Potential Issues and Prospective Solution for Sentiment Techniques

Arpita, Pardeep Kumar, Kanwal Garg

Abstract*: With fast growing internet feedbacks regarding products, policies, services, people are being generated in large volume via different social networking sites or online portals. Data generated through these sites have great significance to numerous applications. But this bulk of content generated need to be assessed using appropriate mining techniques to extract information which can then proven to be useful. Sentiment analysis is the process of analyzing a text to know attitude of writer towards a text to be positive, negative or neutral. At one side sentiments are proven to be beneficial, on the other hand evaluation process of sentiment analysis is vulnerable to various challenges. These challenges act like a hurdle in interpreting the accurate intimation of sentiments and identifying their suitable polarity. This paper has put forward an overview of challenges faced in sentiment analysis and relevant techniques of some researchers acknowledging those issues. Probable solutions for various issues given by varied analysts have also been discussed in the paper.

***Index Terms*: Feature selection, Opinion Mining, Objectivity, Subjectivity Detection.**

I. INTRODUCTION

One of the biggest source of big data in today's era are the social media sites. Social media brings people together so that they can generate ideas, share information or their experiences with products, they have used with each other. This accumulation of data creates a huge unstructured big data which is analyzed for making decisions. Big data from such place is considered as a great source of real time estimation due to its high frequency of creation and low cost integration[32]. From past few years to assess the feelings of social media users towards a subject, a common method called sentiment analysis or opinion mining is increasingly been used. It is defined as the process of analyzing a text to know attitude of writer towards a subjective text to be positive, negative or neutral [19]. Therefore, in the domain of sentiment analysis, first major step is to identify subjective content out of accumulated big data. Feature selection is a process which is best suited for the purpose. Primary goal of feature selection in addition to identification of subjective content is to enhance the performance of classifier by reducing the feature vector, selecting only useful and pertinent features. It involves removal of redundant, irrelevant and noisy features. Reduction in feature vector ultimately leads to reduction in attributes, which result in improving the accuracy of classifier, and reduces the time needed for classification.

Though, feature selection finally result in reducing the number of attributes in a dataset which is also the ultimate goal of dimensionality reduction, yet these two approaches are different from each other in context that, feature selection on one hand works by mere inclusion and exclusion of attributes from the dataset without changing them, while dimensionality reduction on other hand do so by creating new combinations of attributes. Further, feature selection follows the process of feature extraction. The main objective of feature extraction is to transform data into well represented format based on their features [13]. There are some steps which need to be followed in advance for accomplishing the task of feature selection and extraction which includes case Normalization, Tokenization, Stemming process etc. The task of feature selection can then be assisted by a process of providing weight-age to each individual feature based on frequency of occurrence of that particular feature in the text. This process is commonly known as feature weighing.

Many feature selection algorithms have been used successfully for categorization of opinion bearing text. Information Gain [17], Mutual Information [26], Term Strength[37] and Chi-Square Statistics [12] are some of the data mining methods used for feature selection but considering volume of data generated through social media cites, use of optimization techniques like Genetic Algorithm [35] and Ant Colony Optimization [33], for feature selection have been proven to be more efficient method in recent years.

II. BACKGROUND

Large amount of work has been centered around the issue of sentiment analysis at different dimensions.

In the beginning era of sentiment analysis, Hatzivassiloglou and Wiebe [9] in their work attempted to identify features stipulating usage of subjective language but the work presented could only help with recognition of opinion bearing documents and significance of opinion used was still a mark of question. Pang et al. [28] and Tan and Zang[34] worked on classification of text based on overall orientation of document and faced the problem of

Revised Manuscript Received on December 22, 2018.

Arpita, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, India.
Pardeep Kumar, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, India.
Kanwal Garg, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, India.

contextual polarity. Hence, Nasukawa and Yi [25] and Wilson et al. [36] came up with the solution by carrying out their research at phrase level. The method included understanding of semantic of text at sentence level through various axioms. Also, Wilson et al. [36] in their research prompted issue of neutral word recognition as sometimes non neutral words also appear in text as neutral words. The issue was later on contemplated by Ramanujam et al. [30] in hadoop environment for analysis of semi-structured big data generated by customer reviews regarding a product. For sentiment classification at final step Ramanujam et al. [30] used lexicon based approach while machine learning approach was used over hadoop framework by Liu et al. [20]. Mullen and Collier [24] worked on complex data from diverse source. Their major focus was on text representing the tone of entire document which aroused problem of thwarted expectation. Cambria et al.

[5] and Jain and Katkar [11] propounded an approach for word level sentiment analysis which failed in comprehending implicit sentiment and sarcasm. Tan and Zhang [34] and Povoda et al. [29] explored language independent preprocessing and classification using machine learning approach. Ahmad et al. [2], Lin et al. [18], Zheng et al. [38] and Abualigah et al. [1] in their results they have manifested that meta-heuristic approach for feature selection provide better result in terms of precision and computational time as compared to traditional methods.

Henceforth, it can be concluded that ultimate goal of sentiment analysis is to evaluate opinion bearing text as positive, negative or neutral. Various researchers have worked in this direction merging numerous attributes with protocols used for classification so as to handle shortcoming of each other and take advantage by working in collaboration. Also, it can be seen that use of meta heuristic techniques in performing feature subset enhance the performance metrics. However, the primary obstacle which have been faced by researchers in carrying out this job were to understand the hidden meaning of text.

III. POTENTIAL ISSUES

Though the task of text categorization for sentiment analysis seem an easy job compared to traditional topic based classification, as only three classes need to be considered, yet it face numerous challenges. A challenging aspect of sentiment analysis, that seems to distinguish it from traditional topic-based classification is that, while topics are often identifiable by keywords alone, sentiments can be expressed in a more subtle manner. Some of potential issues induced as a result of this are:

A. Subjectivity Detection.[9]

Sentiment analysis at very first level face problem of distinguishing subjective content from objective con- tent. Objective content basically comprise of the facts like “Sun rises in the east” while all the opinion bearing content like “One plus has good camera quality” can be categorized as subjective content. As opinion mining is centered to analysis of sentiments or opinion therefore extraction of subjective content is the key step for sentiment analysis.

B. Implicit Sentiment and Sarcasm.[11]

Sometimes, explicitly text doesn’t carry any negative opinion bearing idiom but in all its orientation is negative for example “How can anyone buy these shoes?” Though, there is no negative word in the example taken here yet overall orientation of sentence is negative. Thus, identification of semantic is more significant in opinion mining than detection of syntax.

C. Thwarted Expectation.[24][25]

Occasionally, the context is set deliberately by author which finally is to be refused for example “This play should be creative. Theme sounds motivational, actors are marvelous and the cast in support is good as well. Still, it could not hold up.” Last sentence in this example changes the context of entire passage here. Hence, in such cases the text representing tone of entire document need to be considered for determining the polar- ity of document in end.

D. Word Knowledge.[3]

However, researchers have given abundant of resources (such as SentiNet [23], SentiNet3 [5], ANEW [21], SentiStrength [31]) for inducing the intensity of opinion bearing idioms but here the problem lies in the limitation of lexicon based approach with the scope of resources available. Hence, the method may fail in cases where words considered for analysis are not defined already in any of the resource. For example “Krish is Frankenstein”. Here, meaning of word “Frankenstein” need to be known for defining opinion of sentence.

E. Contextual Dependency.[15]

In some areas the words are same but the context in which they are used makes the difference. For example

“I hate love stories.”

I do not like the movie “I hate love stories.”

Here both sentences use exactly same words but the context with which words used are totally different. In first statement person in general puts forward his dislike for love stories while second statement is regarding a particular movie with the name “I hate love stories.” Such understanding for correct analysis is very crucial.

F. Entity Identification.[34]

Recognition of semantic association between sentiment expression and the subject regarding which an opinion is generated, plays a major role for enhancing the accuracy of overall analysis. For example “BMW is better than Mercedes”. In this sentence the senti- ment is positive for one domain i.e. BMW and nega- tive for another i.e. Mercedes. In such scenarios rather than classification of entire document towards one po- larity, it is very important to examine the data fragment thoroughly and identify polarity of text with respect to each domain used.

G. Negation.[4]

There are a few adverbs which when used with an idiom, reverse its polarity.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

For example "Apple is not good." Here the word "not" is an operator which is reversing the positive polarity of "good". Such adverbs are known as minimizers and have a negative impact on an opinion. Even worse is that, at some places use of these minimizers intensify, instead of negating the polarity [36]. For example "The show was not only good but outstanding". Despite of the use of minimizer here polarity of sentence is positive.

H. Scope of Negation.[6][25]

Beside, understanding of negation words it is very critical to analyze the scope to which negated operator has its impact. For example "I do not like driving a scooter but i like driving car". Here, dislike for driving is only limited to car and is not carried forward to car. Hence, the scope of operator "not" is till scooter only. Such cases should also be under consideration while mining sentiments for better outcome.

I. Identification of Neutral Words.[30][36]

Very often words with non neutral prior polarity appear in text as neutral words. For example "Dayal Singh Public college belongs to Saraswati Trust". Although, word "trust" has prior polarity as positive but in the sentence used above it is used in neutral oriented word. Subsequently, categorization of such words to positive class, bring a state of inaccurate classification.

J. Language Dependency.[29][10]

Most of the tools, resources, lexicons, libraries etc. for sentiment analysis are developed for only English language and implementation of these resources to other languages generate inaccurate results. Consequently, the opinionated text in other language is prevented from being utilized.

IV. PROSPECTIVE SOLUTION

This section briefly provide a survey on previous work on challenges of sentiment analysis.

A. Subjectivity Detection.

Hatzivassiloglou and Wiebe [9] have stated that dif- ferent variants of adjectives (i.e. grad-able adjectives, dynamic adjectives and adjectives which are semanti- cally oriented) has a considerable impact on determi- nation of subjective content. [9] have studied effect of these variants and have proposed a trainable method by statically merging two different benchmarks of grad- ability while an approach of minimum cuts in graph have been used by Pang and Lee [27] for subjectivity detection.

B. Implicit Sentiment Sarcasm.

Several rules have been compiled by Maynard and Greenwood [22] in their research work for handling the situation when it is known about the presence of sarcasm by tokenization via hashtag(e.g. What a wonderful day..!!#sarcasm). Although, they were successful in achieving 91% precision for detection of sarcasm with hashtag combination yet determining the scope of sarcasm was still a point of vulnerability in their work(For e.g. I am not happy that I woke up with a fight today. #great start #sarcasm. Here sarcasm is limited to "great start" only but "#sarcasm" to end of statement may five an illusion of association with com- plete sentence). Later on Gidhe and Ragha have cen- tered their research to discernment of sarcasm without #sarcasm using Multilayer Perceptron Back Propaga- tion [8]. However, deducing sarcasm scope is still a challenging job to accomplish.

C. Thwarted Expectation.

Pang et al. have put forward that the issue of thwarted expectation can be handled by evaluating that whether the text is on topic or not[28] and a while later Support vector machine modeling was used by [24] for study- ing the text representing tone of whole document [24] which helped in resolving issue of thwarted expectation to some extent.

D. Entity Identification.

Nasukawa and Yi have demonstrated that the identification of entity in whose respect sentiment is been ex- pressed need examination of semantic relation among the subject and sentiment expression and therefore, syntactic parser was introduced in their course of findings[25] . Nasukawa and Yi carried out this work with precision of 95% whilst Kanayama and Nasukawa performed an on demand analysis of narrow domain with precision of 96%. An annotated corpora have been used for helping the lexicon to be expanded automatically.

E. Negation.

Di Caro and Grella contended with the fact that with- out consideration of interaction between words, po- larity of sentence cannot be evaluated as there are some words like "not", "hardly" which may change the meaning of words following them. In the wake which an approach of dependency parsing was propounded so as to deal with the such negation operators [7]. Nevertheless, as given by [36] intensification of polarity with usage of such minimizer operators is still a topic of interrogation.

F. Identification of Neutral Words.

Wilson et al. noticed that usage of non neutral words as neutral affect accuracy of sentiment analysis and so a two step model was introduced for dealing with the problem [36]. At first step, several cues were used to distinguish neutral words from polar ones and then orientation of polar words was determined using subjectivity lexicon with prior polarity. A while later Ramanujam et al. evaluated another model for saving neutral words from categorized as positive or negative. The model given was implemented in MapReduce environment for better analysis on semi-structured data [30].

G. Language Dependency.

It has been noticed that mostly the lexicons or tools put forward for sentiment analysis have been focusing on English language itself but then there are few researchers who have demonstrated their words towards some other language like Tan

and Zhang [34] have given a Chinese Corpus for examining the reviews of Chinese document at a Chinese blogging site and Kaur and Gupta [16] have focused on inspection of Punjabi reviews. Similarly, Povoda et al. [29] have used optimization technique for exploring language independent pre-processing and classification and Hogenboom et al. [10] have come up with lexicon based sentiment analysis with multi- lingual support.

Even though a large amount of work has been done focusing on varied challenges for sentiment analysis but there are still some areas with scope of improvement and a few areas which still need attention. For example, despite of adding maximum words in lexicons but still there can be words which are still naive for classifier keeping the problem of "word knowledge" at its place. Then, regardless of the fact that languages other than English have been pointed by researchers but still there are still numerous languages which still need to be considered. Furthermore, determining the scope of sarcasm and intensification of polarity with use of negation operators are still a big challenge for researchers. And with fast exposure of web to people, increase of complexity in sentences is leading to many more challenges which ultimately make the job of sentiment analysis even more troublesome. Therefore, more future re- search could be dedicated to these challenges.

V. CONCLUSION

Sentiment analysis for various applications in today's world is a job of priority for coping up with high competitive environment. Many areas can be benefited by optimum analysis of sentiments. For extracting useful information from opinionated data set, use of data mining approaches is proven to be an advantageous practice. But, the actual benefit of opinion mining can only be attained by getting accurate and precise results which becomes even more difficult with emerging high volume data. Such large volume of data and subtle semantics of opinionated text exposes opinion mining to several challenges.

Various challenges and their presumable solutions given by different researchers have been highlighted in the article. It can be concluded from the study that, though analysts have worked on many problems successfully yet a large number of areas still need attention for better results. Like language dependency, determining scope of negation, intensification of polarity with negation operators, word knowledge are some of the issues which researchers can take into consideration in future. Also, hindrance of performance due to presence of complexity in sentence is a big challenge for sentiment analysis which need constant attention.

REFERENCES

1. Abualigah, L. M., Khader, A. T., and Hanandeh, E. S. (2018). A new feature selection method to improve the document clustering using particle swarm optimization algorithm. *Journal of Computational Science*, 25:456– 466.
2. Ahmad, S. R., Bakar, A. A., and Yaakub, M. R. (2015). Metaheuristic algorithms for feature selection in senti- ment analysis. In *Science and Information Conference (SAI), 2015*, pages 222–226. IEEE.
3. Astya, P. et al. (2017). Sentiment analysis: Approaches and open issues. In *Computing, Communication and Automation (ICCCA), 2017 International Conference on*, pages 154–158. IEEE.
4. Benamara, F., Cesarano, C., Picariello, A., Recupero,D. R., and Subrahmanian, V. S. (2007). Sentiment anal- ysis: Adjectives and adverbs are better than adjectives alone. In *ICWSM*. Citeseer.
5. Cambria, E., Olsher, D., and Rajagopal, D. (2014). Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *Twenty- eighth AAAI conference on artificial intelligence*.
6. Councill, I. G., McDonald, R., and Velikovich, L. (2010). What's great and what's not: learning to clas- sify the scope of negation for improved sentiment anal- ysis. In *Proceedings of the workshop on negation and speculation in natural language processing*, pages 51–59. Association for Computational Linguistics.
7. Di Caro, L. and Grella, M. (2013). Sentiment analysis via dependency parsing. *Computer Standards & Inter- faces*, 35(5):442–453.
8. Gidhe, P. and Ragha, L. (2017). Sarcasm detection of non# tagged statements using mlp-bp. In *Advances in Computing, Communication and Control (ICAC3), 2017 International Conference on*, pages 1–4. IEEE.
9. Hatzivassiloglou, V. and Wiebe, J. M. (2000). Effects of adjective orientation and gradability on sentence sub- jectivity. In *Proceedings of the 18th conference on Com- putational linguistics-Volume 1*, pages 299–305. Associ- ation for Computational Linguistics.
10. Hogenboom, A., Heerschop, B., Frasincar, F., Kay- mak, U., and de Jong, F. (2014). Multi-lingual support for lexicon-based sentiment analysis guided by seman- tics. *Decision support systems*, 62:43–53.
11. Jain, A. P. and Katkar, V. D. (2015). Sentiments analy- sis of twitter data using data mining. In *Information Pro- cessing (ICIP), 2015 International Conference on*, pages 807–810. IEEE.
12. Jin, X., Xu, A., Bie, R., and Guo, P. (2006). Machine learning techniques and chi-square feature selection for cancer classification using sage gene expression profiles. In *International Workshop on Data Mining for Biomedi- cal Applications*, pages 106–115. Springer.
13. Joseph, S., Mugauri, C., and Sumathy, S. (2017). Sen- timent analysis of feature ranking methods for classifica- tion accuracy. In *IOP Conference Series: Materials Sci- ence and Engineering*, volume 263, page 042011. IOP Publishing.
14. Kanayama, H. and Nasukawa, T. (2006). Fully auto- matic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 conference on em- pirical methods in natural language processing*, pages 355–363. Association for Computational Linguistics.
15. Kasthuriarachchy, B. H., de Zoysa, K., and Pre- marathne, H. (2015). Context-aware sentiment clas- sification. In *Advances in ICT for Emerging Regions (ICTer), 2015 Fifteenth International Conference on*, pages 276–276. IEEE.
16. Kaur, A. and Gupta, V. (2014). Proposed algorithm of sentiment analysis for punjabi text. *Journal of Emerging Technologies in Web Intelligence*, 6(2):180–183.
17. Lee, C. and Lee, G. G. (2006). Information gain and divergence-based feature selection for machine learning- based text categorization. *Information processing & management*, 42(1):155–165.
18. Lin, K.-C., Zhang, K.-Y., Huang, Y.-H., Hung, J. C., and Yen, N. (2016). Feature selection based on an improved cat swarm optimization algorithm for big data classification. *The Journal of Supercomputing*, 72(8):3210–3221.
19. Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:627–666.
20. Liu, B., Blasch, E., Chen, Y., Shen, D., and Chen, G. (2013). Scalable sentiment classification for big data analysis using naive bayes classifier. In *Big Data*,

2013 IEEE International Conference on, pages 99–104. IEEE.
21. Martínez-Cámara, E., Martín-Valdivia, M. T., Urena- López, L. A., and Montejo-Ráez, A. R. (2014). Senti- ment analysis in twitter. *Natural Language Engineering*, 20(1):1–28.
22. Maynard, D. and Greenwood, M. A. (2014). Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *LREC 2014 Proceed- ings*. ELRA.
23. Mishra, P., Rajnish, R., and Kumar, P. (2016). Senti- ment analysis of twitter data: Case study on digital india. In *Information Technology (InCITe)-The Next Genera- tion IT Summit on the Theme-Internet of Things: Con- nect your Worlds, International Conference on*, pages 148–153. IEEE.
24. Mullen, T. and Collier, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the 2004 conference on em- pirical methods in natural language processing*.
25. Nasukawa, T. and Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language process- ing. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77. ACM.
26. Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326.
27. Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.
28. Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.
29. Povoda, L., Burget, R., Dutta, M. K., and Sengar, N. (2017). Genetic optimization of big data sentiment analysis. In *Signal Processing and Integrated Networks (SPIN), 2017 4th International Conference on*, pages 141–144. IEEE.
30. Ramanujam, R. S., Nancyamala, R., Nivedha, J., and Kokila, J. (2015). Sentiment analysis using big data. In *Computation of Power, Energy Information and Com- muincation (ICCPEIC), 2015 International Conference on*, pages 0480–0484. IEEE.
31. Smailović, J., Grčar, M., Lavrač, N., and Ž nidaršič, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information sciences*, 285:181–203.
32. Sohangir, S., Wang, D., Pomeranets, A., and Khosh- goftaar, T. M. (2018). Big data: deep learning for finan- cial sentiment analysis. *Journal of Big Data*, 5(1):3.
33. Tabakhi, S., Moradi, P., and Akhlaghian, F. (2014). An unsupervised feature selection algorithm based on ant colony optimization. *Engineering Applications of Artificial Intelligence*, 32:112–123.
34. Tan, S. and Zhang, J. (2008). An empirical study of sentiment analysis for chinese documents. *Expert Sys- tems with applications*, 34(4):2622–2629.
35. Uğuz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algo- rithm. *Knowledge-Based Systems*, 24(7):1024–1032.
36. Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Rec- ognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human lan- guage technology and empirical methods in natural lan- guage processing*, pages 347–354. Association for Com- putational Linguistics.
37. Xu, Y. and Chen, L. (2010). Term-frequency based feature selection methods for text categorization. In *Genetic and Evolutionary Computing (ICGEC), 2010 Fourth International Conference on*, pages 280–283. IEEE.
38. Zheng, L., Wang, H., and Gao, S. (2018). Sentimental feature selection for sentiment analysis of chinese online reviews. *International journal of machine learning and cybernetics*, 9(1):75–84.

AUTHORS PROFILE

Arpita is pursuing Ph.D. in Computer Science & Applications from Department of Computer Science & Applications, Kurukshetra University, Kurukshetra. She completed her M.Tech in Computer Science & Engineering from Deenbandhu Chotu Ram University Science and Technology, Murthal, Sonepat. Her research area is Sentiment Analysis.

Dr. Pardeep Kumar is an Associate Professor at Department of Computer Science & Applications, Kurukshetra University, Kurukshetra. He holds an experience of 22 years. He pursued his Ph.D. from DCSA Department, Kurukshetra University, Kurukshetra. His area of research includes Optimization, Cloud Computing, Privacy Preservation, Soft Computing, Machine Learning etc. He has published more than 60 papers in National and International journals.

Dr. Kanwal Garg is an Assistant Professor at Department of Computer Science & Applications, Kurukshetra University, Kurukshetra. He holds more than 18 years of experience. He received his Ph.D. from GJU S&T, Hisar. His area of research includes Big Data, Web Mining, Data Stream, OLAP Cubes, Data Mining and Warehousing. He has published more than 80 papers in National and International journals.