Hybrid Models for Aspects Extraction without Labelled Dataset

Wai-Howe Khong
Faculty of Computing and Informatics
Multimedia University, Malaysia.
swordmasterex@hotmail.com

Lay-Ki Soon
School of Information Technology
Monash University Malaysia
soon.layki@monash.edu

Hui-Ngo Goh
Faculty of Computing and Informatics
Multimedia University, Malaysia.
hngoh@mmu.edu.my

Abstract

One of the important tasks in opinion mining is to extract aspects of the opinion target. Aspects are features or characteristics of the opinion target that are being reviewed, which can be categorised into explicit and implicit aspects. Extracting aspects from opinions is essential in order to ensure accurate information about certain attributes of an opinion target is retrieved. For instance, a professional camera receives a positive feedback in terms of its functionalities in a review, but its overly high price receives negative feedback. Most of the existing solutions focus on explicit aspects. However, sentences in reviews normally do not state the aspects explicitly. In this research, two hybrid models are proposed to identify and extract both explicit and implicit aspects, namely TDM-DC and TDM-TED. The proposed models combine topic modelling and dictionary-based approach. The models are unsupervised as they do not require any labelled dataset. The experimental results show that TDM-DC achieves F_1-measure of 58.70%, where it outperforms both the baseline topic model and dictionary-based approach. In comparison to other existing unsupervised techniques, the proposed models are able to achieve higher F_1-measure by approximately 3%. Although the supervised techniques perform slightly better, the proposed models are domain-independent, and hence more versatile.

1 Introduction

Opinion holds positive or negative view, attitude, emotion or appraisal on entity. An entity can be a product, person, event, organization, or topic. Aspect, also known as feature, is the various distinctive attributes on the entity itself (Liu, 2010, 2012). For example, for a product review on mobile phone, the mobile phone is the entity and its aspects may include battery life, design, screen size, and charging time. Being able to identify the specific aspects of an opinion target is crucial as it gives more accurate analysis of the opinion. Aspects can be categorised into explicit and implicit aspects. Explicit aspect is explicitly stated in the review while the latter is not. For instance, given this review *"This is an affordable smartphone with a very long battery life."*, *battery life* is explicitly stated with the associated opinion but the aspect of *price* is implicitly denoted by *affordable*.

Most of the existing research works focus on explicit aspects identification and extraction (Hu and Liu, 2004). Few models have been proposed to identify implicit aspect from the dataset using supervised or semi-supervised approaches (Fei et al., 2012; Wang et al., 2013; Xu et al., 2015). Supervised approaches requires annotated training dataset, which is laborious to label. Furthermore, models produced by supervised model are domain-dependent. Supervised models need to be trained with domain-specific dataset. To the best of our knowledge, unsupervised approach has yet to be proposed to identify both explicit and implicit aspects. Hence, in this research work, the main objective is to propose unsupervised models, which are domain-independent, and able to extract both explicit and implicit aspects, without using any labelled training dataset.

The remainder of this paper is organised as follows: Section 2 discusses some relevant related works. Section 3 presents the proposed models. The experimental setup and results are discussed in Section 4. Finally, the paper is concluded in Section 5.

2 Related Work

Aspect extraction for opinion mining has three main approaches, namely the supervised, semi-supervised and unsupervised approach. Models

Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), pages 63–68
Hong Kong, November 3, 2019. ©2019 Association for Computational Linguistics

from supervised approach are trained using annotated corpus. The resultant models are normally domain-dependent. In other words, a supervised model trained in one domain often performs poorly in another domain. An example of supervised approach uses Lexicalized Hidden Markov Models (HMM) to learn patterns to extract aspects and opinion expressions through part-of-speech and surrounding contextual clues in the document (Jin et al., 2009). Jakob and Gurevych used Conditional Random Fields (CRF) to train review sentences from different domains for domain independent extraction (Jakob and Gurevych, 2010). Toh and Su trained their Sigmoidal Feedforward Neural Network (FNN) with one hidden layer with a training set to predict the aspect categories (Toh and Su, 2015). Repaka, Palleira et al. used Linear Support Vector Machine (SVM) model with Bag-of-Words (BoW) as features and trained it using the multi-class classification method (Repaka et al., 2015). Table 1 summarises the techniques used, whether it extracts implicit aspect or otherwise, and their limitations.

3 Proposed Models

In this research, two domain-independent models are proposed to identify and extract both explicit and implicit aspects. The proposed models are topic dictionary model - direct combine (TDM-DC) and topic dictionary model - topic extended with dictionary model (TDM-TED). Both TDM-DC and TDM-TED combine topic modelling and dictionary-based approach to identify the aspects from a given corpus. Every review is segmented into sentences. Each sentence is used as a document. Part-of-speech (POS) tagged documents are used for topic modelling. Stop words and unused part-of-speech (for example, determiner and conjunction) are filtered from the POS-tagged documents. For dictionary-based approach, noun and opinionated word pairings are extracted as candidate aspects, which is notated as $< N, Ow >$, where N represents the noun extracted from the dataset and Ow represents the opinionated word associated with the noun N in the corpus. Nouns are identified using TreeTagger. Opinionated words are identified using word sense disambiguation (WSD) and sentiment tagging. Sentiment tags are obtained using SentiWordNet (Baccianella et al., 2010). The pairings are identified through pairing noun and opinionated words in the same sentence segment. $< N, Ow >$ notations are then extended using different sets of semantic relations from the dictionary. Extended noun list consists of hypernym and hyponym of nouns, while extended opinion list consists synonym and antonym of the opinionated words. The extended lists are constructed to enlarge the pool of nouns and opinionated words, which in return increases the coverage of aspect candidates. Eventually, each model generates a ranked list of candidate aspects. The details of TDM-DC and TDM-TED are presented in the following subsections.

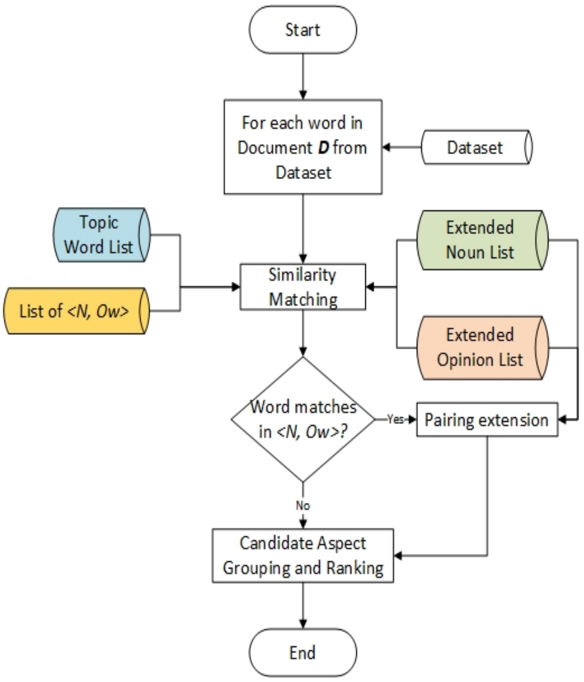

Figure 1: The process of TDM-DC grouping and identifying candidate aspects.

3.1 Topic Dictionary Model - Direct Combine (TDM-DC)

TDM-DC is a direct search and match of words from the given dataset to the words generated from both models. As shown in Figure 1, every word from the document will be matched with the words in four generated lists, which are the topic word list, $< N, Ow >$ notation list, extended noun and extended opinion lists. For topic model, it will find a match of word *w1* from document D in topic model T, if a match is found in topic *T1*, every word in topic *T1* will be extracted and labelled with the same aspect as *T1*. For dictionary model, it will find a match of word *w1* from document D in $< N, Ow >$ notation list P. If *w1* is matched

Table 1: Summary of literature review.

Models	Approach	Explicit Aspect	Implicit Aspect	Limitation
HMMs (Jin et al., 2009)	supervised	Yes	No	laborious data pre-processing step
CRF (Jakob and Gurevych, 2010)	supervised	Yes	Yes	dependant on labelled data
Dictionary Based	supervised	Yes	Yes	highly dependent on dictionary definitions
FNN (Toh and Su, 2015)	supervised	Yes	No	requires a variety of features
Linear SVM (Repaka et al., 2015)	supervised	Yes	No	does not work on sentences without noun
Double Propagation (Hu and Liu, 2004)	semi-supervised	Yes	Yes	only identify adjectives
PSWAM (Liu et al., 2013, 2015)	semi-supervised	Yes	No	does not identify implicit aspects
Topic Model (Blei et al., 2003)	unsupervised	Yes	No	groups unrelated words together
Word2Vec (Mikolov et al., 2013) (Pablos et al., 2015)	unsupervised	Yes	No	require representative seed words

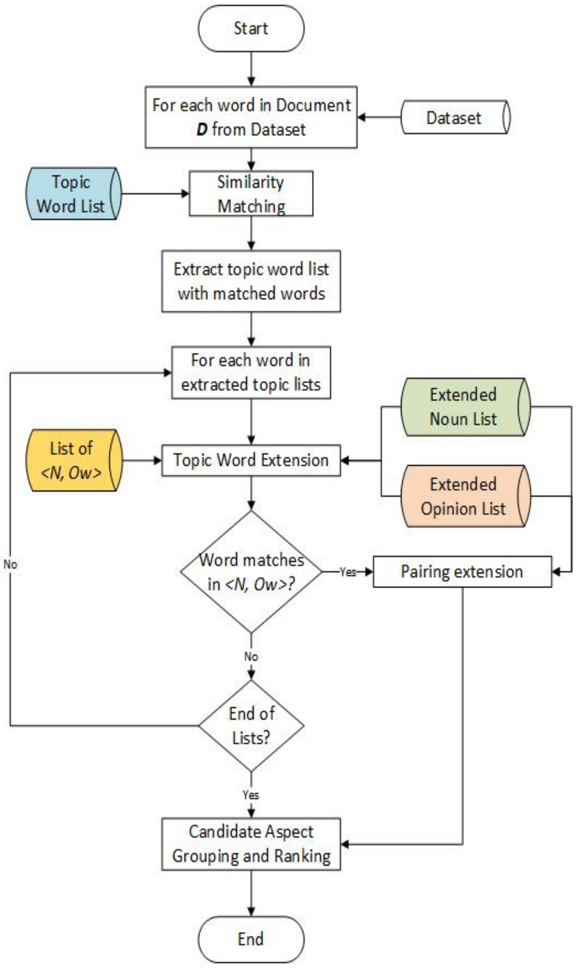

Figure 2: The process of TDM-TED grouping and identifying candidate aspects.

with a notation, all its extended noun *Pn* and opinion *Po* words will be labelled with the same aspect, as its parent $< N, Ow >$ notation *P1*. It will also search from the extended list, *Pn* and *Po* and extract all the matched words. To reduce duplicate entry of the same word (same word, with same aspect and same POS), duplicates will be eliminated from the final list, after aggregating all the candidate words from all models. TDM-DC ranks candidate aspect list as follow:

1. If a word from the document is matched with a word from the topic model, extract the candidate aspect of the topic model and add a count equivalent to the number of words in the topic.

2. If a word from the document is matched with a word from the $< N, Ow >$ notation, extract the candidate aspect of the notation and add two counts to the candidate aspect because there are two words in the pair.

3. If a duplicate match is found in both $< N, Ow >$ notation list and extended list, it will not add to the count for the candidate aspect.

4. With the parent $< N, Ow >$ notation from the matched notation, add the count for every words matching the parent notation in the extended word lists. Duplicates are excluded in the process.

5. If there is a match in the extended list, extract the word's parent candidate aspect, and add one count to it.

6. Aggregate all the matched candidate aspect count in a ranked list of candidate aspects identified for the provided document.

3.2 Topic Dictionary Model - Topic-Extended (TDM-TED)

Similar to TDM-DC, this proposed model, as illustrated in Figure 2 will search and match the similar words in a document. TDM-TED is different from TDM-DC where it will directly search for similar words in the topic word list and indirectly on other word lists. If a word in document *D*, *w1* is matched in a single topic *T1*, all its words will be extracted from the topic. Then, for every word list in the topic *T*, it will search for its matched word in the $< N, Ow >$ notation list *P*, together with its extended words from both Extended Noun *Pn* and Extended Opinion *Po* list. Furthermore, for every words in Topic *T*, it will also directly search for its match in both Extended Noun *Pn*, and Extended Opinion *Po* lists. Finally, similar to TDM-DC, duplicate entries will be removed from the aggregated list of words. TDM-TED ranks candidate aspect list as follow:

1. If a word from the document matched with a word from the topic model, extract the candidate aspect of the topic model and add a count equivalent to the number of words in the topic. Extract the list of words in that topic.

2. For every words in the topic, if there is a match from the $< N, Ow >$ notation, extract the candidate aspect of the notation and add two counts to the candidate aspect for every candidate notation found because there are two words in the pair.

3. If a duplicate match is found in both $< N, Ow >$ notation list and extended list, it will not add to the count for the candidate aspect.

4. With the parent word from $< N, Ow >$ notation, add the count for every words matching the parent notation in the extended list to the parent's candidate aspect. Each candidate aspect extracted from the candidate extended word lists will add a count. Duplicates are excluded in the process.

5. For every words in the topic, if there is a match in the extended list, extract the word's parent candidate aspect, and add one count to it.

6. Aggregate all the matched candidate aspect count in a ranked list of candidate aspects identified for the provided document.

4 Experimental Design

The dataset used for this experimentation was downloaded from SemEval-2015 Task 12: Aspect Based Sentiment Analysis [1]. It contains multiple complete reviews breakdown into pre-labelled sentences with potentially out of context sentences about Restaurants. Their aspect category contains both entity labels (e.g. Restaurant, Service, Food) and attribute labels (e.g. prices, quality). To evaluate against other existing models, the entity and attributes of the entity are notated together to form the aspect tuple for the restaurant dataset. Data pre-processing steps have been implemented on the dataset prior to constructing the models, which include POS-tagging using TreeTagger and word sense disambiguation (WSD). Sentiment tagging is subsequently carried out to assign sentiment tags to every word based on SentiWordNet (Baccianella et al., 2010). For the weighted sentiment on SentiWordNet, sentiment with the largest weight and with the matched POS attached to the word sense were taken into account. For example, given a row in SentiWordNet of $< a, 0.5, 0.125, living\#3 >$, *a* is the part-of-speech of the word (living), *0.5* is the positive weight and *0.125* is the negative weight and *living#3* is the word sense. Since *0.5* is more than *0.125*, the word will be considered as positive. In case of same weight, it will be tagged as neutral. For example, *living#a#3* will be tagged as *living#p* where*p* represents the positive sentiment for the word *living* in that sentence.

For words that are not included in SentiWordNet, they were checked against a compiled list of opinion lexicon (Hu and Liu, 2004) to determine the sentiment polarity of a word. The words were then tagged as *p, g* or *n* respectively, where *p* represent positive, *g* represents negative and *n* represents neutral. As sentiment tagging assigns sentiment on a word-by-word basis, a sentence with negation (e.g. no, not, never etc.) will give the opposite sentiment instead. To solve this, the sentiment of opinionated words are flipped if a negation

[1] http://alt.qcri.org/semeval2015/task12/

word is detected in the sentence. Once data preprocessing is completed, LDA was implemented. Baseline LDA model was chosen because it outperforms complex models of LDA when there is more than two hundred reviews (Moghaddam and Ester, 2012). Complex models of LDA include topic models which are built using phrases or grammatical dependencies (Moghaddam and Ester, 2012). The topics in the resultant model come from the labels provided by the dataset. In other words, the number of topics are set based on the number of labels from the dataset. For the dictionary model, WordNet [2] and Wordnik [3] are used to extract words in the selected semantic relations.

5 Results and Discussion

Precision, recall and F_1-measure are used to evaluate the experimental results. Due to space limitation, only F_1-measure is presented in this paper. Table 2 shows that among the four models, which include two baseline models and two proposed models, TDM-DC has the highest score. The performance of TDM-DC and TDM-TED are very close, with TDM-DC leading on all three columns of comparison. This is unexpected as TDM-TED generates more candidate aspects compared to TDM-DC. Dictionary-based approach has the lowest score. Dictionary-based approach is good in generating a vast amount of candidate aspects using semantic relations. However, if the defined relations are lacking or none to be found, it will highly affect the candidate aspect count.

Table 2: Baseline model and topic dictionary model F_1 comparison by percentage

Model	F_1	Explicit F_1	Implicit F_1
Topic Model	55.80	57.47	32.53
Dictionary Model	54.67	56.56	21.71
TDM-DC	**58.70**	**60.51**	**33.15**
TDM-TED	58.34	60.15	32.34

The performance of TDM-DC and TDM-TED are also compared against other existing models based on F_1-measure obtained in identifying implicit and explicit aspects, as presented in Table 3. NLANGP represents Sigmoidal Feedforward Network (FNN) with one hidden layer implemented by Toh et. al. (Toh and Su, 2015). It is the best supervised approach for both datasets. UMDuluthC uses Linear Support Vector Machine (SVM) Model for both dataset (2015). Finally, V3 uses Word2Vec to identify the aspect from both dataset, which is the only unsupervised approach used on this dataset (Pablos et al., 2015). By comparing with the baseline approach, which is a Support Vector Machines (SVM) with a trained linear kernel (Pontiki et al., 2015), most approaches outperform it excluding V3. For TDM-DC and TDM-TED, both proposed models are able to outperform UMDuluthC by a small margin, but lost to NLANGP model; both constrained (using only the provided training set of the corresponding domain) and unconstrained approaches. Baseline, NLANGP and UMDuluthC run on supervised classification, which require labelled datasets, while TDM-DC and TDM-TED use unsupervised approach. Among the unsupervised approaches, the proposed TDM-DC and TDM-TED outperform V3 by more than 10%.

Table 3: F_1 comparison by percentage with other approaches. * indicate unconstrained systems.

Approach	Model	F_1
Supervised	NLANGP	$62.68^\star$
Supervised	NLANGP	61.94
Unsupervised	**TDM-DC**	**58.70**
Unsupervised	**TDM-TED**	**58.34**
Unsupervised	Topic Model	55.80
Unsupervised	Dictionary Model	54.67
Supervised	UMDuluthC	57.19
Unsupervised	V3	$41.85^\star$
Supervised	Baseline	51.32

6 Conclusion

The main strength of the proposed models is its ability in identifying both explicit and implicit aspects without any labelled dataset. Although the result is not the best when compared to state-of-the-art supervised approach, it is a huge step forward for unsupervised approach in identifying both explicit and implicit aspect. In future, the proposed models will be experimented on opinions that have more implicit aspects to verify its effectiveness at a greater measure.

Acknowledgements

This work is funded by Fundamental Research Grant Scheme (FRGS) by Malaysia Ministry of Higher Education (Ref: FRGS/2/2013/ICT07/MMU/02/2).

[2] https://wordnet.princeton.edu/
[3] https://www.wordnik.com/

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Geli Fei, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2012. A dictionary-based approach to identifying aspects im-plied by adjectives for opinion mining. In *24th international conference on computational linguistics*, page 309.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Niklas Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1035–1045. Association for Computational Linguistics.

Wei Jin, Hung Hay Ho, and Rohini K Srihari. 2009. A novel lexicalized hmm-based learning framework for web opinion mining. In *Proceedings of the 26th annual international conference on machine learning*, pages 465–472. Citeseer.

Bing Liu. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:627–666.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Kang Liu, Heng Li Xu, Yang Liu, and Jun Zhao. 2013. Opinion target extraction using partially-supervised word alignment model. In *IJCAI*, volume 13, pages 2134–2140.

Kang Liu, Liheng Xu, and Jun Zhao. 2015. Co-extracting opinion targets and opinion words from online reviews based on the word alignment model. *IEEE Transactions on knowledge and data engineering*, 27(3):636–650.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Samaneh Moghaddam and Martin Ester. 2012. On the design of lda models for aspect-based opinion mining. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 803–812. ACM.

Aitor García Pablos, Montse Cuadros, and German Rigau. 2015. V3: Unsupervised aspect based sentiment analysis for semeval2015 task 12. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 714–718.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495.

Ravikanth Repaka, Ranga Reddy Pallelra, Akshay Reddy Koppula, and Venkata Subhash Movva. 2015. Umduluth-cs8761-12: A novel machine learning approach for aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 742–747.

Zhiqiang Toh and Jian Su. 2015. Nlangp: Supervised machine learning system for aspect category classification and opinion target extraction. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 496–501.

Wei Wang, Hua Xu, and Xiaoqiu Huang. 2013. Implicit feature detection via a constrained topic model and svm. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 903–907.

Hua Xu, Fan Zhang, and Wei Wang. 2015. Implicit feature identification in chinese reviews using explicit topic mining model. *Knowledge-Based Systems*, 76:166–175.