Analysis of Customer Reviews for Opinion Feature Summary

Varsha Sarnikar, Prof.Pankaj Agarkar

M.E. Research Scholar, P.G. Department of Computer Engineering, D.Y.Patil School of Engineering, Maharashtra, India
Assistant Professor , P.G. Department of Computer Engineering , D.Y.Patil School of Engineering , Maharashtra, India.

Abstract - The idea of opinion mining or sentiment analysis is to process a set of search results for a given entity, generating a list of attributes which are termed as opinion features of that entity. The system proposed helps in identifying the opinion 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 specific 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. These opinion features contribute to summarizing product reviews which evaluates all the features.

Key Words: Opinion mining, Opinion Feature, Natural Language Processing, Information Search & Retrieval

1. INTRODUCTION

Human life is packed with emotions and opinions. We cannot visualize the world exclusive of them. Emotions and opinions play important role nearly in all human actions. They direct the individual's life by influencing the way we think, what we do and how we take certain decision. Text mining is an integrative technique used in different fields like machine learning, information retrieval, statistics, and computational linguistics. Web mining is a sub discipline of text mining used to mine the semi structured web data in the structure of Web Content mining, Web Structure mining and Web Usage mining. Opinion mining also referred as sentiment analysis which is a process of analyzing user's opinion or sentiment about particular topic or a product or problem. A topic can be any entity such as news, event, product, movie, etc [8]. Opinion mining is a research domain in Text mining, Natural Language Processing, and Web mining discipline [9]. The objective of Opinion Mining is to make computer able to identify and express emotions. An idea, view, or temperament based on emotion instead of a fact is called sentiment. Figure 1 has the hierarchy of Data Mining and the categories of how Opinion Mining is formed under the branch. Opinion mining also referred as sentiment analysis which focuses on analyzing public opinions, sentiments and temperament towards entities such as services, products and their features. Opinions expressed in textual reviews are generally scrutinized on various dimensions. Opinion mining generally works at two levels document level and feature level.

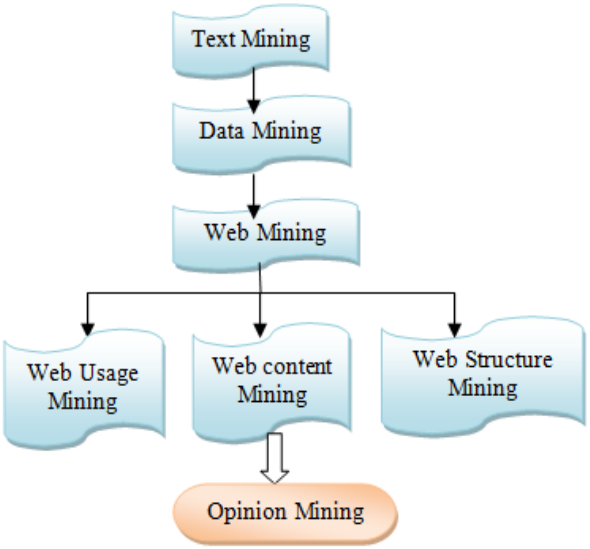


Figure 1.Hierarchy of Data Mining

With the tremendous growth of social media such as reviews, forum discussions, blogs, twitter comments and postings in social networking sites on the web are used by individuals and organizations for decision making. Different attributes of an entity on which opinions are expressed are often referred as opinion feature or feature and the orientation of such opinions is called as polarity of the opinion.

A major research area in this domain is of opinion feature identification and extraction which has already been addressed and various techniques such as Natural

Language Processing techniques and modeling techniques are proposed. In real life reviews syntactic rules which are used in NLP do not work well as these reviews lack formal structure, modeling techniques which implement semantic rules are used for coarse grained analysis.

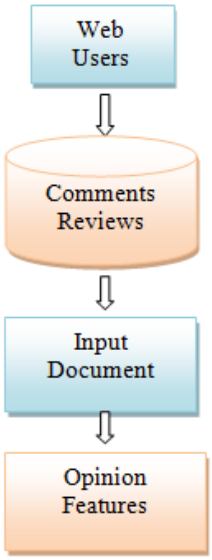


Figure 2.Workflow of Opinion Feature Extraction

2. RELATED WORK

Zen Hai and C Yang proposed a model for 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. These opinion features contribute to summarizing product reviews which evaluates all the features [1].

Vasileios Hatzivassiloglou and Jance Wiebe have studied the effect of adjectives on predicting the subjectivity of opinions. They have proposed a method for predicting subjectivity of opinions at sentence level by a supervised classification method. This method first classifies the adjectives according to their orientation afterwards each adjective is assigned a label depending on its appearance frequency in the corpus. This label is applied depending on the highest applicable conjunctions threshold. Later gradability is decided which classifies the adjectives into gradable and non gradable. Lastly depending on orientation and gradability subjectivity of the sentence is determined. Disadvantage of this model is that it is limited to sentence level classification it is not extended to document level [2].

Machine learning algorithms are being used by authors Pang and Lee which are Naïve Bayes, Maximum entropy classification and support vector machine have been used for text categorization. Naive bayes classification method classifies the overall document by bayes rule where in MaxEnt there are no assumptions made regarding relationships between features. In support vector machine the problem of text categorization is converted into optimization problem by considering support vectors. Disadvantage of this technique is that accuracy obtained is not better over the traditional text categorization techniques. The advantage of the system is it performs better than human baseline [3].

Bo Pang and Lillian Lee have proposed sentence level subjectivity detector to identify the sentences in a document as either subjective or objective. Document level polarity classification proposes minimum cut framework which discards the objective sentences in the document [4].

Ryan McDonald and Kerry Hannan [5] have proposed a structured model for classifying sentiments at different levels of granularity like document level, sentence level or word level which is also referred as fine to coarse sentiment analysis. The simplest approach includes having separate system for each level of granularity. The proposed model has the advantage of building the single model for all granularity levels. Labeling is done by MIRA algorithm which works at document and sentence level by applying a weight vector to each label. Disadvantage of this model is its performance is not stable for longer documents.

Lizhen Qu and Georgiana Ifrim [6] have proposed a model based on regression method which is used for prediction of review ratings from a sparse text pattern. This technique proposes an algorithm for estimating opinion scores from regression method. Disadvantage of this model is domain dependent attributes may not give same performance as domain independent attributes. Advantage of this model is it overcomes the problem of sparsity faced in n-gram models.

Yessenalina and Cardie [7] proposed a compositional matrix space model which works for phrase-level sentiment analysis. One of the advantage of the proposed model is that by learning matrices for words, the model can handle unseen word compositions (e.g., unseen bigrams) as far as the component unigrams have been learned.

A method proposed by E. Cambria and D.Olsher works on two level affective reasoning to mimic the integration of conscious and unconscious

3. PROPOSED WORK

3.1 SYSTEM ARCHITECTURE

Opinion mining is a process of automatic extraction of knowledge from the opinion of others about some particular entity. The idea of opinion mining or sentiment analysis is to process a set of search results for a given entity, generating a list of attributes of that entity and aggregating opinions. The purpose to design this system is to propose a model in order to provide accurate results for feature extraction from online reviews on different real world dataset in training and testing set. This helps in breaking down the overall rating or review of an entity into feature specific review. Here the system must be holding characteristics such as system should be user friendly, responsive, informative. It should be designed such flexible so that any future modifications can be easily adapted by the system. For example now in the system non-noun opinion features or noun phrases are not considered which can be made part of the future expansion of this system. Such fine grained topic modeling approach can be employed in order to identify domain independent features. Furthermore good use of IEDR extracted opinion features to summarize online reviews of products or services.

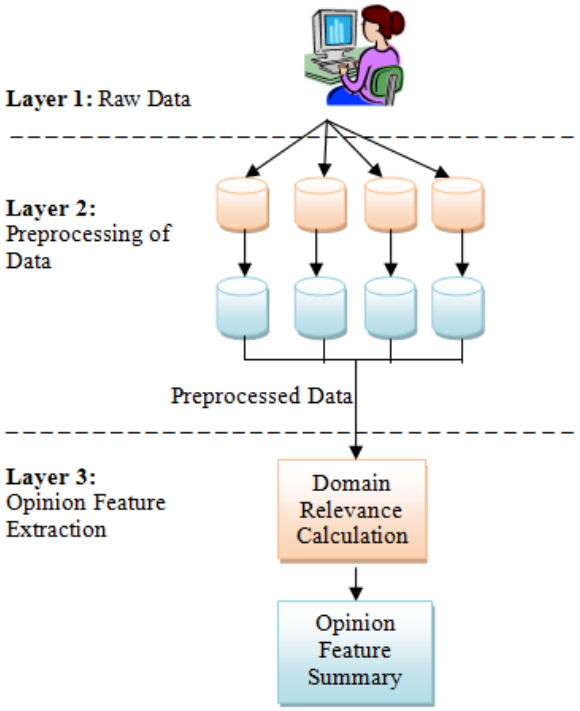


Figure 3. System Architecture

This proposed system focuses on opinion feature identification and extraction based on feature filtering criterion. Which utilizes the candidate features from both the domains that is domain dependent corpus as well as domain independent corpus [10][11]. To overcome the disadvantages faced in existing architecture as domain independent review corpus is not considered while determining opinion features for the product or service. In existing approach infrequent features are not considered while doing candidate feature extraction later from these opinion features are selected. Infrequent features are not prominently visible in domain dependent review corpus to make them a part of candidate feature so that if frequency with which a particular feature appears in review corpus is though less then also it should be considered while extracting opinion features.

4 .IMPLEMENTATION DETAILS

A .work breakdown structure mainly focuses on following areas:

1. Module 1: Load Input File containing reviews about Product.

2. Module 2: Preprocessing of Input Data.

3. Module 3: Candidate Feature Extraction.

4. Module 4: Opinion Features.

B. Module 1: Load Input File containing reviews about Product.

This module consists of uploading an input file which contains reviews or comments about the product which will be combination of opinioned text and non opinioned text. As further processing happens with this data only. Data collection is assumed to done already which gives this input document

C. Module 2: Preprocessing of Input Data

This preprocessing module consists of series of operations as tokenization, stemming and stopwords removal. Preprocessing is most important module of this system as the input document contains unopinieted text which does not take part in the feature extraction process presence of such data in the feature extraction process reduces the performance as unopinieted data also participates in the process of feature extraction. Tokenization, in this phase tokens from the input document are identified and are separated from the rest of the document. Later stemming

takes place which is applied on tokenized data which further processes the data eliminating the unnecessary text from the input data. which take part in the candidate features extraction which will be the potential opinion features of the product.

D. Module 3: Candidate Feature Extraction

Candidate features are the ones from which final opinion features will be obtained. Here domain dependent and domain independent features are classified depending on domain relevance score. For each candidate feature its domain relevance score i.e Intrinsic Extrinsic Domain Relevance score is calculated this score gives the classification for each and every candidate feature depending on this frequent and infrequent features are classified.

E. Module 4: Opinion Features

Domain Relevance gives the opinion features which belong to product domain and independent of domain. Finally summary of all opinion features which belong to a domain and are independent is presented to the user.

5. MATHEMATICAL MODEL

Following is the relevant mathematics related to the proposed system and that is represented by set theory.

Let I represents dataset of reviews or comments proposed against the product. R represents the recognized opinion features set, F1 represents function that performs preprocessing, F2 represents function that performs domain relevance calculation, F3 represents a function that performs opinion feature extraction. Input is mapped to the output which is shown in the following vein diagram.

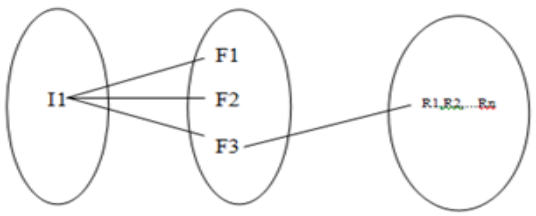


Figure 4. Venn Diagram

Let the system be S where,

S= {I, R, F}

Where I = Set of Input

R=Set of Output

F=Set of Function Implementation

where

I = {I1, I2, I3,},

F = {F1, F2, F3},

R = {R1, R2, Rn}.

Following diagram shows flow of data in the system.

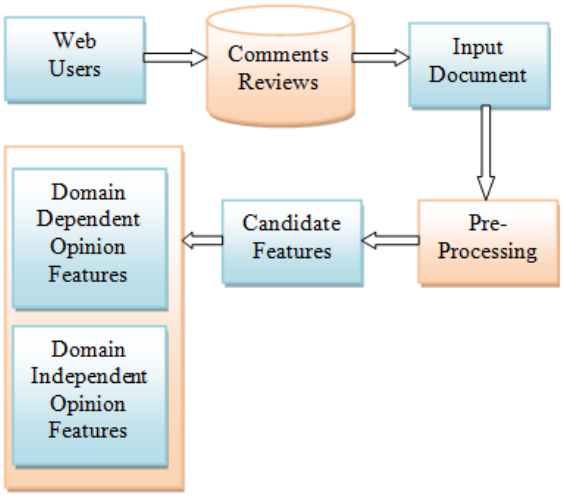


Figure 5. Block Diagram

Here in this system first the user uploads the input file, this input file is then passes through the phase of preprocessing where unwanted data is removed which actually improves the system performance as at the end for feature extraction phase only data that participates in the feature extraction process. This preprocessed data is then passed onto the next step where candidate feature identification takes place, candidate features are the potential opinion features. After this domain relevance scores are estimated this classifies the features into domain dependent and domain independent features. These are the collection of all opinion features which belongs to the domain as well as independent of domain.

6. RESULT AND DISCUSSION

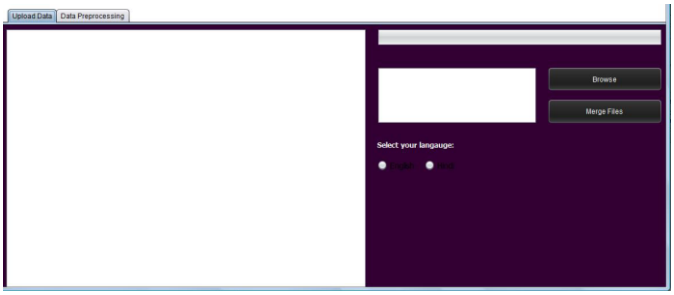
6.1 DATASET

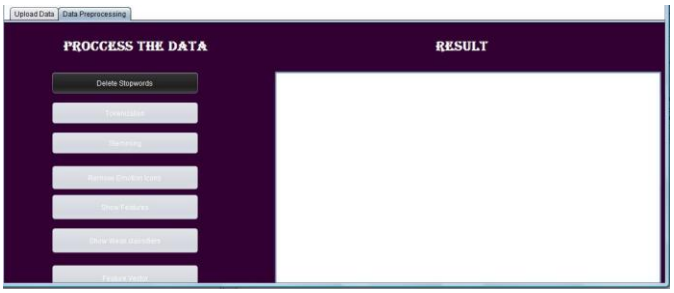
In this paper we are considering comments and reviews from the different sites which is already collected and

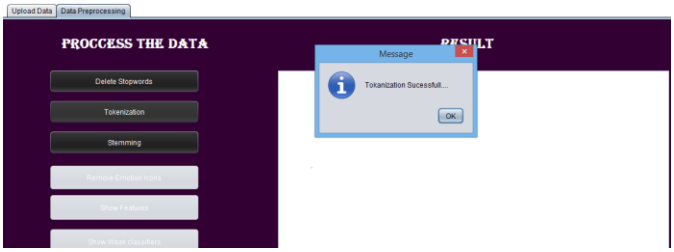
saved in the review dataset. These reviews and comments are considered for experiment.

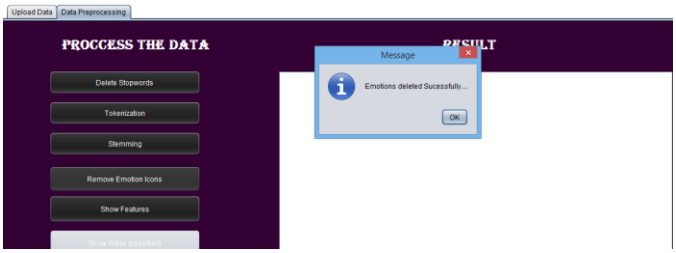
6.2 RESULTSET

In the first module we are going to consider reviews collected from different sites as input data. In second module.









After data cleaning which involves tokenization, stemming and stopwords removal operations the structure of the file is expected to be reduced to the much smaller size with most of the words remaining to be used as features for feature extraction process. Later domain relevance estimation should take place which categorizes the features into domain dependent opinion features and domain independent opinion features. At last we are expecting frequent as well as infrequent features to participate in the overall product review. Actual experimental results by considering reviews from different sites for finding domain dependent as well as domain independent features concludes the product summary.

7. CONCLUSION

We propose a feature identification system for addressing the challenges of feature based opinion mining such as identification of domain independent features of an entity,

8. AKNWOLEDMENT

The satisfaction achieves after successful completion of any task would be incomplete without mentioning those people who are responsible to complete that task. I'm grateful to many persons who contributed to the completion of this research. Particularly I wish to express my deepest gratitude towards Prof. Pankaj Agarkar my honorable guide and PG coordinator, Prof. Soumitra Das Head Of Department and Dr. S.S.Sonavane Director of D.Y.Patil School Of Engg. Lohegaon, Pune, for providing comments, information, and review of this report. Lastly I would like to thank all my friends who have shared their knowledge with me during my research work.

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 trnsactions on knowledge and data engineering, Vol. 26, NO. 3, MARCH 2014.

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

[3] 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.

[4] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," Proc. 42nd Ann. Meeting on Assoc. for Computational Linguistics, 2004.

[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] 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.

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

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

[9] 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.

[10] 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.

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

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

BIOGRAPHIES

M.E Research Scholar, PG Department of Computer Engineering. B.E in Computer Engineering from B.A.M.U university, Maharashtra.

Prof. Pankaj Agarkar is working as an Asst. Prof. in PG Department of Computer Engineering. He has 19 years of teaching experience and he has published many papers in international and national journals.

