Domain Independent Approach for Extraction of Features in Customer Reviews

Nilesh Shelke **Dr. S. P. Deshpande**

Abstract —**Individuals as well as institutions are paying increasing attention to sentiment analysis. Companies are interested in what bloggers are saying about their products. Politicians are interested in how different news media are portraying them. The state of the art in sentiment analysis focuses on assigning a polarity or a strength to subjective expressions (words and phrases that express opinions, emotions, sentiments, and so on) in order to decide the objectivity-subjectivity orientation of a document or the positive/negative/neutral polarity of an opinion sentence in a document. Very first step involved in this process is to extract the features from the reviews. Much research has been done on extracting the features from domain dependent reviews. This research work addresses domain independent approach for feature extraction from the given reviews.**

Key Words* — O*pinion mining, POS tagger,

I. Introduction

The Web contains a wealth of opinions about products, politicians, and more, which are expressed in newsgroup posts, review sites, and elsewhere. As a result, the problem of "opinion mining" has seen increasing attention over the last many years. Sentiment analysis aims to use automated tools to detect subjective information. It is the extraction of people's opinions, appraisals and emotions toward entities, events and their attributes [1]. Given a set of evaluative text documents *D* that contain opinions (or sentiments) about an object, opinion mining aims to extract attributes and components of the object that have been commented on in each document $d \in D$ and to determine whether the comments are positive, negative or neutral. Little work has been done on the processing of opinions until only recently. Yet, opinions are so important that whenever one needs to make a decision one wants to hear others' opinions. This is not only true for individuals but also true for organizations.

Opinions can be expressed on anything, e.g., a product, a service, a topic, an individual, an organization, or an event. The general term *object* is used to denote the entity that has been commented on. An object has a set of *components* (or *parts*) and a set of *attributes*. Each component may also have its sub-components and its set of attributes, and so on. Thus, the object can be hierarchically decomposed based on the *part-of* relationship.

To understand the notion, consider the following review related of the iphone:

"(1) *I bought an iPhone a few days ago.* (2) *It was such a nice phone.* (3) *The touch screen was really cool.* (4) *The voice quality was clear too.* (5) *Although the battery life was not long, that is ok for me.* (6) *However, my mother was mad with me as I did not tell her before I bought it.* (7) *She also thought the phone was too expensive, and wanted me to return it to the shop. …* "

The question is: what we want to mine or extract from this review? The first noticeable thing is that there are several opinions in this review. Sentences (2), (3) and (4) express positive opinions, while sentences (5), (6) and (7) express negative opinions or emotions. Further it can also be noticed that the opinions all have some targets or objects on which the opinions are expressed. The opinion in sentence (2) is on the iPhone as a whole, and the opinions in sentences (3), (4) and (5) are on the "touch screen", "voice quality" and "battery life" of the iPhone respectively. The opinion in sentence (7) is on the price of the iPhone, but the opinion/emotion in sentence (6) is on "me", not iPhone. This is an important point. In an application, the user may be interested in opinions on certain targets or objects, but not on all (e.g., unlikely on "me"). Finally, we may also notice the sources or holders of opinions. The source or holder of the opinions in sentences (2), (3), (4) and (5) is the author of the review ("I"), but in sentences (6) and (7) is "my mother". With this example in mind, we now formally define the sentiment analysis or opinion mining problem. We start with the opinion target. In general, opinions can be expressed on anything, e.g., a product, a service, an individual, an organization, an event, or a topic. We use the term *object* to denote the target entity that has been commented on. An object can have a set of *components* (or *parts*) and a set of *attributes* (or *properties*).

Each component may have its own sub-components and its set of attributes, and so on.

Definition (**object**)**:** An *object O* is an entity which can be a product, topic, person, event, or organization. It is associated with a pair, *O*: (*T*, *A*), where *T* is a hierarchy or taxonomy of *components* (or *parts*) and *sub-components* of *O*, and *A* is a set of *attributes* of *O*. Each component has its own set of sub-components and attributes.

II. Literature Survey

Many researchers have become interested in sentiment analysis, as more people learn of the scientific challenges posed, and the scope of new applications enabled, by the processing of subjective language. The papers studied in [2] are a relatively early representative sample of research in the area. Major deficiency of the previous work is that it only focuses on detecting the overall sentiment of a document, without performing an in-depth analysis to discover the latent topics and the associated topic sentiment. In general, a review can be represented by a mixture of topics. For instance, a standard restaurant review will probably discuss topics or aspects such as food, service, location, price, etc. Preliminary work on this issue has been proposed in [3]. In this section, we discuss others' related works for feature extraction and opinion words extraction. Hu and Liu [4] proposed several methods to analyze customer reviews.

To sufficiently capture information from various aspects, [5] proposed an aspect-based segmentation algorithm to first segment a user review into multiple single-aspect textual parts, and an aspect-augmentation approach to generate the aspect-specific feature vector of each aspect for aspect-based rating inference.

Joint Aspect/Sentiment model (JAS) to extract aspects and aspect dependent sentiment lexicons from online customer reviews in a unified framework has been proposed in [6].

The extracted sentiment lexicons are applied to a series of aspect-level opinion mining tasks, including implicit aspect identification, aspect- based extractive opinion summarization, and aspect-level sentiment classification [7].

To sufficiently capture information from various aspects, [8] proposed an aspect-based segmentation algorithm to first segment a user review into multiple single-aspect textual parts, and an aspect-augmentation approach to generate the aspect-specific feature vector of each aspect for aspect-based rating inference.

III. The Proposed Techniques

Major steps for doing Aspect (Feature) Oriented Sentiment Analysis are the following:

I. Identify product features.

II. Identify opinions regarding product features.

III. Determine the polarity of opinions.

IV. Rank opinions based on their strength.

This research paper deals with first step only i. e. identifying the features of the product that customers have expressed opinions on (called *opinion features*) and rank the features according to their frequencies that they appear in the reviews.

Given review was inputted to the POS module. A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural'. Parser package available on standford university web site has been used. This software is a Java implementation of the log-linear part-of-speech taggers.

After POS tagging is done, we need to extract feature that are nouns or noun phrases as nouns are the probable candidates of having features.

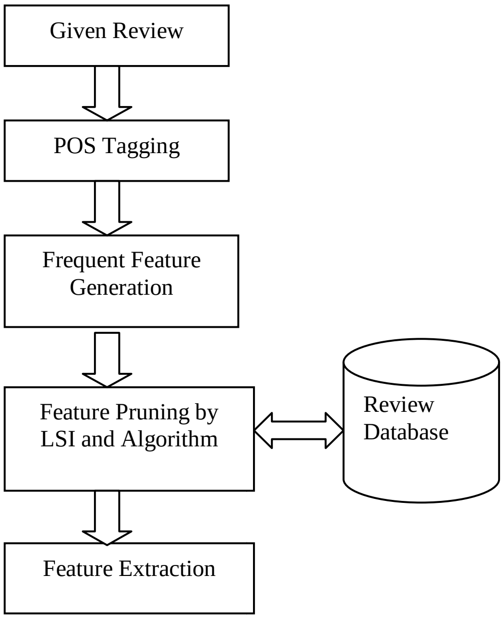

Given Review
POS Tagging
Frequent Feature Generation
Feature Pruning by LSI and Algorithm
Review Database
Feature Extraction

Fig. 1. Extraction of Features from the given reviews

Earlier much of the work in this direction has been done by using Supervised Algorithms like SVM, Maximum Entropy, Naïve Bayes Model. We have attempted here Latent Semantic Indexing (LSI) extracting the product features from the product reviews for domain independent reviews.

Latent Semantic Analysis (LSA) [9] is an approach to automatic indexing and information retrieval that attempts to map documents as well as terms to a representation in the so called latent semantic space. LSA usually takes the (high dimensional) vector space representation of documents based on term frequencies as a starting point and applies a dimension reducing linear projection. The specific form of this mapping is determined by a given document collection and is based on a Singular Value Decomposition (SVD) of the corresponding term/document matrix. The general claim is that similarities between documents or between documents and queries can be more reliably estimated in the reduced latent space representation than in the original representation. The rationale is that documents which share frequently co-occurring terms will have a similar representation in the latent space, even if they have no terms in common. LSA thus performs some sort of noise reduction and has the potential benefit to detect synonyms as well as words that refer to the same topic. In many applications this has proven to result in more robust word processing.

Latent Semantic Indexing (LSI) is a common technique in natural language processing area. Fig. 2 shows how LSI

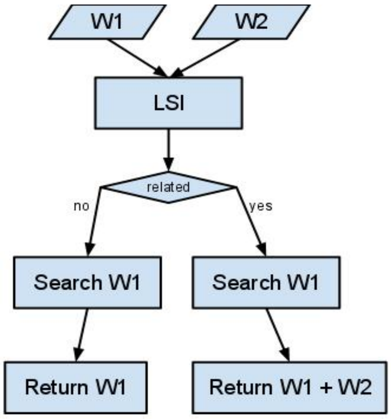


Fig. 2 Effect between LSI and keyword search. W stands for a document.

works by comparing the pure key-word-based search. Latent Semantic Indexing (LSI) is an indexing and retrieval method that uses a mathematical technique called Singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. LSI is based on the principle that words that are used in the same contexts tend to have similar meanings. -

For example, Paris and Hilton are associated with a woman instead of a city and a hotel, Tiger and Woods are associated with golf.

Regular Keyword Search vs. LSI

By using regular keyword search, a document either contains the given word or not, and there is no middle ground.

LSI adds an important step to the document indexing process. LSI examines a collection of documents to see which documents contain some of those same words. LSI considers documents that have many words in common to be semantically close, and ones with less words in common to be less close.

An LSI Example

If we use LSI to index a collection of articles and the words "program" and "code" appear together frequently enough, the search algorithm will notice that the two terms are semantically close. A search for "program" will therefore return a set of articles containing that phrase, but also articles that contain just the word "code". LSI does not understand the word distance, but by examining a sufficient number of documents, it knows the two terms are related. It then uses that information to provide an expanded set of results with better recall than a plain keyword search. The diagram below describes the effect between LSI and keyword search. W stands for a document.

IV. Experiments

We have conducted experiments on the customer reviews of five electronics products.

1. Digital Camera: Canon G3
2. Digital Camera: Nikon Coolpix 4300
3. Cellular Phone: Nokia 6610
4. Mp3 Player: Creative Labs Nomad Jukebox Zen Xtra 40GB
5. DVD Player: Apex AD2600 Progressive-Scan DVD Player

All the reviews are from amazon.com.

Table 1: Experimental Evaluation

Sr. No	Name of Products	Manual Features	Features extracted using proposed technique
1	Digital Camera: Canon G3	79	70
2	Digital Camera: Nikon Coolpix 4300	96	81
3	Celluar Phone: Nokia 6610	67	59
4	Mp3 Layer: Creative Labs Nomad Jukebox Zen Xtra 40GB	57	63
5	Dvd Player	49	41
	Average	67	62.8

Conclusion

The objective is to extract the product features from large number of customer reviews of a product sold online. Our experimental results indicate that the proposed techniques are very promising in performing their tasks. We believe that this problem will become increasingly important as more people are buying and expressing their opinions on the Web. Further work is to extract the sentiment polarity about these features expressed in the review.

References

[1] Hsinchun Chen, David Zimbra, "Artificial Intelligence and Opinion Mining," IEEE Intelligent Systems, ISSN 1541-1672, published by the IEEE Computer Society, Vol. 25 (3), pp 74-80, Oct. 2009.

[2] Esuli Baccianella Stefano, Esuli Andrea, Sebas-tiani Fabrizio, "SentiWordNet 3.0: An Enhanced Lexical Re-source for Sentiment Analysis and Opinion Mining". In Proceedings of the 7th Conference on Language Resources and Evaluation, pp 2200-2204, 2010.

[3] Chenghua Lin, Yulan He, Everson R., Ruger, "Weakly Supervised Joint Sentiment-Topic Detection from Text", Published in IEEE Transactions on Knowledge and Data Engineering,Vol. 24 (6), p.p. 1134 – 1145, June 2012.

[4] M.Hu and B. Liu, "Mining and summarizing customer reviews," in Proceedings of the Tenth ACMSIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04), pp. 168–177, August 2004.

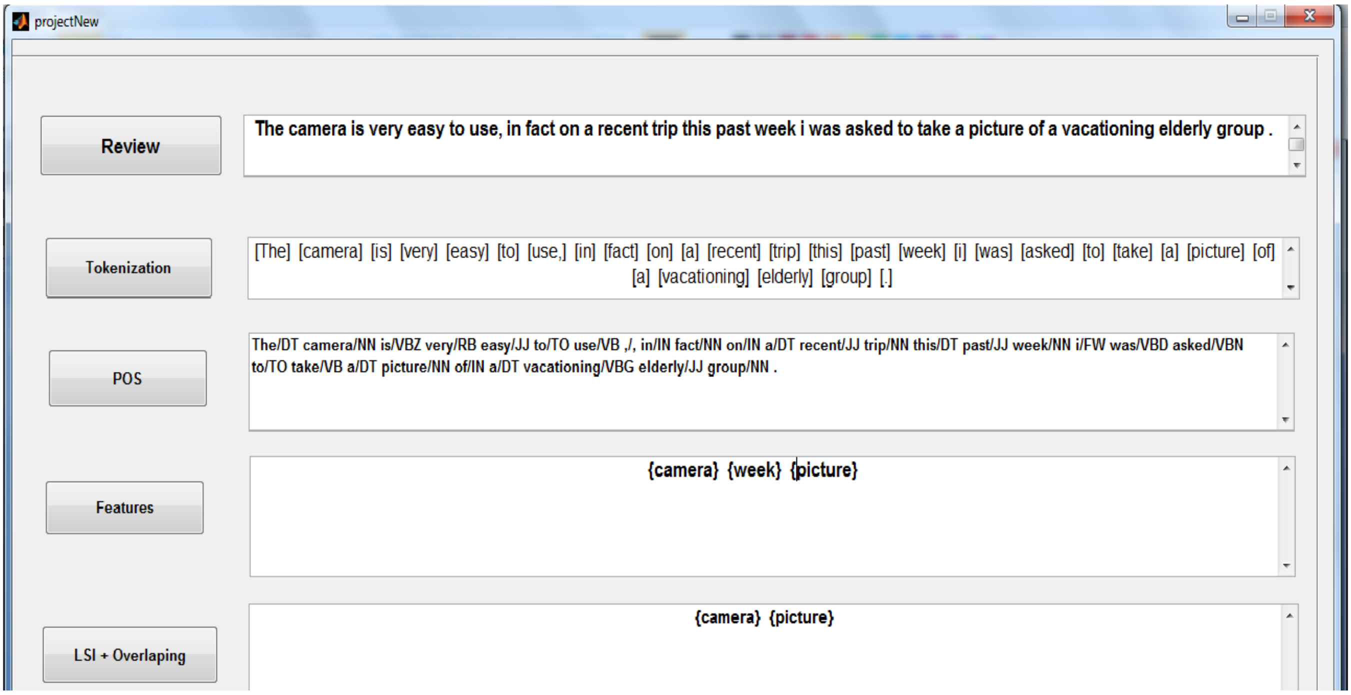
[5] Jingbo Zhu, Chunliang Zhang, Matthew Y. Ma, "Multi-Aspect Rating Inference with Aspect-Based Segmentation", IEEE Transactions on Affective Computing, pp. 469-480. 2012.

[6] H. Xu Xueke, Cheng Xueqi, Tan Songbo, Liu Yue, Shen Huawei, (2013), 'Aspect-level Opinion Mining of Online Customer Reviews", China Communications, IEEE publications, p.p. 28-41.

[7] Jingbo Zhu, Member, Huizhen Wang, Muhua Zhu, Benjamin K. Tsou (2011), "Aspect-Based Opinion Polling from Customer Reviews", published in IEEE Transactions on Affective Computing, Vol. 2 (1), p.p. 37-49.

[8] Jingbo Zhu, Chunliang Zhang, Matthew Y. Ma, (2012), "Multi-Aspect Rating Inference with Aspect-Based Segmentation", IEEE Transactions on Affective Computing, p.p. 469-480.

[9] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. Indexing by latent semantic analysis. Journal of the American Society for Information Science (1990).



The camera is very easy to use, in fact on a recent trip this past week i was asked to take a picture of a vacationing elderly group .

[The] [camera] [is] [very] [easy] [to] [use,] [in] [fact] [on] [a] [recent] [trip] [this] [past] [week] [i] [was] [asked] [to] [take] [a] [picture] [of] [a] [vacationing] [elderly] [group] [.]

The/DT camera/NN is/VBZ very/RB easy/JJ to/TO use/VB ,/, in/IN fact/NN on/IN a/DT recent/JJ trip/NN this/DT past/JJ week/NN i/FW was/VBD asked/VBN to/TO take/VB a/DT picture/NN of/IN a/DT vacationing/VBG elderly/JJ group/NN .

{camera} {week} {picture}

{camera} {picture}

Fig. 3. Matlab GUI for Extraction of Features from the given reviews

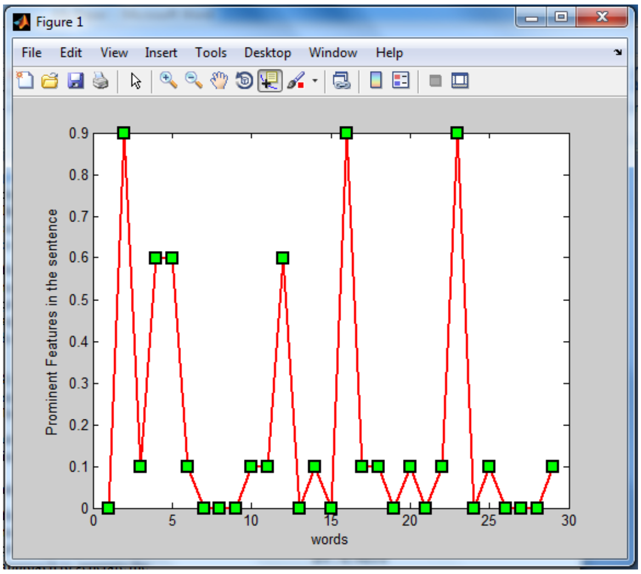


Fig. 4. X-Axis show the words in the given review and Y axis shows their probability of being Product Features. Threshold of 0.6 is kept

AUTHOR'S PROFILE

Nilesh Shelke had done MCA, M. Phil.(Comp. Sci), M.Tech(CSE). He is research scholor at S.G.B. Amravati University (INDIA). He is having 14 years of Experience in Education field. He is currently working in PIGCE, Nagpur(India). He had published many research papers in international and national journals and papers. He is member of ISTE, CSI etc.

Dr. S. P. Deshpande is currently working as Associate Professor at Post Graduate Department of Computer Science & Technology, MCA at Shree H. V. P .Mandal's DCPE, Amravati since last 15 years. He has published 25 papers in various national & International conferences & 10 papers in International journals. He has guided more than 100 students at Post Graduate level. His interest of research is DBMS, Data Mining, Web based technologies, AI and NLP. He holds Doctorate Degree from Sant Gadge Baba Amravati University, Amravati