Analysis of Research Paper Titles Containing Covid-19 keyword using Various Visualization Techniques

[*]Dr. Mangesh Bedekar, Dr. Sharmishta Desai
School of Computer Enginering & Technology,
Dr. Vishwanath Karad MIT World Peace University,
Kothrud, Pune, Maharashtra, India 411038
[*]mangesh.bedekar@mitwpu.edu.in,
sharmishta.desai@mitwpu.edu.in

Abstract. It's been around two years from the outbreak of the coronavirus, thus labeled as Covid- 19, and there has been an explosion of literature being published by research scholars related to work done on Covid-19. Covid- 19 as a keyword has been mentioned in the titles of most of these papers. It was thought to analyse the number of papers and the titles of papers which include Covid- 19 in the title of the research papers. The various combinations of other words like, prefixes, suffixes, N-gram combinations with the keyword Covid- 19 in the titles of these papers were also analysed. The research publication repositories analysed were: IEEE Explore, ACM Digital Library, Semantic Scholar, Google Scholar, Cornel University etc. The domains of research publication title analysis were restricted to computer science / computer engineering related papers. As the term labeling the corona virus outbreak as Covid-19 was labeled in 2020, the timeline of title analysis was restricted from 2019 till December 2021.

The term Covid-19 is also one of the most searched terms in most of these research repositories as is evident from the search suggestions offered by them. Considering the usefulness of Bag of Words and N Gram algorithm in analytics and data visualization, a methodology is proposed and implemented based on bag of words algorithm to do prefix and suffix words analysis. This methodology is working correctly to state different prefix and suffix words used by various researchers to demonstrate significance of their titles.

Methodology based on N Gram analysis is found effective to find topic on which most of the researchers have done work. Word Clouds are generated to demonstrate different buzz words used by researchers in their respective paper titles. These are useful for providing visualization of the data if it is in big size.

Keywords: Covid-19, Bag of Words, N Gram, Research Titles, Word Cloud, Tag Cloud

1 Introduction

Covid-19 has dramatically affected each and everyone's life worldwide. The economic and social disruption caused by the pandemic is shocking. Tens of millions of people are at risk of falling into extreme poverty. Millions of enterprises are facing existential threat. Almost every sector is affected by the Covid Pandemic. Similarly, work done by researchers in the domain of Covid-19 and related domain has increased tremendously. It was thought to analyse the number of papers and the titles of papers which include Covid-19 in the title of the research papers. The various combinations of other words like, prefixes, suffixes, N-gram combinations with the keyword Covid-19 in the titles of these papers were also analysed. As the term labelling the corona virus outbreak as Covid-19 was labelled in 2020, the timeline of title analysis was restricted from 2019 till date.

This paper is intended to analyze the research done by different people in this corona year under computer domain. To analyze the research title effectively, a combination of Bag of words and N-Gram algorithm. Bag of words algorithm is used to extract keywords from text which can be used for doing analysis further. N Gram algorithm is used to find frequency of sequence of words in a text.

Titles containing "Covid" or "Covid-19" or "Corona" as a keyword were collected from different major digital libraries like ACM, IEEE, Semantic Scholar, Google Scholar etc. Analysis is done to identify keywords that are mostly preferred as suffix and prefix with keyword "Covid or corona". N-Gram analysis on titles is done to analyze on which topic most of the researchers have done work. Word Clouds are generated to demonstrate different buzz words used by researchers in their respective paper titles to highlight significance of their research.

1.1 Statistics of Papers from each Research Repository

Table 1 – Statistics of number of Papers from each Repository

Publisher	Number of Papers		
	2018-19	**2019-2020**	**2020-21**
IEEE Computer Society of India	125	585	420
ACM	58	597	300
Cornel University	0	1	1190
Semantic Scholar	36	8930	9560
Google Scholar	1880	27600	81700
Mendeley	517	185873	26553

It can be observed in the Table-1, research related to Corona started majorly from the end of year 2019 when the naming nomenclature for Corona disease was finalized, “Covid-19”. It can be observed that there are more papers on Google or Mendeley database and on Semantic Scholar, in decreasing order. Google Scholar has the highest number of papers as it contains papers from all domains (including medical and technical). Year 2019-20 is the year when a lot of research on Corona was published. Data collected till December-2021.

1.2 Comparative Analysis of Number of papers across these repositories

Comparative analysis of year wise papers published under different publishers is shown in Fig.1. It shows that papers were reflected in the databases of the repositories considered from the year 2018-19.

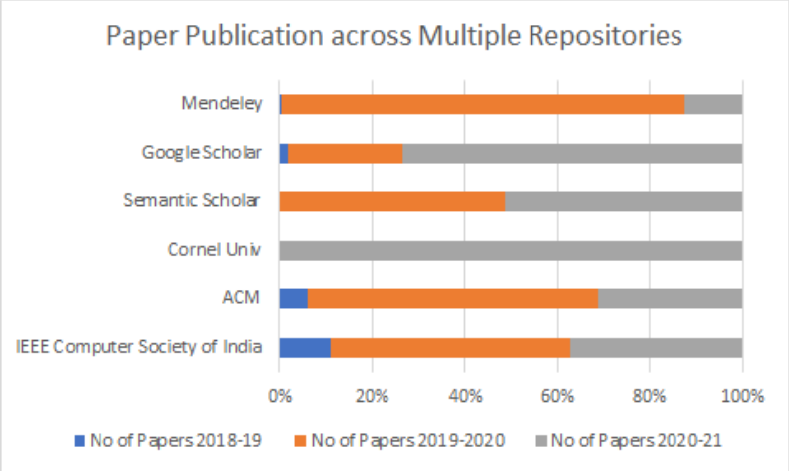

Fig. 1. Paper Publication across Multiple Repositories

2 Literature Review

For visualization of big size of data and for critical decision making, different techniques are useful like Bag of Words, N Gram Analysis and Word Cloud or Tag clouds. Algorithms like Bag of Words and N Gram analysis are used by authors in literature text mining and analysis. In [9] the authors have classified human actions using visual N gram algorithm. Bag of visual words is also used to improve the system. The problem of non-vocabulary conversion of human actions to words is removed by using proposed method. Graph based N gram method is followed to improve the algorithm. Word based atten-tion model which is based on pair of words designed by using N gram model in [10]. In [11] authors have proposed a method based naïve Bayes and Support Vector machine for text classification. N gram method is used to capture order of words. Bag of n words representation is used to represent text document.

The problem high dimension vector space is removed in [12]. Constant feature space is proposed based on standard alpha-bet. It has avoided the use of document vocabulary space and NLTK tools. It helped to reduce the dimension of vector space. In [13] a method is proposed to organize and reuse safety instructions using Ontology method. This method can be co-related with the method used for research papers titles analysis.
Arabic Text Classification by using N gram method is illustrated in [14]. Frequency statistics are calculated using proposed method. In [15] authors have proposed a methodology for image classification using Bag of Words algorithm. Pooling and coding methods are proposed and evaluated for image features extraction.
Image classification method using graph-based Bag of Words algorithm is proposed in [16]. Image representation using bag words algorithm is proposed by [17]. Authors have used Inverse Document Frequency (IDF) to optimize Bag of Words algorithm.
Word Cloud are studied and extended by many authors in literature. Authors have created word cloud for smart-phone named "Pediacloud" for visualizing links between text and images [17]. Capabilities of Tag Cloud or Word Clouds [18] are elaborated in detail in [19].

3 Word clouds / Tag Clouds Visualization of the Titles

Word clouds/Tag Clouds are useful for representing textual data so that visualization will be better. It is useful for provid-ing attention to data and keywords if data is in big size. Word Clouds or Tag Cloud, are useful in decision making and for critical observation [19]. By writing scrappers in python using Beautiful Soap libraries, paper titles were successfully ex-tracted related to Covid. Extracted titles are stored in file database. Title's pre-processing is done to remove punctuation marks, stopwords etc. Keywords are extracted by applying bag of words algorithm. Combinations of keywords are ex-tracted further by using N Gram algorithm. Python scripts were written to extract 2-, 3- or 4-gram text from the extracted titles. This process is demonstrated in Fig.2

Visual representation of different keywords used by researchers in their titles is shown by drawing word clouds [18] as given in Fig.3(a) and Fig.3(b).

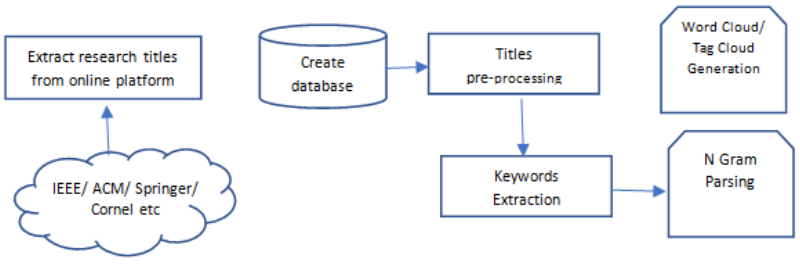

Fig. 2.Visualization of Research Titles using Word/Tag Cloud

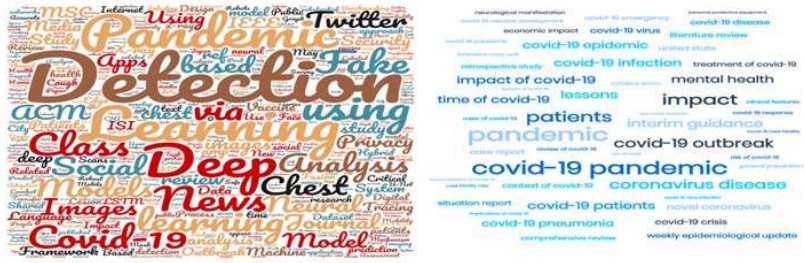

Fig. 3. (a) Word Cloud for Cornel University Repository; (b) Word Cloud for Google Scholar Repository.

4 Prefix and suffix word analysis and visualization

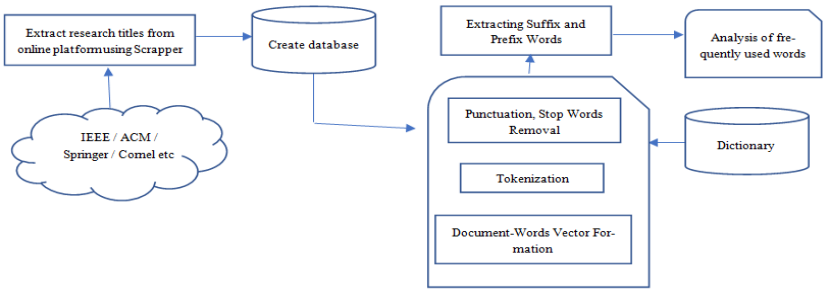

Fig. 4. Methodology followed for Prefix and Suffix Analysis

Bag of Words algorithm is used for text processing and extracting features from documents. For analyzing the titles of the research papers stating Covid-19, Suffix and prefix word analysis is done to understand author's tendency of using keywords before and after "Covid". For doing this analysis, a methodology based on Bag of Words algorithm is proposed as depicted in Fig.4 above.

Further analysis of frequently used prefix and suffix words with 'Covid-19' in different research titles is shown in Table-2. It can be observes that maximum researchers have used 'pandemic' or 'outbreak' words to show significance of their titles. Some have done research on spread on fake news with respect to Covid.

Table 2 –Frequently used Suffix and Prefix with “Covid” Keyword

Most Frequently Used Suffix	Most Frequently Used prefix
pandemic	Impact
epidemic	Context
outbreak	Review
pneumonia	Novel
treatment	Response
infection	Prevention
Research	Effects
fake news	Model

5 Tree Map Visualization of the Titles

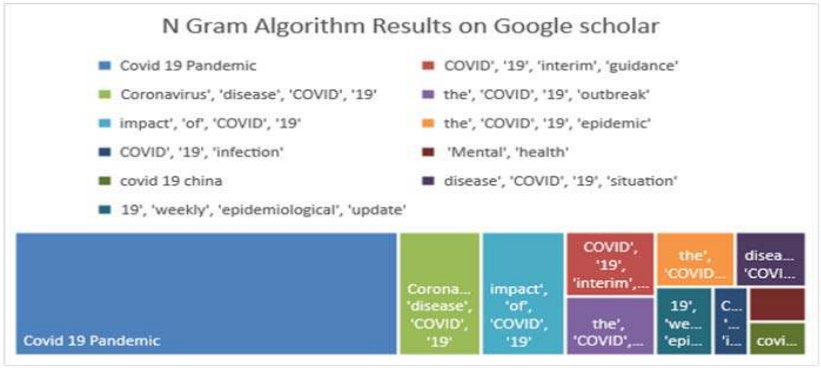

Fig. 5.Tree Map for N Gram Visualization on Google Scholar

Tree maps are used to visualise data using rectangles of different sizes and colours. Size of triangle signifies its contribution towards whole dataset. These results show that on which parameter of corona, maximum research work is done. To draw Tree Map, N Gram analysis is applied to extract combination of various keywords used in research titles. Tree map is drawn on Google Scholar Database shown in Fig.5.

6 Comments and Discussion

N gram analysis has shown that Covid-19 pandemic word is used by most of the researchers among different publication repositories. Computer Specific researchers have used words like Fake news detection or sentiment analysis or classifica-tion in their titles. These publications may be objectively considered and critically reviewed whether they as truly contribut-ing to the body of knowledge or are just following the hype of using popular keywords in the Title of the paper and thus piggybacking on the keyword Covid-19.

7 Conclusion

Along with major social or economic disruption, analysis is done to find how Covid has affected research domains and the titles of research papers being communicated by researchers. In Year 2019-20, maximum papers were reflected with the title on Covid-19 under all major publishers. N-gram analysis was done as well to check which words are frequently pre-ferred by researchers in their title. Observation using Prefix and suffix analysis, suggests more than 70% researchers have used Pandemic, Outbreak prefix in their research paper titles. Visual Representation is possible using Word Clouds or Tag clouds or Tree maps. For using these visualization techniques correctly, Bag of Words and N gram algorithm can be used for required keyword extraction.

References

1. https://scholar.google.com/
2. https://www.semanticscholar.org/
3. https://www.library.cornell.edu/arxiv
4. https://dl.acm.org/
5. https://ieeexplore.ieee.org/Xplore/home.jsp
6. https://www.mendeley.com/
7. https://www.wordclouds.com/
8. https://colab.research.google.com/
9. Ruber Hernández-Garcíaa, Julián Ramos-Cózarb, Nicolás Guilb Edel, García-Reyesc, Hichem Sahli, “Improving Bag-of-Visual-Words model using visual n-grams for human action classification”, Expert Systems with Applications, Volume 92, February 2018, Pp. 182-191
10. I. Lopez-Gazpioa M. Maritxalara M. Lapatab E. Agirre, ”Word n-gram attention models for sentence similarity and inference”, Expert Systems with Applications, Volume 132, 15 October 2019, Pp. 1-11
11. Bofang Li, Zhe Zhao, Tao Liu, Puwei Wang, Xiaoyong Du, “Weighted Neural Bag-of-n-grams Model: New Baselines for Text Clas-sification”, Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, December 11-17 2016, Pp. 1591–1600
12. Fatma Elghannam, “Text representation and classification based on bi-gram alphabet”, Journal of King Saud University - Computer and Information Sciences, Volume 33, Issue 2, February 2021, Pp. 235-242
13. S. Zhang, F. Boukamp, J. Teizer, “Ontology-based semantic modeling of construction safety knowledge: Towards automated safety planning for Job Hazard Analysis (JHA)”, Autom. Constr., 52 (2015), Pp. 29-41
14. L. Khreisat, “Arabic text classification using N-gram frequency statistics a comparative study”, Proceedings of the 2006 International Conference on Data Mining (2006), Pp. 78-82
15. Chong Wang, Kaiqi Huang, “How to use Bag-of-Words model better for image classification”, Image and Vision Computing, Volume 38, June 2015, Pp. 65-74
16. Fernanda B.Silva1,Rafael de O.Werneck Siome, Goldenstein, Salvatore Tabbone, Ricardo da S.Torres, “Graph-based bag-of-words for classification”, Pattern Recognition,Volume 74, February 2018, Pp. 266-285

17. Qun LI, Hong-gang ZHANG, Jun GUO, Bir BHANU,LeAN, “Improving bag-of-words scheme for scene categorization”, The Journal of China Universities of Posts and Telecommunications, Volume 19, Supplement 2, October 2012, Pp. 166-171
18. Bjørnar Tessem, Solveig Bjørnestad, Weiqin Chen, Lars Nyre, “Word cloud visualisation of locative information”, Journal of Location Based Services, Volume 9, Issue 4, 2015.
19. Jason R. C. Nurse, Ioannis Agrafiotis, Michael Goldsmith, Sadie Creese & Koen Lamberts, “Tag clouds with a twist: using tag clouds coloured by information’s trustworthiness to support situational awareness”, Journal of Trust Management, 2: 10. Dec 2015. DOI 10.1186/s40493-015-0021-5