Semisupervised Learning Based Opinion Summarization and

Classification for Online Product Reviews

**Prof. Rupali Kale, Sujith Nair, Onkaar Sawant, Apresh Pandit, Sanal Pillai**

*Abstract*—There are many users who purchase products through E-commerce websites. Through online shopping many E-commerce enterprises were unable to know whether the customers are satisfied by the services provided by the firm. This boosts us to develop a system where various customers give reviews about the product and online shopping services, which in turn help the E-commerce enterprises and manufacturers to get customer opinion to improve service and merchandise through mining customer reviews. Project presents a semi supervised approach for mining online user reviews to generate comparative feature-based statistical summaries that can guide a user in making an online purchase.

Keywords- summary,opinion,e-commerce

# Introduction

Nowadays, several websites are available on which a

variety of products are advertised and sold. Prior to making

a purchase an online shopper typically browses through

several similar products of different brands before reaching

a final decision. This seemingly simple information retrieval

task actually involves a lot of feature-wise comparison and

decision making, especially since all manufacturers advertise

similar features and competitive prices for most products.

However, most online shopping sites also allow users to

post reviews of products purchased.There are also dedicated

sites that post product reviews by experts as well as end

users. These user reviews if appropriately classified and

summarized can play an instrumental role in influencing a

buyers’ decision.

Project presents a semi supervised approach for mining online user reviews to generate comparative feature-based statistical summaries that can guide a user in making an online purchase.

In this paper, we apply a multistep approach to the

problem of automatic opinionmining that consists of various

phases like Preprocessing, Semantic feature-set extraction

followed by opinion summarization and classification.

The rest of the paper is organized as follows. Section 2

explores related work in the area of opinion mining, Section 3

describes the strategy used for opinion mining,

Finally, we conclude and discussthe scope for future work in this field.

# Related Work

Classification and summarization of online blog reviews are very important to the growth of E-commerce and social networking applications. Earlier work on automatic text summarization has mainly focused on extraction of sentences that are more significant in comparison to others in a document corpus. The main approaches used to generate extractive summaries are (1) combinations of heuristics such as cue words, key words, title words or position (2) lexical chains, and (3) rhetorical parsing theory.[1]

However, it is important to note that the task of summarizing online product reviews is very different from traditional text summarization, as it does not involve extracting significant sentences from the source text. Instead, while summarizing user reviews, the aim is to first of all identify semantic features of products and next to generate a comparative summary of products based on feature-wise sentiment classification of the reviews that will guide the user in making a buying decision. In the authors have demonstrated that traditional unsupervised text classification techniques like naive Bayes, maximum entropy, and support vector machine do not perform well on sentiment or opinion classification and pointed out the necessity for feature-oriented classification. Thus, recent research work in opinion mining has focused on feature based extraction and summarization.[3]

Opinion mining from users’ reviews involves two main tasks—(1) identification of the opinion feature set and (2) sentiment analysis of users’ opinions based on the identified features.

It has been observed that nouns and noun phrases (N and NP) frequently occurring in reviews are useful opinion features, while the adjectives and adverbs describing them are useful in classifying sentiment.[4]

In order to extract nouns, noun phrases, and adjectives from review text, parts-of-speech (POS) tagging is performed. However, all nouns and noun phrases are not useful in mining and cannot directly be included in the feature set. So, the feature set is subsequently extracted using approaches that involve frequency analysis and/or use of domain knowledge as is discussed next.[4]

Various methods exist in the literature to associate features with their corresponding descriptors. Hu and Liu proposed the nearby-adjective heuristic. Although this method is simple and fast, it may result in inaccuracies. So, supervised approaches to determine association have been proposed in recent years such as syntactic dependency parsing and syntactic tree templates.[2]

# Proposed Opinion Summarizer and Classifier:

A] Working:

We generated an opinion review database by crawling some popular websites that categorically post product reviews by actual users. Our product opinion summarizer has three main phases. These phases are

(1) preprocessing phase,

(2) feature extraction phase, and

(3) opinion summarization and classification phase.

These phases are briefly described next.

Preprocessing Phase:

Online blog reviews posted by users frequently contain spelling errors and incorrect punctuation. Our next phase—the feature-extraction phase—requires parts-of- speech tagging which works at the sentence level. Thus, it becomes important to detect end of sentences. So, in this phase we performed basic cleaning tasks like sentence boundary detection and spell-error correction. Sentences normally end with punctuations like period (.), question mark (?), or exclamation mark (!). Sometimes bloggers overuse the “?” and “!” symbols for emphasis. For example, a blogger may post a review that says

“It’s surprising that the eBook reader does not have a touch screen !!!!”

In such cases we conflate the repetitive punctuation symbols to a single occurrence (i.e., “!!!!” is replaced by a single “!”).

Several other considerations arise during the Preprocessing phase. The period (.) requires to be disambiguated as it may mean a full stop or a decimal point or an abbreviation (e.g., “Dr.,” “Ltd.”). Sometimes a single sentence straddles multiple lines as the user presses unnecessary return keys. In such cases we apply the sentence merge rules as proposed by Dey and Haque [1]. After sentence boundary detection, we perform spell-error correction using a word processor.

Feature Extraction Phase:

In this phase we extract opinion features from the pre-processed review text obtained From the previous phase. We treat frequently occurring nouns (N) and noun phrases (NP) as possible opinion features and associated adjectives describing them as indicators of their opinion orientation.

We perform parts-of-speech (POS) tagging on the review sentences using the Link Grammar Parser [3]. The Link Grammar Parser is a well-known and efficient syntactic parser for English language. First, we extract all nouns (N) and noun phrases (NP) tagged by the Link Grammar Parser and identify the frequently occurring N and NP as possible opinion features. By frequently occurring N and NP we mean those Ns and NP which occur at least five times in the users’ reviews. We do not extract frequent itemsets from review sentence database using the Apriori based approach, since this method mines frequent features using a BOW (bag-of- words) approach and does not take into account the order in which the words of a phrase occur. Moreover, mining in this way would require ordering besides compactness and redundancy pruning [4, 8]. We also do not use the seed-set expansion approach as it would require prior domain knowledge to specify a seed set. Instead we generate a frequent feature set using the multiword approach.

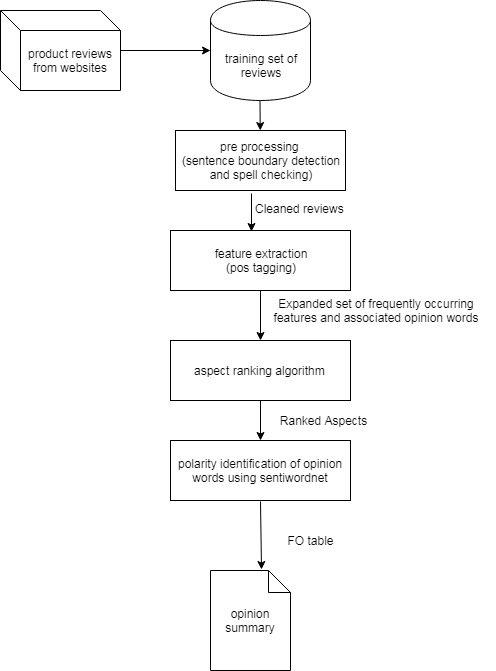
Opinion Summarization and Classification Phase:

In the previous phase we extracted opinion features, adjectives describing them, and any modifiers if present. We also generate a statistical feature-wise summary for each product which enables comparison of different brands selling similar products. In order to determine the sentiment polarity of an adjective describing an opinion feature we make use of SentiWordNet which is a lexical resource for opinion mining. SentiWordNet assigns three normalized sentiment scores: positivity, objectivity, and negativity to each synset of WordNet. Let us revisit the review sentence:

“The processor[.n] is[.v] significantly faster[.a], and the text[.n] is[.v] clear[.a].”

In this example, the SentiWordNet scores assigned to the appropriate usage of adjective clear is indicated as (P:0.625; O: 0.375; N: 0). Since the value of the positive polarity is highest, the adjective “clear” can be assigned a positive polarity. In this way, we generate a feature-orientation table (FO table) that records the opinion features and their corresponding descriptors of positive and negative polarities. The Table 1 shows the FO table entries for some of the features of product “Tablet.” The FO table, thus generated, enables us to generate feature-wise summary of a product or comparative summaries of different brands of similar products.

Fig 1.1: System Architecture Diagram [1]



B] Aspect ranking algorithm for identifying important aspects in review:

a) Terms used in Algorithm

•D={r1, r2,r3…rn} be the set of reviews.

•Ak= {a1, a2, a3,…… an} be the set of aspect

•Ca,D is the number of times aspect term a occurs in review

dataset D.

•Pa is the number of comments in the positively labeled set

with aspect term a. .

•|P| is the number of comments in the positively labeled set.

•Na is the number of comments in the negatively labeled set with aspect term a.

•|N| is the number of comments in the negatively labeled set.

•Va,D is the feature value for aspect term a in review dataset D.

•Let Φ = set of positive words

Φ= {P1, P2, P3 ... Pn}

•Let ψ = set of negative words

ψ ={N1, N2, N3 ... Nn}

•(𝛷)=probability of Φ

•𝑃 (𝜓)=probability of ψ

•ω weight of aspect a

b) Algorithm Steps

•Calculate the value of aspect a, given by

Va,D= Ca,D\*log2(|P|/ Pa)−Ca,D∗log2(|N|/ Na)

= Ca,D\*log2((|P| Na /Pa|N|)

= Ca,D\*log2(Na / Pa)

•Calculate the occurrence probability of each positively opinionated word.

n

α = ∑ P (Φi) ∗W(Φi)

𝑖=1

•Calculate the occurrence probability of each negatively opinionate word.

n

β = ∑ P (ψi) ∗W(ψi)

𝑖=1

•Calculate weight,

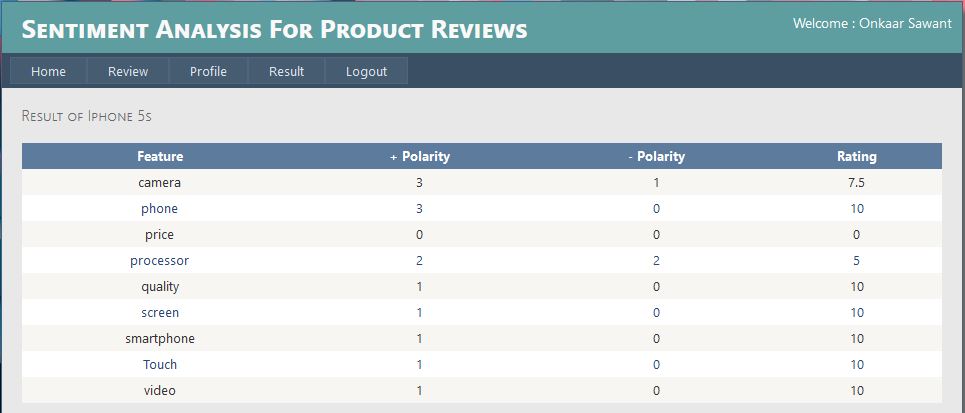
D

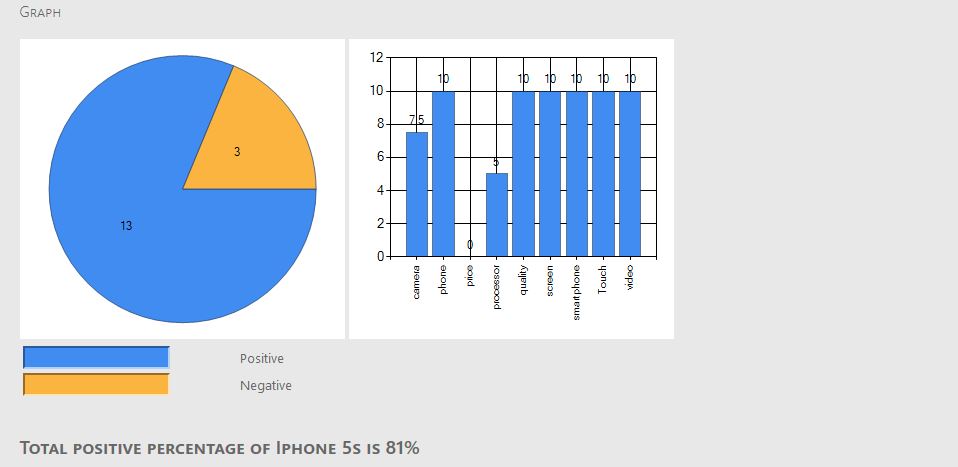
𝜔 = Va,D − ∑(α – β)

i=1

C] Results and Analysis

We were able to get the overall sentiment of a product review fetched online. The output included what percentage of the review was positive or negative, including the key feature words which lead to the decision taken by the system.

****



# CONCLUSION

In this paper we developed a semi supervised approach for ranking product aspects. SENTI WORDNET score is used to classify opinion on each aspect. Ranking algorithm generate a list of ranked aspects according to its importance. Result shows that proposed system gives appropriate output for various product reviews. In future work a larger data set can be used for in depth analysis and fuzzy models could be used for classifying review sentiments based on descriptive adjective (accelerators/deaccelators) word identification and linguistic hedges.

#### .

##### References

[1] L. Zhao and C. Li, “Ontology based opinion mining for movie reviews,” in Proceedings of the 3rd International Conference on Knowledge Science, Engineering and Management, pp. 204–214, 2009.

[2] A. Balahur, Z. Kozareva, and A. Montoyo, “Determining the polarity and source of opinions expressed in political debates,” in Proceedings of the 10th International Conference on Intelligent Text Processing and Computational Linguistics, vol. 5449 of Lecture Notes in Computer Science, pp. 468–480, Springer, 2009.

[3] Y. H. Gu and S. J. Yoo, “Mining popular menu items of a restaurant from web reviews,” in Proceedings of the International

Conference onWeb Information Systems andMining (WISM’11), vol. 6988 of Lecture Notes in Computer Science, pp. 242–250, Springer, 2011.

[4] M.Hu and B. Liu, “Mining and summarizing customer

reviews,” in Proceedings of the 10th ACMSIGKDD International

Conference on Knowledge Discovery and Data Mining (KDD

‘04), pp. 168–177, August 2004.

[5] M. A. Jahiruddin, M. N. Doja, and T. Ahmad, “Feature and opinion mining for customer review summarization,” in Proceedings of the 3rd International Conference on Pattern Recognition and Machine Intelligence (PReMI ‘09), vol. 5909 of Lecture Notes in Computer Science, pp. 219–224, 2009.

[6] S. Shi and Y. Wang, “A product features mining method based

on association rules and the degree of property co-occurrence,” in

Proceedings of the International Conference on Computer Science

and Network Technology (ICCSNT ’11), pp. 1190–1194,

December 2011.

[7] S. Huang, X. Liu, X. Peng, and Z. Niu, “Fine-grained product

features extraction and categorization in reviews opinion mining,” in Proceedings of the 12th IEEE International Conference on Data

MiningWorkshops (ICDMW’12), pp. 680–686, 2012.

[8]C.-P.Wei,Y.-M.Chen,C.-S.Yang, and C. C.Yang, “Understanding what concerns consumers: a semantic approach to product feature extraction fromconsumer reviews,” Information Systems and e- Business Management, vol. 8, no. 2, pp. 149–167, 2010.