ACCENTS Transactions on Information Security, Vol 1(3)
ISSN (Online): 2455-7196

Research Article

Extraction of key/title/aspect words from document using wordnet

Sheikh Muhammad Saqib*, Tariq Naeem and Khalid Mahmood
Institute of Computing and Information Technology, Gomal University, Dera Ismail Khan, Pakistan

Abstract
Changeling research problems are associated with opinion mining but it has also potential benefits as defines by different researchers. In opinion mining, key source for inputs are documents. These are either stored document or from different sites. In opinion mining, if documents are aspects or key words based, then results will be high accurate. Now issue is that how these key words can be accessed? Here we proposed a framework, which can determine key/ title or aspects words by creating a similarity path matrix of words. We use SentiWordNet and WordNet to calculate such matrix. These words not only helpful for document searching but also are useful for clustering, and lexicon making.

Keywords
Opining mining, Aspect based sentiment analysis, SentiWordNet, WordNet.

1.Introduction
In opinion mining field, when we want to analyze any documents i.e. Summary of document or sentiment analysis of document, besides all other efforts first of all it is very important to find out that either selected document belongs to required key words or aspects or entities. Most of the work has been done by aspect based analysis [1-3]. Support Vector Machine SVM can recognize and analyze the aspects for sentiment classification [16], Supervised Machine Learning algorithm focuses on mining relevant information from reviews has been done for aspect based sentiment analysis [17, 18]. The work [19] has shown the hybrid classification for aspect based classification. To check either, selected document is relevant to required aspect; first of all we will extract aspects from selected document and then compare those aspects with required aspects. How we can extract such aspects, there is lot of methods in literature proposed by different researchers. In [4] a novel rule-based approach to extract an aspect from reviews of product has been proposed. In [5] unsupervised approaches to find out polarity of an aspect from different domains have been proposed. In [6, 10, 11] summarization of aspects with respect to sentiment score has been proposed. Some researchers have explored aspect from a document with respect to subjective words [12, 13] means a word which have no polarity can be considered as an aspect.

*Author for correspondence

A machine learning techniques has been used to find out rating estimation of an aspect for reviews. Information distance was the asset for such rating [7]. Researchers have extended the aspect based opinion mining which was for physical products [8]. These extended works have used more complex natural language processing (NLP) based rules for online tourism product reviews [9]. A supervised learning algorithm has been used to find out important aspects and predictions with respect to reviews of such item [14].

Logic Programming, particularly answer set programming (ASP), has been used to elegantly and efficiently implement the key components of syntax based aspect extraction [15]. Aspect detection is more important so it must be identified accurately. Subjective words considered as aspect [12] requires more effort to find sentiment score for all words. Machine learning, NLP based rules, ASP can also provide better solution for such issue but aspects/topics are the key words of a paragraph and can be determined with less efforts. Words found by proposed framework not only be the aspects but also relates to key/title words, which can be helpful for document searching or clustering.

2.Framework key words extraction
Here we will take the objective words, because these neutral words which have minimum positive or negative concepts. In opinion mining, objective words can be categorized as:

1. A word which have greater objective score from sum of positive and negative scores. ObjectiveScore > 1 – (PosScore + NegScore)
2. A word which have zero positive and negative scores.
3. A word which have no positive, negative and even no objective score i.e. not belongs to synset [21, 22].

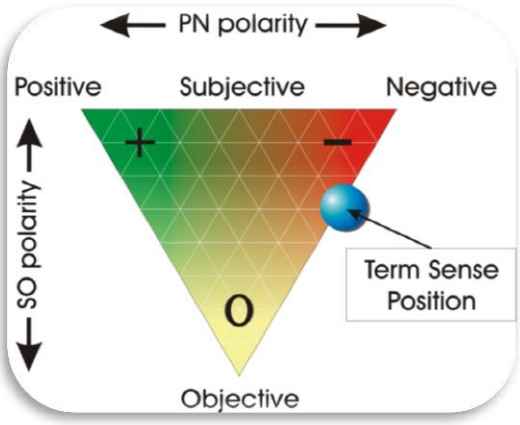

Figure 1 The graphical representation SENTIWORDNET for representing the opinion-related properties of a term sense [22]

From above three points, 2nd and 3rd points related to objective words are very important to extract title/key words also shown in *Figure 1*. But these objective words should be NN (NNS) or NNP (NNPS). Their description is shown in following *Table 1*.

Table 1 Part-of-speech (POS) tags

S no	Tags	Description
1	NN	Noun, singular or mass
2.	NNS	Noun, plural
3.	NNP	Proper noun, singular
4.	NNPS	Proper noun, plural

NN, NNS and NNP are the most important POS for identifying keywords. In Table: there are four POS, complete table is can be downloaded from internet [20]. These words can be used in different senses, there positive and negative scores may be different in different senses. It is important to calculate the sum of positive and negative score from all sense of each tag.

Then filter all those words who have accumulative zero positive and negative score. Now there will be long list of these words. Next process is to create a matrix of these words. First column will determine the frequency of words and remaining columns will represent the similarity distance of the words as shown in flowing *Table 2* where F means frequency and D means similarity path.

Table 2 Sample of similarity path matrix

	Frequency	W1	W2	W3	W4
W1	F1	D1W1	D1W2	D1W3	D1W4
W2	F2	D2W1	D2W2	D2W3	D2W4
W3	F3	D3W1	D3W2	D3W3	D3W4
W4	F4	D4W1	D4W2	D4W3	D4W4

Where F1 to F4 means frequency of 1st to 4th words. D2W1 means similarity path between word2 and word1, D3W2 means similarity path between word3 and word2 and so on.

Now take only that row which have maximum value in 1st column, sort this row and take all words with similarity distance greater than 0.08 or half list of the sorted words. Whole process for proposed framework is shown in following *Figure 2*.

Suppose I am taking following paragraph for extracting key words, their chunks and tags are shown in *Figure 3*.

"I had such a great time staying at Hotel Selva Candida. You know all of the annoying rules and limitations in other hotels? Well this one is like staying with family and they really try to work out any need, no matter the time! I checked out early in the morning and they turned on the cappuccino machine for me anyways! It's things like that that make a stay in a hotel worth it and make you feel comfortable for business and leisure! Thanks very much!"

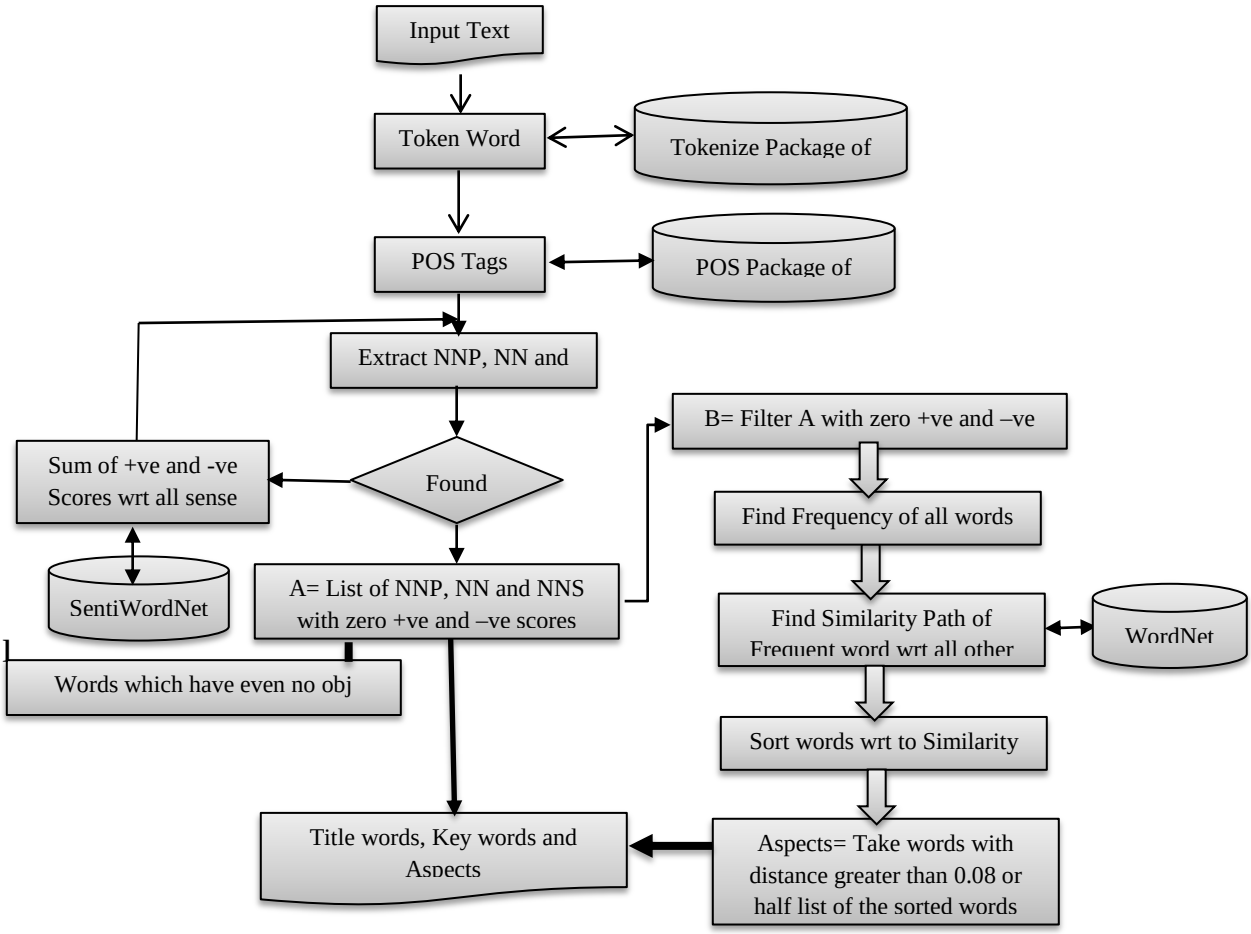

Figure 2 Flow of proposed framework

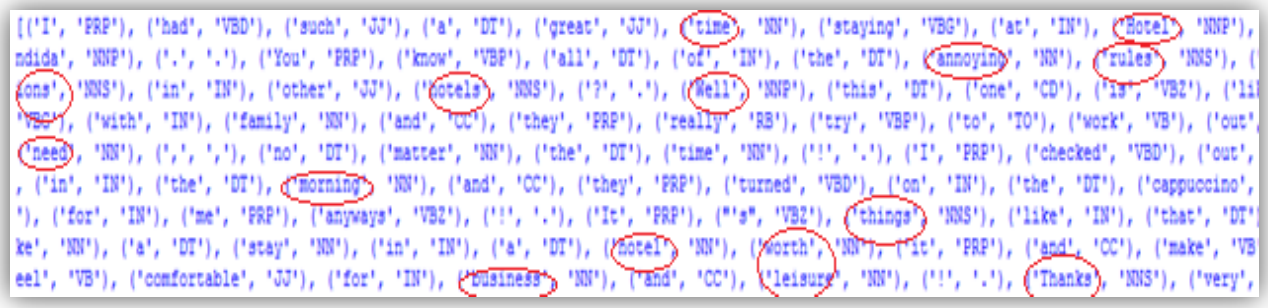

Figure 3 POS tags

Chunks and Tags can be obtained using word_tokenize and pos_tag packages of nltk repectively [21]. Positive and Negative scores of each word in different senses can be obtained using sentiwordnet package of nltk [22]. Following table, *Table 3* depicting the positive and negative scores of each NN (NNS), NNP (NNPS) in five senses. All words have not same number of sense, that is why words 1, 6,7,9,10,11 have 5 senses, words 2,3,4,5 has 1 sense and word 5 have 2 senses etc.

Table 3 Negative positive scores of NN, NNP in different senses

SNO	Words	Tags	+Ve –Ve Score In 1st Sense		+Ve –Ve Score In 2nd Sense		+Ve –Ve Score In 3rd Sense		+Ve –Ve Score In 4th Sense		+Ve –Ve Score In 5th Sense	
1	time	NN	0.0	0.0	0.0	0.0	0.5	0.0	0.75	0.0	0.0	0.0
2	hotel	NNP	0.0	0.0								
3	selva	NNP	0.0	0.0								
4	candida	NNP	0.0	0.125								
5	annoying	NN	0.0	0.25	0.0	0.0						
6	rules	NNS	0.25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	limitations	NNS	0.25	0.0	0.375	0.125	0.0	0.25	0.0	0.0	0.0	0.0
8	hotels	NNS	0.0	0.0								
9	well	NNP	0.0	0.0	0.0	0.0	0.0	0.125	0.0	0.0	0.0	0.0
10	family	NN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	need	NN	0.125	0.125	0.375	0.25	0.0	0.0	0.0	0.375	0.25	0.25
12	matter	NN	0.125	0.25	0.0	0.0	0.0	0.0	0.0	0.125	0.0	0.25
13	time	NN	0.0	0.0	0.0	0.0	0.5	0.0	0.75	0.0	0.0	0.0
14	morning	NN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		
15	cappuccino	NN	0.0	0.0								
16	machine	NN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
17	things	NNS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.125	0.25
18	make	NN	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
19	stay	NN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.375
20	worth	NN	0.0	0.0	0.75	0.0	0.0	0.0				
21	business	NN	0.0	0.0	0.0	0.0	0.0	0.0	0.375	0.0	0.0	0.0
22	leisure	NN	0.0	0.0	0.0	0.0						
23	thanks	NNS	0.125	0.0	0.0	0.125	0.0	0.0				

Now take only those words which have zero negative and positive scores. No need to take those words which little bit positive or negative concept. From above *Table 3*, take words of serial number 2, 3,8,10,14,15,16 and 22. Now make similarity path matrix of these words by finding frequency of each words and similarity path [21][22] of each word with all other words. *Table 4* is showing a similarity path matrix.

Table 4 Similarity path matrix of objective words

SNO	Words	Frequency	Hotel	Selva	Family	Morning	Cappuccino	Machine	Leisure
1	selva	1	0.07	1.0	0.1	0.09	0.07	0.07	0.08
2	hotel	2	1.0	0.07	0.07	0.08	0.08	0.14	0.07
3	family	1	0.07	0.1	1.0	0.1	0.07	0.07	0.09
4	morning	1	0.08	0.09	0.1	1.0	0.08	0.08	0.25
5	cappuccino	1	0.08	0.07	0.07	0.08	1.0	0.08	0.07
6	machine	1	0.14	0.07	0.07	0.08	0.08	1.0	0.07
7	leisure	1	0.07	0.08	0.09	0.25	0.07	0.07	1.0

From *Table 4*, it is clear that we will take 2nd row because word ‘hotel’ on this row have maximum frequency. Before going to further processing, first of all it is very necessary to clear the question that “if more than two words have same maximum frequency?” To solve this question first determine maximum relationships of the words which can be obtained by taking sum of all similarity path in each

row. Suppose we re-arrange above example as: "I had such a great time staying at Hotel Selva Candida. You know all of the annoying rules and limitations in other hotels? Well Selva one is like staying with family and they really try to work out any need, no matter the time! I checked out early in the morning and they turned on the cappuccino machine for me anyways! It's things like that that make a stay in a hotel worth it and make you feel comfortable for business and leisure! Thanks very much!"

Table 5 Similarity path matrix of objective words with two words of maximum frequency

S.no.	Words	Frequency	Hotel	Selva	Family	Morning	Cappuccino	Machine	Leisure	Sum
1	selva	2	0.07	1.0	0.1	0.09	0.07	0.07	0.08	1.55
2	hotel	2	1.0	0.07	0.07	0.08	0.08	0.14	0.07	2.51
3	family	1	0.07	0.1	1.0	0.1	0.07	0.07	0.09	1.57
4	morning	1	0.08	0.09	0.1	1.0	0.08	0.08	0.25	1.76
5	cappuccino	1	0.08	0.07	0.07	0.08	1.0	0.08	0.07	1.53
6	machine	1	0.14	0.07	0.07	0.08	0.08	1.0	0.07	1.65
7	leisure	1	0.07	0.08	0.09	0.25	0.07	0.07	1.0	1.7

Now in above *Table 5* row 1 and 2 have same maximum frequency. Now in last column there is sum of similarity path which depicts relationships of the words. Row-1 have 1.55 score and row 2 have 2.51 score, it means row 2 have close relationships between words. So take row-2 for further processing as shown in following *Table 6*.

Table 6 Row with maximum frequency and sum

Words	Frequency	Selva	Family	Morning	Cappuccino	Machine	Leisure	Sum
hotel	2	0.07	0.07	0.08	0.08	0.14	0.07	2.51

Table 7 is showing the sorted words wrt to similarity path. *Table 8* is representing those words which have value greater or equal to 0.08. *Table 9* is showing the title/key words of the document.

Table 7 Sorted words wrt similarity path

Words	Selva	Family	Leisure	Morning	Cappuccino	Machine
hotel	0.07	0.07	0.07	0.08	0.08	0.14

Table 8 Words with value higher or equal to 0.08.

Words	Morning	Cappuccino	Machine
hotel	0.08	0.08	0.14

Table 9 Title/Key words

Hotel	Morning	Cappuccino	Machine

In above example there is nothing any NN or NNP which does not belong to synset of WordNet. These words also are appended with list of key words.

Now to check maturity of this work, we have applied this algorithm on different document as shown in *Table 10*, to find out title/key words. Here we will also determine those words which have no positive, negative or objective scores.

Table 10 Text of documents

Documents	Text
Document-1	"I spent a week in the hotel and I could appreciate the staff friendliness and service. The environment is perfect, a mix of tradition with modern services"
Document -2	The staffs were amazing. The hotel is easy to get to, and in a nice out of the way spot in a tranquil suburb which is close enough to quickly get anywhere in Rome, but out of the way enough to have peace and relaxation. The staffs were amazing and attentive, the value was great and the breakfast was superb. We will be coming back, and recommend this hotel highly.
Document -3	QMobile Noir X25 Price in Pakistan, Spec & Reviews. Stylish mobile in affordable rates is the forte of QMobile, leading local mobile phone manufacturer in the country. Talking about the new QMobile x25, the device comes in simple features targeting average users. The specs and features of the QMobile X25 are not extravagant and there is no scope of further improvement. QMobile X25 comes in a dual SIM facility that

Documents	Text
	allows you to keep professional and personal affairs separate. It sports a 4 inches HQ display screen and impressive image transmission quality. It comes in a dual core processor 1.3 GHz, enough to make you thrilled and excited, packed with Android 4.4.2 KitKat. It comes in a 512MB of RAM, internal built in storage memory of 4 GB extendable up to 32 GB. The other significant features of QMobile X25 include geo tagging, Google Search, Gmail, Maps, torch that helps you overcome darkness, and speakerphone.
Document-4	Hp Envy is an Elite Class Full Featured Laptop and always comes with Genuine operating system. Its comes with Solid type Metal body and there are options of Detachable & Convertible Laptops. Envy Series starts from Core i5 and ends in Core i7, Envy always has an Eye catching designs, The performance is Superb, Option to buy with Touch Screen, Price Starts from Rs 85000 for a New Laptop, Backlit keyboard is mostly there, Option to buy with HD and Full HD Display, Beats Sound is also there, 5th Gen is now introduced in Hp Envy Series The Hp Envy 14 & Hp Envy 15 is the most famous of all Hp Envy Laptops.
Document-5	"A camera is an optical instrument for recording images, which may be stored locally, transmitted to another location, or both. The images may be individual still photographs or sequences of images constituting videos or movies. The word camera comes from camera obscura, which means dark chamber and is the Latin name of the original device for projecting an image of external reality onto a flat surface. The modern photographic camera evolved from the camera obscura. The functioning of the camera is very similar to the functioning of the human eye"

After analyzing above documents, their resultant attributes have been shown in following table, *Table 11*.

Table 11 Documents with resultant attributes

Document	Maximum frequency words	Similarity path		Words with 0.08 or higher path	Words not belongs to synset of wordnet
Document -1	Week	Hotel	0.08	Week Hotel Mix Staff Environment	---
		Mix	0.08		
		Staff	0.1		
		Environment	0.11		
Document-2	Hotel	Tranquil	0.0	Breakfast Suburb Relaxation	Tranquil Superb
		Superb	0.0		
		Staff	0.07		
		Rome	0.07		
		Breakfast	0.08		
		Suburb	0.09		
		Relaxation	0.1		
Document-3	Mobile	Storage	0.06	Mobile Spec Phone Scope Inches Hq Search Maps Gb	Qmobile Noir Kitkat X25 Geo Sim Gmail
		Pakistan	0.07		
		Manufacturer	0.07		
		Country	0.07		
		Specs	0.07		
		Processor	0.07		
		Android	0.07		
		Speakerphone	0.07		

Document	Maximum frequency words	Similarity path		Words with 0.08 or higher path	Words not belongs to synset of wordnet
		Spec	0.08	Built	
		Phone	0.08	Users	
		Scope	0.08	Mobile	
		Inches	0.08		
		Hq	0.08		
		Search	0.08		
		Maps	0.08		
		Gb	0.09		
		Built	0.1		
		Users	0.11		
		Mobile	1.0		
Document-4	Hp	Superb	0.0	Hp	I5
		Laptop	0.05	Operating	I7
		Type	0.07	Metal	Superb
		Keyboard	0.07	Body	Backlit
		Operating	0.08	Series	Hd
		Metal	0.08	Gen	
		Body	0.08	Full	
		Series	0.09	Rs	
		Gen	0.09		
		Full	0.1		
		Rs	0.2		
Document-5	Camera	Latin	0.06	Videos	Obscura
		Sequences	0.07	Camera	
		Word	0.07	Functioning	
		Projecting	0.07	Location	
		Videos	0.08	Movies	
		Functioning	0.1		
		Location	0.13		
		Movies	0.13		

3.Conclusion and future work

Searching of documents is not only part for search engine, but also very important for Sentiment analysis, text summarization etc. Aspect based sentiment analysis also requires those documents related to same aspects. How these documents are identified, there is need of some words which can determine those documents which are related to these key words. Instead of comparing these key words with all documents, just compare these key words with key words of each document. How key words of each document can be obtained? Here author suggest a frame work using similarity path matrix to find out the key/ title or aspects words from given documents. Key words of five documents from *Table 10* has been shown in *Table 12*. These key words will be very beneficial for making clusters. In future work, we will find a method to create clusters using these key words.

Table 12 Key/Title/Aspects words of each document

Documents	Key/Title/Aspects words
Document-1	Week, hotel, mix, staff, environment
Document-2	hotel, breakfast, suburb, relaxation, tranquil, superb
Document-3	qmobile, noir, kitkat, x25, geo, sim, gmail, mobile, spec, phone, scope, inches, hq, search, maps, gb, built, users, mobile
Document-4	i5, i7, superb, backlit, hd, hp, operating, metal, body, series, gen, full, rs
Document-5	videos, camera, functioning, location, movies, obscura

Acknowledgment
None.

Conflicts of interest
The authors have no conflicts of interest to declare.

References
[1] Chinsha TC, Joseph S. A syntactic approach for aspect based opinion mining. In IEEE international conference on semantic computing (ICSC) 2015 (pp. 24-31). IEEE.
[2] Qi L, Chen L. Comparison of model-based learning methods for feature-level opinion mining. In proceedings of the IEEE/WIC/ACM international conferences on web intelligence and intelligent agent technology 2011 (pp. 265-73). IEEE Computer Society.
[3] Thakur D, Singh J. The SAFE miner: a fine grained aspect level approach for resolving the sentiment. In third international conference on computer, communication, control and information technology (C3IT) 2015 (pp. 1-6). IEEE.
[4] Poria S, Cambria E, Ku LW, Gui C, Gelbukh A. A rule-based approach to aspect extraction from product reviews. In proceedings of the second workshop on natural language processing for social media (SocialNLP) 2014 (pp. 28-37).
[5] Zhu J, Wang H, Tsou BK, Zhu M. Multi-aspect opinion polling from textual reviews. In proceedings of the 18th ACM conference on information and knowledge management 2009 (pp. 1799-802). ACM.
[6] Hu M, Liu B. Mining and summarizing customer reviews. In proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining 2004 (pp. 168-77). ACM.
[7] Long C, Zhang J, Zhut X. A review selection approach for accurate feature rating estimation. In proceedings of the 23rd international conference on computational linguistics: posters 2010 (pp. 766-74). Association for Computational Linguistics.
[8] Liu B. Web data mining: exploring hyperlinks, contents, and usage data. Springer Science & Business Media; 2007.
[9] Marrese-Taylor E, Velásquez JD, Bravo-Marquez F. A novel deterministic approach for aspect-based opinion mining in tourism products reviews. Expert Systems with Applications. 2014; 41(17):7764-75.
[10] Gamon M, Aue A, Corston-Oliver S, Ringger E. Pulse: mining customer opinions from free text. In international symposium on intelligent data analysis 2005 (pp. 121-32). Springer Berlin Heidelberg.
[11] Zhuang L, Jing F, Zhu XY. Movie review mining and summarization. In proceedings of the 15th ACM international conference on information and knowledge management 2006 (pp. 43-50). ACM.
[12] Riloff E, Wiebe J, Wilson T. Learning subjective nouns using extraction pattern bootstrapping. In proceedings of the seventh conference on natural language learning at HLT-NAACL 2003 (pp. 25-32). Association for Computational Linguistics.
[13] Zagibalov T, Carroll J. Unsupervised Classification of sentiment and objectivity in Chinese text. In third international joint conference on natural language processing 2008 (pp. 304-11).
[14] Jeyapriya A, Selvi CK. Extracting aspects and mining opinions in product reviews using supervised learning algorithm. In 2nd international conference on electronics and communication systems (ICECS) 2015 (pp. 548-52). IEEE.
[15] Liu Q, Gao Z, Liu B, Zhang Y. A logic programming approach to aspect extraction in opinion mining. In international joint conferences on web intelligence (WI) and intelligent agent technologies (IAT) 2013 (pp. 276-83). IEEE.
[16] Kirange DK, Deshmukh RR. Aspect based sentiment analysis SEMEVAL-2014 task 4. Asian Journal of Computer Science and Information Technology. 2014; 4(8): 72 - 5.
[17] Gupta DK, Ekbal A. IITP: supervised machine learning for aspect based sentiment analysis. Proceedings of the 8th international workshop on semantic evaluation (SemEval) 2014 (pp. 319–23).
[18] Brychcın T, Konkol M, Steinberger J. Uwb: machine learning approach to aspect-based sentiment analysis. In proceedings of the 8th international workshop on semantic evaluation 2014 (pp. 817-22).
[19] Brun C, Popa DN, Roux C. Xrce: Hybrid classification for aspect-based sentiment analysis. In proceedings of the 8th international workshop on semantic evaluation 2014 (pp. 838-42).
[20] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html. Accessed 26 May 2016.
[21] Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In LREC 2010 (pp. 2200-4).
[22] Esuli A, Sebastiani F. Sentiwordnet: a publicly available lexical resource for opinion mining. In proceedings of LREC 2006 (pp. 417-22).