ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Feature Extraction from Informal Text for Opinion Mining

Namrata Adhao *, A.G.Phaktkar
Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India.
namrata0388@gmail.com

Abstract

With the rapid development of web, most of the customers express their opinions on various kinds of entities, such as products and services on web. These reviews provide useful information to customers for reference. These reviews are also valuable for merchants to get the feedback from customers and improve the qualities of their products or services. However, the contents are stored in mostly either unstructured or semi-structured format. We are trying to improve mining approach to mine product features, opinions from Web opinion sources for informal text. The extracted feature-opinion pairs and sentence-level review source documents are modeled using a graph structure.

Keywords: Data mining, opinion mining, text mining, feature identification.

Introduction

Opinions are central to almost all human activities and are key influencers of our behavior. When we need to make a decision we often seek out the opinions of others.

With the rapid development of e-commerce, most customers express their opinions on various kinds of entities, such as products and services. These reviews not only provide customers with useful information for reference, but also are valuable for merchants to get the feedback from customers and enhance the qualities of their products or services. Reviews generally involves specific product feature along with opinion sentence. Many times reviews may be quite lengthy it is hard for the customers to analyze them through manual reading any make an inform decision to purchase a product.

A large number of reviews for may make harder for individual to evaluate quality of a product. In some cases, customers or persons may naturally attract to read a few re- views for making a decision regarding the product and services. Similarly, manufacturers also want to read the reviews for identification about strengths and weakness of products and services provided to customers. And also improve the quality of products or services. Since, most of the reviews are stored either in unstructured or semi-structured format, if the reviews could be processed automatically and presented in a summarized form highlighting the product features and users opinions would be a great help for both customers and manufacturers.

In this paper, we propose a mining approach to extract product features and opinions from review documents. As observed in [12][1], most product features can be found by exploiting local information and their Parts-Of-Speech (POS). Therefore, the proposed approach implements the feature extraction mechanism as a rule-based system. An information component contains < f, m, o >where f represents a feature generally identified as a noun phrase, o represents an opinion expressed over f generally identified as adjective, and m is a modifier generally used to model the degree of expressiveness of o. We have extracted feature and opinion pairs and resource documents as a graph. HITS [9] algorithm is applied for each feature-opinion pair for feasibility analysis with respect to the underlying corpus.

The remaining paper is structured as follows. Brief review of the existing opinion mining systems is represented in section 2. Section 3 presents detail of the programmers
design of the proposed system. Finally, section 4 Result and discussion of the paper. Section 5 concludes the discussion with possible enhancements to the proposed system.

Related work

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

Lot of works has been done in this area. Extract positive or negative opinion words by Turney. Identify feature-opinion pairs together with the polarity of each opinion [7]. The approaches to mine opinions at different levels of granularities including documents [20], sentences [8] and words [6]. In [20], Turney proposed an algorithm to classify a review as positive or negative, which applies POS analysis to identify opinion phrases in review documents and uses PMI-IR algorithm [21] to identify their semantic orientations. Feature-based opinion mining is also proposed considering above facts [7], [13], [15].

In [13], the authors have proposed a supervised pattern mining method, which identifies product features from pros and cons sections of the review documents in an automatic way. In [16], the design of OPINE system based on an unsupervised pattern mining approach is presented, which extracts explicit product features using feature assessor and web PMI statistics. In [10], the authors have proposed a pattern mining method in which patterns are described as a relationship between feature and opinion pairs. In [18] double propagation approach is used to extract opinion words and features using a seed opinion lexicon. Since a complete opinion is always expressed in one sentence along with its relevant feature [11], the feature and opinion pair extraction can be performed at sentence-level to avoid their false associations. Classification of document according to formal and informal style in [2]. Feature opinion pairs are extracted from formal text and reliability score generated form web opinion sources [1]. Various methods of feature-based opinion mining used for feature extraction and refinement, which includes rule-based methods and NLP [7], [9], ontology-based methods [5], and statistical methods [20]. Liu [6] proposed a system using association rule mining which extract features from review data. The system selects frequent terms and then extracts features by measuring the similarities between selected terms. The main problem of this method is that the system only considers the information from
the term itself, for example, term frequency, which does not reflect the relationship between a feature and its related opinion information. Ding [10] proposed rule based system for feature extraction method. This method extracts a relatively large number of features compared with the amount of review data. For example, it generates 189 features from 50 reviews for digital cameras.
The main reason for the extraction of so many features is that terms that have the same or similar meanings are not considered as the same features. For example, some words have same meaning like ’photo,’ ’picture,’ and ’image’ all have the same meaning; however, they are considered as different features simply because they are different words. Consequently, this system could not provide proper summary information for the product. This problem is solve in FEROM in that the number of features are reduced by merging words that have similar meanings using the semantic similarity between features and then providing reliable summary information for the product based on the merged features. Aciar [13] proposed a feature extraction method that uses ontology for opinion mining. Although this method worked well semantically, the main problem is the maintenance of the ontology to address the constant expansion of the review data. In this system, the ontology is manually constructed and when new features are added it must be updated. In addition, a concept that is defined in the ontology is only able to be classified. Thus, it is necessary to construct an automatic system to avoid continued intervention.

Programmer’s design

The purpose of the analysis is to extract, organize, and classify the information contained in the required documents. The proposed method is based on object- oriented approach to software development. In this section, we present the architecture and functional detail of the proposed opinion mining system to identify feature-opinion pairs. Figure 1 presents the complete architecture of the proposed opinion mining system, which consists of different functional components.

Review Document
In this module the crawler retrieves reviews document from sources such as web. Then Locate and download the reviews.

Review Cleaning
After that review document is processed to review cleaning or filtering. Filtering process, filter out or remove noisy review.

Classification of Review
After removing noisy review classify the remaining data review according to formal and informal style [2]. Filtered review document are divided into manageable record size chunk.

Data Preprocessor

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

Filtered review document are divided into manageable record size chunk. This is assign as input for document preprocessor to Parts of Speech tag (POS) to each word, like Stanford Parser [21]. It converts each sentence into set of dependency relationship between pair of words.

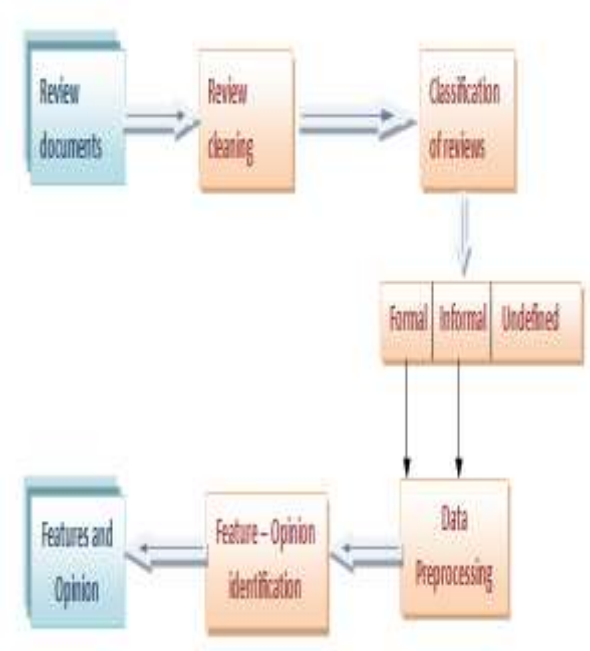

Figure1. Architecture of proposed opinion mining system.

Feature and opinion identification

In Feature and opinion identification module we represents the dependency relations between a pair of words w1 and w2 is as relation type(w1,w2), in which w1 is called head or governor and w2 is called dependent or modifier. This may be direct or indirect Relation type id. In direct, one word depends on other directly and in indirect on through other word or both of them depends on third word indirectly. As information component is defined as < f, m, o >. This module represents rule based system for formal text as in [1].

For informal text for example we are in a dependency relation R, if there exists a abbrev(w1, w2) relation such that , POS(w1) = NN,POS(w2) = JJ , and w1 and w2 are not stop-words then w2 is assumed to be an opinion and w1 as an feature.

Following steps for InformalTextMining system:

where, Input = Informal review sentences

1. Identification of the Informal sentence reviews.
2. Add it to data preprocessor
3. Apply Parts of speech tagging.
4. Identification of noun, verb, adverb.
5. Apply rules as in[1] if reviews are not informal.

Large number of noun, verb, adjective are extracted which gives features and opinion represented as undirected graph as shown in figure 2.

3.6. Mathematical Model

Our system consist of set S,

S = {I, O, Sc, Fu, R, F, Op, P}

Where,

I – Set of Inputs
O – Set of outputs
Sc – Success state
Fu – Failure state
R – Review documents
F – Features
Op – Opinion
P – Feature-opinion pair

Set of Inputs:

I = { R }

Where,

R = {r1, r2, r3, ……., rn } = set of reviews

We can have input function

I() =R

$\forall r \in R$,

Where, $r \neq \phi$

$I() = R \rightarrow Cust$

Add (R);

Where,

Add (R) = Formal + Informal + undefined

Classify (R);

POS_TAG (Classify (R));

FEATURE_EXTRACT ();

RESULT(F, Op);

Success State:

Sc = { F, Op}

i.e. identified features and opinion

Failure State:

Fu = { } or {Ø}

Set of Outputs:

O = { F ,Op } = Sc

Where,

F = { f1, f2, f3, ……., fn} ,

i.e. Set of features. And

Op = { Op1, Op2, Op3, ……., Opn }

i.e. set of opinion.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

3.7. Dynamic Programming and Serialization

Mapping and dependencies

A Dependency Map allows us to visualize the critical cross-project dependencies throughout the duration of the program.

If R then F
$F \rightarrow R$
If R then Op
$Op \rightarrow R$
Features and opinions are dependent on reviews taken from customers.
Where,
P is pair of features and opinions i.e.
$P = \{P1, P2, \ldots, Pn\}$;
Where,
P1 (f1, o1), Pn (fn, on).

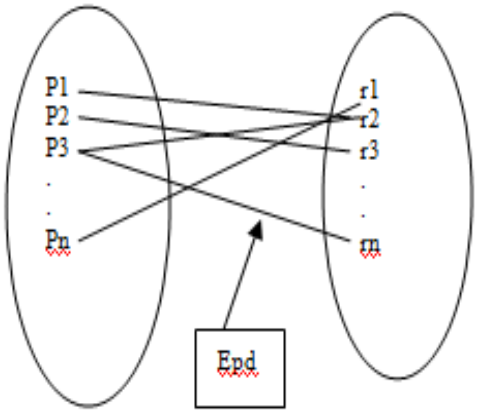

Figure 2: Mapping Dependencies

Results and discussion

Results can be evaluated using standard Information Retrieval (IR) metrics Precision and Recall respectively. Results are given in Table 3, by considering different products. That is features and opinions which is representation of noun and adjectives.

Performance Analysis

Table.1.Classification of online reviews

Sr. No.	Product Name	Total No. Of reviews	Formal Reviews	Informal Reviews	Undefined reviews
1	Nikon	66	10	54	2
2	Sony	106	6	98	2
3	Sony-cyber-shot-dsc-wx300	16	0	16	0
4	Canon	33	5	26	2

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

Table.2. Calculation of Precision and recall value of our system

Sr.No	Product Name	Features	Opinions
1	Nikon	Camera	good
2	Nikon	Picture	Nyc
3	Nikon	price	high
4	Nikon	camera	compact
5	Nikon	pic	bright

Sr. No.	Product Name	Total number of feature available(Fa)	Total number of feature extracted(Fe)	Total number of correct features(Fc)	Precision (Fc /Fe)	Recall (Fc/Fa)
1	Nikon	60	58	52	0.896	0.87
2	Sony	100	95	90	0.94	0.9
3	Sony-cyber-shot-dsc-wx300	18	17	13	0.76	0.72
4	Canon	32	30	27	0.9	0.84

Conclusion

In this paper, we have presented system for opinion mining for informal text and identify feature-opinion pairs from review documents. Our system is able to identify and extract feature and opinion pairs along with the source documents.

References

1. *Ahmad Kamal, Muhammad Abulaish , Tarique Anwar, Mining Feature-Opinion Pairs and Their Reliability Scores from Web Opinion Sources. WIMS '12, June 13-15, 2012 Craiova, Romania. Copyrightc 2012 ACM 978-1-4503-0915-8/12.*
2. *Fadi Abu Sheikha and Diana Inkpen, Learning to Classify Documents According to Formal and Informal Style , LiLT Submitted, March 2012, Published by CSLI Publications.*
3. *J. Allan, C.Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, Toronto, Canada, 2003.*
4. *A. Andreevskaia and S. Bergler. Mining wordnet for fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, EACL '06, pages 20-216, 2006.*
5. *E. Breck, Y. Choi, and C. C. Identifying expressions of opinion in context. In Proceedings of the 20th international joint conference on Artificial intelligence, Menlo Park, CA, USA, pages 2683-2688,2007.*
6. *M. Hu and B. Liu. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '04, pages 168-177. ACM, 2004.*
7. *S. Kim and E. Hovy. Determining the sentiment of opinions. In Proceedings of the 20th international conference on Computational Linguistics, Geneva, Switzerland, pages 1367-1373, 2004.*

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

8. *J. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM), 604- 632, 1999.*
9. *N. Kobayashi, K. Inui, and Y. Matsumoto. Extracting aspect-evaluation and aspect-of-relations in opinion mining. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, pages 1065-1074, 2007.*
10. *B. Li, L. Zhou, S. Feng, and K. Wong. A unified graph model for sentence-based opinion retrieval. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, pages 1367-1375, 2010.*
11. *F. Li, Y. Tang, M. Huang, and X. Zhu. Answering opinion questions with random walks on graphs. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore, pages 737-745, 2009.*
12. *B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In Proceedings of the 14th international conference on World Wide Web, Japan, pages 342-351, 2005.*
13. *J. Otterbacher, G. Erkan, and D. Radev. Using random walks for question-focused sentence retrieval. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, pages 915-922, 2005.*
14. *B. Pang and L. Lee. Opinion mining and sentiment analysis. Found. Trends Inf. Retr., 21-135, January 2008.*
15. *A. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, Canada, pages 339-346, 2005.*
16. *G. Qiu, B. Liu, J. Bu, and C. Chen. Expanding domain sentiment lexicon through double propagation. In Proceedings of the 21st international joint conference on Artificial intelligence, San Francisco, CA.*
17. *G. Qiu, B. Liu, and C. Chen. Opinion word expansion and target extraction through double propagation. Association for Computational Linguistics, 2010.*
18. *Q. Su, X. Xu, H. Guo, Z. Guo, X. Wu, B. Zhang, X. and Swen, and Z. Su. Hidden sentiment association in chinese web opinion mining. In Proceeding of the 17th international conference on World Wide Web, Beijing, China, pages 959-968, 2008.*
19. *P. D. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toe. In Proceedings of the 12th European Conference on Machine Learning, EMCL '01, pages 49-502. Springer-Verlag, 2001.*
20. *L. Zhuang, F. Jing, and X.-Y. Zhu. Movie review mining and summarization. In Proceedings of the 15th ACM international conference on Information and knowledge management, CIKM '06, pages 43-50. ACM, 2006.*
21. *Marie-Catherine de Marneffe and D.Manning, Stanford typed dependencies manual. Revised for Stanford Parser v.1.6.9 September.*

Author Biblography

Namrata Adhao
Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India
Email: namrata0388@gmail.com

Anupama Phaktkar
Assistant professor, Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India
Email: agphaktkar@pict.edu