Review on Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis

Yasir Ali Solangi[1], Zulfiqar Ali Solangi[2], Samreen Aarain[1], Amna Abro[1], Ghulam Ali Mallah[1], Asadullah Shah[3]
yasir_solangi@yahoo.com, zulfs@hotmail.com, arainsam08@gmail.com, abroamna5@gmail.com
Ghulam.ali@salu.edu.pk

[1]Shah Abdul Latif University, Khairpur, Sindh, Pakistan
[2]Colleges and Institutes Sector Royal Commission for Jubail Saudi Arabia
[3]Kulliyyah of Information and Communication Technology, International Islamic University, Malaysia

Abstract: **As the majority of online networking on the Internet, opinion mining has turned into a fundamental way to deal with investigating such huge numbers of information. Different applications show up in an extensive variety of modern areas. In the interim, opinions have various pronunciations which bring along investigate challenges. The research challenges make opinion mining a dynamic research region recently. In this paper, Natural Language Processing (NLP) techniques for opinion mining and sentiment analysis are reviewed. Initially NLP is reviewed then briefed about its common and useful preprocessing steps also. In this paper opinion mining for various levels are analyzed and reviewed. At the end issues are identified and some recommendation are suggested for opinion mining and-sentiment-analysis.**

Key words: ***natural language processing, NLP, Opinion mining, sentiment analysis, toolkits***

I.ⁿ INTRODUCTION

With the tense development of client created messages on the Internet, to reduce valuable data consequently from submitted documents get interests from analysts in different fields, specifically groups of Natural Language Processing (NLP).

Moreover, so many applications of opinion mining, like “product pricing”, “competitive intelligence”, “market prediction”, “election forecasting”, “nation relationship analysis, risk detection in banking systems” manage high focusing from industrial, manufacturing and trade areas[1] furthermore social media is growing, online review sites, like Amazon, Twitter and other like them providing the huge corpora which are vital assets for educational research. It has both Interests to the both scholarly community and industry to improvement of opinion mining. There' sentiments can be critical when it's a great opportunity to settle on a choice or pick among numerous choices. At the point when those decisions include important assets (for instance, investing cash to purchase items or services) individuals frequently depend on their companions' past happenstances. As of not long ago, the fundamental wellsprings of data were companions and concentrated magazine or sites[2]. Presently, the Social Networking Sites (SNS) gives new apparatuses to productively make and offer opinions with everybody associated with social networking sites via Web 2.0 enable individuals to share valuable data and opinions about the products or services purchased online[3]. However, the shared data and opinions are unstructured may include emotions, sentiments, characteristics, numbers, dates, and facts. These days, it is the center of attention for majority of researchers and scientists to gather and capture popular sentiments about the socials, political growths, business procedures, and the products liking or disliking from the business world. Hence, as an outcome of it, they can conceivably promote and forecast the financial share of the market for the business. The subsequent developing fields are opinion mining and sentiments analysis[2]. In modern e-businesses, the merchants invest on opinion mining and sentiment analysis research then utilize the findings to enhance client relationship administration and suggestion frameworks through positive and negative clients’ input. As a result, the findings may also help distinguish and prohibit "flares" (excessively warmed or hostile dialect) in social correspondence and upgrade antispam frameworks within the organization[4]. More, the businesses can evaluate and foresee open states of the public thoughts towards their products and services, and reputation [5].

II.ⁿ METHODOLOGY

The different steps involved in research methodology are described as:

978-1-5386-7966-1/18/$31.00 ©2018 IEEE

- Different NLP techniques were reviewed
- Various published material was collected regarding NLP Techniques, opinion mining and sentiment analysis.
- Fundamental information regarding NLP Techniques, opinion mining and sentiment analysis were reviewed.

III. NLP TECHNIQUES FOR TEXT PREPROCESSING

There are some preprocessing stages required for structuring the text, extracting features, segmentation, tokenization, Parts of Speech (POS) tagging and parsing in opinion mining.

Tokenization is an essential strategy for most NLP tasks. It parts a sentence or archive into tokens which are words or expressions. For English, it is minor **to** part words by the spaces, but some extra information should be need to consideration, such as "opinion phrases", named elements.

1. In tokenization, some simple words, like "the", "a", could not useful or help full that's why these words will be removed because of little help, give minimal helpful data. As a major method, numerous tokenization devices are accessible, for example, many tokenization tools are existing as a fundamental technique like "Stanford Tokenizer", "Open NLP Tokenizer"

2. For Chinese, Japanese or different dialects which don't have express the limit markers of the word, tokenization isn't as minor as English and segment the word is required. The segmentation of the word is a consecutive marking issue.

3. Conditional Random Fields (CRFs)[6] is connected to this issue & beat concealed "Markov models" & most extreme entropy "Markov models". Now days, "word embedding" and "deep learning" based methodologies have been connected for the segmentation off Chinese word. Many tools are available, for example, "ICTCLAS, THULAC and Stanford Segmenter"

4. POS labeling and parsing are strategies that investigate the lexical and syntactic data. POS labeling is utilized to decide the comparing POS tag for each word. Like word division, it is additionally a successive labeling issue. The POS labels, for example, descriptive word, thing, are very useful in light of the fact that opinion words are normally modifiers and opinion targets (i.e., entities and aspects) are things or mix of things. While POS labeling gives lexical data, parsing acquires syntactic data. Parsing produces a tree which speaks to the linguistic structure of a given sentence with the comparing relationship of various constituents. Contrasting with POS labeling, parsing gives wealthier structure data. As the closeness and significance among word division, POS labeling, and parsing, some methodologies are proposed to manage these errands all the while.

IV. AVAILABLE TOOLS FOR NLP

There are several tools have been introduced which deal with Chinese word segmentation and tokenization, some are joined in powerful toolkits which are as under:

1. Fudan NLP tool in JAVA Language [3] is the open toolkit source for Chinese NLP, these techniques supports dependency parsing, coreference resolution, word segmentation, named entity recognition, POS labeling and so on[7].

   "https://code.google.com/archive/p/fudannlp"

2. Language Technology Platform (LTP) tool in C++ NLP system is open source for lexical analysis ("word segmentation, POS labeling, named entity recognition"), syntactic parsing and semantic parsing ("word sense disambiguation, semantic role labeling") modules[8].

   "http://www.ltp-cloud.com/intro/en/"

3. Niu Parser tool in C++ language is a Semantic Analysis and Chinese Syntactic Toolkit, which supports the segmentation of the word, dependency parsing, named entity recognition, constituent parsing, semantic role labeling and POS labeling[9].

   "http://www.niuparser.com/index.en.html"

4. Gensim Python offers open source libraries for supporting large-scale corpora via distributed parallel versions of online Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), Random Projection, Hierarchical Dirichlet Process and word2vec[10].

   "http://radimrehurek.com/gensim/"

5. The Stanford CoreNLP is a common platform for NLP all types of elementary jobs i.e. parsing,

coreference resolutions, POS tagging, named entity recognition, and advanced sentiment analysis [55]. This toolkit is supportive in analysis of Arabic, French, Spanish, Chinese, and German languages than NLTK and Open NLP platforms via adopting different modules CRFs, deep learning, and maximum entropy[11].

"http://stanfordnlp.github.io/CoreNLP"

V. OPINION MINING AND SENTIMENT ANALYSIS

Opinion mining and sentiment analysis have target to extract the sentiment orientation of writings. Generally, opinion mining can be separated into 03 levels: document level and sentence level and fine-grained level.

LEVEL-I

To decide over all extremity of the text in document which fuses different sentences. This endeavor makes the assumption that select a lone sentiment target is discussed in one text of document, similar to text level opinion mining. The task of the sentence level is in like manner a course of action issue anyway focuses on each sentence in the documents. It chooses if a sentence imparts positive, negative or fair presentation. Subjectivity arrange is another endeavor at the sentence level, which isolates emotional and target parts of document.

LEVEL-II

The above assignments give careful consideration to extricating the detail data of opinions, for example, opinion target and opinion holder. For instance, "The screen of this cell phone is great." communicates a positive edge to viewpoint "screen" of substance "cell phone". Same the outcome may get negative or neutral.

LEVEL-III

Fine-grained opinion mining manages this issue past order methods. From the document level to the fine-grained level, the complexities and issues are expanding. The absence of clarified corpora at the finer level exacerbates it. Generally, regulated techniques beat unsupervised strategies. Notwithstanding, necessities of clarified corpora are not generally fulfilled, particularly for fine-grained level opinion mining which pushes specialists for create semi-regulated or unsupervised techniques.

VI. ISSUES AND RECOMMENDATIONS

Although world has rapidly growth so it could be good to say that social media is performing basic role to grow up the industries, trades, companies and other sectors, online marketing or business which are growing rapidly For example anyone who by the cell phone online he must reviews the comments and feedback of the other byers also, same way give the feedback also to the company mostly. This feedback and online purchasing is growing up rapidly. Nevertheless, when we are focusing in the backside of this whole method like to give feedback to submit review, comments etc. it appears huge of questions. But in this paper I reviewed the NLP techniques and text preprocessing techniques for opinion mining and sentiment analysis, during review it has analyzed that when anyone who submit his feedback online or comment his feedback online so not sure that every this is submitting accurately, reason is this that in the backside of the NLP programing it is fix that jus noising the text remove the non-meaningful words like "a", "the" and just jump on the meaningful worked. But this is not good for the users because mostly user not known properly writing and submitting their feedback or asking different issues regarding brand or other things. So how his issue could be solved, how he will get the proper answer of his question. That's why I suggesting that modify the NLP techniques as well as all other concerned techniques through that communication gape should be removed and each will get their desire information accurately and easily.

VII. CONCLUSION

In this paper, Natural Language Processing (NLP) techniques and most useful toolkits of NLP that deal with tokenization and Chinese word segmentation were reviewed. Furthermore, opinion mining and sentiment analysis also reviewed. During the review it has been analyzed that huge of NLP techniques are available for opinion mining and sentiment analysis. The fundamental point of opinion mining and sentiment analysis is extracting presence of sentiments from the given writings. To process the given undertaking feeling mining could be separated into three dimensions: document level, sentence level, and fine-grained level. In over three dimensions have own procedures and function respectively. For example, level one. It chooses if a sentence imparts positive, negative or fair presentation. Subjectivity arrangement is another endeavor at the sentence level, which isolates abstract and target parts of given text in report or document. Level 2 which perform the task to evaluate the opinions either opinion may be the positive, negative or neutral. After completion of all levels company or user could be able to get the feedback on screen but not sure either screening result is as per user

requirement, recommendation, suggestion or interrogation. Because in the back hand for text reprocessing some techniques are designed which consisting some checks for tracking or sorting the data, because of some limited checks screening result could not be properly or feedback, comments or review also could not be properly screened or submitted. That's why there is some need to little modification in the backhand NLP techniques or other relevant techniques as user or beneficiary can give and receive proper data of information.

VIII. REFERENCE

[1] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Inf. Fusion*, vol. 36, pp. 10–25, 2017.

[2] C. Sea, "A u s t r a l i a," no. April, pp. 15–21, 2013.

[3] Y. A. Solangi, Z. Solangi, A. Raza, S. A. Aziz, and M. Syarqawy, "Social Commerce in e-business of Pakistan : Opportunities , Challenges and Solutions," in *International Conference for Information and Communication Technology for the Muslim World ICT4M 2018*, 2018.

[4] Y. Kai, Y. Cai, H. Dongping, J. Li, Z. Zhou, and X. Lei, "An effective hybrid model for opinion mining and sentiment analysis," *2017 IEEE Int. Conf. Big Data Smart Comput. BigComp 2017*, pp. 465–466, 2017.

[5] P. Gupta, A. Goswami, S. Koul, and K. Sartape, "IQS-intelligent querying system using natural language processing," *Proc. Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2017*, vol. 2017–Janua, pp. 410–413, 2017.

[6] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[7] Ahmed, S. F., Desa, H., Azim, F., Surti, A., & Hussain, W. (2013, April). Remote access of SCADA with online video streaming. In Computer Science & Education (ICCSE), 2013 8th International Conference on (pp. 270-274). IEEE

[8] W. Che, Z. Li, and T. Liu, "Ltp: A chinese language technology platform," in *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, 2010, pp. 13–16.

[9] J. Zhu, M. Zhu, Q. Wang, and T. Xiao, "Niuparser: A Chinese syntactic and semantic parsing toolkit," *Proc. ACL-IJCNLP 2015 Syst. Demonstr.*, pp. 145–150, 2015.

[10] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," 2010.

[11] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.