Aspect-based Sentiment Analysis on a Large-Scale Data: Topic Models are the Preferred Solution

M.Taimoor Khan, Mehr Durrani, Kamran H. Khan, Armughan ali and Shehzad Khalid

***Abstract* – Topic models are successfully used for text analysis to identify product aspects and their associated sentiments. There are various extensions of topic models that focus on specific problems of Aspect-based Sentiment Analysis. The hybrid and semi-supervised models are used to improve on the model accuracy at the cost of some training or expert guidance. Knowledge based topic models are very popular in other research areas having recently applied to Natural Language Processing. The model uses the large volume of data to get intuition from, which is used to improve on the accuracy of the model. Automatic knowledge based models learn in human like manner, having a never ending learning mechanism. The models are evaluated through topic coherence where a better model produces more coherent topics. Performance has been an issue for topic models as the inference techniques require higher number of iterations to converge. With the newly introduced sub-domains of Sentiment analysis i.e. bias analysis, emotion analysis, influence analysis and information leakage etc. the topic models are expected to evolve.**

***Index Terms* – Aspect based opinion mining, aspect-based sentiment analysis, machine learning, sentiment orientation, topic models**

I. INTRODUCTION

People prefer to seek others' opinions before making a decision. Automatic techniques applied on the huge review data available online can serve the purpose efficiently. Automatic techniques are used to process the content and extract useful information from it that can be used for decision making. It helps users to invest sensibly while manufacturers get feedback on their products to improve. According to the surveys, conducted by BusinessWeek (2008) and Comscore / Kelsey group (2007), more than 70% of people consult online reviews before making a purchase and found them to be honest. Sentiment analysis is the aggregation of public opinion towards useful information, also known as opinion mining.

The outcome of sentiment analysis can be a positive or negative label for a product, as a binary class classification problem. A neutral class is also considered in multi-class classification. The results can be converted into a summary which is hard to produce automatically, due to the NLP challenges. Summaries are preferred for human users as they convey more information than numbers. Sentiment analysis is in active research focus on social domains, with practical applications available for commercial domains.

M. Taimoor Khan and Dr. Shehzad Khalid, Bahria University, Islamabad. Mehr Durrani and Armughan Ali COMSATS Attock, Kamran H. Khan (Lecturer) University of Haripur. Email: taimoor.muhammad@gmail.com.
Manuscript received June 27, 2015; revised on September 10 and November 25, 2015; accepted on December 30, 2015.

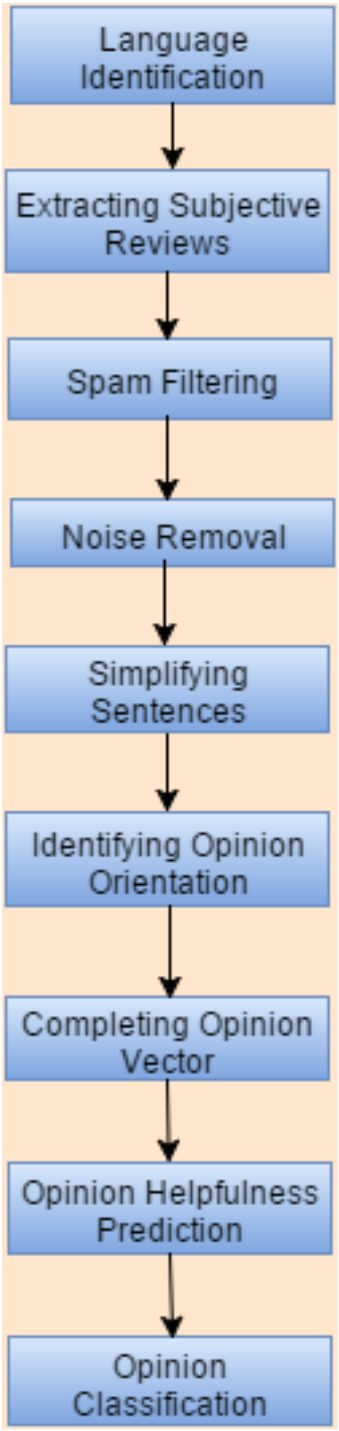

Fig 1: Flow model of aspect based opinion mining

ISSN – 1999-4974

ABSA is a two-step process i.e. aspect extraction and sentiment analysis with aspects as targets. Mainly Three types of approaches are used in the literature, which are Lexicon based, machine learning and statistical evaluation based techniques. Lexicon based techniques benefit from an external dictionary e.g. WordNet to identify the strength and polarity of sentiment words. It doesn't require training data and doesn't face the problem of overfitting but miss the context too. However, they are ineffective towards specialized domains where sentiments are context dependent. For example "Long batter time" is a positive sentiment for a Laptop but "Long start-up time" is a negative sentiment. Learning techniques require labeled data which is expensive and time consuming to produce. Therefore, they are mostly preferred for sensitive domains where high accuracy is desired. Frequency and relation based techniques are trapped by the variation in NLP as identifying all possible scenarios in which an aspect may exist is impossible considering the richness of natural language. Statistical evaluation techniques e.g. Topic Models favor the scenario as they require abundance of data to produce stable results and do not require any manual tuning. Topic models are based on LDA, drawing the evaluations from within the corpus and therefore, provide contextual categorization. For example, a topic model may put VGA and Card together in a topic for being highly co-related in the Electronics domain.

Online sources are more suitable for ABSA where the sentiment data can be found on various blogs, forums and other review and social media websites. The content is fresh and is produced by amateur authors from different cultures, locations, religion etc. Blogs and forums are rich in information where discussions in favor or against an opinion can help find the reasons for it. However, since they are not professional authors, all sorts of mistakes and inconsistencies are expected in this review. The data is to be pre-processed to filter noise and stem words to their roots and fix other problems so that the accuracy of the technique can be maintained. Some commonly faced problems are over or under utilization of capitalization, swearing, spelling mistakes, shortened words, slang etc. Topic models extract latent topics from a large collection of documents. Unlike classification, it assigns many topics or labels to a document. It favors aspect extraction as there are multiple product aspects discussed together in a single document. In order to avoid global topics and focus on local topics as aspects, the window size or unit of analysis is to be as small as a phrase. The words appearing together under a topic are synonyms and near-synonyms that used to represent an aspect or sentiment.

In Section 2, the nature of subjective data is reviewed. Section 3 reviews the statistical topic models used in the study. Section 4 provides a discussion while Section 5 gives a conclusion on the use of topic models for aspect based sentiment analysis.

II. NATURE OF SUBJECTIVE DATA

Opinion or sentiment mining is the subjective analysis of users' opinions towards various issues or products. It's different from text categorization, where relevant fragments of text are grouped together. In subjective analysis, the words that convey subjective sense e.g. *adverbs* and *adjectives* have higher importance than other words. Text categorization considers the documents to be identical if they have high frequency of co-occurring words. However, this is not true about opinion mining, where a single word can change the whole context of the scenario. For example, saying *"I am happy with my new iPhone"* is the complete opposite of saying *"I am not happy with my new iPhone"*, even though all the words other than "not" are the same. In order to perform sentiment analysis the opinion or sentiment vector has to be identified as given in Eq.1:

$$OpinionVector = (e_j, a_{jk}, h_i, t_l, so_{ijkl}) \quad (1)$$

The opinion vector possess the necessary information required to process a single opinion. Where in Eq.1 e_j is the target entity, a_{jk} is the opinion holder or the person who is of the mentioned opinion. Identifying the opinion holder is important if the author shares opinions of the other people they know. t_l represents the time of the opinion. Normally this information is not mentioned in datasets, however, it is very important to find the popularity graph of a product, as people tends to change their opinions with time based on the recurring events.

A positive aspect about sentiment analysis is its dependency on the content produced by general public. This content is available in huge volume on social media websites, blogs, forums and product review websites. The structural information available on these platforms can be used to improve and verify the accuracy of the model. Structural information as also referred to as meta-information available with the content. For example, *likes* and *shares* on *Facebook*, *tweets*, *retweets* and *trends* on *Twitter* etc. Other meta-information includes author, date, author's location, education, political views etc.

Associating an opinion to its target aspect of the given entity is easy with simple sentences. However, natural languages tends to have all sorts of variations, which makes it harder for the analysis model to associate sentiment with their respective entities and aspects. Comparative sentences have the aspects of multiple products, discussed in comparison to each other. Compound sentences have many aspects and their sentiments discussed together using connectives. The sentiment modeling techniques used for aspect / sentiment extraction is adversely affected by these types of sentences, as it doesn't consider semantics in a sentences. The sentences that are hard to evaluate are ignored or broken down into few simple sentences at the pre-processing stage.

ISSN – 1999-4974

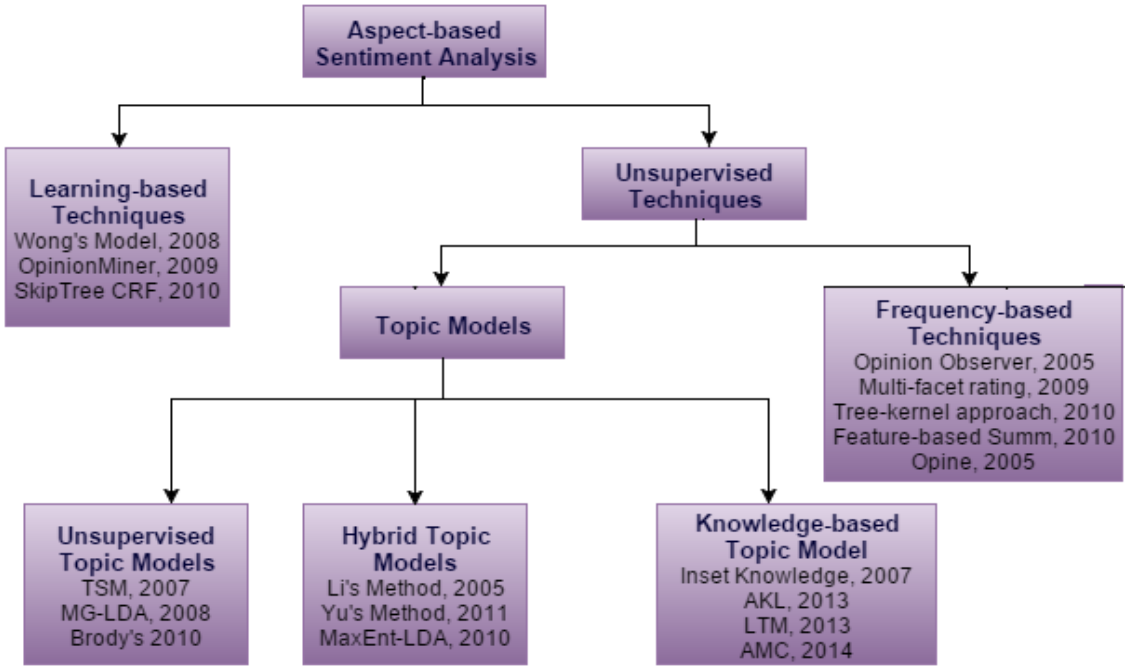

Fig 2: Taxonomy of topic extraction models for ABSA

III. REVIEW OF TOPIC MODELS

Despite of ignoring semantics in text, topic models perform better than other techniques and are preferred for text analysis. The unsupervised topic models are efficient, cheap and can be directly applied to fresh data, but have low accuracy as they produce many incoherent topics as well. These incoherent / incorrect topics does not have strong semantic relation among its words to convey meaningful information. Considering the benefits of the unsupervised topic models, most of the research work is focused on improving their accuracy. Topic models are provided with a small set of seed aspects, on which the model grows to extract many other aspects with higher accuracy. The domain specific seed aspects are provided by domain expert in an unsupervised manner. The hybrid models approached the problem with a solution that needs training a classifier on a small labeled data to evaluate better hyper-parameters for topic models. There are also some knowledge based models proposed that improve the accuracy of topic extraction for big data consisting my many domains. In a product domain a topic represents an aspect, where the topic words are the synonyms and near synonyms used to refer to it. The literature covered for aspect based topic modeling is shown in Fig2.

The topic models used for Aspect-based sentiment analysis are extended differently to improve their performance with subjective text analysis. The extensions of topic models can be categorized as Supervised, Unsupervised, Semi-supervised, Hybrid Models, Transfer learning and Knowledge-based topic models. Some of the most frequently used techniques are discussed.

A. Hybrid Topic Models

Hybrid topic models combine probabilistic topic models with a supervised classifier using a small labeled dataset. Initially the classifier is used to generate appropriate starting values for the topic model hyper-parameters. With the hyper-parameters initialized, topic models is applied to the unseen data in an unsupervised manner. MaxEnt-LDA, a hybrid model is used to discover aspects and sentiments and the separation is performed through syntactic patterns [1]. Multinomial distributions are performed for the extracted word to indicate whether it is aspect or sentiment word, while it learns initial values from training data. More hybrid models are used for reliable separation between aspects and sentiments by training the model with a labeled dataset [20, 21].

MaxEnt has separate bins for aspects and sentiments to prevent them from falling into the same topic, which doesn't make sense in this problem domain. They are further distributed into general and specific groups. General aspects be the frequent aspects found with many other products while specific group holds the domain specific aspects. The symmetric Dirichlet prior parameter β possess a value predicted by the Maximum Entropy classifier. The model predicts several multinomial from the value of β which are background model along with the general and specific models for both aspects and opinions. The words in the corpus are placed in any of the categories based on its type that is entity or aspect and then how frequently it occurs to decide between general and specific. Maximum entropy model is applied to the feature vector having values that draws the separation among various categories. The indicator variables are used to support the decision making process.

ISSN – 1999-4974

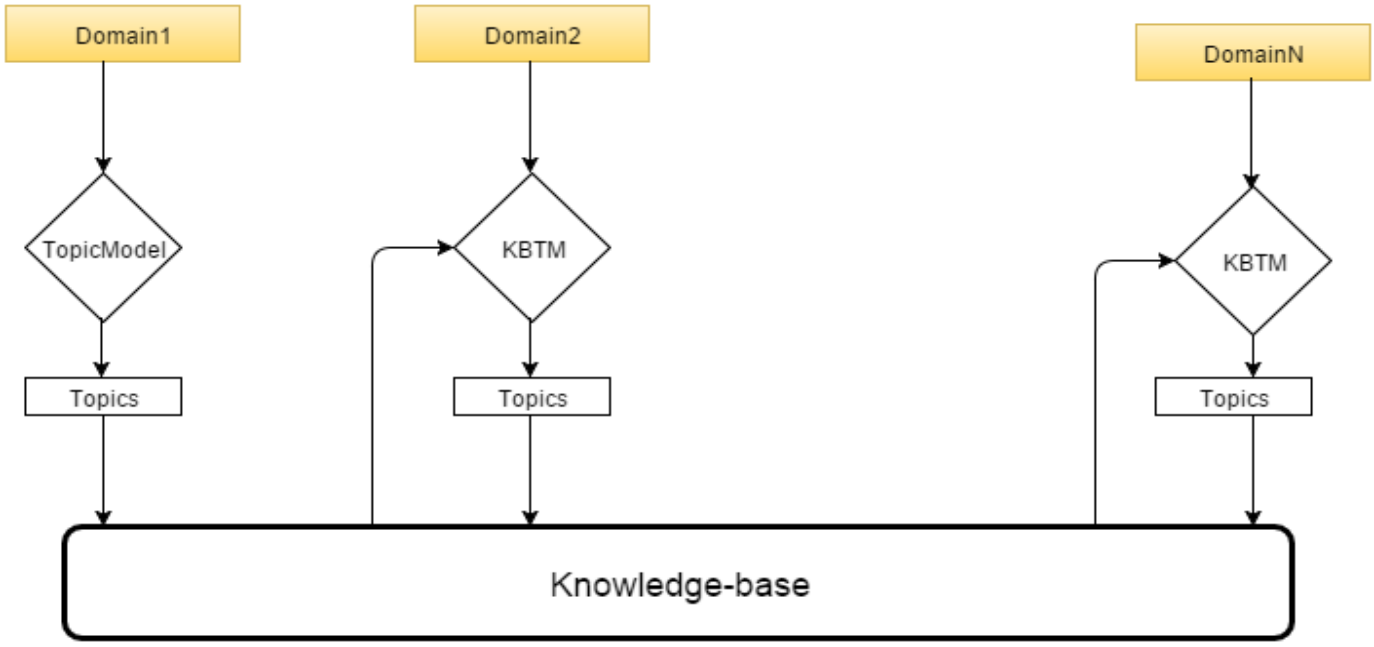

Fig 3: Automatic knowledge-based topic model using complex network

B. Unsupervised Topic Models

The unsupervised topic models extract topics and their distribution by using LDA. However, the results are usually not satisfactory having many incoherent topics. Multi-grain LDA or MG-LDA [2] is proposed for a special review type documents. These review documents are expected to have ratable aspects separate from the review content. The ratable aspects are used as intuition into the model which finds other aspects from the review content with higher accuracy. The new aspects are explored based on their co-existence with the known aspects. A common problem with topic models is that of having products and entities grouped together in topics. The entities or products are more widely used and therefore influence the topic models with higher probabilities. In order to address this problem, MG-LDA operates in two steps. It is applied to the corpus initially to identify global topics which are assumed to be products and brands. The model is applied again at the sliding window to focus on local topics which have high chances of domain specific aspects as topics. LDA also face the problem of over fitting with rise in the size of the corpus. It incorporates expectation maximization algorithm to estimate the hyper-parameters for the model.

S. Brody et al in [3] proposed another variation of LDA to extract sentiments and their target aspects. The model is applied at the sentences as documents to avoid global topics. It extract clusters which are validated for the suitable number of clusters in the given domain. Topics are converted into clusters having sentences assigned to it. With the help of two connectivity matrices the sentences and their clusters are identified.

C. Knowledge-based Topic Models

Knowledge-based LDA are used for aspect extraction having a knowledge module. Instead of domain specific seed aspects, two sets of rules are provided to it as input. The rules are also domain specific and are provided by human expert following a semi-supervised approach. The sets of rules are must-link and cannot-link that decides which of the aspect-terms can co-exist under the same topic or in an aspect cluster for a given domain. It empowered the model to decide where to place an extracted aspect term by incorporating its knowledge. A problem identified with this model is that of being two strict for its rules which limited the use of knowledge to certain scenarios. The rules were relaxed in [22] to can-set and cannot-set. Instead of having a fixed one-to-one relationship for co-existence, it has two pools of words where a word in can-set have high probability of co-existence with any other word in the same pool for an aspect. Chen et al. [24] proposed the concept of automatic knowledge-based LDA that learns knowledge automatically without any user intervention. The model learns adaptively, is independent of the nature of domains to which it is applied and can scale up to a large dataset consisting of multiple domains. Fig3 shows an automatic knowledge-based topic model that maintains a knowledge-base after performing each task. The knowledge learnt is be used for relevant future tasks. These models are also called Lifelong learning models. They exploits the huge size of data and its variety of formats to its own advantage for verifying its knowledge [37]. It benefits from the grey region overlapping among various domains.

IV. DISCUSSION

Sentiment analysis is an area of diversified research fields including machine learning, natural language processing, and language identification and text summarization. Most of its issues are related to NLP which are quite complex and in research focus. Sentences that do not possess any subjective orientation are discarded as objective sentences having factual information. However, both aspects and sentiments can be implicit as well. For example, *"I charge my phone once in few days"* is apparently an objective sentence having no sentiments but there is an

ISSN – 1999-4974

implicit praise for long battery life. Exploring implicit details are very hard and require a lot of contextual information. The NLP issues discussed affect all Sentiment analysis techniques, however, the learning based techniques are more vulnerable to it.

Opinion orientation has a context inclined towards psychology and linguistics. Complex networks can help in resolving context through preserving sequence and by associating words in a sentence, sentences in a paragraph and even paragraphs in an article. Sentiment analysis has its roads crossed with many different research areas and therefore, its problems are to be addressed with solutions coming from areas other than machine learning.

Sentiment analysis and opinion mining is out of its earlier stages and there is a strong need to standardize the datasets and evaluation methodology. Accuracy, area under the curve, precision / recall and F-measure are frequently used for evaluation. The bag-of-words approach do not contain information about context and proximity and therefore, needs to be replaced with concept-centric approach. There are a variety of semantic repositories available online to benefit from e.g. WordNet, SentWordNet, WordNetAffect,

Microblogging (twitter) and transcribed text is unstructured having more noise and therefore, lexicon-based techniques do not perform well. Similarly depending upon the nature of platform structural information can also be incorporated e.g. likes, share, retweets, hashtags etc. Machine learning techniques are more supportive to accommodate structural information e.g. meta-data as non-textual features. ML techniques depend on the feature set to which proximity and context based features can also be added. Transcribed text is also used for sentiment analysis which introduce a new type of textual content. It also contain terms like *"Emm"* and *"Aah"* etc. that doesn't have any meaning but are used while speaking. Similarly sentences are left incomplete and grammar is ignored. This opens new avenues to these techniques to deal with this type of content.

V. CONCLUSION

Aspect based sentiment analysis are more resourceful and therefore, most of the recent work is focused on it. It does not only recommend or non-recommend a product or service but also provide feature/aspect based details about them. However, finding product aspects for a new domain is a problem of its own. Topic models are used in combination with approaches from other research areas to improve on accuracy while keeping the training cost manageable. Knowledge based models are recently introduced to the NLP tasks and is a hot research area. It has a promising future because it doesn't require training or domain experts which are expensive to found and only benefits from the large amount of data. This data is frequently available of social media and other review websites and can be extracted freely. The subjective data on social media is recently referred to as the big social data [31, 34]. New sub-domains of sentiment analysis and opinion mining are identified that are of interest to the research community including bias analysis, threat / danger analysis, emotion analysis, influence analysis and information leakage. Sentiment analysis is used to highlight the health care problems from the perspective of a patient [28].

REFERENCES

[1] W. Zhao, J. Jiang, H. Yan and X. Li, "Jointly modeling aspects and opinions with a maxent-lda hybrid", Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP '10, pages 56–65, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[2] I. Titov and R. McDonal, "Modelling online reviews with multi-grain topic models", proceedings of 17th International conference on WWW, pages 111-120, New York NY, USA, 2008, ACM.

[3] S. Brody and N. Elhadad, "An unsupervised aspect-sentiment model for online reviews", Proceedings of Human language technologies, Annual conference of north America chapter of the association for computational linguistics, pages 804-812, USA 2010.

[4] 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 WWW, pages 171-180, New York, USA, 2007, ACM.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet allocation". J. Mach. Learn. Res., 3:993–1022, March 2003.

[6] Thomas Hofmann. "Probabilistic latent semantic indexing", In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pages 50–57, 1999.

[7] Bing Liu. "Sentiment Analysis and Opinion Mining". Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[8] Pang B. and L. Lee, "Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval" 2: 1-135, 2008.

[9] Chau, M., & Xu, J., "Mining communities and their relationships in blogs: A study of online hate groups", International Journal of Human – Computer Studies, 65(1), 57–70.

[10] Somprasertsri G., and Latitrojwong P., "Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization", J. UCS Number 6, Vol16, pages 938-955, 2010.

[11] Jindal N., and Liu B., "Mining Comparative Sentences and Relations", AAAI Press, AAAI pages 1331-1336, 2006.

[12] Qiu G. et al, "Opinion Word Expansion and Target Extraction through Double Propagation", Computational Linguistics, no1, vol37, pages 9-27, 2011.

[13] Zhai Z. et al, "Grouping Product Features Using Semi-Supervised Learning with Soft-Constraints", {COLING} 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China, edition Chu-Ren Huang and Dan Jurafsky, pages 1272-1280, 2010.

[14] Ding X., and Liu B., "Resolving Object and Attribute Co-reference in Opinion Mining", {COLING} 2010, 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China, Tsinghua University Press", pages 268-276, "Chu-Ren Huang and Dan Jurafsky,2010.

[15] Patella M., Ciaccia P., "Approximate similarity search: {A} multi-faceted problem", J. Discrete Algorithms, Number1, volume 7, 2009.

[16] Bollen, Johan, Huina Mao, and Xiao-Jun Zeng. "Twitter mood predicts the stock market". Journal of Computational Science, 2011.

[17] Bar-Haim, Roy, Elad Dinur, Ronen Feldman, Moshe Fresko, and Guy Goldstein. "Identifying and Following Expert Investors in Stock Microblogs" in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2011). 2011.

[18] Feldman, Ronen, Benjamin Rosenfeld, Roy Bar-Haim, and Moshe Fresko. "The Stock Sonar - Sentiment Analysis of Stocks Based on a Hybrid Approach". in Proceedings of 23rd IAAI Conference on Artificial Intelligence (IAAI-2011). 2011.

[19] Zhang, Wenbin and Steven Skiena. "Trading strategies to exploit blog and news sentiment". in Proceedings of the International Conference on Weblogs and Social Media (ICWSM-2010).

 ISSN – 1999-4974

[20] Griffiths, Thomas L., Mark S., David B., and Joshua T., "Integrating topics and syntax", Advances in Neural Information Processing Systems, 2005. 17: p. 537–544, 2005.
[21] Liu, J. et al., "Low-quality product review detection in opinion summarization", Proceedings of the Joint Conference on Empirical Methods in NLP and Computational Natural Language Learning, 2007.
[22] Andrzejewski, D. and Zhu, X., "Latent dirichlet allocation with topic in-set knowledge", In Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing, pp. 43–48, 2009.
[23] Chen, Z., Mukherjee, A., and Liu, B., "Aspect extraction with automated prior knowledge learning", In Proceedings of ACL, pp. 347–358, 2014.
[24] Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., and Ghosh, R., "Discovering coherent topics using general knowledge". In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, pp. 209–218, 2013.
[25] Cambria, E. and White, B., "Jumping NLP curves: a review of natural language processing research". Computational Intelligence Magazine, IEEE, 9(2), 48-57 2014.
[26] Gangemi, A., Presutti, V. and Reforgiato Recupero, D., "Frame-based detection of opinion holders and topics: a model and a tool". Computational Intelligence Magazine, IEEE, 9(1), 20-30 2014.
[27] Poria, S., Gelbukh, A., Hussain, A., Howard, N., Das, D. and Bandyopadhyay, S., "Enhanced SenticNet with affective labels for concept-based opinion mining". IEEE Intelligent Systems, (2), 31-38 2013.
[28] Khan, M. T., and Khalid, S., "Sentiment Analysis for Health Care". International Journal of Privacy and Health Information Management (IJPHIM), 3(2), 78-91. doi:10.4018/IJPHIM.2015070105 2015.
[29] Katz, G., Ofek, N. and Shapira, B., "ConSent: Context-based sentiment analysis". Knowledge-Based Systems, (84), 162-178 2015.
[30] Guellil, I. and Boukhalfa, K., "Social big data mining: A survey focused on opinion mining and sentiments analysis". In Programming and Systems (ISPS), 2015 12th International Symposium IEEE, 1-10 2015.
[31] Jaafar, N., Al-Jadaan, M. and Alnutaifi, R., "Framework for Social Media Big Data Quality Analysis". In New Trends in Database and Information Systems II. Springer International Publishing 301-314 201.
[32] Machova, K. and Marhefka, L., "Opinion Classification in Conversational Content Using N-grams". In Recent Developments in Computational Collective Intelligence. Springer International Publishing 177-186 2014.
[33] Medhat, W., Hassan, A. and Korashy, H., "Sentiment analysis algorithms and applications: A survey". Ain Shams Engineering Journal, 5(4), 1093-1113 2014.
[34] Nguyen, D.T., Hwang, D. and Jung, J.J., "Time-Frequency Social Data Analytics for Understanding Social Big Data". In Intelligent Distributed Computing VIII. Springer International Publishing 223-228 2015.
[35] Rabade, R., Mishra, N. and Sharma, S., "Survey of influential user identification techniques in online social networks". In Recent Advances in Intelligent Informatics. Springer International Publishing 359-370 2014.
[36] Tang, J., Chang, Y. and Liu, H., "Mining social media with social theories: A survey". ACM SIGKDD Explorations Newsletter, 15(2), 20-29 2014.
[37] Tang, J., Hu, X. and Liu, H., "Social recommendation: a review". Social Network Analysis and Mining, 3(4), 1113-1133 2013.
[38] Tuveri, F. and Angioni, M., "An Opinion Mining Model for Generic Domains". In Distributed Systems and Applications of Information Filtering and Retrieval. Springer Berlin Heidelberg 51-64 2014.
[39] Zhang, L. and Liu, B., "Aspect and entity extraction for opinion mining". In Data mining and knowledge discovery for big data. Springer Berlin Heidelberg 1-40 2014.

 ISSN – 1999-4974