Identification of Features from User Opinions using Domain Relevance

Mr. A. V. Moholkar
ME II Computer
DGOI,FOE, Duand
Savitribai Phule Pune University (MH), India.
Duand, Pune, India
abhijit.moholkar8@gmail.com

Prof. S. S. Bere
HOD & Assistant Professor
DGOI,FOE, Duand
Savitribai Phule Pune University (MH), India
Duand, Pune, India
sachinbere@gmail.com

***Abstract*—** Identification of opinion features from online user reviews is a task to identify on which feature user is going to put his opinion. There are number of existing techniques for opinion feature identification but, they are extracting features from a single corpus [2]. These techniques ignore the nontrivial disparities in distribution of words of opinion features across two or more corpora. This work discusses a novel method for opinion feature identification from online reviews by evaluation of frequencies in two corpora, one is domain-specific and other is domain-independent corpus. This distribution is measured by using domain relevance [12]. The first task of this work is the identify candidate features in user reviews by applying a set of syntactic rules. The second step is to measure intrinsic-domain relevance and extrinsic-domain relevance scores on the domain dependent and domain-independent corpora respectively. The third step is to extract candidate features that are less generic and more domain specific, are then conformed as opinion features. This approach is called as intrinsic extrinsic domain relevance.

***Index Terms*—** *candidate features, Intrinsic Domain Relevance, Extrinsic Domain Relevance, Opinion Feature.*

I. INTRODUCTION

This technique is to identify opinion features from user opinions on any product. These opinions are important role in sale of the product. This work is to extract opinion features from user opinions to identify on which feature users are going to do opinion. There are number of techniques for the identification of these features but, they are operated on a single corpus and ignore nontrivial distribution in word. This work discusses a novel method for mining features in user opinions from two types of corpus one is domain dependent and other is domain independent. A supervised learning approach [2] [3] works well in given domain only but can't retain in other domain. Unsupervised approach [4] [5] [6] will apply some of the syntactic rules for opinion feature identification. Topic modeling approach [7] [8] is to mine generic topics.

One solution is the selection of domain independent corpus such that frequency of feature in user review is more in domain dependent corpus than the domain independent corpus. Let us consider one example containing feature battery. This feature may be present in both mobile domain and vehicle domain. The frequency of feature is high in mobile domain and relatively less in vehicle domain. The feature extraction from two domains is better achieved using novel technique. In novel technique domain relevance score is measured for each domain dependent and domain independent corpus [10] [11]. The measurement of domain relevance score on domain dependent score is termed as intrinsic domain relevance, in other case domain relevance score on domain independent corpus is termed as extrinsic domain relevance. The application of Intrinsic Extrinsic domain relevance on results of previous step yields accurate opinion features from user reviews.

II. RELATED WORK

An Opinion in reviews is analyzed at document, sentence or phrase level for classification of overall subjectivity in a single review. Hatzivassiloglou and Wiebe [12] stated a technique to predict subjectivity. Pang et al [13] stated a technique for classification of reviews in to positive and negative sentiments. This classification is helpful in improvement of product quality also in price reduction of a product. To prevent consideration of nonrelated text pang and lee [14] stated a technique to identify sentence as subjective or objective followed by this it discard objectives to correctly identify opinion features using sentiment classifier. This subjectivity extraction is down through identification of minimal cuts in the graph. Mcdonald et al. [15] studied a global structured model for prediction of sentiments at different granularity levels. The regression method is proposed for prediction of reviews rating from sparse text pattern. Bollegala et al. [17] stated cross domain sentiment classifier using extracted sentiment thesaurus.

Zhang et al. [19] provided rule based sentiment analysis to classify sentiment for text review. Maas et al. [20] presented a technique for document level and sentiment level task classification. The sentiment polarity of an opinion feature is context dependent and domain specific. Wilson et al. [21] proposed a technique for contextual sentiment classifier at phrase level. Yessenalina and Cardie [22] stated a compositional matrix space model for phrase level sentiment analysis. The supervised models [2] [3] discussed above perform well on the given domain but, it is not retained for other domains otherwise transport learning is needed to adopt.

Unsupervised learning approach [4] [5] identifies opinion features by application of different syntactic rules. This may extract incorrect features due to colloquial nature of online reviews. Hu and Liu [10] provided an association rule mining technique for mining frequent itemsets. But it has limitation that it may extract frequent but invalid features and rare but valid

features may be overlooked. Su et al. [13] stated a mutual reinforcement clustering which utilize co-occurrence weight matrix generated from the given review corpus. However precision of this is low due to poor real life clusters. Yu et al. [23] stated an aspect ranking algorithm for identification of opinion features based on probabilistic regression technique. Latent Dirichlet allocation (LDA) [7], stated for aspect based opinion mining. It is a generative three way probabilistic model. They may be poor in dealing with identification of specific feature terms commented explicitly in user reviews. As above discussion, existing approaches deal with identification of features in a single review corpus ignoring their variation in domain independent corpuses. The IEDR approach will identify accurate opinion features from two or more corpuses. In first step it will find a set of candidate features and extract opinion features in second step.

III.PROPOSED SYSTEM

The feature battery in mobile domain is domain specific as it has higher frequency in mobile domain than outside domain. This work identifies Nouns, Noun phrases and adjectives by applying part of speech tagging on user input review. The next step is the extraction of candidate features by application of syntactic rules on output of POST. The extraction of these domain specific candidate features is based on the designing of syntactic rules. Domain relevance score is measured on each domain dependent corpus called intrinsic domain relevance score and on domain independent corpus called extrinsic domain relevance score by application of IDR/EDR algorithm. The candidate features with IDR score greater than user defined intrinsic relevance threshold and EDR scores less than user defined extrinsic relevance threshold are the exact opinion features. These extracted features are more domain specific and less generic features. The identification of opinion features from candidate features is done by application of IEDR algorithm.

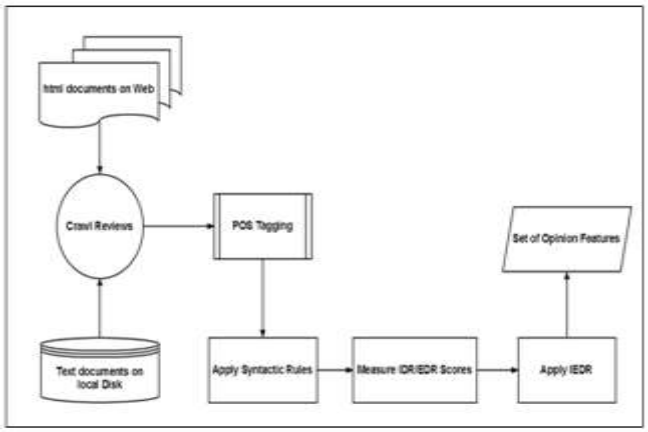

Fig. 1: Architecture of the Proposed System

As shown in architecture diagram user select reviews from either internet in the form of html file or from text file on local system. The part of speech tagging is applied on the collected reviews for classification of nouns, noun phrases, or adjectives. The application of language dependent syntactic rules will find most probable features from the user review.

The features extracted in this phase may be incorrect; to filter irrelevant features intrinsic domain relevance and extrinsic domain relevance scores are measured on domain dependent corpus and domain independent corpus respectively. This domain relevance score represents frequency of relevant feature term in a specific document. The last step is the application of intrinsic extrinsic domain relevance, in IEDR two thresholds are selected called intrinsic relevance threshold and extrinsic relevance threshold. The features with IDR score greater than intrinsic relevance threshold and EDR score less than extrinsic relevance threshold are extracted as an opinion features. They are more domains specific and less generic.

As shown in the figure the opinion feature price which is associated with adjective expensive. In other figure noun feature exterior associated with the verb like. The first step for extracting candidate feature is the construction of dependence tree. The second step is the application of the syntactic rules for candidate feature identification. As shown in table1 there are number of syntactic rules for extraction of candidate features from user review.

Table 1 list out different syntactic rules for English language

Rules	Interpretation
NN->JJ || NN->VBP || NN->PRP || NN->VB || NN->VBZ || NN->VBD || NN->VBG || NN->VBN|| JJ->NN	Extract NN as CF
NNP->JJ || NNP->VBP || JJ->NNP || NNP->PRP || NNP->VB || NNP->VBZ || NNP->VBD || NNP->VBG || NNP->VBN	Extract NNP as CF
NNS->JJ || JJ->NNS || NNS->VBP || NNS->PRP || NNS->VB || NNS->VB || NNS->VBZ || NNS->VBD || NNS->VBG || NNS->VBN	Extract NNS as CF

VI.PROPOSED ALGORITHMS AND ANALYSIS:

A. Algorithms

1) Calculation of Intrinsic/Extrinsic domain relevance
Input: Domain specific/Independent corpus
Output: Domain relevance score (IDR/EDR)
1) For each candidate feature in corpus C calculate $w_{ij.}$
2) Calculate standard deviation s_i.
3) Calculate Dispersion $disp_i$.
4) Calculate Deviation $devi_{ij}$.
5) Calculate Domain relevance dr_{ij}.

2) Identification of Opinion features using IEDR
Input: Domain Review corpus R and Domain independent corpus D
Output: A validated list of opinion features.
1) Find candidate features.
2) For each candidate feature calculate intrinsic domain relevance idr_i on review corpus R.
3) For each candidate feature calculate extrinsic domain relevance edr_i on domain independent corpus D.

4) Candidate features with idr score greater than threshold vale and edr score less than another threshold are conformed as opinion features.
5) Calculate Domain relevance dr_{ij}.

B. Analysis of Algorithms:

Intrinsic Domain Relevance / Extrinsic Domain Relevance: The Intrinsic Domain Relevance and Extrinsic Domain Relevance Algorithms are NP complete type of problems, because they return domain relevance value and executes in polynomial time. These algorithms to find domain relevance values for input reviews.

Intrinsic Extrinsic Domain Relevance: The Intrinsic Extrinsic Domain Relevance Algorithms are NP complete type of problems, because it executes in polynomial time and return candidate features as a output to the user. This algorithm identifies opinion features by selecting two threshold values.

V. EXPERIMENTS

We have applied the IEDR feature extraction into an existing opinion mining technique named iMiner [30], and thus far evaluated its performance using real-world English reviews from two different domains, i.e., mobile and hotels.

A. Corpus Description

The mobile review corpus contains 110 real-life textual reviews collected from a social sites like flip cart, amazon. The hotel review corpus contains 111 reviews crawled from a social sites. The Summary of the four domain review corpora are shown in Table 3. This work randomly selected 5 review corpuses. Two persons manually marked opinion feature(s) expressed in every review sentence in each of the mobile category. A marked opinion feature is considered valid if and only if both annotators highlight it. If only one of the annotators mark an opinion feature, then a third person has a final decision on whether to keep or reject it. A total of 18 opinion features were obtained from the mobile review files. Using the same method, we annotated 19 opinion features from randomly selected hotel review files. The precision and recall are measured by equation 1 and 2. The value for precision is 0.93 and 0.90 for mobile and hotel reviews, respectively.

We also collected 4 domain-independent (generic) corpora from above website, each corpus containing 8 documents. The collected corpora cover domain irrelevant heterogeneous topics containing Vehicle, Laptop, and so on. Summary statistics of the 4 domain independent corpora are shown in Table 2. All documents from the domain review corpora as well as the domain-independent corpora were parsed using the language technology platform (LTP) [31], a Chinese natural language analyzer.

Table 2: Dataset description of the work.

Sr. No.	Domain Dependent Corpus	#reviews	#statements	#features
1	Mobile	110	287	23
2	Hotel	111	331	24
3	Vehicle	98	300	21
4	Laptop	99	290	22

VI. RESULT DISCUSSION

This work calculates precision and recall values for the domains using equation 1 and 2.

Precision= #Correct features / # Retrieved features (1)

Recall= #Retrieved features / #features in domain (2)

Sr. No.	Intrinsic Domain	Extrinsic Domain	# features in domain	# Retrieved features	# Correct features	Precision	Recall
1	Mobile	Hotel	23	21	18	0.93	0.88
2	Hotel	Vehicle	24	21	19	0.90	0.87
3	Vehicle	Mobile	21	17	14	0.82	0.81
4	Laptop	Mobile	22	18	16	0.89	0.82
5	Laptop	Vehicle	23	20	17	0.85	0.87

A. *Precision versus Recall*

The work first extract candidate features from the given review corpus, i.e., mobile and hotel reviews, using the syntactic rules provided in Table 1. Based on the same set of candidates, the precision-recall curve for IEDR is plotted as solid lines in Fig. 2. Note that the best performing vehicle corpus was selected as the domain-independent corpus for both IEDR and EDR. This is perfectly acceptable as precision values at large recall levels are more practical. Across all recall levels, the largest precision gap of IEDR over IDR is 11.90 percent (located at 0.55 recalls). At recall rates larger than 0.5, the best IEDR precision is 93.00 percent, which is 13.00 percent higher than the best IDR precision. The Proposed IEDR thus achieved a large improvement over either IDR or EDR. The best IEDR precision is 91.67 percent for recall rates higher than 0.5, which is 15.06, 16.18, 18.76, and 31.08 percent better than the best precision for LDA, ARM, MRC, and DP methods, respectively. The experimental results demonstrated the effectiveness of our proposed IEDR approach on the mobile review domain. This work further evaluated the IEDR feature extraction performance on a different domain, hotel reviews.

B. *Summary of Evaluated Methods*

Fig.2. Precision-recall curves for mobile feature extraction. Results are generated by plotting precision at each of 5 recall levels.

This graph represents Precision-Recall curve of the Intrinsic extrinsic domain relevance approach for the above four domains.

 ISSN: 2321-8169
3006 - 3009

A fig. 3 represents set of candidate features extracted intrinsic domain relevance and fig. 4 represents set of opinion features extracted via IEDR.

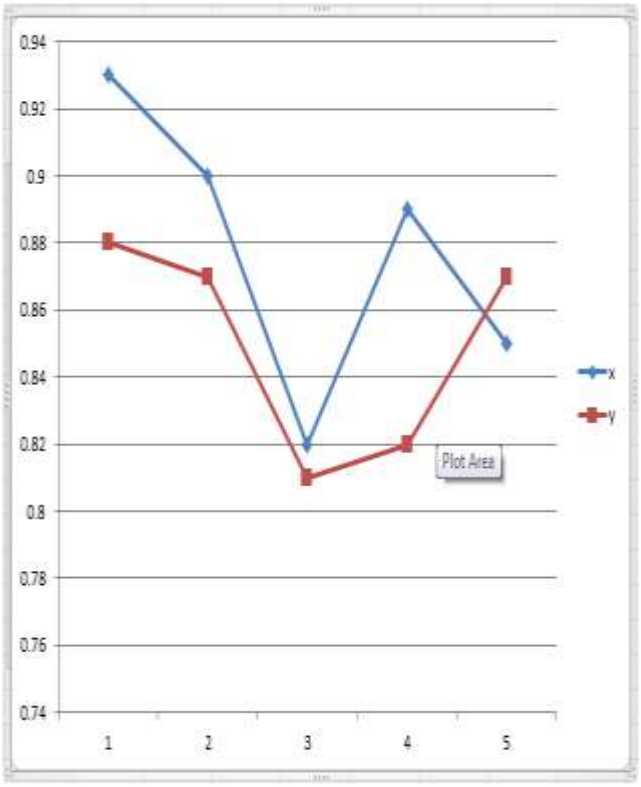

Fig. 2: Precision – recall curve of IEDR approach.

Fig. 3: Set of candidate features extracted via intrinsic domain relevance.

Fig. 4: A set of opinion features ectracted via IEDR.

VI. CONCLUSION:

In this work, a novel based IEDR approach is discussed, which utilizes the disparities in distributional characteristics of features from two corpora; one of them is domain dependent and other is domain-independent. IEDR identify candidate features that are domain specific to the user domain.

The experimental result represents that the IEDR approach is better than IDR, EDR, LDA, ARM, MRC, and DP, in terms of performance as well as opinion mining results.

In this work, the selection of domain-independent corpus in terms of its size and topic reflects the quality of the work. It is found that domain-independent corpora of similar size but topically different from the given review domain will yield better results.

VII. ACKNOWLEDGMENT:

I express great many thanks to Prof. S. S. Bere and Department Staff for their great effort of supervising and leading me to accomplish this fine work. They were a great source of support and encouragement. To every person who gave me something too light along my pathway. I thank them for believing in me.

REFERENCES

[1] 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

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

[3] N. Jakob and I. Gurevych, "Extracting Opinion Targets in a Singleand Cross-Domain Setting with Conditional Random Fields," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 1035-1045, 2010.

[4] G. Qiu, C. Wang, J. Bu, K. Liu, and C. Chen, "Incorporate the Syntactic Knowledge in Opinion Mining in User-Generated Content," Proc. WWW 2008 Workshop NLP Challenges in the Information Explosion Era, 2008.

[5] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion Word Expansion and Target Extraction through Double Propagation," Computational Linguistics, vol. 37, pp. 9-27, 2011.

[6] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," J. Machine Learning Research, vol. 3, pp. 993-1022, Mar. 2003.

[7] I. Titov and R. McDonald, "Modeling Online Reviews with Multi- Grain Topic Models," Proc. 17th Int'l Conf. World Wide Web, pp. 111-120, 2008.

[8] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 168-177, 2004.

[9] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 168-177, 2004.

[10] A. Popescu and O. Etzioni, "Extracting Product Features and Opinions from Reviews," Proc. Human Language Technology Conf. and Conf. Empirical Methods in Natural Language Processing, pp. 339-346, 2005.

[11] V. Hatzivassiloglou and J.M. Wiebe, "Effects of Adjective Orientation and Gradability on Sentence Subjectivity," Proc. 18th Conf. Computational Linguistics, pp. 299-305, 2000.

[12] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment Classification Using Machine Learning Techniques," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 79-86, 2002.

[13] 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.

[14] L. Qu, G. Ifrim, and G. Weikum, "The Bag-of-Opinions Method for Review Rating Prediction from Sparse Text Patterns," Proc. 23rd Int'l Conf. Computational Linguistics, pp. 913-921, 2010.

[15] D. Bollegala, D. Weir, and J. Carroll, "Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus," IEEE Trans. Knowledge and Data Eng., vol. 25, no. 8, pp. 1719-1731, Aug. 2013.

[16] P.D. Turney, "Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews," Proc. 40th Ann. Meeting on Assoc. for Computational Linguistics, pp. 417- 424, 2002.

[17] C. Zhang, D. Zeng, J. Li, F.-Y. Wang, and W. Zuo, "Sentiment Analysis of Chinese Documents: From Sentence to Document Level," J. Am. Soc. Information Science and Technology, vol. 60, no. 12, pp. 2474-2487, Dec. 2009.

[18] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis," Proc. 49th Ann. Meeting of the Assoc. for Computational Linguistics: Human Language Technologies, pp. 142-150, 2011.

[19] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing, pp. 347-354, 2005.

[20] A. Yessenalina and C. Cardie, "Compositional Matrix-Space Models for Sentiment Analysis," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 172-182, 2011.

[21] E. Cambria, D. Olsher, and K. Kwok, "Sentic Activation: A Two-Level Affective Common Sense Reasoning Framework," Proc. 26th AAAI Conf. Artificial Intelligence, pp. 186-192, 2012.

[22] S.J. Pan and Q. Yang, "A Survey on Transfer Learning," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 10, pp. 1345-1359, Oct. 2010.

[23] W.X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly Modeling Aspects and Opinions with a Maxent-Lda Hybrid," Proc. Conf. Empirical Methods in Natural Language Processing, pp. 56-65, 2010.

[24] L. Tesniere, Elements de la syntaxe structurale. Librairie C. Klincksieck, 1959.

[25] F. Fukumoto and Y. Suzuki, "Event Tracking Based on Domain Dependency," Proc. 23rd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 57-64, 2000.

[26] Z. Hai, K. Chang, Q. Song, and J.-J. Kim, "A Statistical Nlp Approach for Feature and Sentiment Identification from Chinese Reviews," Proc. CIPS-SIGHAN Joint Conf. Chinese Language Processing, pp. 105-112, 2010.

[27] W. Che, Z. Li, and T. Liu, "LTP: A Chinese Language Technology Platform," Proc. 23rd Int'l Conf. Computational Linguistics, pp. 13- 16, 2010.

[28] A.J. Viera and J.M. Garrett, "Understanding Interobserver Agreement: The Kappa Statistic." Family Medicine, vol. 37, no. 5, pp. 360- 363, May 2005.

[29] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 342-351, 2004.

[30] Q. Su, X. Xu, H. Guo, Z. Guo, X. Wu, X. Zhang, B. Swen, and Z. Su, "Hidden Sentiment Association in Chinese Web Opinion Mining," Proc. 17th Int'l Conf. World Wide Web, pp. 959-968, 2008.

[31] T. Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence," Computational Linguistics, vol. 19, no. 1, pp. 61-74, Mar. 1993.

Mr. Moholkar Abhijit V. Received his B.E. degree in Information Technology (Distinction) in the year 2013 from Pune University. He is currently working toward the M.E. Degree in Computer Engineering from the Savitribai Phule Pune University, Pune. He has 02 years of teaching experience at undergraduate level. His research interests lies in Data Mining and he has published 4 paper in the same domain.

Prof. Bere Sachin S. received his B.E. degree in Computer Engineering (First-class) in the year 2008 from Pune University and M.Tech Degree (Distinction) of Computer Engineering in2013. He has 08 years of teaching experience at undergraduate and postgraduate level. Currently he is working as Assistant Professor and HOD in Department of Computer Engineering of DGOI, FOE, swami-chincholi, Daund, Pune University. His research paper has been published in IJTITCC, IJISET, and IJRITCC year 2014. His research interests are Digital Image processing.