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

Muhammed Yaseen Morshed Adib
Computer Science and Engineering,
BRAC University
Dhaka, Bangladesh
yaseen.morshed.adib@g.bracu.ac.bd

Sovon Chakraborty
Computer Science and Engineering,
BRAC University
Dhaka, Bangladesh
sovon.chakraborty@g.bracu.ac.bd

Munim Bin Muquith
Computer Science and Engineering,
BRAC University
Dhaka, Bangladesh
munim.bin.muquith@g.bracu.ac.bd

Md Humaion Kabir Mehedi
Computer Science and Engineering,
BRAC University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

Annajiat Alim Rasel
Senior Lecturer, Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
annajiat@bracu.ac.bd

ABSTRACT

Opinion mining, which makes up the majority of online networking on the Internet, has evolved into a crucial strategy for handling such massive amounts of data analysis. Different uses can be seen in a wide range of contemporary settings. However, views have a variety of pronunciations, which makes research difficult. Opinion mining has lately become a vibrant study area due to research obstacles. This study reviews sentiment analysis and opinion-mining Natural Language Processing (NLP) approaches. NLP is first reviewed, and then its typical and practical preprocessing methods are explained. This study reviews and analyzes opinion mining at various levels. The conclusion, difficulties are noted and some suggestions for sentiment analysis and opinion mining are made.

KEYWORDS

NLP, sentiment analysis, natural language processing, Opinion mining

ACM Reference Format:
Muhammed Yaseen Morshed Adib, Sovon Chakraborty, Munim Bin Muquith, Md Humaion Kabir Mehedi, and Annajiat Alim Rasel. 2018. Review on Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Minimizing important data as a result of supplied papers getting interested from analysts in many sectors, particularly groups of Natural Language Processing, is becoming more challenging with the rapid development of client-created communications on the Internet (NLP). Additionally, a lot of opinion mining applications, such as "product pricing," "competition intelligence," "market prediction," "election forecasting," "nation relationship analysis," and "risk detection in banking systems," handle high emphasis from the industrial, manufacturing, and trade areas[1]. In addition, social media is expanding, with online review sites like Amazon, Twitter, and others like them offering vast corpora that are essential resources for academic study. Both the academic community and businesses are interested in the advancement of opinion mining. When it's time to make a decision or choose from a variety of options, feelings may be quite important.

Moreover, People usually rely on their friends' prior experiences when making decisions that involve significant resources (such as investing money to buy goods or services). Prior to recently, friends and specialized magazines or websites were the primary sources of information[2]. Currently, social networking sites (SNS) offer new tools to effectively create and share thoughts with everyone connected to them through Web 2.0, allowing people to exchange insightful information and opinions on the goods and services they buy online[3]. However, the exchanged information and ideas are unstructured and may contain thoughts, feelings, traits, statistics, dates, and facts. The bulk of academics and scientists are now focused on gathering and capturing public opinion regarding social, political, and corporate developments as well as the things people like or dislike from the business sector. As a result, they may be able to advertise and anticipate the financial market share of the company. Opinion mining and sentiment analysis are the next-to-emerging fields[2]. More companies in today's e-businesses are able to assess and anticipate public perceptions of their goods, services, and reputation [4]. In contemporary e-businesses, the merchants invest in sentiment analysis and opinion mining research, using the results to improve client relationship management and suggestion frameworks via both positive and negative customer feedback. The results may therefore aid in differentiating and outlawing "flares"

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

(excessively heated or hostile dialect) in social communication and improve antispam frameworks within the organization[5].

2 METHODOLOGY

The several steps in the research technique are listed as follows:

- Various NLP methods were reviewed.
- The fundamentals of sentiment analysis, opinion mining, and NLP techniques were examined.
- To learn more about NLP techniques, opinion mining, and sentiment analysis, many published materials were gathered.

3 NLP TECHNIQUES FOR TEXT PREPROCESSING

In order to structure the text, extract features, segment, tokenize, tag parts of speech (POS), and parse it, there are certain preprocessing steps needed.

The majority of NLP jobs need the use of the tokenization approach. It divides a phrase or archive into words or expressions called tokens. The use of spaces to separate words in English is small, but some additional information, such as "opinion phrases" and named elements, should be taken into account.

1. Simple words like "the" and "a" could not be useful or helpfully deleted during tokenization since they provide only minimally valuable information. Numerous tokenization programs are available as a primary methodology, for instance, "Stanford Tokenizer" and "Open NLP Tokenizer," which are both common tokenization tools.

2. Tokenization is not as minor in Chinese, Japanese, or other languages that lack word limit signals as it is in English, therefore word segmentation is necessary. The word segmentation is a problem with sequential marking.

3. In relation to this, Conditional Random Fields (CRFs)[6] outperformed both the most severe entropy and disguised "Markov models." For the segmentation of Chinese words, "word embedding" and "deep learning" based approaches have recently been linked. There are several tools, such as "ICTCLAS, THULAC, and Stanford Segmenter."

4. The lexical and syntactic data are examined using POS labeling and parsing techniques. To determine the appropriate POS tag for each word, POS labeling is used. It is also a successive labeling issue, like word division. In view of the fact that opinion words are frequently modifiers and opinion targets (i.e., entities and aspects) are things or mixtures of things, the POS labels, for example, descriptive word, and object, are particularly helpful. While parsing obtains syntactic data, POS labeling provides lexical data. With the help of examining the relationships between distinct parts, parsing creates a tree that testifies to the linguistic structure of a particular phrase. When compared to POS labeling, parsing provides more detailed structural information. Word division, POS tagging, and parsing are closely related processes, hence certain approaches are suggested to handle these tasks concurrently.

4 AVAILABLE TOOLS FOR NLP

Bengali word segmentation and tokenization are dealt with by a number of techniques, some of which are combined into potent toolkits, as follows:

- Using distributed parallel versions of online Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), Random Projection, Hierarchical Dirichlet Process, and word2vec, Gensim Python provides open-source tools to serve large-scale corpora[7].
- For lexical analysis ("word segmentation, POS labeling, named entity identification"), syntactic parsing, and semantic parsing ("word sense disambiguation," "semantic role labeling") modules, the Language Technology Platform (LTP) tool in the C++ NLP system is open source[8].
- Niu Parser is a semantic analysis tool written in the C++ programming language that provides word segmentation, dependency parsing, named entity identification, constituent parsing, semantic role labeling, and POS labeling[9].
- A typical framework for NLP tasks including parsing, coreference resolutions, POS tagging, named entity identification, and sophisticated sentiment analysis is the Stanford CoreNLP [5]. By utilizing several modules CRFs, deep learning, and maximum entropy, this toolkit is more helpful than NLTK and Open NLP platforms in the analysis of Arabic, French, Spanish, Chinese, and German languages[10].

.

5 OPINION MINING AND SENTIMENT ANALYSIS

Opinion mining and sentiment analysis aim to identify the texts' emotional tone. In general, opinion mining may be divided into three levels: fine-grained level, phrase level, and document level.

LEVEL-I

To judge the text's whole extreme in a document that combines many sentences. Similar to text-level opinion mining, this project assumes that a single emotional goal is mentioned in a single text of the document. The work of the sentence level, which concentrates on each sentence in the papers, is similarly a course-of-action issue. It determines whether a statement presents itself in a favorable, negative, or fair manner. Another effort at the phrase level is subjectivity organized, which separates the document's emotional and intended themes.

LEVEL-II

In the aforementioned assignments, the extrication of opinion detail data, such as opinion target and opinion holder, has been carefully considered. For instance, the statement "This mobile phone's screen is fantastic" conveys a favorable opinion about the subject matter "cell phones" The result might be either bad or neutral.

LEVEL-III

Beyond order techniques, fine-grained opinion mining handles

this problem. The complications and difficulties are growing from the document level to the finer level. It is made worse by the lack of finer-level defined corpora. In general, controlled methods outperform unsupervised ones. However, the needs of clarified corpora are not always met, particularly for fine-grained level opinion mining, which forces experts to develop unsupervised or semi-supervised algorithms.

6 ISSUES AND RECOMMENDATIONS

Despite the world's rapid development, it may be accurate to argue that social media plays a fundamental part in children's development. The quickly expanding businesses, trades, industries, and other areas, including internet marketing Anyone who purchases a mobile phone online, for instance, is required to read the evaluations and comments of other purchasers. In a similar vein, customers are encouraged to provide feedback to the corporation. This feedback and internet shopping is expanding quickly. However, while we are concentrating on the method's downside involves asking for feedback in order to submit reviews, comments, etc. However, in this paper, I reviewed NLP techniques as well as text preprocessing techniques for opinion mining and sentiment analysis. During the review, it has been analyzed that when anyone submits his feedback online or comments on his feedback online so not sure that this is submitting accurately, the reason being that in the backside of the NLP programming it is fixed that just cleaning the text removes the non-meaningful words like "a," "the," and just jumps on the meaningful words. However, this is not beneficial for the users because the majority of them do not know how to correctly write and submit their comments or ask various questions about brands or other things. How may his problem be resolved, and how will he obtain the right response to his query. For this reason, we advocate for the modification of NLP approaches as well as any other relevant techniques so that the communication gap may be closed and everyone can quickly and accurately obtain the information they want.

Furthermore, when making decisions that require major financial investments, people frequently draw on the prior experiences of their friends (such as investing money to buy goods or services). Prior to recently, the main information sources were friends and specialized journals or websites. Social networking sites (SNS) now provide new tools to efficiently develop and share thoughts with everyone connected to them through Web 2.0, enabling individuals to communicate intelligent knowledge and opinions about the items and services they purchase online. Thoughts, sentiments, qualities, statistics, dates, and facts could all be included in the unstructured information and ideas that are transmitted. Currently, the majority of academics and scientists are concentrating on gathering and capturing public opinion regarding social, political, and business trends as well as the things that people appreciate.

POS labeling and parsing techniques are used to investigate the lexical and syntactic data. POS labeling is used to establish the proper POS tag for each word. Similar to word division, it is likewise a successive labeling problem. The POS labels, like "descriptive word" and "object," are especially useful because opinion words are typically modifiers and opinion targets (i.e., entities and aspects) are objects or combinations of things. POS labeling offers lexical data, whereas parsing yields syntactic information. Parsing produces a tree that illustrates the grammatical organization of a given phrase by looking at the connections between different elements. Parsing offers more thorough structural information than POS labeling. Parsing and word division are closely related processes.

7 CONCLUSION

This article reviews tokenization and Chinese word segmentation Natural Language Processing (NLP) techniques and the most practical NLP toolkits. Additionally, sentiment analysis and opinion mining were studied. Numerous NLP approaches are accessible for sentiment analysis and opinion mining, which was determined throughout the evaluation. Finding the presence of feelings in the provided texts is the core goal of opinion mining and sentiment analysis. Feeling mining might be divided into three levels to handle the supplied task: document level, phrase level, and fine-grained level. Over three dimensions each have its own protocols and functions. Consider level one. It determines whether a statement presents itself in a favorable, negative, or fair manner. Another attempt at sentence-level subjectivity arranging separates the abstract and intended sections of a text in a report or document. Level 2 is responsible for evaluating views, whether they be good, negative, or neutral. After completing all steps, a corporation or user may be able to see feedback on the screen, but it is unclear whether the screening results are in line with the user's requirements, suggestions, or questions. The screening results could not be correctly screened, and feedback, comments, or reviews could not be properly screened or sent since in the backhand for text reprocessing certain procedures are created that consist of some checks for tracking or sorting the data. The backhand NLP techniques or other pertinent techniques must be somewhat modified as a result so that the user or beneficiary may provide and receive accurate information.

REFERENCES

[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] 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.

[5] 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.

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

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

[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. 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.

[11] Jha, R., Jbara, A. A., Qazvinian, V., Radev, D. R. (2017). NLP-driven citation analysis for scientometrics. Natural Language Engineering, 23(1), 93-130.

[12] RBlodgett, S. L., Barocas, S., Daumé III, H., Wallach, H. (2020). Language (technology) is power: A critical survey of" bias" in nlp. arXiv preprint arXiv:2005.14050.

[13] Alnajjar, K., Hämäläinen, M., Rueter, J., Partanen, N. (2020). Ve'rdd. Narrowing the Gap between Paper Dictionaries, Low-Resource NLP and Community Involvement. arXiv preprint arXiv:2012.02578.
[14] Field, A., Blodgett, S. L., Waseem, Z., Tsvetkov, Y. (2021). A survey of race, racism, and anti-racism in NLP. arXiv preprint arXiv:2106.11410.
[15] Briscoe, T., Copestake, A., Boguraev, B. (1990). Enjoy the Paper: Lexicology. In COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics.
[16] Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R. (2019, June). FLAIR: An easy-to-use framework for state-of-the-art NLP. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations) (pp. 54-59).
[17] Santaholma, M. E. (2007). Grammar sharing techniques for rule-based multilingual NLP systems. In Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA).
[18] Kang, D., Ammar, W., Dalvi, B., Van Zuylen, M., Kohlmeier, S., Hovy, E., Schwartz, R. (2018). A dataset of peer reviews (peerread): Collection, insights and nlp applications. arXiv preprint arXiv:1804.09635.
[19] 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.