendapat mereka mengenai pelbagai perkhidmatan atau produk dengan mudah. Hal ini adalah untuk menyiasat pendapat umun yang dinyatakan dalam ulasan. Analisis sentimen berasaskan aspek (ABSA) dipercayai menerima satu set teks (cth., ulasan produk atau ulasan dalam talian) dan mengenal pasti sasaran pendapat (aspek) dalam setiap ulasan. Sistem analisis sentimen berasaskan aspek kontemporari, seperti pengelompokan aspek, banyak bergantung pada benih berasaskan leksikon dan berlabel secara manual yang disertakan ke dalam model topik. Sistem yang sebelumnya dikembangkan untuk Pengesanan Kategori Aspek atau Aspect Category Detection (ACD) banyak bergantung pada teknik pembelajaran mesin yang diselia. Masalah pengekstrakan aspek tersirat ditangani dengan menggunakan peraturan pra-binaan atau petunjuk pra-label untuk melakukan pengesanan aspek tersirat. Untuk mengatasi isu-isu ini, model probabilistik Bayesian mengusulkan untuk melakukan pengelompokan aspek, ACD, dan vektor teragih untuk pengekstrakan aspek tersirat. Model Bayesian parametrik dan bukan parametrik masing-masing dikembangkan untuk menjalankan data beranotasi dan tidak beranotasi, iaitu; Topic-seeds Latent Dirichlet allocation (TSLDA) dan Hierarchical Dirichlet Process-Collapsed Gibbs Sampling (HDP-CGS). Kumpulan aspek yang dihasilkan menggunakan model Bayesian yang dikembangkan dimasukkan ke dalam vektor teragih yang disarankan (cth:, Skip-gram) untuk pengekstrakan aspek tersirat.

Metodologi yang dicadangkan dinilai menggunakan beberapa set data penanda aras ulasan dalam talian (termasuk set data yang dianotasi menggunakan ulasan yang diambil dari Amazon.com dan TripAdvisor.com). Kami menjalankan sejumlah eksperimen yang mencukupi dan membuat demontrasi keberkesanan kaedah yang kami cadangkan. Khususnya, bagi data yang dianotasi, model TSLDA membuktikan bahawa penggabungan penyisipan kata adalah lebih baik dengan penggunaan benih dan leksikon yang dibuat secara manual untuk menjaga keteraturan semantik. Model HDP-CGS berkesan dengan membuat perbandingan dengan kedua-dua model parametrik dan bukan parametrik dari segi koheren topik. Akhirnya, Vektor Teragih secara perbandingannya lebih berpotensi dalam pengekstrakan aspek tersirat tanpa menggunakan kaedah berasaskan peraturan.

DOMAIN-INDEPENDENT BAYESIAN MODEL FOR ASPECT CATEGORY DETECTION AND DISTRIBUTED VECTOR FOR IMPLICIT ASPECT EXTRACTION

ABSTRACT

The development of Web 2.0 has improved peoples' ability to share their sentiments, or opinions, on various services or products easily. This is to investigate the public opinions that are expressed within the reviews. Aspect-based sentiment analysis (ABSA) deemed to receive a set of texts (e.g., product reviews or online reviews) and identify the opinion-target (aspect) within each review. Contemporary aspect-based sentiment analysis systems, like the aspect grouping, rely predominantly on lexicon-based and manually labelled seeds that is being incorporated into the topic models. The previously developed systems for Aspect Category Detection (ACD) rely mostly on supervised machine learning techniques. The problem of implicit aspect extraction is being addressed using either pre-constructed rules or pre-labelled clues for performing implicit aspect detection. To cope with these issues, Bayesian probabilistic models proposed to perform the aspect grouping, ACD, and distributed vectors for implicit aspect extraction. Parametric and non-parametric Bayesian models are developed to conduct both the annotated and non-annotated data, that are; Topic-seeds Latent Dirichlet allocation (TSLDA) and Hierarchical Dirichlet Process-Collapsed Gibbs Sampling (HDP-CGS), respectively. The yielded aspect groups using the developed Bayesian models fed into the advised distributed vector (i.e., Skip-gram) for implicit aspect extraction. The proposed methodology evaluated using several online reviews benchmark datasets (including datasets annotated using reviews retrieved from Amazon.com and TripAdvisor.com). We conducted an adequate

amount of experiments and demonstrate the effectiveness of our proposed methods. Specifically, for the annotated data, TSLDA model proves that the incorporation of word embeddings is outperformed by the utilization of manually crafted seeds and lexicons to maintain the semantic regularities. HDP-CGS model is effective by making comparisons with both parametric and non-parametric models in terms of topic coherence. Finally, Distributed Vectors are comparatively promising for the implicit aspect extraction using no rule-based methods.

CHAPTER 1

INTRODUCTION

1.1 Overview

Sentiment analysis (or opinion mining) is a paramount field of research in natural language processing, information retrieval, data mining, and Web mining. It is a computational study of people's opinions, sentiments, moods, and emotions. In recent years, sentiment analysis research and applications have been widely witnessed in management science and social science due to their importance for both business and society. Before the advent of the internet, many of us asked our friend's opinion to recommend a specific product before purchasing it. The opinion used to be collected as a form of surveys or opinion polls. Currently, after the explosive growth of Web reviews, if someone wants to buy a product, there is no need to ask for a friend's opinion, because there are many public forums of reviews and discussions on the Web about the product.

Manually finding and examine the opinion reviews on the Web is a formidable task because of the proliferation of diverse reviews. Thus, sentiment analysis has emerged which not only can treat the diversity of the opinion reviews, but also identify and extract user opinions (Cambria et al., 2017).

Do, et al. (2019) argued that the study of aspect-based sentiment analysis is feasible at three levels – document, sentence, and entity or aspect.

Document-level depends on the sentiment expressed in the whole document, whether it is positive, negative or neutral (Ray & Chakrabarti, 2019). For example, given a product review, document-level system expresses the overall polarity of the product, whether it is positive, negative or neutral. This level of analysis is insufficient as it deals with the document as a whole, which expresses the polarity of the product by assuming each document represents one entity/product. This suggests a new level of analysis which can deal at sentence-level, because document-level is not applicable for multiple entities or products.

Sentence level expresses the sentiment as a sentence and determines the polarity based on each sentence. As it is a level of subjectivity analysis, C. Lin & He (2009) distinguish sentences that express factual information (aka objective sentences) than the sentence that conveys opinions called subjective sentences. Subjective sentences usually express the sentiment towards an entity/product. Even though this level of analysis is more specific in terms of analysis, a more fine-grained level is required to express the sentiment of an entity.

Aspect level sentiment analysis, also known as aspect-based sentiment analysis, is more fine-grained than document- and sentence level sentiment analysis. It deals with analysis of extracted text uploaded online to summarize opinions expressed on entities and aspect/features of entities (also called target). Example of entities include topics, services, issues, persons, products, organizations, or events. Product entity “Laptop” may consist of several aspects e.g. “CPU”, “screen” and “keyboard”. (L. Zhang et al., 2018). In “Restaurant” reviews, people usually complain about the price of the food e.g., “it is expensive.” Some opinionated reviews are ambiguous and have an implicit aspect in them, therefore stating something as expensive can refer to the implicit aspect of the price of food.

Aspect-level sentiment analysis as a whole has several utmost subtasks, such as aspect grouping (B. Liu, 2012; L. Zhang et al., 2018), opinion target[1] extraction (OTE), aspect category detection (ACD), and sentiment polarity (SP) (Pontiki et al., 2016). Figure (1.1) shows the different levels of text granularity, and aspect-level subtasks.

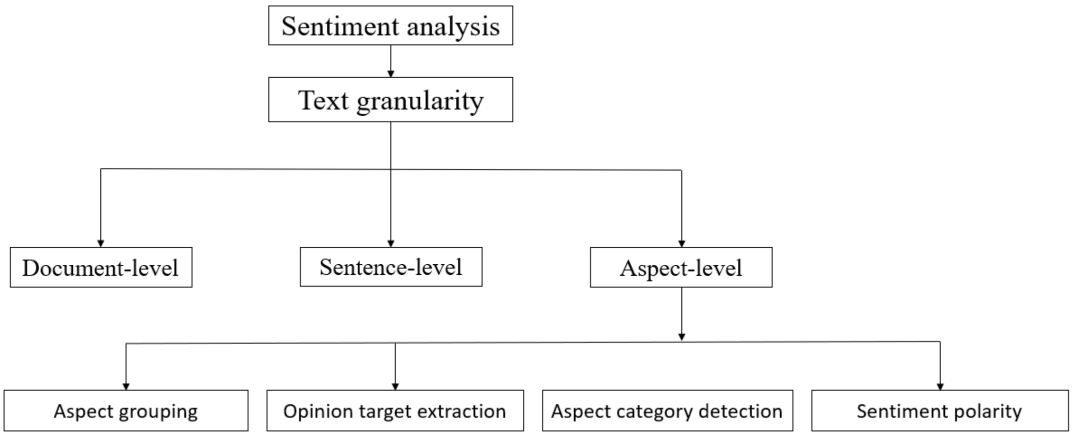

Figure 1.1 Granularity of text reviews in sentiment analysis and aspect-level subtasks.

Aspect grouping is to group semantically related aspect-terms (or expressions) into categories also known as aspect/aspect-term categorization. As the aspects may be expressed using synonymous words and phrases, a mechanism to group these words and phrases into category is needed (Einea et al., 2019; Ray & Chakrabarti, 2019; Q. Xu et al., 2020). In the domain of digital camera, the aspect-terms "photo" and "picture" are synonymous aspects. Aspect grouping is quite a crucial and critical subtask because the aspect terms are synonymous in a particular domain, but they are antonyms in other domains. Example, the aspect terms "movie" and "picture" are

[1] In this work "opinion-target" and "aspect term" are used interchangeably.

synonymous in movie reviews, but they are not in camera reviews. Where the aspect-term "photo" is more likely to be synonymous to "picture" while "video" is synonymous to "movie." Thus, the use of lexicon-based methods might not be eligible to group the aspect synonyms in some domains, because it required semantic analysis rather than a syntactic analysis. The aspect-term "design" has the synonymous aspect-term "appearance" but are antonyms in the lexicon-based methods (e.g., WordNet). Besides, determining which expression belongs to which opinion-target can be dependent on the domain's application need (or user's application need). In phone reviews, one may want to study battery as a whole (one opinion-target), or as more than one opinion-target (e.g., battery weight, and battery life). For this, figure (1.2) gives three examples of expressed opinions towards the aspect "graphic quality" of "laptop" as an entity. Example of the graphic aspect-category "graphic quality" may include several aspects/opinion-targets e.g., "graphics," "graphic card" or even "graphic quality" itself as shown in figure (1.2).

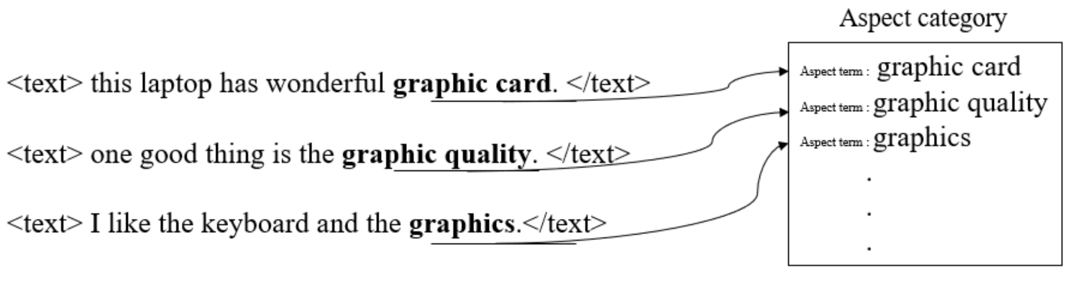

Figure 1.2 Aspect grouping on Laptop reviews.

Figure (1.3) represents the three other subtasks of aspect-based sentiment analysis. Opinion target extraction (OTE) is used to identify and extract the aspect in the opinionated reviews. Aspect category detection (ACD) is conducted to detect the aspect category discussed in each sentence, and sentiment polarity (SP) is to classify the identified aspect-terms based on the polarity; positive, negative or neutral. In

SemEval-2014 Restaurant dataset, the task of OTE is to extract the aspect term “staff” in the sentence review “but the staff was horrible to us.” For the ACD, the aspect category “service” has to be detected for the presented sentence and it is related with the aspect term and the sentiment polarity of the aspect which is “negative.”

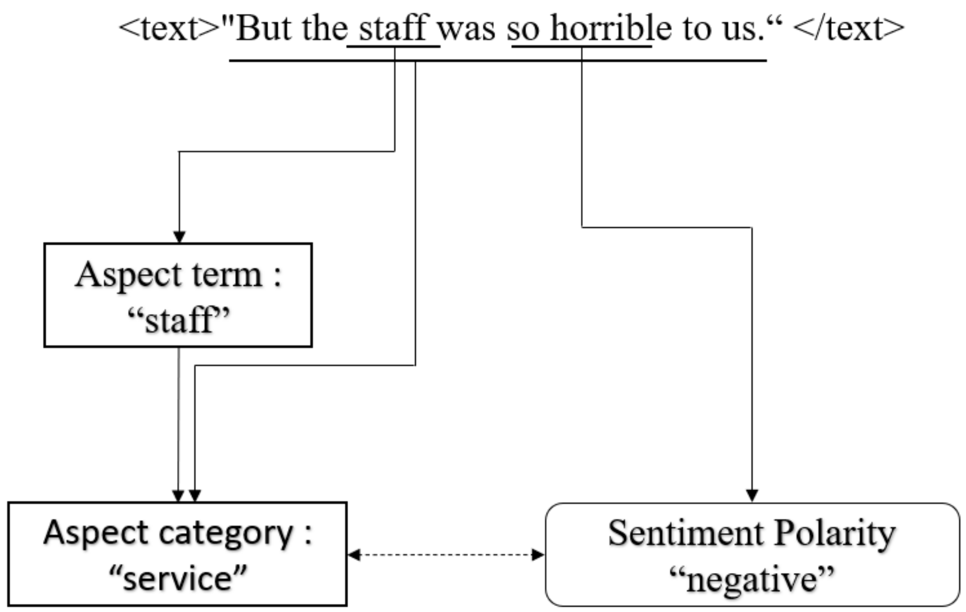

Figure 1.3 The aspect-based analysis subtasks in a sample sentence from SemEval-2014 Restaurant dataset.

In terms of aspects, there are two types, explicit and implicit aspects. In Web reviews, explicit aspects are the phrases within the sentence that are explicit and can be recognised easily. Meanwhile implicit aspects are the implied phrases within the sentence that cannot be identified easily. (Afzaal, Usman, & Fong, 2019; Asghar et al., 2019; Ganganwar & Rajalakshmi, 2019; Gobi & Rathinavelu, 2019; Henríquez et al., 2019; Y. Wang et al., 2019). Explicit aspect is the aspects that are mentioned as nouns or noun phrases in a review/sentence, e.g., “staff” in the review “but the staff was so horrible to us.” Implicit aspect is the aspects that not explicitly mentioned in the sentence, but it is implied in-context e.g., “price” in the sentence “This isn't a value joint.” The implied word is not explicitly mentioned in the review because the implicit aspect can be indicated by the encounter with the whole sentence. Whereas in some sentences, the implicit aspect “inexpensive" is an indicator of the aspect “price” and “weight” are implicit aspect term in the sentence “My HP is very heavy.”

1.2 Motivation

The exponential growth of Social Web services has enlarged the user-generated contents on the World Wide Web. Generating people’s contents, ideas, and opinions have become an interesting topic not only for customers but for industries too. Those opinions give such an excellent opportunity for the industries to capture their customer’s satisfaction regarding their products, and it helps the people to decide on their product selection. For any customer or industry, those generated contents are not machine-processable. To enable those contents to be more practical, aspect-based sentiment analysis models can be proposed to extract the relations between the opinion-targets (aspect-terms) and the polarity values related to them.

For each occurrence of aspect terms in a sentence, the assumption of estimating the polarity of that identified aspect is relatively more straightforward than the assumption that can be made if there are multiple aspects. This suggests that a methodology can be used to group aspect terms into appropriate category (aspect grouping). The designated groups can be used to rate the most important aspects accurately. The other benefit of aspect grouping will be to avoid the problem of considering multi-aspect that refers to the same aspect as different aspects. Besides, online reviews are not always labelled. A class label is absent in online reviews (“text data”) found on public platforms like Amazon or TripAdvisor, which leaves the current methods lacking. The best example for this is the online reviews in SemEval-2014 (Restaurant and Laptop) domain data. Here, the restaurant corpus (figure 1.4) is annotated into five aspect categories, including “price” and “food,” following which the annotated aspect term in the opinionated sentence is further assigned to an aspect topic, that is “food” in this example. Figure (1.5) reveals the lack of aspect topic (“aspect-category”) for the annotated aspect-terms in the laptop corpus (“SemEval-2014”).


```
<sentence id="813">
      <text>All the appetizers and salads were fabulous, the steak was mouth
watering and the pasta was delicious!!!</text>
      <aspectTerms>
         <aspectTerm term="appetizers" polarity="positive" from="8" to="18"/>
         <aspectTerm term="salads" polarity="positive" from="23" to="29"/>
         <aspectTerm term="steak" polarity="positive" from="49" to="54"/>
         <aspectTerm term="pasta" polarity="positive" from="82" to="87"/>
      </aspectTerms>
      <aspectCategories>
         <aspectCategory category="food" polarity="positive"/>
      </aspectCategories>
   </sentence>
```


Figure 1.4 Annotated aspect category in the SemEval-2014 Restaurant dataset.


```
<sentence id="2339">
      <text>I charge it at night and skip taking the cord with me because of the
good battery life.</text>
      <aspectTerms>
         <aspectTerm term="cord" polarity="neutral" from="41" to="45"/>
         <aspectTerm term="battery life" polarity="positive" from="74"
to="86"/>
      </aspectTerms>
   </sentence>
```


Figure 1.5 Unannotated aspect categories in the SemEval-2014 Laptop dataset.

In aspect-based sentiment analysis, implicit aspect identification (or extraction) is the most critical and challenging task (B. Liu, 2017; Mowlaei, et al. 2020; Q. Xu, et al. 2019). It is referring to extract aspect terms that is implicitly expressed within the sentence, rather than explicit terms. The implicit aspects can be implied as a single term within the review that refers to the actual aspect term as in the sentence "it is an expensive food." The term "expensive" is an implied term that refers to the aspect "price." Also, the implicit aspects can be implied in the overall meaning of the sentence as the aspect "price" expressed the price of the food in "the place is a great bargain." The identification relied on the overall meaning of the review.

1.3 Problem Statement

Majority of the presented aspect analysis methods is to conduct annotated aspect categories. This research, however, was conducted to address aspect grouping and aspect category detection using both annotated and non-annotated data, also the extraction of the implicit aspects on online reviews.

The task of aspect grouping is to extract and group the synonym aspects (e.g., price, cost) into similar groups. Almost no attempt was previously presented to address the aspect grouping on unannotated aspect categories except for the clustering algorithms used (e.g., k-means) that was integrated with a feature representation method (e.g., BoWs). It ignores the semantic regularities of the aspect-terms and it is aspect category.

Topic models and dictionary-based methods are two primary methods used for the task of aspect grouping in online reviews. The Bayesian topic models used (e.g., LDA), suffer from two main limitations. They ignore the order of the words as they rely on the frequency-based distributions i.e., word and topic distributions. That in return ignore the semantics of the aspects and do not encounter the less frequent aspects. To cope with those limitations, a common-sense knowledge (e.g., SenticNet3) is integrated into the sampling algorithms of the topic models or by introducing weighting schemes to filter irrelevant aspect terms. Besides, some of the topic models introduced a distribution constraint for aspect grouping (e.g., must-link, cannot-link rules for word to topic distribution) (Z. Chen et al., 2013a; Ekinci & Omurca, 2017; Hai et al., 2017; Khalid et al., 2018; Kiritchenko et al., 2014; Poria, Chaturvedi, et al., 2016; Ray & Chakrabarti, 2019; Santosh et al., 2016; Shams & Baraani-Dastjerdi, 2017; Q. Xu et al., 2020). Those methods seem to be domain-dependent, because they are aimed to suit a particular domain, and required man intervention because they are manually articulated.

The other method for aspect grouping is dictionary-based that needs regular updates to keep up with newly published terminology (Tao & Fang, 2020). However, it is domain-dependent, as aspect-terms are synonyms in one domain but are antonyms in another.

Apart from being the existing aspect grouping methods are domain-dependent for the reasons mentioned above. The presented topic models are not suitable for non-annotated aspect groups (i.e., raw online reviews have no class label), because the parametric models required a specific number of parameters, specifically the number of topics. That made the current topic models deficient toward non-annotated dataset, as in SemEval-2014 Laptop datasets.

The task of detecting the aspect category for each sentence (i.e., aspect category detection) is mostly approached as a sub problem based on the other aspect-based analysis approaches. It was approached using similarity measures conducted to find the aspect category for a sentence based on an extracted aspect using dependency rules and clustering algorithms. Others used the classification methods alongside with natural language processing techniques. A supervised classification method was used for the aspect category detection (Ghadery et al., 2019; Hoang et al., 2019; Y. Li et al., 2019; Movahedi et al., 2019; Tay et al., 2018). Support vector machine was also proposed as a supervised machine learning (Álvarez-López, et al. 2016; Kiritchenko et al., 2014). They are relatively more accurate than the rest of the methods, but they require a labeled datasets for the training of the model, which is laborious.

Implicit aspect extraction is mostly addressed in tandem with the explicit aspects, extracted using topic models, clustering algorithms and co-occurrence matrixes. By considering the opinion words in an opinionated sentence that has no explicit aspects, the opinion word will be mapped to an extracted explicit aspect using other methods, like the rule-based methods for explicit aspect extraction (Benkhelifa et al., 2019; Poria et al., 2015; Poria, Cambria, et al., 2016; Poria, Chaturvedi, et al., 2016; Wan et al., 2016). But on the other hand, the implicit aspects extracted based on hand crafted rules and pre-stated lexicons comprises a set of implicit aspects clues used to identify the implied expressions (Cruz, et al. 2014). Manually constituting rules for the extraction of implicit aspects is a robust way of identifying the implied expressions in a particular domain data, but it is domain dependent as new rules are required for different domains.

Having the above problems, this thesis is looking to address the following research questions:

1. What is the best way to maintain semantic representation in the LDA topic model?
2. How to prepare an unsupervised model to perform aspect-based analysis in non-annotated aspect-categories?

3. How to detect the implied aspects using no manually crafted rules or co-occurrence methods?

1.4 Research Objectives

This research seeks to propose an efficient Bayesian probabilistic model to conduct aspect grouping, and aspect category detection in both annotated and unannotated corpora. Note that, the proposed models are being evaluated using the annotated data. Consequently, the distributed vector is the proposed method for the extraction of implicit aspects relying on the generated aspect categories using Bayesian models. To achieve that, this research is conducted to achieve the following objectives:

i. To propose word embedding that is to be integrated to the conditional distribution and sampling algorithm of LDA model to solve the lagging of semantic representation.

ii. To propose a hierarchical model that has its topic parameter adjustable to variant number of aspects and can fit the non-annotated aspect-categories.

iii. To propose a distributed vector that was trained using the generated aspect categories in Bayesian models for the extraction of implicit aspects.

1.5 Scope

This research is to accomplish ABSA subtasks, aspect grouping, aspect category detection and implicit aspects extraction. The aspect grouping, and aspect category detection are attained to conduct annotated and unannotated categories, and the proposed models for these tasks are being evaluated using annotated categories. Subsequently, the generated aspect categories (group of aspects) in aspect grouping, are used to train a distributed vector for the extraction of implicit aspects. The proposed methodology is meant for the product reviews that were written in the English language only.

1.6 Research Contributions

The contribution of this work is in line to cope with the existing shortcomings in the current methods earlier stated in the problem statement of this chapter. Those were presented for aspect grouping, aspect category detection, and implicit aspect extraction.

The Bayesian probabilistic models (aka topic models) are extendable approaches extensively used throughout the aspect-based analysis subtasks. It is distribution biased using manually crafted seed, and common sense knowledge to maintain the semantic representations. Hence these are domain-dependent mechanism and time inefficient. Therefore, we have developed the TSLDA model that is built on the premises of the LDA model. The developed model maintained the semantic representations using word embedding model integrated into the two components of the model: i) conditional distribution, and ii) sampling algorithm. For the first part we have introduced the word embedding model to generate a topic seed based on the aspect categories in the corpus. These topic seeds are a vector representation of the semantically closest terms for the given aspect category (e.g., “price,” “money,” and “pay” are the topic seeds for the aspect category “price” in TripAdvisor: Hotel dataset). These topic seeds are integrated into the conditional distribution of the model; therefore the conditional distribution is strengthened by the adding the third hyperparameter that is the topic to seeds distribution along with the two original distributions; word to topic distribution, and topic to document distribution. For the sampling algorithm, the word embedding is also being integrated in the collapsed Gibbs sampling algorithm of the model to guid the distribution of the topics that generate more coherent topics than the original sampling algorithm. The word embedding is integrated in the sampling algorithm using similarity measure and threshold score to decide the distribution of the aspect terms in the topics (or categories). The developed model is used to address two subtasks in the aspect-based sentiment analysis: aspect grouping, and aspect category detection. The achievement of the subtasks relied on the word to topic distribution of the model, and that is aspect to groups in our case. The aspect category detection concerning the document to topic distribution, is by considering the documents as the sentence reviews that belong to a single category. The parametric model requires a number of topics i.e., hyper-parameter to be known prior to the modelling (or training).

The central ideation of the proposed non-parametric model (i.e. HDP-CGS) is towards accommodating the level of complexity embedded in the data. The challenge perceived is in carrying out joint mining for the aspect topic (i.e. category) and aspect terms within a single sentence review. This is attributable to it possibly having multiple aspects belonging to either the same or different aspect topic (category). As those reviews are non-annotated, the supervised methods are rendered lacking due to their requirement for large-scale annotated data. Here, the multivariate Gaussian distribution proposed drawing the aspect topic (category) in HDP-CGS model in consideration of the embedding space of Wikipedia word embedding. In the advanced topic model (i.e. HDP) that is the base model for our proposed non-parametric model, there is a need to represent the Dirichlet process (DP). DP is sort of a hierarchical distribution that interpret the two levels of distributions: base level (or Base distribution), and descendent distribution. The base distribution is to draw the number of topics for an infinite number of topics that is performed using the Chinese restaurant process topic representation. For the descendent distribution that given documents sampled using multinomial distribution. The sampling of the proposed model is done via the collapsed Gibbs sampling algorithm (CGS). It is duly introduced to condition the aspects to topics (categories). The whole model (HDP-CGS) is presented to conduct two subtasks in the aspect-based analysis; aspect grouping and aspect category detection on un-annotated data.

Consequently, the generated aspect groups in the introduced Bayesian model (i.e. TSLDA, and HDP-CGS), the extraction of the implicit aspects is attained. The generated aspect groups being encoded into vectors, each vector comprises of the aspect-terms for each category (e.g. "Food" category in the SemEval-2014 Restaurant dataset). The encoded aspects into vectors are fed into Skip-gram model (distributed vectors (DV)) (Mikolov, et al. 2013a). Skip-gram is a neural network representation of input, hidden and projection layers. The identification of the implicit aspects relied on the similarity measure that depends on the trained Skip-gram model. We find the implicit aspects based on the similarity values between the implicit aspect and their related aspect-category.

Thesis contributions:

1) A domain-independent parametric model (i.e., TSLDA) is an unsupervised model which maintains semantic representations presented for aspect grouping, and aspect category detection on annotated aspect categories.

2) A domain-independent non-parametric model (i.e., HDP-CGS) with a leveraged wiki-knowledge presented to conduct various topics for aspect grouping, and aspect category detection in unsupervised categories.

3) A distributed vector model (i.e., Skip-gram) reliant on the generated aspect categories using Bayesian models for the extraction of implicit aspects.

1.7 Thesis Organization

The organization of this work is as follows:

Chapter 1 provides an overview of the granularity of the text in sentiment analysis, and the ABSA subtasks required to understand the differences between the document-level, sentence-level, and aspect-level, and aspect-level worth doing. Besides, we have stated the motivations that directed us to this work. The problem statement is also stated, in which we have briefly stated the state-of-the-art problem of the current methods for aspect grouping, aspect category detection (ACD), and implicit aspects extraction. Correspondingly, we have mentioned the research questions and the objectives of this thesis as well as the scope of the work and research contributions that we developed to cope with the current methods.

Chapter 2 provides a brief background of the probabilistic machine learning and the neural network models. Then a survey was carried out on the current methods for the aspect grouping, ACD, and implicit aspect extraction in online reviews. Topic modelling and the dictionary-based methods were used for the aspect grouping. The topic models were then divided into models relied on the LDA for aspect grouping, and other methods relied on the LDA combined with knowledge-based methods. However, the approaches that were being previously proposed for aspect category detection and implicit aspect extraction were divided into: Unsupervised, Supervised, and Semi-supervised. Following that is the discussion of the presented methods and why they were deficient.

Chapter 3 demonstrates the proposed methodology for the aspect grouping, aspect category detection and the identification of implicit aspects. Topic-seeds Latent Dirichlet Allocation (TSLDA) as a Bayesian probabilistic model proposed for the annotated aspect-categories. While Hierarchical Dirichlet Process-Collapsed Gibbs Sampling (HDP_CGS) proposed for aspect-terms from unannotated corpora. Besides, distributed vectors that have been trained with the generated aspect-categories were used for implicit aspect extraction.

In **Chapter 4,** we have implemented the proposed methodology of TSLDA model. The model is an improved parametric topic model relied on the traditional LDA topic model. Its improvements were to maintain the semantic representations in topics. The model was presented to achieve the aspect grouping and aspect category detection using annotated aspects. The chapter presented the experimental design that is to select the optimal hyperparameter values, and the performance of the model against the standard and the improved sampling algorithm. Then it was compared with the previous models for the aimed subtasks.

In **Chapter 5**, we have implemented the proposed methodology of HDP-CGS model. It is a non-parametric model relied on the advanced HDP topic model. The model is presented to achieve the aspect grouping and aspect category detection using un-annotated data. The chapter investigated the performance of the model for the topic coherence and for the aimed subtasks in aspect-based analysis.

In **Chapter 6**, the extraction of the implicit aspect was elaborated. We have evaluated the performance of the presented Skip-gram model for the extraction of the implicit aspects using the generated aspect topic (category) from the introduced Bayesian models and compared the performance of the model with the current methods for the same task.

Finally, **Chapter 7** summarizes the achievements of the research objectives, research conclusion, and future work.

CHAPTER 2

BACKGROUND AND LITERATURE REVIEW

2.1 Introduction

This chapter examines the literature that serves as the thesis' theoretical foundation and inspiration. It introduces the essential background and fundamental materials on probabilistic models and neural networks. This chapter also covers the current efforts for aspect grouping, aspect category detection, and the extraction of implicit aspects. Finally, it covers a comprehensive discussion and summary of the chapter.

2.2 Background

2.2.1 Probabilistic models

The probabilistic model is a joint distribution of hidden variables z and observed variables x. As in equation 2.1.

$$P(z, x) \tag{2.1}$$

The inference for unknowns is achieved through the posterior, the conditional distribution of the hidden variables given the observations as in the following equation:

$$P(z|x) = \frac{P(z, x)}{P(x)} \tag{2.2}$$

The conditional probability of hidden variable z given the observed variable x equals the joint distribution of hidden variable z given the observed variable x divided by the marginal distribution of the observations. In topic modelling models, the denominator is not tractable. The posterior inference is approximated to resolve the denominator difficulty. In other words, what is the probability of hidden variable z given the observed variable x.

Two main methods were used to approximate the posterior inference. I.e., Monte Carlo MCMC and Variational Inference VI. Further, the MCMC approximates the posterior using the sampling algorithm in Gibbs sampling GS and Collapsed Gibbs Sampling CGS, whereas the VI approximates the posterior inference using the optimization algorithm such as Kullback-Leibler KL divergence.

In the latent Dirichlet allocation, the sampling algorithm is being used to approximate the posterior inference. It relies on the Dirichlet Distribution (DD), which is document-topics mapping. In probability distributions, however, the DD is a continuous probability distribution (or categorical distribution) that gives occurrence of different possible outcomes — parameterized by α of positive reals, represented by $Dir(\alpha)$. Also, it is a multivariate probability distribution of a list of vectors with unknown values because either the values have not occurred yet or insufficient knowledge about the value (Strawderman, 2001).

However, LDA is a generative probabilistic model for collecting distinct observable pieces of data such as text corpora. LDA has performed statistical evaluations on word co-occurrence to mine and group popular words in topics. Topics are considered concepts that are used to represent their top words. In sentiment analysis and particularly aspect-based analysis, topics are used to represent the sentiment category or product aspect.

The standard LDA presented in Blei et al. (2003) emphasizes the assumption of exchangeability. In that, the random variables are independent and identically distributed. Concerning the latent parameters are conditionally distributed. However, the distribution of the tokens in the corpus relies on the bag-of-words. For a word in the essential collection of discrete data, defined to be a token from a collection of vocabulary $\{1, \ldots., V\}$. The word is represented using a unit-basis vector. One for the current word in the vector and zero for all the other components. So that a $V-vector$ represent vth word w in the vocabulary. $w^v = 1$ for the current word and $w^u = 0$ for the all the others, thus $u \neq v$.

For the given sequence of N words within a document denoted by $w = (w_1, w_2, \ldots\ldots, w_N)$, and the n^{th} word in the sequence denoted by w_n. However, a corpus is a collection of M documents denoted by $D = \{w_{1,} w_{2,} \ldots\ldots\ldots, w_M\}$.

To infer the number of topics in a corpus, the generative process for the LDA is represented in the algorithm (2.1).

Figure (2.1) shows the plate notation for the standard LDA. The boxes were representing replicates or repeated entities. For instance, the document is represented with the outer plate, words positioned in the inner plate in the given document, and each of which is associated with the topic. However, the number of documents denoted by M, and N are the number of words in the document.

Var

α is the parameter of the latent Dirichlet distribution per-document topic-distributions.

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

θ_i is the topic distribution for each document i.

φ_k is the word distribution for each topic k.

$z_{i,j}$ is the word distribution for the $j-th$ word in a document i,

$w_{i,j}$ is the particular word (only observed component).

Choose $\boldsymbol{\theta_i} \sim \boldsymbol{Dir(\alpha)}$, where $\boldsymbol{i} \in \{\mathbf{1}, \ldots \ldots, \boldsymbol{M}\}$, Note: $\boldsymbol{Dir(\alpha)}$ is Dirichlet distribution

Choose $\boldsymbol{\varphi_k} \sim \boldsymbol{Dir(\beta), where\ k} \in \{\mathbf{1}, \ldots \boldsymbol{K}\}$ and $\boldsymbol{\beta}$ typically is sparse

For each of the word positions $\boldsymbol{i, j, where\ \ i} \in \{\mathbf{1}, \ldots \ldots, \boldsymbol{M}\}, \boldsymbol{and\ j} \in \{\mathbf{1}, \ldots, \boldsymbol{N_i}\}$

Choose a topic $\boldsymbol{z_{i,j}} \sim \boldsymbol{Multi(\theta_i)}$.

Choose a word $\boldsymbol{w_{i,j}} \sim \boldsymbol{Multi}\left(\boldsymbol{\varphi_{z_{i,j}}}\right)$.

Note: $\boldsymbol{Multi\ is\ the\ Multinomial\ Dirstribution}$

Algorithm 2.1 LDA generative process[2]

[2] https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

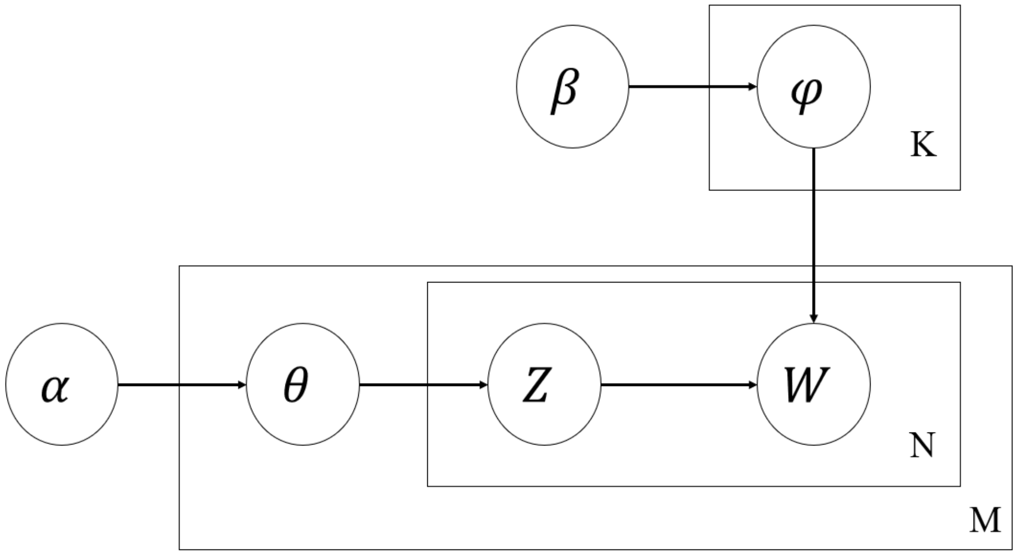

Figure 2.1 LDA graphical model[3]

The latent topics' discovery is accomplished using the multinomial distribution, while the word to topic distribution is intractable to be accomplished using the multinomial distribution. Therefore, to address the word to topic distribution, the Gibbs sampling algorithm being used.

In LDA, the number of topics is specified before the sampling of the words or observed tokens. In LDA, the number of topics is specified in advance. This kind of allocation lacks when it comes to the corpora without the class label. To this issue, non-parametric models emerged. HDP (Teh, et al., 2006) is a non-parametric model that relied on the Dirichlet Process rather than on the DD as in LDA. Practically, DP is a probability distribution whose range is itself a set of probability distributions. The probability distribution of the random variables of one another has relied on the DP.

[3] https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation#/media/File:Smoothed_LDA.png

In the non-parametric model like HDP, the DP is specified by the base distribution H and a hyperparameter α that is concentration parameter. The DP draws distributions around the base distribution much similar to the distribution of the normal distribution draws the real numbers around the means.

Practically, in HDP, three different perspectives; one based on the 'stick-breaking construction,' one based on 'Po'lya urn model or Chinese Restaurant Process (CRP),' and one based on a 'limit of finite mixture model.' Also, HDP is composed of multiple levels of DP. Each of these mentioned perspectives has a specific interpretation in HDP.

Dirichlet Process (DP) has been first proposed to generate the number of topics automatically. Base distribution G, which is called the concentration parameter used to specify the Dirichlet process.

The intuition behind the CRP is that the restaurant with an infinite number of tables (topics), and customers walk in to choose a table to sit at (each customer is aspect-word). The tables are chosen according to the concentration parameter because if there is no limit for the chosen tables, the model will never converge..

2.2.2 Neural networks

As the word embedding models "distributed vectors" are neural network approach to learn-quality of the embedded words, we have explained the simple artificial neuron (unit), in which there are input values $\{x_1, \ldots, x_k\}$, for each input, there is a unique weight value $\{w_1, \ldots, w_k\}$; the inputs are mapped into a scalar output y. Then, the activation function f simulating the neuron, whether it will fire or not, figure (2.2) shows a simple example of a neural network.

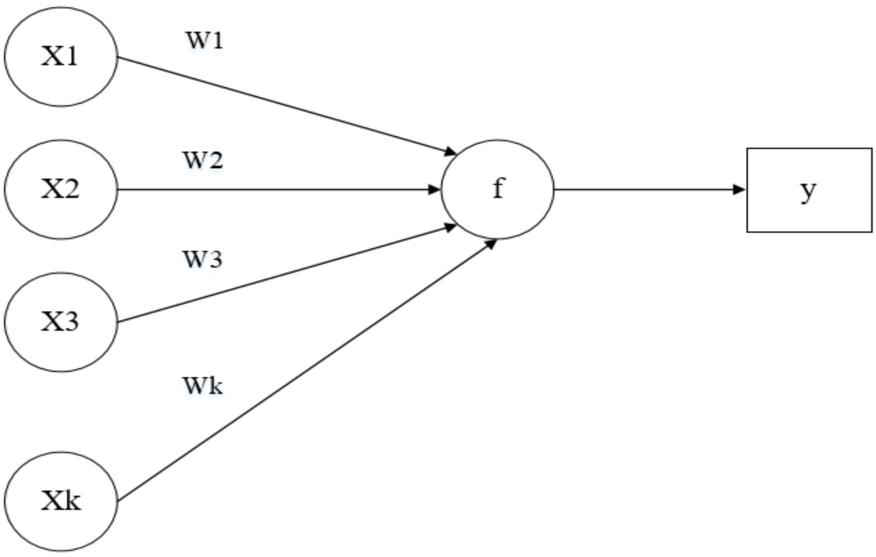

Figure 2.2 An artificial neural network[4]

The neural network can be articulated as follows:

$$y = f(u), \tag{2.3}$$

Where u is a scalar number, which is the net input of the neuron. u is defined as follows:

$$u = \sum_{i=0}^{K} w_i x_i. \tag{2.4}$$

Using vector notation, we can write

$$u = W^T X \tag{2.5}$$

[4] https://www.computing.dcu.ie/~humphrys/Notes/Neural/single.neural.html