Domain IndependentApproach for Detecting Domain Relevant Opinion Words and Opinion Targets

Aye Aye Mar
University of Information Technology
Yangon, Myanmar
ayeayemar@uit.edu.mm

Nyein Thwet Thwet Aung, Su Su Htay
Faculty of Information Science
University of Information Technology
Yangon, Myanmar
nyeinthwet@uit.edu.mm, suhtay@uit.edu.mm

***Abstract*— Opinion words come in various forms and different opinion words are occurred in different domains. An opinion word for one domain may not always be the opinion word for all domains. Relevant opinion words extraction for a given domain is a key task of opinion mining.The polarity of an opinion word can change depending on domain.The polarity of an opinion word may be positive in one domain but it may be negative or neutral opinion (no opinion) in some domains. Moreover, it also depends on the Opinion targetfound together with it even in the same domain.To solve these problems, a domain independent approach is proposed for detecting relevant opinion words of a given domain and the respective Opinion targets of these opinion words.In the proposed system, maximum relevancy (MR) method is introduced to find the domain relevant opinion words. The results of the MR methodare applied as the initial seeds of double propagation approach for finding more opinion words and opinion targets. Dependency relations between opinion words and opinion targets are designed to propagate opinion words and opinion targets.This systemmay be applied in other domains and mayextract the domain relevant opinion words together with their polaritiesand the corresponding opinion targets automatically and effectively.**

Keywords—opinoin mining; sentiment analysis;domain relevancy; opinion word;maximum relevancy; double propgagation

I. INTRODUCTION

Due to the growth of using internet, the amount of data on the web is increasing more and more. Basically, Data on the web is categorized into two types: objective content and subjective content. The objective contents are the facts and information describing about the entities characteristics. The subjective contents mean opinions, emotions, sentiments, feelings and attitudes about entities. An entity may be a product, hotel, event, organization and an individual. Generally, the objective contents are described by the companies and organizations and the subjective contents are written by the customers and users.

Opinion Mining also called sentiment analysis involves an area of NLP, computational linguistics and text mining, and refers to a set of techniques that deals with the data about opinions and tries to gain valuable information from them. Opinion mining has its applications in every domain such as customer products, services, financial services, and healthcare to political elections and social events [1].

Researcher extracts opinions from reviews in different levels: document level, sentence level, and feature level called aspect level. Document level assumes that all the contents in the document describes just about an entity. Some users describe about two or more entities in the same document. So, researchers did sentence level opinion mining to take account the opinion target intended by users. Opinion target may be on the whole entity or on features of the entity. There are also some users who write their opinions about an entity in detail. For example, the sentence "the ***camera*** of the phone is amazing". In such a comment, It is need to do feature level called finer-grained level opinion mining in order to know opinion target, ***camera***. Getting to know opinion target and opinion polarity is important so that manufacturers and other customers can know which features of the product that the users likes and dislikes. In this paper, finer-grain feature level opinion mining is done to seek detailed opinions.

In finer-grained feature level opinion mining, feature extraction is the main step.To analyse the finer-grained level opinions, this system extracts both opinion words known as opinion modifiers and feature also called aspects. Relevant feature extraction from unstructured data is still a major challenge because the polarity of an opinion word can change depending on domain and context

[2][3][4][15].This system also deals with this challenge.

Extracting the opinion words together with their polarity is a problem. The polarities of some opinion words are changed depending on domain and context [4].Opinion words lexicon and dictionaries can be used to extract opinion words. However, they cannot be obtained for each domain. Although pure unsupervised feature extraction techniques generalized and itdoes not require labelled data and domain independent,it cannot perform well in some domains which have domain dependent opinion words.

Although Double propagation is a state-of-the-art technique, it has a problem when it is applied to extract people's opinions on features of an entity. This method undergoes a great of noise in large corpora and misses important features in small corpora.Double propagation assumes that aspects are nouns/noun phrases and opinion words are adjectives. For large corpora, this method extracts some adjectives which are not opinion words, e.g., "entire" and "current" [8].

To solve these problems, a domain independent approach is proposedwhich will extract automatically the relevant opinion words for a given domain and the respective features of these opinion words. The basic idea of this approach is to extract the opinion words which are very relevant with domain as the initial opinion words and search other opinion words and the respective features using double propagation strategy. In this paper, the words "features","aspects" and "opinion targets" are used alternatively.

In the remainder of this paper, Section II gives the related work of opinion words and opinion target extraction, Section III presents the proposed system and concludes in Section IV.

II. Related Work

Opinion mining has been prominently studied by many researchers during the last years. Extensive works have been done on feature extraction and opinion classification.Although many different approaches have been proposed to find features for finer-grained feature level opinion mining, only a few papers consider domain relevancy issue to extract opinion words and opinion targets.The same opinion word may have opposite orientations indifferent domains. For example, "suck" usually indicates negative sentiment, e.g., "*This camera sucks*," but it can also imply positive sentiment, e.g., "*This vacuum cleaner really sucks*."[4].

Hu and Liu [6] have attempted earlyto detect features by using frequent occurring noun phrases. It has to be considered thattheir research was the source of aspect extraction from reviews. They extracted features only in the form of noun phrases by using association rule mining (ARM) based on Apriori Algorithm. They filtered incorrect frequent featuresvia compactness pruning and redundant pruning. Their approach performed wellfor high frequent features but not for low frequent features.This approachconsiders relation between relevant opinion words and features so it is hoped that both high frequent and low frequent features is found.

A.Bagheri et al. [5], proposediterative bootstrapping algorithm which is originally based on Pointwise Mutual Information (PMI)for aspect detection for sentiment analysis of customer reviews. Their method was unsupervised and domain independent. However, they extracted adjectives with specified POS patterns which are considered as opinion words.All of the adjective could not be opinion words.Alternatively, opinion words do not always appear in adjective form and they can be in other POS patterns such as verb, adverb and noun forms. In feature extraction, domain dependent opinion words are needed to be considered for effective opinion classification of the reviews.On the other hand, the polarities of some opinion words change depending on the domains. Though they have positive meaning in one domain but may be negative meaning in another domain.In this system, relevant opinion wordsare firstly detected using maximum relevancy between opinion words and discriminative class and apply these relevant words as the initial seeds for double propagation.

Qiu et al. [7] emphasized on opinion lexicon expansion and target extraction. They proposed a propagation approach to extract opinion words and targets iteratively by using identified relations between targets and opinion words. The relations were identified syntactically based on dependency grammar.They also proposed a novel method for assigning opinion polarityto new extracted opinion words. They assumed thatreviewershad the same opinion on the same target in the same review unless the review did not contain contrary word such as "but" and "however".The weakness of their system is that it will not perform well in some review containing the sentence such as "At first, I think that the battery of this branch is good. Gradually, I am getting to know that the battery is terrible".Another weak point is that they assumed that the same opinion word had the same polarity in one domain corpus. This assumption conflicted in some cases. Battery and processing time are the features of the same product in the same domain. However, long battery is positive opinion but long processing time is negative opinion.

Agawal and Mittal [9] proposed optimal feature selection for sentiment analysis. They extracted unigram and bigram features. To filter the extracted features, Information Gain (IG) and Minimum Redundancy Maximum Relevancy (mRMR) were investigated. Both of these selection results

wereevaluated using Boolean Multinomial Naïve Bayes (BMNB) and SVM for sentiment classification. Their experimental result presented that mRMRwas better than IG for selecting feature for sentiment analysis.

Zhang et al. [8]improved double propagation using part-whole and no patterns toextract and rank product features. Double propagation has the weakness of dealing noises in large corpus and missing important features in small corpus. Generally, double propagation would detect all nouns near opinion words as features.Therefore, theyadded these two patterns to remove the incorrect features. They also considered other feature relevancy and frequency to filter the features. Nevertheless, they did not attempt to filter irrelevant the opinion. They focused only to get the correct and relevant features in their system. In this system,relevant opinion words are detected by applying maximum relevancy approach and the correspondingrelevant features are extracted by using dependency relation with opinion words.Our approachcan be applied for all domainswhich have class labeled datasets and can detect relevant domain dependent opinion word.

This work is closely related to Qiuet al. [11] which expanded domain sentiment lexicon through double propagation. In their work, they used seed sentiment lexicon to extract initial sentiment words and features. In this system, initial opinion words are extracted by considering relevancy scores (mutual information scores) between words and the discriminate target class.The words in the form of adjective and adverb which have high relevancy scores are extracted as the initial seeds for double propagation.

Mostly related approach with this system is [12]. In their work, Sun and Hua proposed an M-Score algorithm for domain-independent opinion target extraction. This algorithm was derived from pointwise mutual information algorithm. They extracted opinion target using Conditional Random Field (CRF)model with feature templates. To extract the seed set, M-Score algorithm is applied and then propagates candidate opinion words and opinion targets using bootstrapping approach. Word frequency and noun pruning algorithm were applied to filter the opinion targets. The main difference with their work and this system is that this system firstly adoptsonly the domain relevant opinion words using maximum relevancy method. Dependency relations between opinion words and features are also considered to propagate the opinion words and targets.This important information was not considered in their approach.

Hai et al. [13] proposed an approach which extracts the opinion words which are high relevant with the given domain and low relevant with another domain using the intrinsic-domain relevance and extrinsic-domain relevance scores. The key difference between their approach and this system is that they firstly used syntactic dependency rules to generate a list of candidate features and then computes domain relevancy scores.However, in this system, the basic assumption is that dependency relations between opinion words and opinion targets may be similar but the domain relevancy scores of opinion words may not be similar. Therefore, domain relevancy scores are firstly computed to detect the high domain relevant opinion words. Baste and Vaishnavi[14] proposed Opinion feature extraction approach using domain-related corpus and domain-independent corpus. Their approach is very similar and slightly differentwith [13].

In [15], Garg et al.presented context dependent word polarity problem in depth with literature survey. They concluded that there is still a need to consider noun and verb as context dependent words. In this system, both opinion words and opinion targets are considered to detect the polarity of context dependent opinion words.

III. PROPOSED SYSTEM

Fig. 1 gives the architectural overview of the proposed systemfor domain relevant opinion word and target extraction.The proposed approach first extract opinion words which havehigh relevancy scores with the discriminate target class of a given domain. Then, these are used as the initial seeds for finding more opinion words and features via double propagation method. Dependency relations between opinion words and featuresare designed to propagate opinion words and features.

A. Maximum Relevancy Method

The assumption is that opinion words are associated with the whole entity or a feature of the entity.The system computes the maximum relevancyscoresof opinion words and thediscriminate classes using the equation (1).The opinion words which have high relevancy scores for each class are extracted as the initial opinion words. These initial relevant opinion words are firstly used to find the corresponding relevant opinion targets via dependency relations between opinion words and opinion targets.

$$\max D(S, c), D = \frac{1}{|S|}\sum_{x_i \in S} I(x_i, C) \qquad (1)$$

S is the set of features and I(xi , C) is the mutual information between each feature and the discriminate class. Using this maximum relevancy method, only high relevant opinion words together with their polarities are firstly detected from the reviews [10].

B. Dependency Relations

Opinion word and features are propagated based on dependency relations between features and the opinion words. Standford Parser Tool is used to parse

the sentences. The dependency relationsare based on the following intuitions:

- An Opinion word is usually found together with more than one feature.
- Alternatively, a feature is usually co-occurred with more than one opinion word.
- On the other hand, each word in some common noun phrases can be a feature e.g. phone sound recorder, phone sound volume etc.

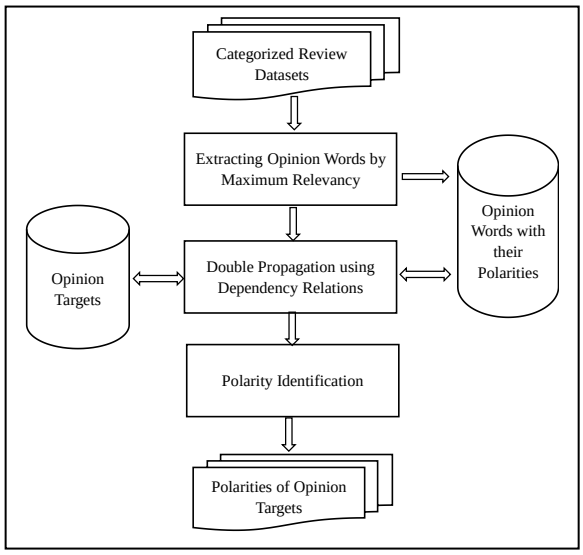

Fig. 1. The Proposed System Design

C. Double Propagation Method

Opinionwords and features are expanded using double propagation strategy. The initial seeds for double propagation arethe extracted opinion words which have high relevancy scores computed by equation (1). For example,the opinion word “good” is one ofthe opinion words which havehigh relevancy scores among the words.Let “Portable”bethe opinionword which is not commonly used by reviewers and haslow relevancy score. This system applies double propagation approach to seekthe opinion wordswhich do nothave high relevancy scores (mutual information scores with the class) but are relevant opinion words.In addition to this, this approach detects the corresponding features via initial relevant opinion words and dependency relations between an opinion word and a feature. Opinion words and the corresponding opinion targets are propagated based on the designed dependency relations. It is assumed that extracted opinion targets are also relevant because of detecting them by usingdependency relations between relevant opinion words and opinion targets candidates.

The***battery*** is **good**. (“battery”isextracted by “good”)

The***battery***is **portable**.(“portable” is got by “battery”)

The ***adapter***is **portable**.(“adapter”isgot by “portable)

D. Defining the Polarity of Opinion Targets

Finally, the system assigns the polarities of each opinion target based on the co-occurrence of positive and negative opinion words. If the opinion target is occurred more often with the positive opinion words than the negative opinion words, then its polarity is assigned as the positive. If not, it is assigned negative polarity. If both of these conditionsare not matched, then its polarity is assigned neutral (no opinion).

IV. Conclusion

A domain independent approachis proposed to deal with the problem of extracting opinion words and opinion targets which are relevant with the domain. Initial opinion words are high relevant by considering high mutual relation with the given domain.The assumptionsof dependency relations may supportto propagate more relevant opinion words and opinion targets.In this system, initial relevant opinion words together with their polarities are extracted by computing relevancy scores using the class labeled datasets .This system can beimproved by considering an automatic method to detect these wordsusing raw datasets. It is assumed that this system will become the effective task for seeking opinion in every domain. Because of detecting opinion words and opinion targets via maximum relevancy and dependency relations, this system can be reused in other domains with categorized reviews datasets.

References

[1] Ganeshbhai, Solanki Yogesh, and Bhumika K. Shah. "Feature based opinion mining: A survey." Advance Computing Conference (IACC), 2015 IEEE International. IEEE, 2015.

[2] Ravi, Kumar, and Vadlamani Ravi. "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications." Knowledge-Based Systems 89 (2015): 14-46.

[3] Agarwal, Basant, and Namita Mittal. Prominent Feature Extraction for Sentiment Analysis. Springer, 2016.

[4] B.Liu, “Sentiment Analysis and Opinion Mining”, Synthesis Lectures on Human Language Technologies, vol.5,no.1,pp. 1-167, May 2012.

[5] Bagheri, Ayoub, Mohamad Saraee, and Franciska De Jong. "Care more about customers: unsupervised domain-independent aspect detection for sentiment analysis of customer reviews." Knowledge-Based Systems 52 (2013): 201-213.

[6] Hu, Minqing, and Bing Liu. "Mining opinion features in customer reviews." AAAI. Vol. 4. No. 4. 2004.

[7] Qiu, Guang, et al. "Opinion word expansion and target extraction through double propagation." Computational linguistics 37.1 (2011): 9-27.

[8] Zhang, Lei, et al. "Extracting and ranking product features in opinion documents." Proceedings of the 23rd international conference on computational linguistics: Posters. Association for Computational Linguistics, 2010.

[9] B.Agawal, N.Mittal , “Optimal Feature Selection for Sentiment Analyis” , Springer, 2013.

[10] H. Peng, Hanchuan, Fuhui Long, and Chris Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy." IEEE Transactions on pattern analysis and machine intelligence 27.8 (2005): 1226-1238.

[11] Qiu, Guang, et al. "Expanding Domain Sentiment Lexicon through Double Propagation." IJCAI. Vol. 9. 2009.

[12] Yongmei, Sun, and Huo Hua. "Research on Domain-independent Opinion Target Extraction." International Journal of Hybrid Information Technology 8.1 (2015): 237-246.

[13] Hai, Zhen, et al. "Identifying features in opinion mining via intrinsic and extrinsic domain relevance." IEEE Transactions on Knowledge and Data Engineering 26.3 (2014): 623-634

[14] Baste, Vaishnavi S. "9 Opinion Feature Extraction via Domain Relevance."

[15] Garg, Sonal, and Dilip Kumar Sharma. "Sentiment Classification of Context Dependent Words." Proceedings of International Conference on ICT for Sustainable Development. Springer Singapore, 2016.