

FEATURE BASED ANALYSIS IN OPINION MINING VIA DOMAIN RELEVANCE

Varsha Machindra Sarnikar, Prof. Pankaj Agarkar

Computer Department

Dr.D.Y.Patil School of Engineering (Affiliated to Savitribai Phule

Pune University)

Pune, India

varsha4684@gmail.com , pmagarkar@gmail.com

Abstract: This paper helps in identifying the features from online reviews by applying feature filtering criterion. Existing opinion feature extraction techniques are mainly based on mining patterns from a single review corpus which is most of the times dependent review corpus.

Identifying candidate features which are from both corpora i.e. domain dependent and domain independent, this is captured by a measure called Domain relevance. Features extracted from this are relevant to a domain. For each extracted candidate feature its respective Intrinsic Domain Relevance and Extrinsic Domain Relevance values are estimated. These values are compared with threshold and are identified as best candidate features.

Keywords: Opinion mining, opinion feature, Information search and retrieval, Natural language Processing.

1. INTRODUCTION

Opinion mining often referred as sentiment analysis mainly focuses on analyzing people's opinion about entities which are termed as products, services, organizations, events and people. These sentiments are expressed in the form of text review on different blogs, forums and social networking sites. Looking at the tremendous growth in social media on the web

individuals and organizations are using these contents in business for decision making. Each site typically contains huge amount of opinionated text which is not easily interpreted in long blogs and forum. The average human reader faces difficulty in identifying relevant sites and extracting and analyzing opinionated text. Automated sentiment analysis is thus need of the time [1].

Recently we have witnessed that opinionated text on social media is helping businesses to reshape and influence public sentiments and emotions which makes great impact on social and political systems. However opinionated data is not only what we get from different sites or forums which is also termed as external data. Organizations have their internal data too which is in the form of customer feedback collected from digital media or surveys conducted by the organization. Because of such sentiment analysis applications industrial activities have been flourished in recent times. Sentiment analysis applications have reached almost all domains of different businesses across the globe from consumer products, medicinal and finance related services to public events and political elections or exit polls made after elections.

Generally from applications which are implemented in real life from them many application oriented research paper have been published. A sentiment model based on opinion mining was proposed to forecast sales performance. In 2010 reviews were used to rank products and vendors. Twitter sentiments were used to predict election results.

2. RELATED RESEARCH

As discussed above in this paper real life applications are major part of the reason why opinion mining is a popular research area now a days. It also faces many challenges as a Natural Language Processing research topic. Additionally there was little research before year 2000 in the area of NLP or linguistics, as there was very little data available in digital form since then it has become a most active research area in NLP. It is also a major researched area in data mining, web mining and data retrieval. It is spreading from the domain of computer science to management studies. Generally opinion mining is done at Document level and Feature level.

2.1 Lexicalized HMM's Learning Framework for Web Mining:

A machine learning framework using lexicalized HMM's was proposed which integrates linguistics features. This strategy falls into the category of feature level opinion mining, feature level focuses on deciding product features being commented on and the polarity of opinions for each feature. It has two main tasks recognition and classification [2].

In this technique given a specific product, the system first makes the decision on product related entities and opinion related entities from the opinions collected from reviews and then extraction of opinion sentences which describe each already identified product entity and finally determines of opinion polarity for each recognized product entity. This technique differs from its early approaches which has focused mainly on natural language processing techniques, this lexicalized HMM's (Hidden Markov Model) technique integrates linguistic features into automatic learning. Traditional Hidden Markov Model used in Part of Speech (POS) tagging and Named entity recognition (NER) problem. The motive of POS tagging is the process of marking up the words in a text i.e. corpus as corresponding to a specific part of speech such as noun and verb. The job of NER is identifying and classifying person's name, location name, organization name. For determining opinion orientations following algorithm steps are used:

Algorithm:

Input: Tagged Opinion Sentence

Output: Polarity of Opinion Sentence.

1. A loop is run through the product entity in given opinion sentence.
2. An initial opinion orientation is marked.
3. A search is performed for negation words.
4. If no negation words are found then polarity is marked if negation word is found then initial
polarity is changed.
5. Conjunctions and prepositions are checked against opinion words.
6. Accordingly polarity is changed.

2.2 CRF Based Approach for Opinion Target Extraction:

Technique used in this approach mainly focuses on extracting individual instances of opinion targets from sentences which contain an opinion expression. This technique is modeled as sequence segmentation and labeling task. The CRF algorithm receives input as sequence of tokens for which prediction of sequence of labels is done. Possible labels are represented as IOB scheme where I is identification of target O for non-target tokens and B for target. Sentences are modeled as a linear chain CRF which is based on an undirected graph where each node corresponds to a token in a sentence and edges connects adjacent tokens [3].

2.3 An Unsupervised Information Extraction System OPINE:

In this technique a particular product is given along with a set of reviews. Opine builds a model for important product features and their evaluation by reviewers and their relative quality across products. OPINE solves the opinion mining task and outputs a set of product features wherein each feature is accompanied by association opinions which are ranked based on the strength of opinions [4]. This output is then used for further opinion summary. This technique mainly focuses on three tasks

- For a particular given product first system identifies product features.
- Identification of opinions mentioned for specific product features.
- Identifying opinion polarity.

OPINE is basically built on top of KnowItAll. It's a domain independent information extraction system. In this a Naïve Bayes classifier is used which produces output which is a probability associated with each fact. OPINE extracts explicit features for a given particular product. Recursively parts and properties of the product feature are searched throughout the set of reviews. Once identified features are extracted using association rule mining and finally opinion words are used to find the infrequent features. Extracted features orientation is then tracked by a technique called relaxation labeling amongst the potential opinion words. In this technique a label is set to each and every object through an iterative procedure

2.4 Structured Models for Fine-to-course Sentiment Analysis:

In this technique a structured model is proposed which mainly focuses on classification of sentiment. It significantly reduces the classification error. This classification basically is done at two levels Document and Sentence. As in most of the reviews by customers there are ambiguities in sentiments which create problems in classifying the sentiment into a particular category. Here Sentence-Document model is proposed where a system is given with a document containing sentences which creates labels for each document as well as sentence. Where documents are analyzed for polarity and sentences are analyzed as a joint label of both document and sentence. Models in this type of work belong to a broad class of global structured models these are typically trained with structured learning algorithm. HMM is one of the earliest structured learning algorithm which is recently followed by discriminative learning approaches such as CRF. In this structured linear classifiers are used which assigns each label a weight. The problem of finding highest scoring labels for all sentences given at a particular document level is solved by applying Viterbi's algorithm [5]. Viterbi's algorithm returns the k-best label for each document. When a customer wants to have cohesive and informative summary this joint label can be used to reduce classification error.

2.5 Mining and Summarizing Customer Reviews:

In this technique the problem addressed is generating feature based summary of customer reviews. This is performed in three steps

1. Mining Product Feature
2. Identifying Opinion Sentences
3. Summarizing the result.

Its major phases are as shown in figure 1. The input to the system is product name and review page. The output is summary of reviews. Product features are usually expressed in terms of nouns or a noun phrase that's why part of speech is a crucial task. NLP parser is used to split text into sentences and produce part of speech for each word. Next is the frequent feature identification. Association miner is run which is based on Apriori algorithm. Each resulting frequent itemset is possible feature, generated frequent itemset are candidate features. Redundant features are pruned. After this opinion words are extracted these extracted opinions are then identified for their semantic orientation. The semantic orientation of a word indicates the direction that the word deviates from the norm for its semantic group. At last final feature based summary is generated according to the frequency with which each candidate feature appearance in the review [6].

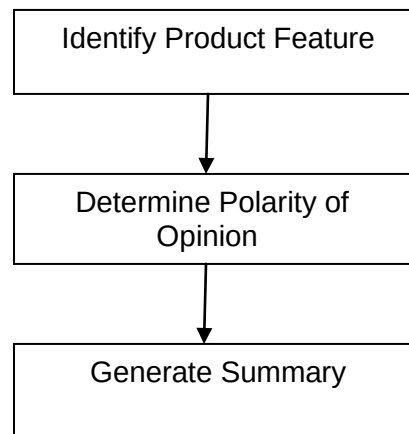


Fig 1. Major phases in mining and summarizing reviews

3. SYSTEM WORKFLOW & ALGORITHM

The system developed for extracting candidate features uses the workflow pattern as shown in Fig.2.

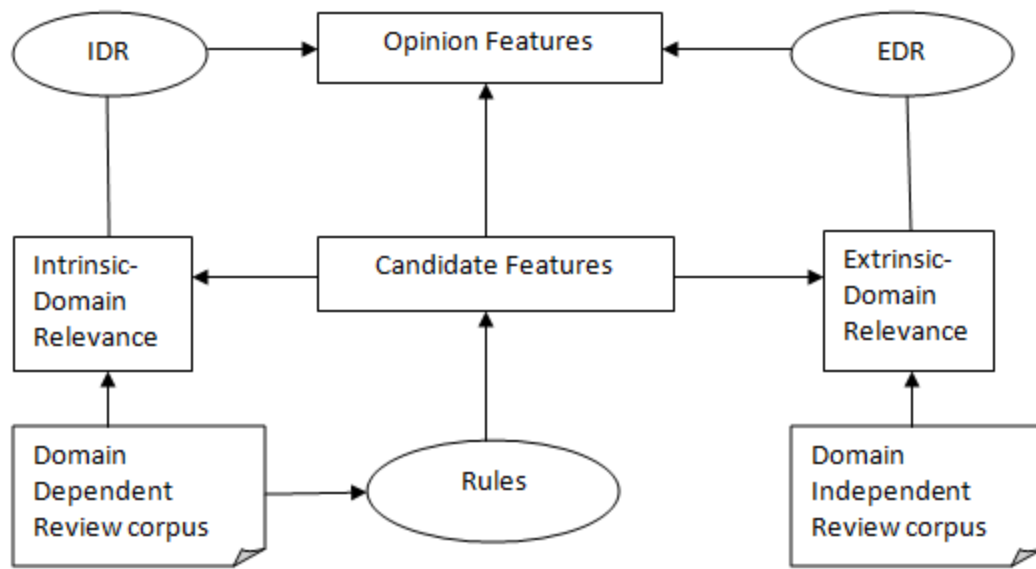


Fig 2. IEDR workflow

Step 1: A set of syntactic dependence norms are used to extract candidate features.

Step 2: IEDR measures are applied to determine corpus specific opinion features.

Step 3: Candidate features with value exceeding threshold values of Intrinsic Domain Relevance (IDR) as well as Extrinsic Domain Relevance (EDR) are selected for opinion extraction.

Proposed IEDR approach considers all possible variations along with domain dependent and domain independent corpus. Existing approach considers only domain dependent corpus. This helps in determining candidate features from independent corpus of reviews. Thus extraction of candidate features operates in following steps [7].

Step 1: To determine syntactic format of each sentence in a corpus.

Step 2: Syntactic rules are applied to the determined dependence structure.

Step 3: Candidate features are extracted.

4. CONCLUSION & FUTURE SCOPE

This study examined opinion mining via domain driven opinion mining which can be applied to different commercial domains in order to yield more useful results. These case studies show effective and efficient ways in the domain of business. In each case study reviewed domain knowledge is implemented in addition to the opinion mining techniques. Areas of future study

can expand the scope to other domains such as agriculture, medical applications and engineering.

To conclude I would like to note that in order to maintain an appropriate size of the paper I had to limit the number of referenced studies. I therefore apologize to the authors of papers which were not cited in this work. In future fine grained theme modeling which also identifies implicit features can be implemented for feature extraction. Furthermore more precision giving opinion extraction approaches needs to be exercised for better results.

ACKNOWLEDGEMENT

I wish to express my deep sense of gratitude towards my guide and PG Coordinator Prof. Pankaj Agarkar for his constant motivation and guidance in completing my task with admirable excellence. I would also like to thank my Head of Department Prof. Soumitra Das as well as our Director Dr.S.S.Sonavane and last but not the least departmental staff members for their support.

REFERENCES

- [1] B. Liu, "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1-167, May 2012.
- [2] W. Jin and H.H. Ho, "A Novel Lexicalized HMM-Based Learning Framework for Web Opinion Mining," *Proc. 26th Ann. Int'l Conf. Machine Learning*, pp. 465-472, 2009.
- [3] N. Jakob and I. Gurevych, "Extracting Opinion Targets in a Single and Cross-Domain Setting with Conditional Random Fields," *Proc. Conf. Empirical Methods in Natural Language Processing*, pp. 1035-1045, 2010.
- [4] S.-M. Kim and E. Hovy, "Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text," *Proc. ACL/COLING Workshop Sentiment and Subjectivity in Text*, 2006.
- [5] R. McDonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar "Structured Models for Fine- to-Coarse Sentiment Analysis," *Proc. 45th Ann. Meeting of the Assoc. of Computational Linguistics*, pp. 432-439, 2007.
- [6] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," *Proc. 10th ACM SIGKD Int'l Conf. Knowledge Discovery and Data Mining*, pp. 342-351, 2004.
- [7] Zhen Hai, Kuiyu Chang, Jung-Jae Kim, and Christopher C. Yang , "Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance", *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, VOL. 26, NO. 3, MARCH 2014