

**A NOVEL ASPECT TAXONOMY AND ASPECT
EXTRACTION METHODOLOGY FOR SCHOLARLY
BOOK REVIEWS**

Wickramarathna Wengappuli Arachchige Chathur Sajeewan Basuru

(158207C)

Degree of Master of Science

Department of Computer Science & Engineering

University of Moratuwa

Sri Lanka

March 2019

**A NOVEL ASPECT TAXONOMY AND ASPECT
EXTRACTION METHODOLOGY FOR SCHOLARLY
BOOK REVIEWS**

Wickramarathna Wengappuli Arachchige Chathur Sajeewan Basuru

(158207C)

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree
Master of Science in Computer Science and Engineering

Department of Computer Science & Engineering

University of Moratuwa

Sri Lanka

March 2019

DECLARATION

“I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:

Date:

Name: W.W.A.C.S. Basuru

The above candidate has carried out research for the master’s dissertation under my supervision.

Signature of the supervisor:

Date:

Name of the supervisor: Dr. Surangika Ranathunga

Abstract

Many people decide on the quality of a product based on its online reviews, which is also the most commonly used method when purchasing books from online book stores. Compared to other products, a scholarly book is one of the most difficult products to purchase online since customers have limited access to its internal content. Therefore, a customer has to go through multiple reviews in order to get insight on the book. However, the sheer volume of online reviews makes it difficult for a human to process and extract all the meaningful information in order to make an educated purchase. As a result, a requirement for a sentiment analysis system for scholarly book reviews are much needed at this stage. A more accurate opinion of the book can be obtained through aspect-based summarization. This type of summarization of opinions is critical for scholarly book reviews since content, organization, and other features interpret whether the book can be recommended to a customer at a certain education level.

Compared to sentiment analysis on reviews of products/services such as movies or restaurants, there is no well-defined research in aspect extraction or aspect-based sentiment analysis of scholarly book reviews. Not surprisingly for this domain, there is no well-defined aspect taxonomy or an annotated dataset available to extract aspects or to identify aspect categories. Compared to other domains, identifying aspects of book reviews is difficult since aspects such as the quality of the book or the discussed topics always appear implicitly in reviews.

The main contribution of this research is to identify potential aspects and an aspect taxonomy for scholarly book reviews. We also present a (1.) dependency rule-based unsupervised model for aspect extraction, which works better than state-of-the-art unsupervised methods, and (2.) a clustering-based aspect category identification method. Both of these are important first steps for aspect-based sentiment analysis.

The aspect taxonomy for scholarly book reviews is a hierarchical model. Book and Author have been identified as the first level of the taxonomy. Readability, content, worthiness and price, are the next level of aspect taxonomy under the book aspect category. Author expertise has been identified as an aspect category under author. In order to validate the aspect taxonomy, an unsupervised aspect extraction and clustering algorithm is proposed. An existing dependency rule-based aspect extraction algorithm is improved by adding new rules that extract aspects from book reviews. Two existing clustering algorithms for aspect clustering are merged to obtain a new clustering algorithm to discover the categories of aspect terms. The clustering algorithm is able to find the semantic similarity of aspect terms, while considering the sharing words between aspect terms, and groups similar aspects in to a one cluster. After successfully generating an annotated corpus for the scholarly book reviews in the computer science domain with Cohen's kappa statistics of 0.76, the dependency rule-based aspect extractor was able to extract both implicit and explicit aspects with precision 76.04%, recall 75.99% and overall F1-score 76.02%. The proposed semantic similarity based aspect clustering algorithm identifies the aspect in the following categories; book, author, readability, content, worthiness, price and author expertise with rand-index 14.41%, V-measure 36.29%, homogeneity 66.18% and completeness 25%.

Keywords: Aspect based sentiment analysis, Dependency rules, Aspect taxonomy, Clustering, Semantic similarity, Stanford dependency parser, GloVe

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisors Dr. Surangika Ranathunga for the continuous support given for the success of this research. This would not have been a success without your tremendous mentorship and advice from the beginning.

I would like to thank all staff from the Department of Computer Science and Engineering for their kindness expressed in all occasions. Special thanks to my loving family: my parents, wife, and sisters. Your encouragement always motivated me to do my best. I would also like to thank all my friends who encouraged me all times.

TABLE OF CONTENTS

DECLARATION	i
Abstract	ii
ACKNOWLEDGEMENT	iii
TABLE OF CONTENTS	iv
LIST OF FIGURES	vii
LIST OF TABLES	viii
LIST OF ALGORITHMS	viii
LIST OF ABBREVIATIONS	ix
1 INTRODUCTION	1
1.1 Background	1
1.2 Problem and Motivation	2
1.3 Research Objective	3
1.4 Research Contribution	3
1.5 Structure of Thesis	4
2 LITERATURE SURVEY	5
2.1 Sentiment Analysis	5
2.1.1 Document Level Classification	5
2.1.2 Sentence Level Classification	6
2.1.3 Aspect Level Classification	6
2.2 Defining Aspect Categories and Taxonomy creation	7
2.3 Aspect Extraction and Categorization	7
2.3.1 Unsupervised Methods	8
2.3.1.1 Rule-based methods	8
2.3.1.2 Clustering based methods	9
2.3.1.3 Frequency or statistical methods	10

2.3.1.4	Bootstrapping methods.....	10
2.3.2	Supervised Methods.....	11
2.3.2.1	Conditional Random Field (CRF) based Approach.....	11
2.3.2.2	Maximum Entropy Based Approaches	13
2.3.2.3	HMM based Approaches.....	13
2.3.2.4	Tree based Approaches	14
2.3.2.5	Support Vector Machine based approaches	14
2.3.2.6	Limitations of supervised learning approaches	15
2.3.3	Deep learning for aspect extraction	15
2.3.4	Semi-Supervised Methods	16
2.3.4.1	Dependency parser based methods	16
2.3.4.2	Lexicon based methods	16
2.3.4.3	Graph based methods	16
2.3.5	Discussion.....	17
2.4	Current trends towards the sentiment analysis of book reviews	18
2.4.1	Association rule based techniques	19
2.4.2	Collaborative filtering based techniques.....	19
2.4.3	Web mining techniques.....	20
2.4.4	Opinion mining and Sentiment analysis at document level.....	20
2.4.5	Sentiment analysis at aspect level.....	22
2.5	Aspect categories for Scholarly Book reviews	22
2.6	Data Sources and Preprocessing.....	25
2.7	Vector representation of words	27
2.8	Summary	28
3	METHODOLOGY	30
3.1	Data extraction and preparation	30

3.2	Aspect taxonomy for scholarly Book reviews	31
3.3	Scholarly book review corpus annotation	33
3.4	End to end system architecture of Aspect extraction and clustering.....	36
3.5	Dependency rule-based aspect extraction	38
3.5.1	IAC Lexicon.....	38
3.5.2	Dependency rules proposed by Poria et al. [1]	38
3.5.3	Proposed new rules	42
3.6	Clustering-based aspect categorization	43
3.6.1	Vector representation of aspects	44
3.6.2	Similarity measure calculation.....	45
3.6.3	Merging constraints	47
3.6.4	Clustering Algorithm	47
3.6.5	Novelty of the proposed algorithm	49
4	EXPERIMENTAL SETUP AND RESULTS	50
4.1	Evaluations on Aspect extraction	50
4.2	Evaluations on discovering aspect categories	54
5	DISCUSSION.....	61
5.1	Aspect taxonomy	61
5.2	Aspect extraction.....	62
5.3	Aspect clustering	62
6	CONCLUSION AND FUTURE WORK.....	64
	REFERENCES	65

LIST OF FIGURES

Figure 2.1 : Abstract architecture of Aspect based sentiment classifier [3]	7
Figure 2.2 : Sample review [47] [48].....	26
Figure 3.1 : Distribution of star rating in AI and Machine learning amazon book reviews	31
Figure 3.2 : Hierarchy of Aspects for scholarly book reviews	32
Figure 3.3 : Distribution of test data (100 reviews)	33
Figure 3.4 : Annotated review for identifying sentence phrases containing aspects and opinions using Brat annotation tool	34
Figure 3.5 : Aspect annotation process	35
Figure 3.6 : Aspect distribution over each aspect category (total aspects=1453).....	36
Figure 3.7 : Pipeline for extracting explicit and implicit aspects from Scholarly book reviews.	37
Figure 3.8: Semantic similarity matrix	45
Figure 4.1 : Rand Index of aspect clustering for different distance thresholds (μ) ...	55
Figure 4.2 : V-measure of aspect clustering for different distance thresholds (μ)	56
Figure 4.3 : Homogeneity of aspect clustering for different distance thresholds (μ)	56
Figure 4.4: Completeness score of aspect clustering for different distance thresholds ($\mu = 0.03$)	56
Figure 4.5 : Cluster distribution for different distance thresholds	57
Figure 4.6 : Precision-Recall curve for different cluster percentages ($\mu = 0.03$).....	58

LIST OF TABLES

Table 2.1 : Appraisal documentation provided by Lee et al. [46]	24
Table 2.2 : JSON Field description [47] [48]	26
Table 3.1 : Statistics about sentences with aspects	36
Table 4.1 : Aspect extraction performance	50
Table 4.2 : Extracted explicit and implicit aspects	51
Table 4.3: Dependency rule relation with extracted aspects.....	52
Table 4.4: Number of aspects extracted by each dependency rule	53
Table 4.5: Aspect clustering performance for different distance threshold values (DT-distance threshold, N- no of clusters, RI-rand Index, Ho-homogeneity, Co- completeness, V- V measure)	57
Table 4.6 : Quality of dependency rule-based aspects after aspect clustering.....	58
Table 4.7: Aspect extraction and categorization performance after selecting top 60% of the generated clusters ($\mu = 0.03$)	59
Table 4.8: Example aspect clusters generated from the clustering algorithm ($\mu =$ 0.03, top 60% clusters).....	59

LIST OF ALGORITHMS

Algorithm 3.1 : Clustering for Aspect Discovery	48
---	----

LIST OF ABBREVIATIONS

ABSA	Aspect Based Sentiment Analysis
CINAHAL	Cumulative Index to Nursing and Allied Health Literature
CNN	Convolutional Neural Network
CRF	Conditional Random Field
GFST	Generalized Aspect Sentiment Tree
HMM	Hidden Markov Model
IAC	Implicit Aspect Clue
KNN	K-Nearest Neighbors
LDA	Latent Dirichlet Allocation
MEDLINE	Medical Literature Analysis and Retrieval System Online
MLE	Maximum Likelihood Estimation
MV-RNN	Matrix Vector Recurrent Neural Network
NER	Named Entity Recognition
NLP	Natural Language Processing
NPMI	Normalized Pointwise Mutual Information
OWA	Ordered Weighted Averaging
PAS	Positional Aggregation based Scoring
POS	Part of Speech
RNN	Recurrent Neural Network
RNTN	Recursive Neural Tensor Network

SDM	Sequential Dependence Model
SemEval	International Workshop on Semantic Evaluation
SSWE	Sentiment Specific Word Embedding