IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-4 E-ISSN: 2347-2693

Aspect-Opinion Identification and Classification Using Custom Heuristic Rules

T. U. Kadam[1*], P. Kaur[2]

[1]Dept. of CSE, Jawaharlal Nehru Engineering College, Aurangabad, India
[2] Dept. of IT, Jawaharlal Nehru Engineering College, Aurangabad, India

[*]*Corresponding Author: tanvikadam76@gmail.com, Tel.: +919730729711*

Available online at: www.ijcseonline.org

Received: 16/Mar/2018, Revised: 24/Mar/2018, Accepted: 12/Apr/2018, Published: 30/Apr/2018

***Abstract*—** Users opinion about different entities forms the huge repository of data over internet. It is highly impossible to accurately monitor and find what actually a user wants to say about an entity from this large amount of data. Data analyst these days concentrates on finely analyzing opinions about particular entity and for this reason the extraction of aspects and its corresponding opinions of that entity are important. This work concentrates on identifying aspects and their corresponding opinion from the provided user opinions which helps to obtain fine grained knowledge about the entity. To obtain aspects and related opinions custom heuristic rules are created by using regular expression on the parts-of-speech tagging. The created rules are provided to Stanford natural language processing (SNLP) classifier and finds association of aspect words and opinion words from the opinion corpus. The classification is done by SNLP classifier and Naïve Bayes (NB) classifier. Identification of aspects and aspect specific opinions are accurately obtained using custom heuristic rules applied over SNLP compared to NB.

***Keywords*—** Aspects, Opinions, Heuristic Rules, SNLP, NB.

I. INTRODUCTION

The services provided by social media are increasing day by day through which user can easily share and use the information on various entities. User makes their decision to purchase any product or to plan a travel to some location based on the opinions of the other users who already purchased that product and who travelled the same place. It is highly impossible for the user to make an informed decision about the entities due to the large amount of scattered information present over the web. This job seems tricky for manufacturer and sellers also to finely monitor the opinions provided by the customers. For this reason, not only mining the customer generalized reviews but also mining the aspects and opinions from that customer reviews are important to gain accurately monitor the opinions and gain the maximum knowledge from opinion.

In real world, entity carries several aspects and the aspects which shows certain characteristics or attributes for that entity. For example, aspects for hotel entity can be staff, room, and hotel and the opinion for the aspects are classified into positive, negative classes. People write negative review about 'room service', 'food' and 'wifi facility' of the hotel and appraising by writing 'helpful', 'best' and 'amazing'. Providing the generalized sentiments for that review is not sufficient to acquire fine-grained knowledge but identifying aspects and opinions and using the association of aspect-opinion to identify the sentiments will result in higher accuracy. Extracting every important aspect and its corresponding opinion from the provided review helps user to achieve knowledge in more productive way. Identifying aspects and opinions, finding the association of aspect-opinion and use this to find the sentiment polarity is one of the important topics in research area. The work in this paper concentrated on identifying aspect words and corresponding opinion words from the opinion considering hotel and smart phone as entities and shows the association of opinions for the corresponding aspect. The aspects and opinions are identified using regular expression syntax on parts of speech tags of SNLP. Classification is carried out for each opinion separately from the opinion document. The paper is organized into six sections: section I contains the introduction of aspects and opinions of entities, section II contains the related work of opinion identification and classification, research gap obtained and describes the approach of this work. Section III contains the details of dataset used for the research, section IV contains proposed methodology and heuristic rules created, section V contains results and discussion and section VI concludes the research work with future scope.

II. RELATED WORK

Most of the work in this research has focused on classifying reviews and extracting aspects however, still faces many challenges. For achieving useful information, the available

enormous data of social media projects various opportunities. The useful information or knowledge gained provides certain actionable patterns inside the data, which benefits the users, market analyzers, business persons. Multimodal aspect-opinion mining model (AOM) on user submitted reviews and photos from BBCNews, tripadvisor and flickr is a probabilistic generative model implemented on multivariate NB or Bernoulli classifier to classify opinions proposed in [1]. Decision tree C4.5 algorithm compared with NB for dataset related with direct marketing campaigns of Portuguese banking institution and found higher accuracy for C4.5 algorithm in [2]. The classification was to predict whether the clients will subscribe the term deposit. F. Chong et al. in [3] considered method for automatically identifying noun phrases (NP) for event monitoring on Twitter. Pang and lee in [5] get modified by introducing Recursive neural tensor network and the Stanford sentiment Treebank including labels for 215,154 phrases in the parse trees of 11,855 sentences. The sentiment detection for positive/negative classification is increased by 5.4% [4]. Variants of NB and support vector machine classification techniques are implemented by including bigram features. For short snippet sentiment task NB performed better compared to SVM but for longer documents showed the opposite results [6].

Moghaddam, S. et al [8] extracted adjectives from movie review dataset and then computing the frequency of each adjective, the research identified opinion polarity of adjectives using NB classifier. By using adjectives near about 10% higher accuracy than pure machine learning techniques is reported. Aspects and their aspect specific opinion words are discovered by applying LDA, by deriving the joint approach of maximum entropy-LDA and the hybrid model of LDA and by using two statistical models from seeds respectively in [9], [10], [11]. A. Mukharjee and B. Liu [11] identified aspects from the opinion dataset used. Considering the work in aspect based opinion mining, M. Hu and B. Liu in [12] suggested the aspects and opinions detection of products from online reviews which then benefits the investment decision making. Q. Liu et al. [13] collected the dataset from [14], which consist of five domains and built their endemic dataset with three domains for computer, wireless router and speaker and proposed an automated rule selection algorithm. Selected rules are used for the feature extraction followed by greedy algorithm which projects on double propagation but it probes to find the error aspects when applied to opinion corpus.

Most existing approaches focus on classifying the opinions by identifying aspect and opinion words and considering only adjectives as opinions rather to focus on other parts of speech tags present in the opinions. Considering only adjectives as opinion words are not sufficient for fine-grained sentiment classification therefore adjectives along with adverb and verbs has to be used for identification of opinion words. In this work custom heuristic rules are created using regular expression on pos tags provided by Stanford pos tagger in order to extract aspects and opinions from the opinion and evaluation is carried out on the basis of identified aspects and opinions. The accuracy achieved by the SNLP with custom heuristic rules is then compared with the NB classification technique.

III. DATASET USED

The dataset is created by collecting user reviews for hotel and smart phone. The 60 opinion documents are provided as an input. The self trained SNLP classifier is applied along with the custom heuristic rules to classify the sentiments. The accuracy of SNLP classifier compared with NB classifier. The opinion lexicon of 2006 positive words and 4789 negative words from [14] are used to train the NB model. Manual examination is carried out to correct the grammar and also stop word removal operation removes noisy, irrelevant data.

IV. PROPOSED METHODOLOGY

A. Custom heuristic rules approach for aspect-opinion identification

The reviews for hotel and smart phone are the input to the system. Each document consists of few aspect and opinion components. Nouns are considered as aspect words and adjective, verbs and adverbs are considered as opinion words. Various works in this area uses adjectives for identification of opinion words but considering only adjectives alone for opinion words identification is not sufficient. Figure 1 shows the proposed architecture of using heuristic rules for aspect and its corresponding opinion identification. The review data collected from social media site has noisy, irrelevant information. To remove this stop word removal process is carried out using standard list of stop words. The stemming operation is performed to maintain the meaning of words.

The noise free opinion document is applied to SNLP classifier where each and every word from the document is tokenized by Stanford linear tokenize method. The pos tagging for each word is done by pos tagging function of Stanford pos tagger. Heuristic rules are created using regular expression on pos tags. The three classifier name entity recognition from Stanford identifies the entities in the opinion document. The aspects and opinions are identified using heuristic rules based on pos tags provided by SNLP toolkit.

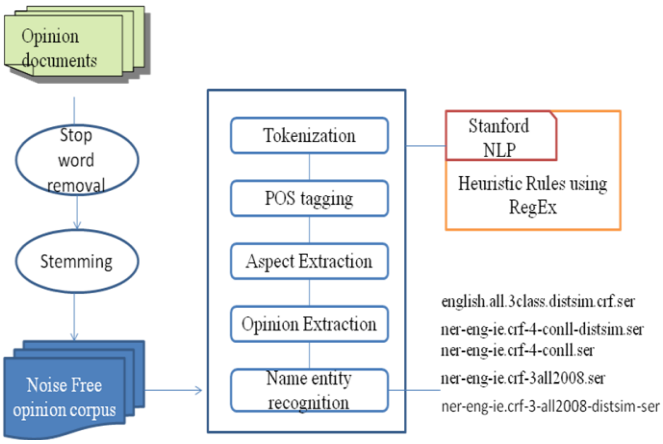

Figure 1. Proposed architecture for aspect-opinion identification and opinion classification

Heuristic rules are generated manually using regular expression for noun (NP), adjective (JJ), adverb (RB) and verb (VB) phrases. Aspect words are noun phrases and opinion words are adjectives, verb and adverbs. The n-gram feature engineering method is use to create thirteen rules including unigram and bigram features. The rules are patterns of pos tags on NP, JJ, RB and VB, formulated using regular expression where four rules for NP and JJ, two for RB and three rules for VB are created. Rules uses alphabetical list of Penn Treebank for technical symbols or expression where words cannot be considered directly. Table 1 shows the heuristic rules created using regular expression on parts of speech tags.

Table 1. Heuristic rules created using regular expression to identify aspect words and opinion words

Identify	Phrases		Heuristic rules
Aspects	***Noun***	NP	(@NP \\ (NP (.*?)))
		NNP	NNP (.*?)
		NNS	NNS (.*?)
		NN	(NP (NN (.*?)))
Opinions	***Adjective***	JJ	DT (JJ (.*?))
		JJ	NP (JJ (.*?))
		JJS	JJS (.*?)
		JJR	JJR (.*?)
	Adverb	RBR	ADJP (RBR (.*?))
		RB	RB (.*?)
	Verb	VBP	VBP (.*?)
		VBG	VBG ()
		VBZ	VBZ ()

1) Identifying noun and adjective phrases using regular expression.

Regular expression syntax is used for creating patterns and RegEx operation of Stanford to match the pattern. Rules for NP pattern as well as their variants such as singular proper noun (NNP), singular noun (NN) and plural noun (NNS) are created. Unigrams created for NP is (@NP (NP) searches for a match after NP along with root NP. Also unigram for NNP look for a match after NNP. Rule created for NNS returns the match found after NNS and created rule for NN search for a match after NN.

Adjectives plays important role as opinion words as it projects the kind of information present in opinion document. Adjectives provide extra information about nouns. Heuristic rules for JJ and adjective superlative (JJS) and adjective comparative (JJR) are created. Bigram feature (NP (determiner DT) (JJ) created for JJ searches for a match after JJ along with DT and NP. Rule for JJ look for a match found after it. Bigram feature (NP (JJ (JJS))) finds the zero or more occurrences of JJ and append the matches found after JJS. One unigram feature for JJR searches the one or more occurrences after JJR.

2)Identifying adverb and verb phrases using regular expression

The opinion document has different combinations of parts of speech other than noun and adjective which provides fine information about the data in opinion document. In this research, adverb (RB) and verb base form (VB) phrases are considered as opinion words to gain more accurate knowledge. Among two feature one unigram for (RB * searches one or more occurrences of RB. Bigram feature (ADJP? (RBR*) searches zero or non zero occurrences of ADJP followed by one or more occurrences of RBR. The Adjectives found in VB, verb gerund (VBG) and third person singular present word (VBZ) are also considered for identifying opinion words. Table 2 shows the aspects and opinions obtained from the example opinion for smart phone entity. For example the opinion is, "IPhone has a super solid stainless steel body surrounded by glass. It is simply the best, more secure among all the smart phones."

The pos tagging by the SNLP classifier is obtained as,
[(ROOT(S**(NP (NNP IPhone))**(VP (VBZ has)(NP**(NP (DT a) (JJ super) (JJ solid) (JJ stainless) (NN steel)(NN body)**)(VP (VBN surrounded)(PP (IN by**)(NP (NN glass)**))))))(. .))))][(ROOT(S(NP (PRP It))(VP (VBZ is)(ADJP (RB simply)(ADJP (DT the) **(JJS best)**)(, **,)(ADJP (RBR more)** (**JJ secure**)(PP (IN among)(NP (PDT all) (DT the) **(NNS smartphones)**))))))(. .))))]

Table 2. Aspects and opinions identified for given example

Raw aspect	Processed aspect	Raw opinion	Processed opinion
Steel	Iphone	more	super
Body		Secure	Solid
smartphones			stainless
Glass			Best

The aspects and opinions identified are stored in raw aspect list (RA), raw opinion list (RO), processed aspect (PA) and processed opinion (PO) list. The raw aspects and raw opinions found here are considered as less strong in analysing the opinion of entity while processed aspects and processed opinions are the stronger aspects and opinions identified which together associates and analyses that entity

opinion in detail. The bold worlds are obtained shows the application of thirteen heuristic rules for given parts of speech. The aspect word 'steel', 'body', 'smart phones', 'glass' and 'IPhone' are obtained as raw aspects and processed aspect respectively. These aspects and opinions are considered as true positive and false positives for the evaluation. The opinion word 'secure' is obtained in raw opinion list and opinion word 'super', 'solid', 'stainless' obtained in processed opinion list. Opinion word more is also identified by RBR rule and considered as one of the aspects.

B. Naïve Bayes text classification

The simple probabilistic classifier based on Bayes theorem identifies the presence of feature or word in one class is related or not related to other class. Naïve bayes works efficiently on large dataset with smaller training set provided. In this research the two classes, positive and negative are provided for sentiment analysis. The most likely class according to naïve Bayes is that class out of all classes which maximizes the product of two probabilities, the prior probability of the class and the product over all positions in the document of both likelihood of the word w in that document given the class to see the positive review times for every position in the document. How likely, that word has been expressed by positive review and same for the negative.

V. RESULTS AND DISCUSSION

Dataset is created using tripadvisor reviews for hotel and the reviews from mouthshut.com for smart phone entity. 60 opinion documents are provided consisting on an average 10 opinion sentences in each document. The development of this research is done using java language. Reviews not showing the factual information about entities is discarded which is achieved by using manual examination. NB is trained with 2006 positive words list and 4789 negative words list.

Stop word removal operation is applied using the standard list of stop words to reduce noisy and irrelevant data from the opinion corpus. Stemming operation is also done using port stemmer operation of Stanford. The experiment is conducted on the dataset reviews of both the entities separately. 13 heuristics rules are created on noun, adjectives and adverb phrases are shown in Table 1. SNLP classification is implemented based on generated heuristic rules and compared the accuracy of aspect and opinion identification for both entities with NB classification. The work here focuses more on correctly identifying aspects and aspect specific opinions and associates them to check they are correctly classified.

Table 3 shows the total number of aspects and its corresponding opinions identified. Due to the maximum occurrences of NNS and NN phrases, unigram feature for NP detects many RA compared to PO, RO and PO.

Table 3. Number of aspects and opinions identified

Entity	#documents	#RA	#PA	#RO	#PO
Hotel	30	348	63	28	266
Smartphone	30	385	121	63	263

Table 4. Result of aspects-opinions identified using SNLP classifier

Entity	SNLP classifier			
	Precision	***Recall***	***accuracy***	***Kappa coefficient***
Hotel	92.13	89.94	88.57	0.68
Smartphone	79.47	90.10	81.20	0.69

Table 5. Results of aspects and opinions identified using NB classifier

Entity	NB classifier		
	Precision	***Recall***	***Accuracy***
Hotel	55.91	86.27	56.61
Smartphone	55.25	86.25	55.96

Table 4 shows the different evaluation metric result for aspect and opinion identification using heuristic rules applied with SNLP classification technique. Also in Table 5 aspect opinion identification results has shown using NB classifier. Utilization of precision, recall and accuracy evaluation metrics are applied to both the classifiers. Here, precision is high when the classifier returns more relevant aspect and opinion words among all the retrieved aspect and opinion words while recall is high when the classifier returns most of the relevant aspects and opinions that have been retrieved over total amount of retrieved aspects and opinions. Overall accuracy is calculated for all true positive and false negative aspects and opinions identified over total number of aspects and opinions identified in the lists of table 3. The level of agreement calculated using Cohen's Kappa is moderate.

For some opinion words stemming operation changes the parts of speech and results in unexpected polarity shift. Hence the inclusion of stemming operation is discarded to maintain the actual polarities and to classify the opinion accordingly.

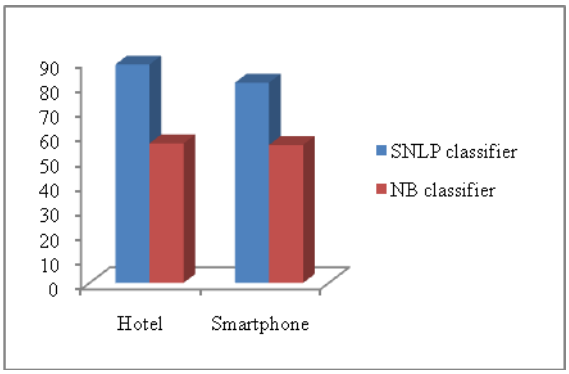

Figure 2. Comparison of evaluation metrics for SNLP and NB classifier for hotel and smart phone entity

As the work aims to identify aspects and opinions as much as possible hence the system focuses to be quantitative or complete than to focus more to be exact or qualitative. The sentiment prediction by SNLP shows greater result due to the

rise in counts of true positive aspects that occurs with the application of heuristic rules. Figure 2 shows the comparison of average precision recall and accuracy for SNLP and NB classifier considering opinion document together for both the entities.

Performance of SNLP is increased than NB due to the aggregation of heuristic rules and the increase in number of true positive aspects in processed aspect and processed opinion list. Recall value here is more in both the classifier due to less inclusion of type II error which leads to obtained more true aspects for predicted and non predicted class. The work in [16] used multinomial NB and multivariate NB where the author finds better result for multinomial NB compared to multivariate NB. In this work, the classification done by SNLP classifier is more accurate compared to NB classification technique. The Accuracy of identifying aspects and opinions using Stanford natural language processing classifier aggregating heuristic rules is 25% more than naïve bayes classifier for both the entities.

VI. CONCLUSION AND FUTURE SCOPE

In this paper, different rules using regular expression are generated to find aspects and opinions on the parts of speech tagging provided by Stanford natural language processing classifier. The effectiveness of aspect-opinion identification and sentiment prediction is highly enhanced using custom heuristic rules. Opinions based on aspects are accurately identified in order to achieve fine-grained knowledge.

In future, the approach is to associate aspects and aspect specific opinions identified from the opinion. More number of heuristic rules is to be added to gain maximum knowledge of aspects and opinions from the entities.

REFERENCES

[1] Quan Fang, Changsheng Xu, Jitao Sang, M. Shamim Hossain and Ghulam Muhammad,"*Word-of-mouth understanding: entity-centric multimodal aspect-opinion mining in social media*," IEEE transaction on multimedia, volume 17.No. 12, pp. 2281-2296, December 2015.

[2] Masud Karim, Rashedur M. Rahman, "*Decision tree and naïve bayes algorithm for classification and generation of actionable knowledge for direct marketing*," Journal of Software Engineering and Applications, 6, pp 196-206, 2013.

[3] F. Chua, W. Cohen, J. Betteridge, E. Lim, "*Community-Based Classification of Noun Phrases in Twitter*," In the Proceedings of the 21st international ACM conference of information and knowledge management (CIKM' 12), pp 1702-1706, 2012.

[4] Richard Socher and et. Al, "*Recursive deep models for semantic compositionality over sentiment a Treebank*," In the Proceedings of the conference on empirical methods in natural language processing, EMNLP'13, 2013.

[5] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, "*Thumbs up? Sentiment Classification using Machine Learning Techniques*," In Proceedings of the EMNLP 2002, pp. 79–86, 2002.

[6] Sida Wang and Christopher Manning, "*Baselines and bigrams: simple, good sentiment and topic classification*," In Proceedings of the 50th Annual Meeting of the Association for computational Linguistics (ACL'12), volume 2, July 08 – 14, pp 90-94, 2012.

[7] B. Liu and L. Zhang,"*A survey of opinion mining and sentiment analysis*," in Mining Text Data. New York, NY, USA: Springer, 2012, pp. 415–463, 2012.

[8] Moghaddam, S., Popowich F., "Opinion *polarity identification through adjectives*", CoRR arXiv: 1011.4623 (2010).

[9] S. Moghaddam and M. Ester, "*On the design of lda models for aspectbased opinion mining*," In the Proceedings of 21st ACM International Conference on Information and Knowledge Management (CIKM'12), pp. 803–812, 2012.

[10] W. X. Zhao, J. Jiang, H. Yan, and X. Li, "*Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid*," in the proceedings of the conference on the Empirical Methods in Natural Language Processing (EMNLP 2010), MIT, Massachusetts, USA, pp. 56–65, 2010.

[11] A. Mukherjee and B. Liu, "*Aspect extraction through semi-supervised modeling*," In the Proceedings of Assoc. Comput. Linguistics, 2012, pp. 339–348, 2012.

[12] M. Hu and B. Liu, "*Mining and summarizing customer reviews*," In the Proceedings of the Tenth ACM SIGMOD International Conference on Knowledge Discovery and Data Mining (KDD 2004), pp. 168–177, 2004.

[13] Qian Liu, Zhiqiang Gao, Bing Liu3 and Yuanlin Zhang, "*Automated Rule Selection for Aspect Extraction in Opinion Mining*," In the Proceedings of Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015), pp 1291-1297, 2015.

[14] Y. Yang, C. Chen, M. Qiu, F. s. Bao, "*Aspect extraction from product reviews using category hierarchy information*," In the Proceedings of the 15th Conference of the European Chapter of Association for Computational Linguistics: Volume 2, Short Papers, pages 675–680, 2017.

[15] Y. Fang, L. Si, N. Somasundaram, and Z. Yu, *"Mining contrastive opinions on political texts using cross-perspective topic model*," In the Proceedings of the Fifth ACM WSDM 2012, pp. 63–72, 2012.

[16] A. McCallum and K. Nigam, "*A comparison of event models for naïve bayes text classification*," In the Proceedings of the ICML/AAAI-98 workshop on learning for text categorization, Madison, WI, pp. 41–48, 26–27 July 1998.

Authors profile

T. U. kadam has completed bachlor of Engineering from Swami Ramanand Thirth Marathwada university, Nanded, Maharashtra in 2012. She is currently pursuing master of engineering in computer science and engineering from Jawaharlal Nehru Engineering college, aurangabad, Maharashtra.

P. Kaur pursued master of engineering in computer science and engineering. She is currently pursuing Ph.D in Remote Sensing and currently working as an associate professor in department of Information Technology in Jawaharlal Nehru Engineering College, Aurangabad, Maharashtra. She has 7 national publications and 2 international publications and has more than 18 years of teaching experience.

