# Semantic Analysis based Customer Reviews Feature Extraction

Ghaidaa A. Bilal Rasha N. Sha`llan

Department of Software, College of Information Technology Babylon University of Babil Ghaidaa.almulla@fulbrightmail.org rasha@itnet.uobabylon.edu.ig

## Abstract

A set of customers' reviews about restaurants has been analyzed syntactically and semantically for deducing syntactic, contextual and semantic features to leverage the textual similarity metrics. In this paper an approach for rule based extracting semantic features from customer's reviews have been proposed. The features were extracted based on the knowledge base, co-occurrence and distributional similarity among the reviews' aspects and descriptors. The approach was applied on the Yelp academic challenges dataset and the results have shown encouraging performance.

Keywords: Aspect Detection, Text Analysis, Features Extraction, Text Mining.

#### الخلاصة

قد تم تحليل مجموعة من آرآء العملاء عن المطاعم نحويا لغويا واستنتاج الخصائص النحوية، والدلالية السياقية ثم الاستفادة منها لتكون مقاييس التشابه. في هذه الورقة قد استخدم طريقة الظهور المشترك وطرق التشابة المعتمدة على معجم البيانات (The Word Net) وطرق التشابه الموزعة لأستخراج خصائص دلالية للاسماء المستخلصة من الاراء وتوصيفاتها (aspects ( aspects)،والتي اشتقت من خلال بناء قواعد نحوية تم تهيئتها لهذا الغرض. النتائج كانت مشجعة جدا للاستثمار في مختلف مجالات تعدين النص. تم استفادة توفيرها من قبل Yelp.com

الكلمات المفتاحية : تحديد الاسماء , تحليل النص, استخلاص الميزات, تعدين النص .

# **1** Introduction

Recent advances in e-Commerce have emerged new challenges for researchers in terms of processing the customers' opinions and feedback on various characteristics of products[Rao and Shah,2015].Mostly, people who are looking for business merchandise, depend on other Customers' reviews as an important resource for making buying decisions [Zheng *et al.*, 2015; Rao and Shah,2015]. However, sifting through large websites-which offer lots of reviews and ratings about specific service-to comprehend the majority opinions about that service is a tedious and time consuming task. Hence, opinion summarization research areas that exploit the incremental accumulation of user-generated content have increased rapidly [Rao and Shah,2015; Gupta *et al.*,2015]. It consists of three stages: the features identification, features tagging and information summarizing [Rao and Shah, 2015].

The big challenge with features mining of text like blogs or reviews is the text noisiness including spelling errors, informal expressions, abbreviations and improper punctuations. Many of the research dealt with the textual data as linguistically correct. Whereas other efforts suggested new approaches towards structuring these noisy data e.g. supervised and unsupervised techniques, and the latter has proved its usefulness to work on domain independent feature extraction [Rao and Shah,2015].In addition, text mining algorithms are considered a significant manner to process, reduce and mine for valuable information from the text in transactions and documents that is not always be necessary to go through all of it [Witten *et al.*,2011].The more extracted valuable information

/features is extracted, the better analysis, understanding and summarizing the document is [Popescu and Etzioni, 2005]. The concept of aspects mining is one of the attempts to extract aspects and analyze its sentiment using the pair aspect – sentiment or aspect – descriptor [Zheng *et al.*, 2015].

The proposed research presents customer reviews text features extraction based on the semantic analysis, syntax and contextual awareness on a set of obtained rule based aspects. Experiments have been tested on the Yelp academic dataset. The proposed algorithm results have shown encouraging performance with regards to the state of the arts methods.

The paper is organized as follows: Section 2 is an outline of related work, section 3 is the proposed algorithm for creating aspects- descriptors pairs and generating rules are discussed. Section 4 presents the experiments and results, and Section 5 discusses conclusions, and identifies future research.

#### 2 Related Works

The problem of detecting semantic similarity in text has led the researchers to give the Opinion analysis much attention recently where Li et al., in [Li et al., 2003] presented a a function that combined three similarity attributes path length, depth, and local density between two words in wordNet or in the Brown Corpus, but they Ignored leveraging the contextual features in their work. Consequently the authors in [Tingting Wei, et al, 2015] employed WordNet to detect the similarity between terms with lexical chains to extract semantic related words from texts, then Bisecting K-means clustering algorithm was performed, while Yadav, et al., in [Yadav et al., 2014] exploited WordNet in their work where all nouns in text file are extracted and then a graph was built to show relations (synonym, hyponym / hyponym and metonymy/ homonymy) among the nouns based on WordNet Ontology. Where the nodes implemented nouns and the arcs implemented relations among these nouns, whereas Lin et al., in [Lin et al., 2015] tried to extract opinion lexicons from reviews and identifying the sentiment polarities of the words based on word vector and matrix factorization. The TF-IDF feature and Cosine function was utilized as similarity metrics. The authors missed the semantic analysis in their research, in which the identification of the relations among the vocabularies might strengthen the similarity process. Besides that, the widespread of lexical taxonomy domain may have irregular densities of links between concepts. While efforts of leveraging text features continue, Li et al., in [Sun and Soergel, 2015] leveraged the terms' relations using word co-occurrence and TF-IDF method to identify a set of hierarchical relations among terms. They tried employing the keywords as concepts source to build text taxonomy, while In [Bharti and Singh, 2015] Top ranked features were selected and union was applied on it while on remaining features sub lists the intersection was applied. The following methods were used to select features term variance (TV) and document frequency (DF) then they applied principal component analysis (PCA) method for reducing dimensions of the features. The authors in[Gupta et al., 2015] adopted a method in which the aspects and it's descriptors of the Customer Reviews were extracted based on syntax rules and clustered based on three features were distributional similarity, co-occurrence and knowledge base. We have modified that method by building a new syntax rules that extracted the aspect - descriptor in a height accuracy and in two directions forward propagation and back propagation. Where the aspect was extracted as e.g. "wooden pool table", the proposed forward propagation has enhanced the identification and extraction for the aspect and it's descriptor to be "pool table" as aspect and "wooden" as descriptor. The semantic approach was applied in "knowledge base" feature only while our method has taken into account the semantic approach in co-occurrence feature where the exactly matching, synonyms and hyponyms of each aspect have treated as one aspect. The results have proved that our method helped the users to perform semantic search for better understanding to the reviews content.

# **3** The proposed system

We have adopted an approach to leverage Semantic –Syntax relations based customers reviews for getting the most importing terms and it's descriptors while gaining features that have helped in analyzing the reviews content semantically as shown in figure 1.



Figure 1: The proposed system

# **3.1** Preprocessing stage

- Sentences detection: The input to this stage is a set of online customers reviews each review document has segmented into sentences .The reviews were filtered and any unrelated information with the customers opinion was removed. This process has taken into account sorting the sentences according to their reviews.
- **Tokenization:** Each sentence for each review has separated in to set of tokens "terms". The separation process has adopted the segregation of each sentence into a set of single tokens (uni-gram) in which each token may identify the sentence and the review that belongs to.
- Stop words and delimiters removing: Stop words are words with less weighting in the reviews with no specific rules to be considered for identification of those words. The researchers can select list of words to be stop words according to their work domain. There are many copies of stop words such as a stop word list that provided by website of Journal of Machine Largening Research; it consists of 571 words. In

this paper, the adopted dataset has suffered from noisiness including spelling errors, informal expressions, abbreviations and improper punctuations. Hence, the list of stop words and delimiters was modified to get rid of the above-mentioned noises and to filter the sentences tokens. A sample of the words that were considered as stop words is: "bla", "\n", "goooood", "where", "a", "the".

• **POS tagging:** in corpus linguistics a Part-Of-Speech Tagger is the operation of encoding the text words as corresponding to a particular part of speech. The tagging operation is contingent on both the words definition and context, i.e. relationship among adjacent words in a phrase, sentence, or paragraph. Taking the following example: "John often gives a book to Mary" that has tagged as: John/NNP often/RB gives/VBZ a/DT book/NN to/TO Mary/NNP. The table 1 shows some of the tags symbols and it's meaning.

Tag	Meaning	Tag	Meaning	Tag	Meaning
CC	Coordinating conjunction	JJS	Adjective, superlative	SYM	Symbol
CD	Cardinal number	NN	Noun, singular or mass	ТО	to
DT	Determiner	NNP	Proper noun, singular	VB	Verb, base form
NNS	Noun, plural	PDT	Pre-determiner	VBD	Verb, past tense
VBG	Verb, gerund or present participle	PRP	Personal pronoun	FW	Foreign word
IN	Preposition or subordinating conjunction	RB	Adverb	VBN	Verb, past participle
VBP	Verb, non¬3rd person singular present	RBR	Adverb, comparative	JJ	Adjective
JJR	Adjective, comparative	RBS	Adverb, superlative	VBZ	5Verb, 3rd person singular present

**Table 1: The Part of Speech Tags Meaning** 

It is clear from the example above that a tagger is basically a classifier where it considers text as input and returns the parts of speech for all its tokens (words) classified as verb, adverb, adjective, and noun ... etc. The online version of Stanford parser that has been used in this research is available at <u>http://nlpstanford.edu:8080/parser/</u>.

• **Chunking:** It is a technique widely used in natural language processing for sentence analysis and constituents (noun groups, verbs, verb groups, etc.) identification. However, it doesn't specifies their internal structure nor their role in the main sentence. It is similar to the concept of lexical analysis in computer languages translators. A unigram chunker simply assigns one chunk tag to each POS tag where in The IOB representation every token is in a chunk or Out of a chunk. Chunks work

just like tags and a chunker is basically a tagger as shown in figure 2 with the sentence: We saw the yellow dog.

We	s a w	t h e	y e l l o w	d o g
PRP	VBD	DT	٦J	NN
B-NP	0	B-NP	I-NP	I-NP

Figure 2	(chunking	process)
I Iguit 2	(Chunking	process

Where DT = B-NP, NN = I-NP, VB = O

Parse tree: A natural language parser is a program that applies the grammatical ٠ structure on sentences. For instance, it separates the sentence words into groups according to its kind – a noun phrase or a verb phrase. The figure 3 and table 2 illustrates an example on the parsing process and the parsing symbols respectively.



Figure 3	(narsing	process)
i igui e J	(pai sing	process

Table2 . Taising Symbols			
S → NP VP	$VP \rightarrow V$		
$S \rightarrow Aux NP VP$	PP -> Prep NP		
NP → Det Nom	N → old   dog   footsteps		
	young		
NP → PropN	$V \rightarrow dog   include   prefer$		
Nom -> Adj Nom	Aux → does		
Nom $\rightarrow$ N	Prep →from   to   on   of		
Nom $\rightarrow$ N Nom	PropN → Bush   McCain		
	Obama		
Nom → Nom PP	Det → that   this   a  the		
VR VNP	Adj -> old   green   red		

Table2 :	Parsing	Symbols

# **3.2** Aspects - Descriptor Extraction Stage or Rules Generation Stage

- **Definition (aspect)** is an entity component, attribute or term in which a review might be expressed explicitly e.g., "restaurant service" or "computer properties". The nuances of aspects can be captured by analysis the sentiment for those aspects. The identification of aspects has performed automatically by syntactic methods where each sentence in every review is scanned for nouns to be considered as aspects.
- **Definition (Descriptor)** essentially noun aspects can be described by adjectives, whether they occur before the noun (e.g. "*Poor customer service*") or after auxiliaries (e.g. "*the texture was nice*") to express the aspect meaning. Participles could also be used as adjectives to the aspects, for example "*The fish was fried*".
- **Definition** (Aspect –Descriptor pair) is a pair consists of aspect and descriptor for that aspect, e.g. "*the waitress was rude*", the pair is (waitress, rude). Sometimes the aspect may have more than one descriptor e.g. "*The hot soup was delicious*", then the pairs will be (soup, hot) and (soup, delicious). However, if no aspect has been recognized but there is a descriptor has presented in the sentence, e.g. "*They were friendly*" then a rule is built for treating this case by extracting the pair as (business, friendly) where business refer to the whole service provided by the business that the reviewer talks about.

## **3.2.1 Rules based Aspects – Sentiments Pairs**

A set of syntactic rules has been built to identify and extract pairs of aspects and descriptors for each review based on generated syntactic rules as illustrated in algorithm1. It follows two directions in generating the rules: backward exploration and forward exploration.

<i>Nam</i> : rules generating algorithm
nput: List of parsed sentences
<b>Dutput</b> : Set of aspect-descriptor pairs
Step 1: do while for each sentence [1] in all reviews
Step 2: if chunk of token [1] is "NP" and tag is ("NN" or "NNS" or
"VBG") then save token[I] as aspect
1) Step 3: Search backwards:
a. If tag of token $[J] = "JJ"$ or "VBN" or "VBG" and If it negative or
positive then save token[J] as descriptor.
=I+1 // forward propagation
check If token[I] is "JJ" or "VBN" or "VBG" and If it negative or positive then
ave token[I] as descriptor
Else
f tag of token[I] is "PRP" then save it as "For business" to be aspect and return
o step 6
Else if token[I] is one of Connectivity Tools such as "but" or "while" etc. then
reat it as new sentence and return to step 2
End for
End while
R) $Fnd$

# **Algorithm1: Rules Generating**

#### Journal of Babylon University/Pure and Applied Sciences/ No.(3)/ Vol.(25): 2017

**Backward Exploration**: in this part of the algorithm, chooses the candidate pair starting from detecting the noun phrase in each sentence to select the elected aspect. Next backward search goes back to the beginning of the sentence looking for any nominee descriptors that is located in front of the aspect. As described in the mentioned algorithm, the descriptors should be an adjective or past participle, e.g." *They had prepared a delicious chicken*", where the extracted pair is (chicken, delicious).

**Forward Exploration**: In this direction, the tokens located behind the elected aspect are tracked to extract new descriptors. Typically, the picked descriptors presents the token located after one of the auxiliary verbs, **e.g.** "The waitress was rude". The aspect-descriptor pair has extracted as (waitress, rude).

The checking continues if a conjunctive exists such as but, and, while, which, ... etc. However, it is treated as a new sentence, where the identification and detecting process of (aspect-descriptor) pair will be repeated. **e.g.** "nice texture but the service was bad", the aspects-descriptors pairs that would extracted are (texture , nice) and (service , bad). Additionally, the proposed algorithm has treated the case of presenting a pronoun in the sentence via considering the pronoun as the elected aspect with the caption "for business". This caption reflects the reviewer's opinion on the whole services **e.g.** "It was amazing", the aspect-descriptor pair here has acted as (for business, amazing).

#### 3.2.2 Normalization

In order to extract the features with high accuracy, all aspects must be transformed into uniform case. This mean all words "aspects " and "descriptors" have transformed into capital letters or into lower letters. In this stage the transformation has been performed is to the lower letters, where if words have differed by the letters case small or capital, after the transformation, they will have been treated similarly e.g. food and Food have transformed into food.

#### 4.3 Semantic based feature extraction Stage

#### 4.4.1 Context of aspect based similarity

the literarture showed that the aspects, which co-occur in the same context, are mostly related and belong to the similar group [Gupta *et al.*, 2015]. The consideration of the semantic occurrence of the compared aspects is missed. Hence, in this paper at first all synonyms and hyponyms for each aspect have been identified and repetition among these synonyms and hyponyms have been deleted. Then the context information for all sentences in the review have been gathered into a context vector, that is used for comparison the semantic co-occurrence of the aspects and their synonyms and hyponyms with all the other aspects have presented in the same review. The association strength for each two aspects in the context vector has been measured by the Point wise Mutual Information (PMI), in which the frequency of the two aspects that appear in the reviews together has compared to their frequencies separately, as shown in equation 1.

$$PMI(x, y) = \log\left(\frac{p(x, y)}{p(x) \cdot p(y)}\right) \tag{1}$$

The computation process of the aspect's context similarity has presented in algorithm 2.

*Name* : Context similarity of the aspects algorithm **Input** : Record of Aspects **Output** : Record of Aspects with PMI value Begin **Step1**: For I= first Aspect to the last Aspect begin 1)Take aspect [1] with all other aspects 2) For J = I + 1 to the last Aspect 3)Convert aspects to lower case *4)find the synonyms and hyponyms for aspect [I]* 5)Compute Probability of occurrence and co-occurrence of aspect[1] and it's synonyms and hyponyms with all other aspects. 6) Apply equation 2 of each pair for aspects and save in PMI End if End for End

#### Algorithm 2: context similarity of the aspects

#### 4.4.2 External Knowledge base based similarity

The semantic similarity between two aspects has been identified by the Word Net knowledge base developed at the University of Princeton. The Word Net is a large English lexical database in which verbs, nouns, adjectives and adverbs are organized by a diverse of semantic relations into synsets (synonym sets) that indicate to one concept [Pratik et al., 2014]. It has used several semantic relations such as synonymy, autonomy, hyponymy, and so on. These relations can be used for word form relation or for semantic relation as a hierarchy structure, for which the Word Net is regarded as a good tool for natural language processing. The wordNet has provided four types of relations among nouns that may occur. The first relation is hyponym/hypernym relation that denoted as (is-a) relation e.g."Ali is a boy". The second one is meronym/part holonym relations (part-of) e.g."battery is part of mobile". While the third relation is expressed by (member-of) member meronym/member holonym relations that defines the relationship between a two terms one of them denoting the whole and the other denoting a part of, or a member of, e.g. the relation between the head and body. The last type of relation is the (substance-of) meronym/substance holonym relation, which identifies how a word or phrase is used to stand in another word e.g. "The pen is mightier than the sword". A fragment of the concepts (is-a) relations has been shown in figure 4.



#### Figure 4: fragment of the concepts (is-a) relations

Some of basic measures definitions in WorNet :

(1) len(ci,cj): the shortest path length from synset ci to synset cj in WordNet.

(2) depth(ci): the path length to synset ci from the global root entity, and depth (root)=1.

(3) deep\_max: the max depth(ci) of the taxonomy

(4) hypo(c): the hyponyms number for a given concept c.

WordNet semantic similarity measures have been grouped in four classes types: path length based measures, feature based measures, information content based measures, and hybrid measures. At this feature, the shortest path based measure has adopted, where the sim(ci,cj) has considered the closeness ci and cj in the taxonomy as shown in equation 3.

(3)

 $sim_{path}(c_1, c_2) = 2*deep \max - len(c_1, c_2)$ 

where deep\_max is a fixed value. The similarity between c1, c2 is len(c1,c2) from c1 to c2. If len(c1,c2) is 0, simpath(c1,c2) gets the maximum value of 2\*deep\_max. If len(c1,c2) is 2\*deep\_max, simpath (c1,c2) gets the minimum value of 0. Thus, the values of simpath (c1, c2) are between 0 and 2\*deep\_max. The extraction of this feature has presented in algorithm 3.

#### Name : Knowledge Base Based Similarity Input : set of aspects Output : set of aspects with Word Net similarity values . Begin Step1 : For I= first Aspect to the last Aspect 1) Take aspect [1] with all other aspects 2)Compute shortest bath of two aspects Step5 : Save result End

#### Algorthim3: Knowledge Base Based Similarity

#### 4.4.3 Distributional Similarity of Descriptors

The relationship of the aspects could be reflected by their descriptors, where descriptors may provide virtual contexts similarity to the unrelated aspects that neither have co-occurrence in the same contexts nor have relation could be identified by WordNet. To extract this feature a word-to-word similarity - normalized PMI- metric that used in [Gupta *et al* ., 2015] has been adopted to indicate the semantic similarity among the descriptors of two aspects in the all reviews as exhibited in equation 4. Some descriptors which do not reflect their aspects or are considered as common words such as "good", "bad", and so on, were ignored from the comparison because they may affect negatively on the results.

Where,  $A_1$  and  $A_2$  are aspects, d is descriptor, N are the aspects' total number,  $n_d$  the number of aspects that d appears with. The algorithm4 has described the extraction of this feature.

$$sim(A_1, A_2) = \frac{1}{2} \left( \frac{\sum_{d \in A_1} (maxSim(d, A_2) * log(N/n_d))}{\sum_{d \in A_1} log(N/n_d)} + \frac{\sum_{d \in A_2} (maxSim(d, A_1) * log(N/n_d))}{\sum_{d \in A_2} log(N/n_d)} \right)$$
(4)

Name : Distributional Similarity of Descriptors
Input : Set of aspects ' descriptors
Output : Set of Aspects with Maximum Similarity values based on its ' descriptors

Begin

Step1 : For I= first Aspect to the last Aspect

1)Take aspect [I] with all other aspects
2) For J= first Descriptor of Aspect [I] to the last Descriptor
3) For K= first Descriptor of Aspect [I+1] to the last Descriptor
4)Convert Descriptors to lower case
5) Compute Probability of occurrence and co-occurrence of Descriptor[J] and

Descriptor[K]

6) Apply equation 4 on the Descriptors pair and Save the result
7)select max result from the previous saved results for each two aspects
8) Apply equation 5 on max result and save result to be similarity value of the pair of aspects

Algorithm4 : Distributional Similarity of Descriptors

# **5** The experiments and results

The experiments of this research have been implemented using Java platform on NetBeans IDE 8.0.2. This program has used Stanford POS tagger and WordNet Ontology for finding the relations (synonym, hyponyms) between the words and for providing semantic similarity among the aspects. The input of the proposed system is set of online businesses reviews from Yelp website "Yelp Dataset Challenge 2014"<sup>1</sup>. The reviews had been written in informal language where the customer didn't care to the language rules, e.g. one of the customers started his review with" bla bla bla" to indicate the food was bad and some of them didn't care to the spelling e.g. "goooood ", abbreviations, improper punctuations such as " \nOpen " and some adjectives had returned as noun e.g." cold". POS tagger has considered all these outliers as noun. The reviews has been considered to be input to the proposed system are 152 reviews, 1237 sentences have been detected, while aspects have been extracted are 1363 and the total descriptors for all aspects are 1475. Table3 shows a sample of the extracted aspect- descriptor pairs.

The detected sentence	Aspect	Descriptors
The waitress was rude	Waitress	Rude
Very poor customer service	Customer service	Poor
Everything on my wife's plate was stale	Plate	Stale
The chicken was tough and the soup had no	Chicken	Tough
flavor	Soup	Had no flavor
The glass was not cleaned inside or outside for quite some time	Glass	Not cleaned
The parking lot looked busy	Parking lot	Looked busy
The prices are reasonable, and the owners are	Prices	Reasonable
very friendly	Owners	Friendly

 Table 3: Aspect – Descriptors Pairs

The results that have been presented in table3 shows that the ability of generated rules to extract the aspect and descriptors in a high efficiency, where if there are more than one aspect or descriptor in the sentence. A sample of the features extraction has shown in table4.

Aspect1	Aspect2	Word Net	PMI	Distributional Similarity of Descriptors
Waitress	Food	0.090909090909090909	0	0
Waitress	Hostess	0.1111111111111111	0	1
Food	Cashier	0.142857142857142	1.9684489712313	0
Food	Meal	0.333333333333333333	0	1
Hostess	Shrimp	0.1666666666666666	0	0
Plate	Rib	0.333333333333333333	0	0
Service	Restaurants	0	0	0.968576925045153
Service	Meal	0.090909090909090909	0	0.487239038084645
Chicken	Soup	0.1111111111111111	0	0
Sauce	Patty	0.1111111111111111	0	0
Meal	Owners	0	0	0

 Table 4: The Aspects' Features values

# Journal of Babylon University/Pure and Applied Sciences/ No.(3)/ Vol.(25): 2017

It is noted that, the related aspects have a WordNet similarity value ranging from 0.1 and above, while aspects with no clear relation, other features could be used to discover if there is a similarity among them e.g. "service" and" restaurants" the similarity between them has been discovered by distributional similarity of descriptors feature. The case of PMI feature based on occurrence and co- occurrence for the aspects, therefor most of its' values are 0.

# 6 Conclusion and Future Work

This paper has introduced a developed approach for discovering the aspects and extracting the descriptors of these aspects by building a set of syntax that rules have treated all the syntax cases of the sentences and extracted aspects-descriptors pairs purely from reviews as shown in table 3 and then these aspects and descriptors have been analyzed semantically and the features values have been extracted where the resulting values have given an overview of similarity between aspects as its shown in table 4.Currently we work on clustering stage to evaluate the whole system steps. In future it is possible to treat the case in which the aspect has extracted as "for business" and also use these features for clustering these aspects.

#### References

Gelbukh (Ed.): CICLing, 2015, Part II, LNCS 9042, pp. 241-256.

- Gupta, P., Kumar, S., and Jaidka, K., 2015, "Summarizing Customer Reviews through Aspects and Contexts. In: Springer International Publishing Switzerland.
- Hoang,H.H.,Kim,S.N.,Kan,M.Y.,2009,: A re-examination of lexical association measures. In:Proceedings of Workshop on Multiword Expressions (ACL-IJCNLP), pp. 31–39.
- Kristina, T., Dan, K. Manning, C., and Singer, Y.2003, "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network" HLT-NAACL, pp. 252-259.
- Li, S.,*et al.*,2015,:A new method for automatically constructing domain oriented term taxonomy based on weighted word co-occurrence analysis. In :Springer Scientometrics , DOI 10.1007/s11192-015-1571-0.
- Li,Y.H.,Bandar, Z.,McLean, D.,2003,: An Approach for Measuring Semantic Similarity Using Multiple Information Sources. IEEE Trans. Knowledge and Data Eng. 15 (4), 871–882.
- MillerG.A., 1995, WordNet: A Lexical Database for English Comm. ACM, 38(11), 39–41.
- Popescu, A.M., Etzioni, O., 2005, :Extracting product features and opinions from reviews. In: Proceedings of Human Language Technology Conference/ Conference on Empirical Methods in Natural Language Processing, HLT-EMNLP 2005, Vancouver, CA, pp. 339–346.
- Pratik, K. Agrawal, Dr. Avinash. J. Agrawal, 2014,: Opinion Analysis Using Domain Ontology for Implementing Natural Language Based Feedback System. In: I.J. Information Technology and Computer Science, 10.5815.
- Rada,M.,Courtney,C.,Carlo,S.,2006,:Corpus-based and knowledge-based measures of text semantic similarity. In: Proceedings of the 21st National Conference on Artificial Intelligence, vol.1, pp.775–780.
- Rao,A.,Shah,K.,2015,:Model for Improving Relevant Feature Extraction for Opinion Summarization.
- Witten, I., Frank, E., Hall, M., 2011, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco.

Yadav, C.S., et al., 2014,: Semantic graph based approach for text mining. In IEEE.

Zheng, L., Weiping, W., Xiaolong, J., Jiguang, L., Dan, M., 2015, A Word Vector and Matrix Factorization Based Method for Opinion Lexicon Extraction.