Proc. of the Sixth Intl. Conf. Advances in Computing, Communication and Information Technology- CCIT 2018

ISBN: 978-1-63248-149-8 doi: 10.15224/978-1-63248-149-8-12

A Novel Opinion Reason Mining Framework Exploiting Linguistic Associations

[Shehzad Khalid and Muhammad Taimoor Khan]

***Abstract*—Abstract—Aspect-based Sentiment Analysis (ABSA) explores the strong and weak aspects of a product. There are many online platforms that allow users to review commercial products while others to aggregate those opinions across millions of reviews at the aspect level. Such analysis is of high regard to potential customers and manufacturers to make profitable decisions. However, the existing ABSA models do not highlight the reasons behind the strengths and weaknesses of the aspects. Moving a step forward, opinion reason mining explores the reasons for the aspects being appreciated or criticized. We propose opinion reason mining framework ORMFW that uses topic model to generate aspects as groups of aspect terms which are refined using paradigmatic word associations. Polarity is evaluated for each aspect using a dictionary based approach. Furthermore, it incorporates syntagmatic word associations to map the aspects to their respective reason terms against a sentiment polarity. Results on twitter dataset reveal that the proposed ORMFW framework efficiently and effectively identifies the prominent opinion reasons in relation to their aspects.**

***Keywords*—opinion minig, reason mining, sentiment analysis, topic modeling, linguistic analysis**

I. Introduction

Aspect-based Sentiment Analysis (ABSA) identifies users' opinions and aggregates them towards the aspects of the target entity. Aspects are the features of the entity studied and are used as sentiment targets. They are also called attributes and objects. ABSA not only highlight the popular entities but also their strong and weak aspects [1]. With a rising interest in the analysis of online user generated content, ABSA has served well for commercial products over the last decade. Its popularity has outgrown to non-commercial domains including governance, health care, social events, stock market and politics. Its importance can be realized for the fact that its practical applications are available while it is still active in research [2]. The traditional Information Extraction (IE) techniques, having promising results with text analysis are used for aspect extraction. These techniques are extended to focus on aspects as potential information [3]. Aspects exhibit certain aspect like properties i.e. they belong to an entity and are associated to an opinion term. The extended techniques incorporate this behavior of aspects for aspect extraction tasks [4]. On the basis of proximity, opinion reasons are the terms that closely exist with the aspect and opinion terms which suggest the motivation for the aspect having such polarity.

ABSA has been a challenging task as introduced in [5]. It has an automatic aspect extraction mechanism that includes identifying the aspect terms and grouping aspect terms that are synonyms and near-synonyms. The aspect terms are commonly of two types, i.e. explicit aspect terms and implicit aspect terms. Explicit aspect terms are explicitly mentioned e.g. price while implicit aspect terms are implied in a situation e.g. the terms cheap implicitly refers to the aspect cost [6]. In the previous studies, the two have been treated the same as indicating aspect. However, implicit aspect terms reveal more than just referring to the aspect price. For instance, the implicit aspect terms costly or cheap indicate the aspect price and refer to the possible reason behind the negative or positive opinion, respectively [7]. Other aspect types are known or unknown aspect terms, based on the users prior knowledge or labeled training data [8]. Supervised and semi-supervised techniques are not effective towards identifying new aspects as they require to be trained for aspects of specific domains [9]. Aspect terms could be frequent or rare based on mentions in dataset. The frequency and relation based techniques identify frequent aspect terms and filter them using relational templates [3]. Topic modeling is a dominantly used technique for aspect extraction [10]. Following an unsupervised approach, it groups observable words in their respective documents into latent topics based on their co-occurrence association. Unlike other techniques, topic models identify aspect terms and group them together as a single step.

The existing topic modeling based techniques for aspect extraction clusters a mix of explicit and implicit aspect terms [6]. In other words, they treat the two types of aspect terms the same as indicators of the aspect represented by the cluster. Therefore, these techniques restricted to identifying aspects and evaluating their polarity only [1]. Furthermore, most of the research in ABSA is focused on improving the accuracy of sentiment classification. The state-of-the art reason mining based models employ data mining techniques to extract aspects while ignoring the peculiarities of the language syntax [11, 12]. We propose an unsupervised opinion reasoning mining framework ORMFW for efficient retrieval of

Shehzad Khalid
line 1: Bahria University, Islamabad
line 2: Pakistan

Muhammad Taimoor Khan
line 1: Bahria University, Islamabad
line 2: Pakistan

Proc. of the Sixth Intl. Conf. Advances in Computing, Communication and Information Technology- CCIT 2018

ISBN: 978-1-63248-149-8 doi: 10.15224/978-1-63248-149-8-12

opinion reasons from an unknown domain [13]. It can be applied directly to large-scale data where it efficiently and effectively evolves with newly emerging aspects and their corresponding opinions. Linguistic associations i.e., paradigmatic and syntagmatic word associations are used to map aspects to their respective reasons.

II. Literature Review

Probabilistic topic models are extensively used for information extraction to great effect. Although, it doesn't mine the semantics in content but uses co-relations between terms by performing heavy statistical computations [1]. It works out terms co-occurrence probabilities through their arrangement in documents. Topic models are either based on probabilistic latent semantic analysis (pLSA) [14] or Latent Dirichlet Allocation (LDA) [10]. It follows a Bag-of-Words (BOW) approach, considering the importance of terms through their presence while ignoring their position information. The idea doesn't seem convincing for text analysis where words change their means when re-positioned [6]. However for large datasets, it produce better results than other relation and template based techniques. The probabilistic topic models extract topics from documents or generate can generate documents from the topic structure [8]. Since candidate aspect terms are mostly found in the form of nouns and noun phrases, therefore, topic models are extended differently to focus on them only as candidate aspect terms by making use of linguistic rules [9]. One of the advantage of using topic models is that they identify and aggregate aspect terms both at the same time.

Aspect-sentiment joint model is based on pLSA model to extract topics and distribute them further among aspects, positive sentiment and negative sentiment models [15]. MGLDA traverses on both global and local topics at the document and sliding window level [16]. The global topics focus on entities and brands while the local topics capture aspects of them. Reviews with separate pros and cons section are considered to have prominent aspects help in identifying other aspects from regular text. Non-aspect sentiment terms, usually implicit aspect terms, are extracted as aspect terms are not good representatives of the aspect. They are filtered by applying POS-tagger [13]. Two joint LDA models are used for extracting aspects and sentiments; however, the separation between the two is inconclusive [17]. MaxEnt-LDA, a hybrid model discovers aspects and sentiments that are separated through syntactic patterns [18]. Further, hybrid topic models are used for similar aspect extraction tasks [19]. The semi-supervised topic models require expert's intuition to guide the inference technique of the topic model [20]. The user's guidance is incorporated differently in various semi-supervised models [21].

Opinion reason mining is an interesting new extension to research in opinion mining and sentiment analysis. Weakness Finder is built on such architecture to identify product aspects getting negative sentiments and explore the reasons behind their criticism [11]. It extracts product aspects through a morpheme based method with Hownet similarity measure to group aspect terms. Polarity is evaluated for each aspect to identify weak aspects which are notified to the manufacturers. The opinion-aware analytical framework also explores weak aspects of products through aspect extraction and its sentiment analysis [12]. However, do not provide evidence on the reasons why certain aspects are getting bad reviews.

III. Proposed Methodology

An opinion reason mining framework (ORMFW) is proposed that extract domain aspects and associate them to their respective reasons. The framework works as a four components pipeline that is depicted in *Figure. 1*. The first component generate raw topics representing aspects using topic modeling technique. The second component uses language syntax to separate the terms of a topic into two groups of terms i.e., candidate aspect terms *cAspectTerms* and candidate reason terms *cReasonTerms* using language syntax. The third component generate compact aspect terms through evaluating their paradigmatic associations. While the reason polarity is calculated for each aspect through WordNet dictionary and aggregated as sentiments scores for each aspect. The fourth component refines the reason group by evaluating syntagmatic word associations.

Algorithm 1 *OpinionReasonMining*

1: T ← LDATopicModeling(Dataset)
2: for t_i as T do
3: for w ∈ t_i do
4: if POS(w) is Nouns or Noun phrases then
5: $cAspectTerms_i \leftarrow t_i$
6: else
7: $cReasonTerms_i \leftarrow t_i$
8: end if
9: end for
10: end for
11: $aspect_i \leftarrow$ paradEval($cAspectTerms_i$)
12: $reason_i \leftarrow$ syntagEval($cReasonTerms_i$,$aspect_i$)

Working of the ORMFW is shown in **Algorithm 1**. Latent Dirichlet Allocation (LDA) based topic model is used for extracting topics from the given dataset as T, in line 1 [10]. Here T is a set of all topics from t_1 to t_k. The purpose of this component is to represent raw topics as a mix of aspect and reason terms [22]. It helps to avoid evaluating the exhaustive vocabulary of unique words as possible aspect or reason term. In lines 2 - 10 each topic t_i in T is traversed to separate candidate aspect words from candidate reason words. In lines 3 - 9, each word w of topic t_i is labeled with POS tagging using WordNet. In lines 4 - 8, the words having POS as nouns or noun phrases from topic t_i are separated from the implicit aspect terms. Lines 5 and 7 show the grouping the explicit aspect terms as $cAspectTerms_i$ and implicit aspect terms as $cReasonTerms_i$, respectively. The implicit aspect terms

Proc. of the Sixth Intl. Conf. Advances in Computing, Communication and Information Technology- CCIT 2018

ISBN: 978-1-63248-149-8 doi: 10.15224/978-1-63248-149-8-12

not only imply aspects but also reveal more about motivation of the users' opinions. In line 11, the $cAspectTerms_i$ representing an aspect is refined through paradigmatic word associations and is preserved as $aspect_i$. The two words hold paradigmatic association that is used as alternate for each other such as synonyms and near-synonyms. Finally in line 12, syntagmatic word associations are evaluated for aspect terms $aspect_i$ with $cReasonTerms_i$. The reason terms that have strong syntactic association with the respective aspect $aspect_i$ are preserved.

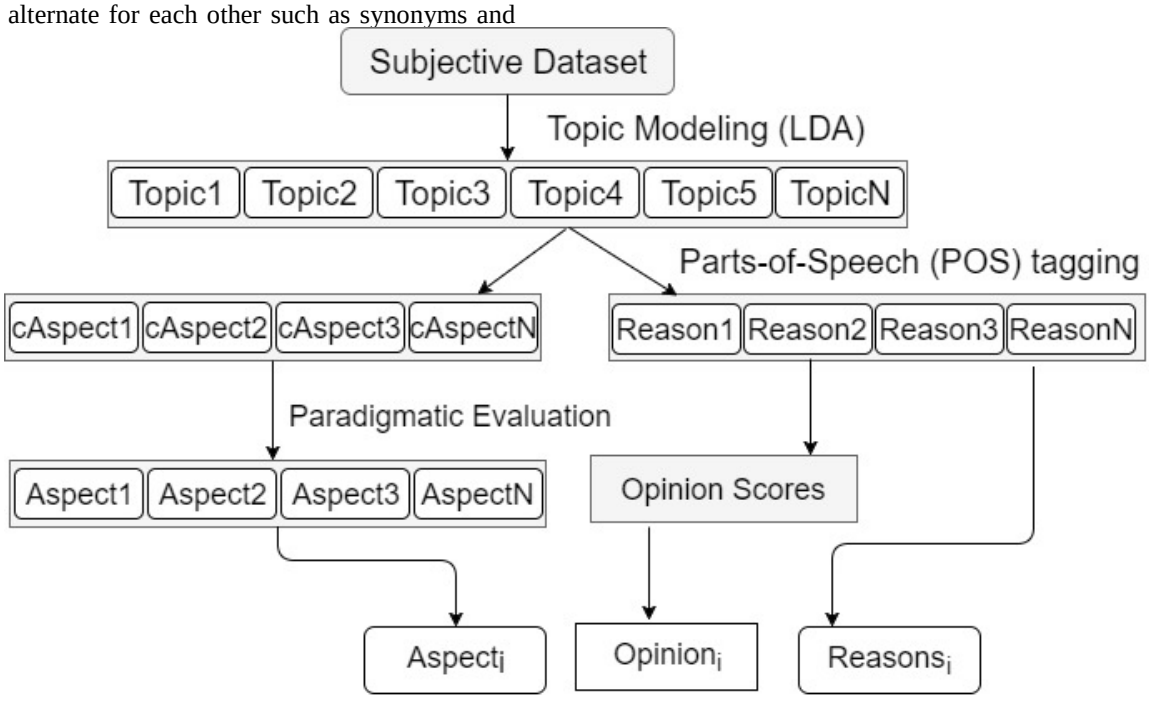

Figure 1: Opinion reason mining work flow using paradigmatic and syntagmatic word associations

A. Generating Raw Topics

LDA is the commonly used approach for topic modeling [9, 10]. It is a probabilistic model that treats data as arising from a generative process which is depicted as topic structures [23]. For a given set of documents, it can predict the topic structure and for a given topic structure it can generate more documents [12]. The documents and its words are the observed variables while the topics and their distributions are the hidden or latent variables [23]. It uses an inference technique to retrieve latent variables from the observable variables by updating the state of the model iteratively. Initially, it assigns words to topics at random, which gives the initial state of the model. Collapsed Gibbs sampling is used as an inference technique following markov chain [24]. The model state improves by iterations and is depicted by the grouping of words into topics [25]. The state of the model is represented by the document-topic and topic-term distributions. Finally, the topics $T = \sum_i t_i$ are generated from the dataset, where each topic hold contextually correlated words.

B. Separating Terms using Syntax

As mentioned earlier, the topic words may not be very coherent and may have intruded, random or imbalanced words that doesn't go well together with the other words of the topic [3]. Therefore, they require further filtering. The T topics extracted in the previous step are grouped as contextually co-related words, irrespective of their POS tags [1]. The POS tags as nouns and noun phrases are referred to as explicit aspect terms while the others i.e., verb, adverb, adjective are referred to as implicit aspect terms. However, implicit aspect terms not only indicate an aspect but provides more information as well. For example, the implicit aspect heavy not only indicate the aspect weight but also give a reason for why it receives negative sentiment [6]. The same goes for expensive, hard to fit, late delivery etc. Each topic ti is separated into two bins as $cAspectTerms_i$ and $cReasonTerms_i$.

C. Refining Aspect Terms

The candidate aspect terms $cAspectTerms_i$ from t_i may not necessarily contain words that hold semantic similarity. The objective is that the terms within a group for representing an aspect should be its synonyms and near synonyms. For example, the words picture, photo, pic, click etc. are used interchangeably for an aspect picture in the camera domain. In order to verify this, paradigmatic word associations are performed for all terms within the $cAspectTerms_i$. For any two terms w_1, $w_2 \in cAspectTerms_i$ the contextual similarity based on paradigmatic association as the context overlap where context is represented by the surrounding terms of w_1 and w_2, respectively. For the sake of simplicity only general context comprising of both left and right side words is considered in this case. It is evaluated for words w_1 and w_2 using Jaccard's Index [26] as,

Proc. of the Sixth Intl. Conf. Advances in Computing, Communication and Information Technology- CCIT 2018

ISBN: 978-1-63248-149-8 doi: 10.15224/978-1-63248-149-8-12

$$J(w_1, w_2) = |w_1 \cap w_2||w_1 \cup w_2| = |w_1 \cap w_2||w_1| + |w_2| - |w_1 \cap w_2|, \quad (1)$$

where $J(w_1, w_2)$ represents the overlap of general context for the terms w_1 and w_2.

The two terms are paradigmatically associated when their $J(w_1, w_2)$ score is high, sharing a common context most of the time. The terms w_i that has higher paradigmatic association with the other words of the same topic is evaluated as $\sum_j J(w_i, w_j)$. In order to retain w_i, it needs to maintain a high contextual overlap with most of the words of the topic or filtered otherwise. It leads to having aspect terms as the refined form of $cAspectTerms_i$. These terms are retained as $aspect_i$ to represent the i^{th} aspect if their value is above t_p while the other terms from $cAspectTerms_i$ are discarded. Table I depicts the improved aspects as better grouping of words when the weakly associated words are removed. Opinion mining is performed for the $cReasonTerms_i$ with the help of *WordNet* giving a score to each term. The score may be positive or negative as the polarity of opinion, while numeric value shows the strength of such opinion [1]. The scores of all terms in $cReasonTerms_i$ is aggregated to give a value that is assigned to $aspect_i$ as the average of all opinions expressed as,

$$opScore_i = \sum_{j=1}^{N} opVal(w_j),\ w_j \in cReasonTerms_i \quad (2)$$

TABLE 1: ASPECT EXTRACTION FROM TOPICS USING PARADIGMATIC ASSOCIATIONS

feeding	production	life story	meal
spoon	music	life	pasta
mouth	guitar	story	diet
food	musician	mother	sugar
bowl	recording	parent	friend
hard	production	real	meal
table	role	death	healthy
boy		father	drink

D. *Evaluation of Reason Terms*

Syntagmatic word associations are explored to identify popular reasons that frequently occur in the context of the aspect terms. Therefore, syntagmatic word associations are evaluated for aspect terms in $aspect_i$ with the corresponding candidate reason terms in $cReasonTerms_i$. Using conditional entropy [27],

$$H(w_{i,r}|w_{i,a} = 1) = -p(w_{i,r} = 0|w_{i,a} = 1)log_2 p(w_{i,r} = 0|w_{i,a} = 1) - p(w_{i,r}=1|w_{i,a}=1)log_2 p(w_{i,r}= 1|w_{i,a} = 1) \quad (3)$$

where $w_{i,a}$ is the known term, so conditional entropy suggests the drop in randomness or uncertainty of having or not having $w_{i,r}$ when we are certain about the presence of $w_{i,a}$. $w_{i,r}$ represent a term from $cReasonTerms_i$ for an $aspect_i$ represented as $w_{i,a}$. So, the objective is to calculate the certainty of having reasons $w_{i,r}$ with the aspect term $w_{i,a}$. Thus the above equation can be modified as,

$$aspect_i \leftarrow Reason_{i,j} = \sum_{j=1}^{N} \frac{H(cReasonTerms_{i,j}|aspect_{i,k})}{K} \quad (4)$$

for all the K terms representing $aspect_i$, the j^{th} reason term from $cReasonTerms_i$ has the highest syntagmatic relevance and is considered a reason for $aspect_i$.

IV. Results and Analysis

The proposed model is evaluated on the twitter dataset, having the general context defined as a sentence. The threshold for retaining the aspect terms t_s is specified to be *0.7*. These threshold value is evaluated experimentally producing better results across multiple datasets. A higher value limits the number of words representing the aspect while a higher value adds noise to it. Low conditional entropy relates to high mutual information and vice versa. The association among the aspect and reason term is evaluated as mutual information gained from knowing the aspect $w_{i,a}$ can be represented as [28],

$$MI(w_{i,r}; w_{i,a}) = H(w_{i,r}) - H(w_{i,r}|w_{i,a}) = H(w_{i,a}) - H(w_{i,a}|w_{i,r}) \quad (5)$$

Thus $MI(w_{i,r}; w_{i,a})$ is the drop in entropy of $H(w_{i,r})$ after we have known $w_{i,a}$ as $H(w_{i,r}|w_{i,r})$ or vice versa. The reasons higher mutual information i.e. above average, with their respect aspect terms are given in *Table II*. The first aspect is labeled as feeding while the possible reasons for it getting positive polarity are using hard bowl, having comfortable crib, and using safe plastic. Similarly, the reasons getting mutual information above average for the aspect meal are free, delicious, healthy and tasty. It shows promising progress in exploring the reasons for aspect terms. It could be beneficial in case where a meal having positive sentiments is preferred by some for being delicious but not healthy and vice versa.

TABLE 2: ASPECT TO REASON ASSOCIATION USING SYNTAGMATIC ASSOCIATIONS

Aspect	reasons
feeding	hard bowl, comfortable crib, safe plastic
production	popular musician, early recording, classical guide
life story	jeannette friend, night, short, dream, real situation
recipe	easy, fun, simple, picture, favorite food
ingredients	natural flavor, taste, organic, hot sauce, fresh
meal	free, delicious, healthy, tasty

V. Conclusion

The research in sentiment analysis and aspect-based sentiment analysis is approaching standard landmarks in its second decade. This has led to the interest of researchers to analyze deeply, giving rise to new research domains under the same umbrella. Opinion reason mining is one such research direction. The proposed approach makes use of statistical techniques and linguistic syntax to generate separate aspects represented by their respective groups of terms. These groups are refined and are mapped to reason terms while their polarity to be positive or negative. The ORMFW based application helps potential

Proc. of the Sixth Intl. Conf. Advances in Computing, Communication and Information Technology- CCIT 2018

ISBN: 978-1-63248-149-8 doi: 10.15224/978-1-63248-149-8-12

users and manufacturers to explore the key reasons that makes some aspects special as compared to others.

References

[1] M. T. Khan, M. Durrani, A. Ali, I. Inayat, S. Khalid, and K. H. Khan, "Sentiment analysis and the complex natural language," Complex Adaptive Systems Modeling, vol. 4, no. 2, pp. 1–19, 2016.

[2] I. Titov and R. McDonald, "Modeling online reviews with multi-grain topic models," in Proceedings of the 17th international conference on World Wide Web, pp. 111–120, ACM, 2008.

[3] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and trends in information retrieval, vol. 2, no. 1-2, pp. 1–135, 2008.

[4] M. T. Khan, S. Yar, S. Khalid, and F. Aziz, "Evolving long-term dependency rules in lifelong learning models," in Knowledge Engineering and Applications (ICKEA), IEEE International Conference on, pp. 93–97, IEEE, 2016.

[5] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168–177, ACM, 2004.

[6] M. T. Khan and S. Khalid, "Multimodal rule transfer into automatic knowledge based topic models," in Multi-Topic Conference (INMIC), 2016 19th International, pp. 1–6, IEEE, 2016.

[7] B. Liu, W. Hsu, and Y. Ma, "Mining association rules with multiple minimum supports," in Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 337–341, 1999.

[8] M. T. Khan, S. Yar, and S. Khalid, "Histogram based rule verification in lifelong learning models," in Multi-Topic Conference (INMIC), 2016, 19th International, pp. 1–5, IEEE, 2016.

[9] D. M. Blei, M. I. Jordan, et al., "Variational inference for dirichlet process mixtures," Bayesian analysis, vol. 1, no. 1, pp. 121–144, 2006.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," the Journal of machine Learning research, vol. 3, pp. 993–1022, 2003.

[11] W. Zhang, H. Xu, and W. Wan, "Weakness finder: Find product weakness from chinese reviews by using aspects based sentiment analysis," Expert Systems with Applications, vol. 39, no. 11, pp. 10283–10291, 2012.

[12] T. Wang, Y. Cai, H.-f. Leung, R. Y. Lau, Q. Li, and H. Min, "Product aspect extraction supervised with online domain knowledge," KnowledgeBased Systems, vol. 71, pp. 86–100, 2014.

[13] S. Brody and N. Elhadad, "An unsupervised aspect-sentiment model for online reviews," in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 804–812, Association for Computational Linguistics, 2010.

[14] T. Hofmann, "Probabilistic latent semantic indexing," in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 50–57, ACM, 1999.

[15] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic sentiment mixture: modeling facets and opinions in weblogs," in Proceedings of the 16th international conference on World Wide Web, pp. 171–180, ACM, 2007.

[16] S. Branavan, H. Chen, J. Eisenstein, and R. Barzilay, "Learning document-level semantic properties from free-text annotations," Journal of Artificial Intelligence Research, pp. 569–603, 2009.

[17] F. Li, M. Huang, and X. Zhu, "Sentiment analysis with global topics and local dependency.," in AAAI, 2010.

[18] W. X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly modeling aspects and opinions with a maxent-lda hybrid," in Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 56–65, Association for Computational Linguistics, 2010.

[19] Y. Lu and C. Zhai, "Opinion integration through semi-supervised topic modeling," in Proceedings of the 17th international conference on World Wide Web, pp. 121–130, ACM, 2008.

[20] A. Mukherjee and B. Liu, "Aspect extraction through semi-supervised modeling," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, pp. 339–348, Association for Computational Linguistics, 2012.

[21] H. Wang, Y. Lu, and C. Zhai, "Latent aspect rating analysis on review text data: a rating regression approach," in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 783–792, ACM, 2010.

[22] M. T. Khan, M. Durrani, S. Khalid, and F. Aziz, "Lifelong aspect extraction from big data: knowledge engineering," Complex Adaptive Systems Modeling, vol. 4, no. 5, pp. 1–15, 2016.

[23] D. M. Blei, "Probabilistic topic models," Communications of the ACM, vol. 55, no. 4, pp. 77–84, 2012.

[24] D. M. Blei and J. D. Lafferty, "Topic models," in In Text mining: classification, clustering, and applications, Srivastava, A. and Sahami, M. (eds), vol. 10, no. 71, 2009.

[25] X. Zheng, Z. Lin, X. Wang, K.-J. Lin, and M. Song, "Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification," Knowledge-Based Systems, vol. 61, pp. 29–47, 2014.

[26] R. Real and J. M. Vargas, "The probabilistic basis of jaccard's index of similarity," Systematic biology, vol. 45, no. 3, pp. 380–385, 1996.

[27] A. Porta, G. Baselli, D. Liberati, N. Montano, C. Cogliati, T. GnecchiRuscone, A. Malliani, and S. Cerutti, "Measuring regularity by means of a corrected conditional entropy in sympathetic outflow," Biological cybernetics, vol. 78, no. 1, pp. 71–78, 1998.

[28] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," Computational linguistics, vol. 16, no. 1,pp. 22–29, 1990.

About Author (s):

Prof. Dr. Shehzad Khalid is a Professor and Head of Department at Department of Computer Engineering, Bahria University, Pakistan. He graduated from Ghulam **Ishaq** Khan Institute of Engineering Sciences and Technology, Pakistan, in 2000. He received the M. Sc. degree from National University of Science and Technology, Pakistan in 2003 and the Ph.D. degree from the University of Manchester, U.K., in 2009. He Heads the Computer Vision and Pattern Recognition (CVPR) research group. His areas of research include Shape analysis and recognition, Motion based data mining and behavior recognition, Medical Image Analysis and text analysis.

Mr. M. Taimoor Khan is a PhD student at Bahria University, Islamabad, Pakistan. He is supervised by Prof. Dr. Shehzad Khalid in using machine learning techniques for text analysis and information extraction. He has received MS degree in Information Technology from Swinburne University, Melbourne, Australia in 2010. He has an undergraduate degree in Computer and Information Sciences (CIS) from PIEAS, Islamabad, Pakistan. His research interests' lies in text analysis, pattern recognition, machine learning and knowledge based models.