

# Feature Ranking in Sentiment Analysis

Maryam K. Jawadwala  
P.G. Student,  
Department of Computer Engineering,  
Thadomal Sahani Engineering College,  
Bandra, Mumbai, India

Seema Kolkur  
Assistant professor,  
Department of Computer Engineering,  
Thadomal Sahani Engineering College,  
Bandra, Mumbai, India

## ABSTRACT

With the rapid expansion of e-commerce over the past 15 years, more products are sold on the Web. More and more people are buying products online. In order to enhance customer shopping experience, it has become a common practice for online merchants to enable their customers to write reviews on products that they have purchased. Some popular products can get hundreds of reviews or more at some large merchant sites. Manual analysis of customer opinions is only possible to a certain extent and very time-consuming due to the multitude of contributions. From the e-commerce perspective, receiving consumer's feedback can greatly improve its strategies in order to increase products of the sector. This research work will present feature wise sentiment analysis of customer review. The goal of feature level sentiment analysis is to produce a feature-based opinion summary of multiple reviews. With summaries of opinions and features of the product, people can make effective decisions in less time. Such mining can be helpful for competitive marketing. Feature extraction can be performed using two approaches. Rule-based algorithm and HAC algorithm. Feature ranking will be done using MAX opinionscore algorithm and opinion score obtained from SentiWordNet.

## Keywords

Sentiment analysis, Opinion mining, Feature ranking, Natural language processing

## 1. INTRODUCTION

Opinion Mining is a field of Web Content Mining that aims to find valuable information out of users opinions. Mining opinions on the web is a fairly new area, and its importance has grown significantly mainly due to the fast growth of e-commerce, blogs and forums. The World Wide Web has grown exponentially in recent years both in terms of size and diversity of the contents provided [1]. It has contributed a very large amount of data termed as user generated content. These new contents include customer reviews, blogs, and discussion forums which expresses customer satisfaction/dissatisfaction on the product and its features explicitly.

Most of the time the customer does not directly indicate the choice in a straight forward manner but does so in sentences which contain the actual reviews along with lines which are general in nature and has nothing to do about the product or opinion. Such sentences are challenging due to many reasons like, user not writing the features explicitly, writing incorrect sentences, omitting punctuation marks and writing grammatical incorrect language [12]. As customer feedback influences other customer decisions about buying the product, these feedbacks have become an important source of information for businesses when developing marketing

strategies and segmenting the customers. The difficulty lies in the fact that majority of the customer reviews are very long and their numbers are also very high which makes the process of distillation of knowledge a very difficult task. Most of the times a user will read a few reviews and will try to make a decision about the product. The chances that a user will end up taking a biased decision about the product are not ruled out. Similarly, manufacturers want to read the reviews to identify what elements of a product affect sales most and what are the features the customer likes or dislikes so that the manufacture can target on those areas. More importantly, the large number of reviews makes it hard for product manufacturers or business to keep track of customer's opinions and sentiments on their products and services.

There are many areas where sentiment analysis can be used as following [12]:

- 1) A company is interested in customer's perceptions about its products and the information may be used to improve products and identifying new marketing strategies. Sentiment Analysis is used to find these customers' perception about product from the thousands of review.
- 2) Tourists want to know the best places or famous restaurants to visit. Sentiment analysis can be used to obtain relevant information for planning a trip.
- 3) By applying sentiment analysis we can detect the users opinion from the posted movie reviews on specialized sites.

## 1.1 Sentiment Classification

There are three types of opinion mining approaches [8].

- [1] Feature level or Phrase level  
In this, for the product, the particular features are classified and for those features, the comments or reviews are taken separately.
- [2] Sentence level  
In this, the comments or reviews are opinionated. The benefit of this approach is in this, the customer can come to know about so many different types of customer's reviews. In this approach, it mainly differentiates between the subjective and objective information. The subjective information is the opinion, which can be negative or positive and the objective information is the fact.
- [3] Document level  
In this the whole document is written for the product, it is written by only one person. So, it is not as useful because the customer will come to know the review of only one customer.

## 2. DATA SOURCE

User's opinion is a major criterion for the improvement of the quality of services rendered and enhancement of the deliverables. Blogs, review sites, data and micro blogs

provide a good understanding of the reception level of the products and services [7].

## 2.1. Blogs

With an increasing usage of the internet, blogging and blog pages are growing rapidly. Blog pages have become the most popular means to express one's personal opinions. Bloggers record the daily events in their lives and express their opinions, feelings, and emotions in a blog (Chau & Xu, 2007). Many of these blogs contain reviews on many products, issues, etc. Blogs are used as a source of opinion in many of the studies related to sentiment analysis (Martin, 2005; Murphy, 2006; Tang et al., 2009).

## 2.2. Review Sites

For any user in making a purchasing decision, the opinions of others can be an important factor. A large and growing body of user-generated reviews is available on the Internet. The reviews for products or services are usually based on opinions expressed in much unstructured format. The reviewer's data used in most of the sentiment classification studies are collected from the e-commerce websites like www.amazon.com (product reviews), www.yelp.com (restaurant reviews), www.CNET download.com (product reviews) and www.reviewcentre.com, which hosts millions of product reviews by consumers. Other than these the available are professional review sites such as www.dpreview.com, www.zdnet.com and consumer opinion sites on broad topics and products such as www.consumerreview.com, www.epinions.com, www.bizrate.com (Popescu & Etzioni, 2005; Hu, B.Liu, 2006; Qinliang Mia, 2009; Gamgaran Somprasertsi, 2010).

## 2.3. Dataset

Most of the work in the field uses movie reviews data for classification. Movie review data's are available as dataset <http://www.cs.cornell.edu/People/pabo/movie-review-dat> a). Other dataset which is available online is multi-domain sentiment (MDS) dataset. (<http://www.cs.jhu.edu/mdredze/datasets/sentiment>). The MDS dataset contains four different types of product reviews extracted from Amazon.com including Books, DVDs, Electronics and Kitchen appliances, with 1000 positive and 1000 negative reviews for each domain. Another review dataset available is <http://www.cs.uic.edu/liub/FBS/CustomerReviewData.zip>.

## 2.4. Micro-blogging

Twitter is a popular micro blogging service where users create status messages called "tweets". These tweets sometimes express opinions about different topics. Twitter messages are also used as data source for classifying sentiment.

## 3. LITERATURE REVIEW

### 3.1 Feature Extraction using Rule-based Algorithm.

The aim of the feature level sentiment analysis is to identify feature-wise good and bad aspects of a given product. This can be a useful practical solution to allow customers to help decide how well a product satisfies his/her needs if they are only looking for few important features in a product and don't care about other features.

The system consists of five major modules [1] shown in figure 1. The working principles of these modules are explained in the following sub-sections.

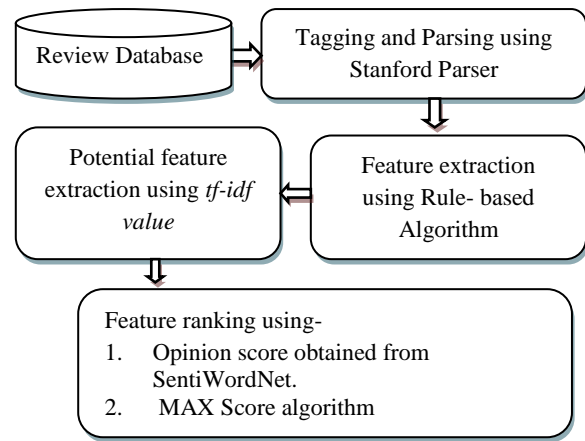


Fig-1: Architecture of the proposed opinion mining system

#### 3.1.1 Document Parser

All subjective sentences are parsed using Stanford Parser, which assigns Parts-Of-Speech (POS) tags to English words based on the context in which they appear.

#### 3.1.2 Feature Extraction

The information component extraction mechanism is implemented as a rule-based system [2], which analyzes dependency tree to extract information components. Each review is expressed in a triplet <F, M, O> where, F is a noun phrase and O is adjective word possibly representing product feature. M represents adverb that act as modifier and used to intensify the opinion.

#### 3.1.3 Potential Feature Extraction

Potential features are extracted using term frequency (*tf*) and inverse document frequency (*idf*). The *tf-idf* value for each noun phrase is calculated using equations 1 and 2.

$$tf - idf(t_i) = tf(t_i) \times idf(t_i) \quad 1.$$

$$idf(t_i) = \log \left( \frac{|D|}{|\{d_j : t_i \in d_j\}|} \right) \quad 2.$$

Where,  $tf(t_i)$  is the number of documents containing  $t_i$   
 $|D|$  is the total number of documents and  
 $|\{d_j : t_i \in d_j\}|$  is the number of documents where  $t_i$  appears.

All those noun phrases having *tf-idf* value above a threshold are considered as relevant features. After that, for each product feature, the list of all opinions and modifiers is compiled, these are used later for polarity determination of the opinion sentences.

### 3.2 Feature Extraction using HAC Algorithm.

Another feature extraction technique is **The High Adjective Count algorithm (HAC)** [3]. The main idea behind the algorithm is that the nouns for which reviewers express a lot of opinions are most likely to be the important and distinguishing features than those for which users don't

express such opinions. The Max Opinion Score algorithm ranks the extracted features using the opinion scores assigned in the HAC.

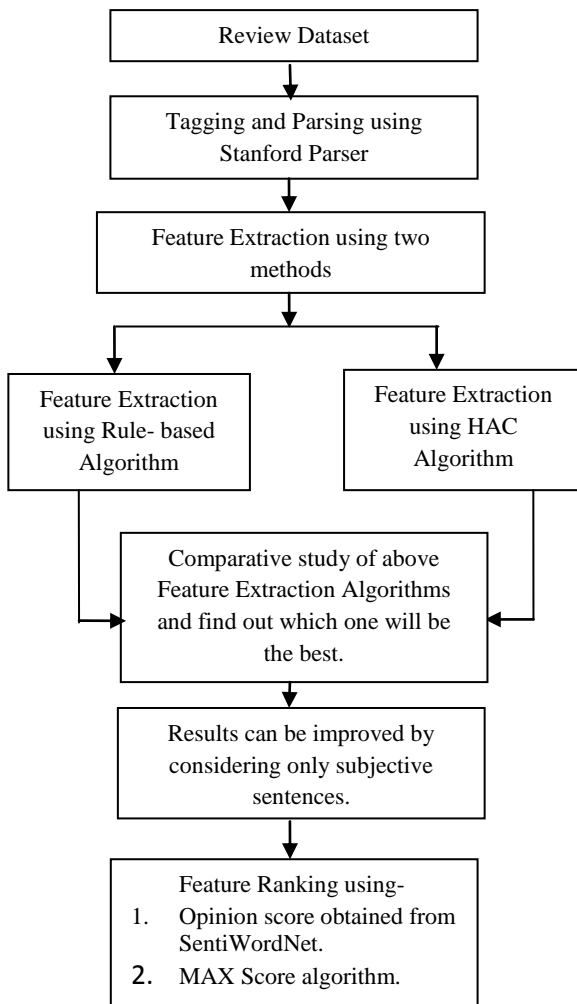
### 3.2 Feature Extraction using H-Mine Algorithm.

In [4] frequent feature extraction is done using H-Mine [5]. Steps for feature extraction using H-Mine are as follows:-

1. Pre-processing of words including removal of stop words and stemming is performed.
2. All subjective sentences are parsed using Stanford Parser, which assigns Parts-Of- Speech (POS) tags to English words based on the context in which they appear.
3. For feature extraction, H-Mine [5] algorithm is used. It is fast and space-preserving frequent pattern mining concept.

### 4. SENTIMENT ANALYSIS SYSTEM.

Figure 2 shows block diagram for system implementation.



**Fig-2: Block diagram for system Implementation.**

Each review sentence is parsed using Stanford parser, which assign POS tags to English words based on the content in which they appear.

1. Battery life is awesome.
2. Although camera is 8 mega pixels, picture quality is not good.
3. Screen size is fantastic.
4. Quality of sound is the best.
5. Good screen size with low cost.
6. Battery life is satisfactory as compared to other android phone.
7. Touch screen is not very impressive.
8. Quality of sound is really good.
9. Camera zoom is not good.
10. Battery life is poor.
11. Touch screen is wonderful.
12. Quality of sound is not good.
13. A decent phone.
14. Low cost phone.
15. Camera does not capture clear picture during night.

**Fig.3: Reviews for Mobile.**

### 4.1 Rule-based Algorithm

Opinion words and product features are related to each other by semantic relations, so each sentence is converted into Stanford Dependency (SD) using Stanford parser [11].SD of reviews shown in figure 3 is shown in figure 4.

1. Battery life is awesome.  
 Batter\_NN life\_NN is\_VBZ awesome\_JJ \_.  
 nn(life-2, Battery-1)  
 nsubj(awesome-4, life-2)  
 cop(awesome-4, is-3)  
 root(ROOT-0, awesome-4)
2. Screen size is fantastic.  
 Screen\_NN size\_NN is\_VBZ fantastic\_JJ \_.  
 nn(size-2, Screen-1)  
 nsubj(fantastic-4, size-2)  
 cop(fantastic-4, is-3)  
 root(ROOT-0, fantastic-4)

**Fig. 4: SD for partial reviews of figure 3.**

Each review is divided into triplet < F, M, O > F- Feature, M- modifier, O- opinion using information component extraction algorithm. It can be implemented using following rules.

**Rule 1:** In a dependency tree T, if there exists a subj(wi, wj) relation such that POS(wi) = JJ\*, POS(wj) = NN\*, wi and wj are not stop-words then wj is assumed to be a feature and wi as an opinion. Thereafter, the relation advmod(wi, wk) relating wi with some adverbial words wk is searched. In case of the presence of advmod relation, the information component identified as <wj, wk, wi> otherwise <wj, -, wi>.

**Rule 2:** In a dependency tree T, if there exists a subj(wi, wj) relation such that POS(wi) = VB\*, POS(wj) = NN\*, and wj is not a stop-word then we search for acomp(wi, wm) relation. If acomp relation exists such that POS(wm) = JJ\* and wm is not a stop-word then wj is assumed to be a feature and wm as an opinion

**Rule 3:** In a dependency tree T if there exists amod(wi, wj) such that POS(wi) = NN\*, and POS(wj) = JJ\* then wi is assumed to be a feature and wj as an opinion. Further if advmod(wi, wk) exists then wk is assumed to be a modifier.

**Rule 4:** In a dependency tree T if there exists NN(wi, wj) such that POS(wi) = NN\*, and POS(wj) = NN\* and Nsubj(wk, wi) exists where POS(wk) = JJ\* or amod(wi, wk) exists where POS(wk) = JJ\* then wj,wi (two words) is assumed to be a feature and wk as an opinion.

Figure 5 represents information components extracted by applying these rules on each SD of reviews shown in figure 4.

1. < (Battery life), - ,awesome > //Apply Rule 4 and Rule 1.
2. < (picture quality), - ,good > //Apply Rule 3 and Rule 4.
3. < (screen size), - ,fantastic > //Apply Rule 4 and Rule 1.
4. < quality, - ,best > //Apply Rule 1.

**Fig. 5: Triplet <F, M, and O> using rule based system.**

Potential features is extracted using term frequency (tf) and inverse document frequency (idf).The tf-idf value for each noun phrase is calculated using equations 1 and 2.

All those noun phrases having tf-idf value above a threshold is considered as relevant features. After that, for each product feature, the list of all opinions and modifiers is compiled. It is used later for polarity determination of the opinion sentences. For example partial list of product feature, opinion and modifiers extracted on mobile review is look like table-1.

**Table-1: A list of extracted features, opinions and modifier.**

Product	Feature	Modifier	Opinion
Mobile phone	Battery life	---	Awesome, satisfactory, poor
	phone	as	Other, decent
	screen	very	Impressive, wonderful
	Quality	really	Good, best

### 4.2 HAC Algorithm

The second method of feature extraction is HAC algorithm. The main idea behind the algorithm is that the nouns for which reviewers express a lot of opinions are important features. List of extracted feature using HAC for threshold value >=2 is look like table-2.

**Table-2: A list of extracted features and opinions.**

Product	Feature	Opinion
Mobile phone	Battery life	awesome, satisfactory, poor
	size	fantastic, low
	phone	android, decent
	screen	touch, wonderful

### 4.3 Improved Algorithm

According to Pang and Lee [9] subjective sentences are expressive of the reviewer's sentiment about the product, and objective sentences do not have any direct or obvious bearing on or support of that sentiment. Therefore, the idea of subjectivity analysis [10] is used to retain segments (sentences) of a review that are more subjective in nature and filter out those that are more objective. This increases the system performance both in terms of efficiency and accuracy.

Improved system divide the review into subjective sentence and objective sentence using objective score obtained from SentiWordNet [6].

$$\text{Objective score} = 1 - (\text{positive score} + \text{negative score})$$

Objective score of each sentence is the average of opinion words present in sentence. It is in the range of 0 to 1. Subjectivity/objectivity analysis is done using following observations –

If (Objective score of sentence > 0.7)

Then sentence is Objective.

Else if (Objective score of sentence <= 0.7)

Then sentence is Subjective.

### 4.3 Feature Ranking

Feature ranking is performed using following steps-

The polarity of extracted opinions for each feature is classified using

1) Opinion score obtained from SentiWordNet [6].

2) Assign polarity manually in the range of [-4, 4], known as MaxOpinionScore algorithm.

The overall weight of a feature is calculated by multiplying the polarity value of the opinion word with the number of sentences which contain that opinion. This is mention in equation 3.

$$TotalWt = \sum_{n=1}^d (WPF - WNF) \quad 3.$$

WPF = Wt of positive features.

WNF = Wt of negative features.

At last feature wise ranking is performed based on overall weight.

## 5. EXPERIMENTAL RESULTS

Review dataset is taken from <http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>. The dataset consists of 2200 reviews of 11 different products. Each product consists of 200 reviews. Available dataset is a XML file. A java program is written to convert XML file to database file. Parsing and tagging is done using Stanford parser.

First method of feature extraction using Rule-based system is implemented. Result of Rule-based algorithm can be improved by considering only subjective sentences of review database. Subjectivity/objectivity analysis is done using SentiWordNet. Objective score of each opinion word is calculated using positive score and negative score obtained from SentiWordNet. Feature ranking is done using opinion score obtained from SentiWordNet. Result is shown in table 3. Feature ranking using MaxOpinionScore algorithm is done for Rule-base algorithm. It manually assigns adjective score in the range of -4 to 4. Result is shown in table 4.

Second method of feature extraction using HAC algorithm is implemented. Result of HAC algorithm can be improved by considering only subjective sentences of review database. Subjectivity/objectivity analysis is done using SentiWordNet. Feature ranking is done using opinion score obtained from SentiWordNet. Result is shown in table 5. Feature ranking using MaxOpinionScore algorithm is done for HAC algorithm. It manually assigns adjective score in the range of -4 to 4. Result is shown in table 6.

Table 3: Comparative study of Rule- based algorithm and improved algorithm, ranking using SentiWordNet.

Sr. No.	Product name	Subjective and Objective sentences				Subjective sentences only			
		(Score using SentiWordNet)				(Score using SentiWordNet)			
		Positive features	Rank	Negative features	Rank	Positive features	Rank	Negative features	Rank
1	Canon-G3	resolution design lense computer software magnisium finish	7.0 6.25 6.0 3.75 3.0	auto mode card period shooting strap	-1.0 -1.668 -2.0 -2.0 -2.5	design lens resolution computer software magnisium finish	6.25 6 5.5 3.75 3.0	auto mode macro period strap card	-1.0 -1.0 -2.0 -2.5 -2.668
2	Canon-PowerShot-SD500	quality balance image photo quality size	17.25 9.0 8.0 3.5 3.0	test video cnet reviewer photographer pocket	-1.5 -2.0 -2.25 -2.612 -3.5	quality balance image photo quality size	15.25 9 5.5 3.5 3.0	test video cnet reviewer photographer pocket	-1.5 -2.0 -2.25 -2.612 -3.5
3	Nikon-Coolpix-4300	battery life mode battery size lense	8.25 5.75 5.25 5.25 3.5	unit	-1.0	battery life mode size battery lense	8.25 5.25 5.25 5.0 3.5		
4	Canon-S100	camera battery case software resolution quality	31.5 8.5 6.0 5.75 5.5 5.5	start picture quality range	-1.5 -4.0 -4.25	camera case software battery resolution quality	28.75 6.0 5.75 5.5 5.5 5.5	start picture quality range	-1.5 -4.0 -4.25

Table 4: Comparative study of Rule- based algorithm and improved algorithm, ranking using MaxOpinionScore.

Sr. No.	Product name	Subjective and Objective sentences				Subjective sentences only			
		(Score using MaxopinionScore)				(Score using MaxopinionScore)			
		Positive features	Rank	Negative features	Rank	Positive features	Rank	Negative features	Rank
1	Canon-G3	resolution quality design lens control	8.0 8.0 6.25 6.0 4.25	macro auto mode period shooting strap	-1.0 -1.0 -2.0 -2.0 -3.0	quality design lens resolution control	8.0 6.25 6.0 6.0 4.25	auto mode macro period card strap	-1.0 -1.0 -2.0 -2.668 -3.0
2	Canon-PowerShot-SD500	quality balance image picture time	16.75 9.0 9.0 7.0 7.0	test video photographer cnet reviewer pocket	-2.0 -2.0 -2.112 -2.25 -3.0	quality balance time picture time	14.75 9.0 7.0 7.0 7.0	video test photographer cnet reviewer pocket	-2.0 -2.0 -2.112 -2.25 -3.0
3	Nikon-Coolpix-4300	camera quality battery life mode setting size	38.25 9.0 8.25 7.25 6.5 6.25	unit	-1.0	camera quality battery life mode size setting	35.25 9.0 8.25 6.25 6.25 5.5		
4	Canon-S100	battery case quality software resolution	11.0 6.0 6.0 5.25 5.0	start picture quality range	-2.0 -4.0 -4.25	battery case quality software resolution	7.0 6.0 6.0 5.25 5.0	start picture quality range	-2.0 -4.0 -4.25

Table 5: Comparative study of HAC algorithm and improved algorithm, ranking using SentiWordNet.

Sr. No.	Product name	Subjective and Objective sentences				Subjective sentences only			
		(Score using SentiWordNet)				(Score using SentiWordNet)			
		Positive features	Rank	Negative features	Rank	Positive features	Rank	Negative features	Rank
1	Canon-G3	Quality design lens image quality magnesium finish resolution	6.5 6.5 6.0 6.0 4.5 4.0	lens unit period shooting strap	-1.5 -2.0 -2.0 -2.5	Camera Design lens image quality magnesium finish resolution	15.5 6.5 6.0 6.0 4.5 2.5	lens unit period strap	-1.5 -2.0 -2.5
3	Canon-PowerShot-SD500	Quality Auto Color sound image quality	26.5 6.0 5.0 4.0 4.0	Noise test photographer	-0.5 -1.5 -2.6	quality auto color sound image quality	24.0 6.0 5.0 4.0 4.0	noise test	-0.5 -1.5
4	Nikon-Coolpix-4300	quality mode battery life size resolution	11 7.0 5.5 4.5 4.0	power slave flash	-0.5 -1.0	quality mode battery life size resolution	11.0 6.5 5.5 4.5 4.0		
10	Canon-S100	battery case quality resolution picture quality	11.0 8.0 6.5 5.5 4.0	start choice	-1.5 -3.0	battery case quality resolution picture quality card	8.0 8.0 6.5 5.5 4.0 4.0	start choice	-1.5 -2.5

Table 6: Comparative study of HAC algorithm and improved algorithm, ranking using MaxOpinionScore.

Sr. No.	Product name	Subjective and Objective sentences both				Subjective sentences only			
		(score using MaxOpinionscore)				(score using MaxOpinionscore)			
		Positive features	Rank	Negative features	Rank	Positive features	Rank	Negative features	Rank
1	Canon-G3	Quality lens design resolution magnesium finish	7.0 6.0 6.0 5.0 5.0	shooting period lens unit strap	-2.0 -2.0 -2.0 -3.0	design lens quality magnesium finish resolution	6.0 6.0 5.0 5.0 3.0	period lens unit strap	-2.0 -2.0 -3.0
2	Canon-PowerShot-SD500	quality light color image quality zoom	26.0 7.0 5.0 4.0 4.0	noise test photographer	-1.0 -2.0 -2.11	quality light color image quality zoom	23.0 7.0 5.0 4.0 4.0	noise test photographer	-1.0 -2.0 -2.11
3	Nikon-Coolpix-4300	camera quality mode battery life size resolution	26.0 11.0 8.0 6.0 5.0 4.0	power slave flash	-1.0 -1.0	camera quality mode battery life size resolution	21.0 11.0 7.0 6.0 5.0 4.0		
4	Canon-S100	camera battery case quality resolution picture quality	23.0 13.0 8.0 7.0 5.0 4.0	start choice	-2.0 -4.0	camera battery case quality resolution picture quality	19.0 9.0 1.0 7.0 5.0 4.0	start choice	-2.0 -3.0

## 6. EVALUATION OF SYSTEM

Manual evaluation has been performed to judge the overall performance of the system. For evaluation of the experimental results, Standard IR performance measures precision, recall and f1-measure are used. The values of the above performance measures are calculated using formula 4, 5 and 6 for canon-G3 product.

**Precision ( $\pi$ ):** the ratio of true positives among all retrieved instances.

$$\pi = \frac{TP}{TP + FP} \quad 4.$$

**Recall ( $\rho$ ):** the ratio of true positives among all positive instances.

$$\rho = \frac{TP}{TP + FN} \quad 5.$$

**F1-measure (F1):** the harmonic mean of recall and precision.

$$F1 = \frac{2\rho\pi}{\rho + \pi} \quad 6.$$

Table 7 and table 9 summarize the performance measure values for existing system and table 8 and table 10 summarize the performance measure values for improved system. From the above experiments, it is clear that the improved sentiment analysis system provide better precision recall and f1-measure values compare to existing system.

**Table 7: Performance measures for HAC algorithm.**

Product	Features	Precision	Recall	F1-Measure
Canon – G3	Quality	72%	40%	52%
	Lens	50%	50%	50%
	Magnesium finish	50%	50%	50%
	Resolution	66%	50%	57%

**Table 8: Performance measures for improved HAC algorithm.**

Product	Features	Precision	Recall	F1-Measure
Canon – G3	Quality	100%	40%	57%
	lens	67%	50%	57%
	Magnesium finish	67%	50%	57%
	Resolution	100%	50%	67%

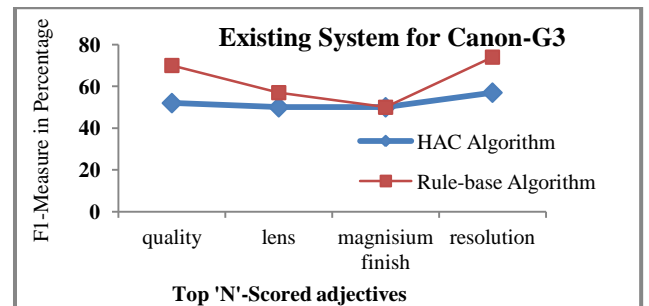
**Table 9: Performance measures for Rule-based algorithm.**

Product	Features	Precision	Recall	F1-Measure
Canon –G3	Quality	78%	64%	70%
	Resolution	74%	74%	74%
	lens	67%	57%	62%
	Magnesium finish	50%	50%	50%

**Table 10: Performance measures for improved Rule-based algorithm.**

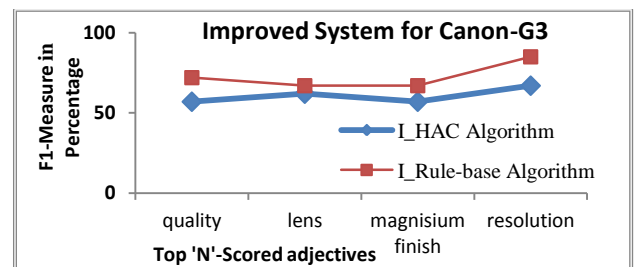
Product	Features	Precision	Recall	F1-Measure
Canon –G3	Quality	86%	60%	72%
	Resolution	100%	74%	85%
	lens	67%	67%	67%
	Magnesium finish	100%	50%	67%

Figures 6 present the f1-measure for different number of features (N), for Canon-G3 review set. From the above experiments, it is clear that the Rule-based algorithm is much better than the HAC algorithm.



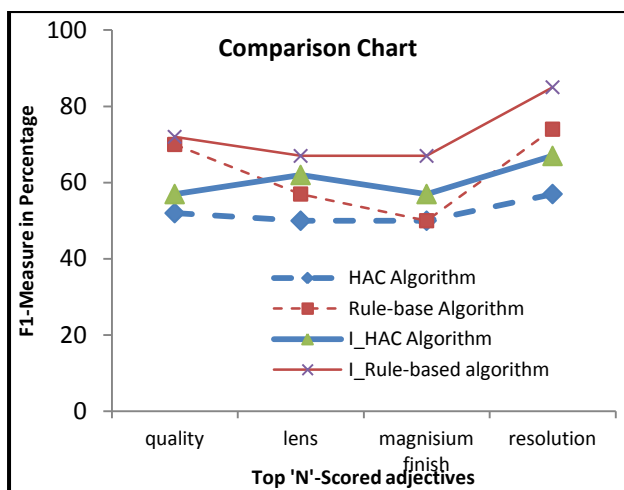
**Fig. 6: Comparison of HAC algorithm and Rule-based algorithm.**

Figures 7 present the f1-measure graph for different number of features, for Canon-G3 review set. From the above experiments, it is clear that the improved system is much better than the existing system. Comparison graph is shown in figure 8.



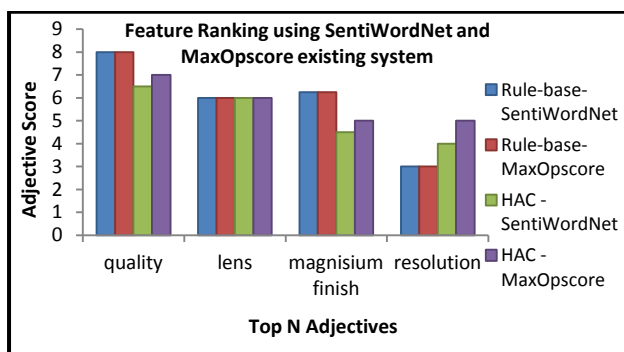
**Fig. 7: Comparison of I\_HAC algorithm and I\_Rule-based algorithm.**



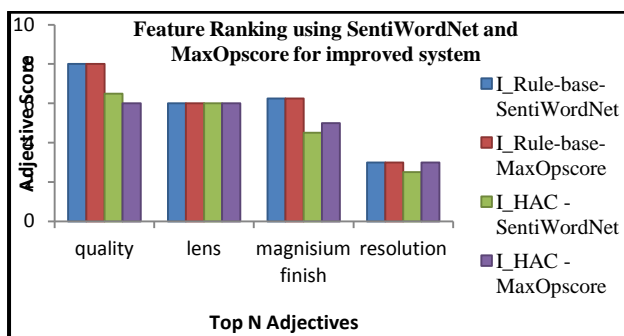


**Fig. 8: Comparison of existing system and improved system.**

Feature ranking is performed using opinion score obtained from SentiWordNet and Maxopinionscore algorithm. Figure 9 and 10 shows MaxOpinionScore algorithm gives better ranking than SentiWordNet.



**Fig. 9: Feature ranking for existing system.**



**Fig. 10: Feature ranking for improved system.**

## 7. CONCLUSION

Sentiment detection has a wide variety of applications in information systems, including classifying reviews, Summarizing review and other real time applications. It is found that different types of features and classification algorithms are combined in an efficient way in order to overcome their individual drawbacks and benefit from each other's merits, and finally enhance the sentiment classification performance.

In future, more work is needed on further improving the performance measures. Sentiment analysis can be applied for new applications. Although the techniques and algorithms used for sentiment analysis are advancing fast, however, a lot of problems in this field of study remain unsolved. The main challenging aspects exist in use of other languages, dealing with negation expressions; produce a summary of opinions based on product features/attributes, complexity of sentence document, handling of implicit product features, etc. More future research could be dedicated to these challenges.

## 8. REFERENCES

- [1] Tanvir Ahmad , Mohammad Najmud Doja, **Ranking System for Opinion Mining of Features from Review Documents**, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012 , ISSN (Online): 1694-0814.
- [2] Tanvir Ahmad , Mohammad Najmud Doja, **Rule Based System for Enhancing Recall for Feature Mining from Short Sentences in Customer Review Documents**, IJCSI International Journal of Computer Science Issues, Vol. 4 No. 06 June 2012, ISSN : 0975-3397.
- [3] Magdalini Eirinaki, Shamita Pisal1, Japinder Singh2, **Feature-based Opinion Mining and Ranking**, Elsevier, Journal of Computer and System Sciences 78 (2012) 1175–1184, 3 November 2011.
- [4] Seyed Hamid Ghorashi, Roliana Ibrahim2, Shirin Noekhah and Niloufar Salehi Dastjerdi , **A Frequent Pattern Mining Algorithm for Feature Extraction of Customer Reviews** IJCSI international Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012 , ISSN , (Online): 1694-0814.
- [5] J. Pei, J. Han, H. Lu, S. Nishio, S. Tang Oand D. Yang, **H-Mine: Fast and space-pre serving frequent pattern mining in large databases**, "IEE TRANSACTIONS, pp. 593-605, 2007.
- [6] A.Esuli, and F.Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining", in Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation, Genova, IT, 2006, pp. 417-422.
- [7] Avani Jadeja, Prof. Indr jeet Rajput, **Feature Based Sentiment Analysis On Customer Feedback: A Survey**, International Journal of Engineering Research & Technology (IJERT) ,Vol. 2 Issue 4, April – 2013, ISSN: 2278-0181.
- [8] Gandhi Vaibhav C. ,Neha Pandya, **Feature Level Text Categorization For Opinion Mining**, International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 5, May - 2013 ISSN: 2278-0181.
- [9] B.Pang, B. and L.Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts", in Proceedings of ACL 2004, 2004, pp. 271-278.
- [10] Alaa Hamuda Mohamed Rohaim Department of Systems and Computers Engineering Al\_Azhar, **Reviews Classification Using SentiWordNet Lexicon**, The Online Journal on Computer Science and Information Technology (OJCSIT), Reference Number: W11-0123.
- [11] Stanford log-linear part-of-speech tagger, 2010, <http://nlp.stanford.edu/software/tagger.shtml>.
- [12] M.Hu, and B.Liu, "Mining and Summarizing Customer Reviews", in Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04), USA, 2004, pp. 168 – 177.