NavaJyoti, International Journal of Multi-Disciplinary Research

Volume 5, Issue 2, February 2021

A Study on Non-Medical Impacts of Covid-19 Using Machine Learning Techniques

[1]Shresta R (182MCA34) and [2] Dr. Irene Getzi S

[1]Student, Dept. of MCA, Jyoti Nivas College

[2]Assistant Professor, Dept. of MCA, Jyoti Nivas College

ABSTRACT

Covid-19 pandemic has made people to face many challenges, It has completely changed the life style of the people, many countries government had taken the decision of emergency lockdown to control the spreading of virus, due to lockdown many lost their jobs and people have suffered lot to survive. It has not only affected to the health of people but also affecting other domains such as social, economic, business, finance etc.. this research aims to study the non-medical impacts of covid-19 using machine learning techniques using covid-19 public media dataset by anacode.

Keywords: Natural language processing, Text mining, Topic modeling, Covid-19, Text classification

I. INTRODUCTION

The covid-19 pandemic has affected lot to the human life's in many different ways, people are not only suffering from the health problem, because of lockdown many people have lost their jobs, IT field is somehow managed by doing work from home, but other works which has to be done in the work place has affected lot. People can't go out to their work place to do their work, both workers and company if facing loss. This research paper aims to study which of the non-medical domains has affected much comparatively to other domains by using the covid-19 public media dataset by anacode and applying machine learning techniques, as the dataset is unstructured LDA (latent dirchlet allocation) algorithm is used for topic modeling.

II. RELATED WORK

Paper 1: This paper presents the overview of text mining and its current research status using six scientific database like springer, science direct, Cambridge, IEEE, SAGE, Wiley. This research process includes papers extraction, retrieve and pre-processing steps, text mining techniques such as clustering, association rule, visualization, Terms frequency. The experimental results includes the frequent keywords in collected articles, the most frequent terms, common topics, articles interrelated to each other. The results indicate that springer article is the richest source of mobile learning in medical education.

Paper 2: In this paper, text classification is done using different machine learning algorithms, the research process includes reading the document, text tokenization, stemming, deleting stop words, vector representation of text, feature selection and feature transformation, learning algorithms. The machine learning algorithms used are SVM and KNN. The experimental result was different for different training data and classifiers.

Paper 3: This paper presents topic modeling of Wikipedia articles data and users tweet data. Two experiments are proposed for topic modeling. This research paper process includes data collecting, data pre-processing, and model training. Pre-processing step includes data tokenization, removing of stop words and stemming word, the data modeling is done by using LDA algorithm. Each experiment resulted with five different topics.

Paper 4: This paper presents topic modeling of cora dataset (collection of 2410 scientific documents), nine documents are randomly selected from cora dataset for topic modeling, By using the latent Dirichlet allocation on the fixed vocabulary, the nine documents are divided the vocabulary into 10 latent topics that remain constant.

III. METHODOLOGY

Text Mining: Uses natural language processing to transform the unstructured data into structured data suitable for information extraction. High quality of information can be obtained from text mining. Before applying different text mining techniques.it has to be started with pre-processing step that is to clean and transform text data into machine usable format. Pre- processing usually involves tokenization, changing all text to lower case, removing stop words, stemming, part of speech tagging etc... After pre-processing, text mining algorithms can be applied.

The text data can be organized in three different ways such as:

- **Structured data** : is an organized data and clearly defined and easily searchable, which can be used for analysis and machine learning algorithms,
- **Unstructured data:** it is not organized in a pre-defined format or pre-defined data model, difficult to search, difficult to analyze, it is not suitable to information extraction or knowledge discovery, some of the machine learning algorithms should be used to transform unstructured data into machine and human understandable format.
- **Semi-structured data:** the data is informative but does not reside in a relational database. Some of the organized properties makes it to easily analyze and understand. But it has no strict structural framework, the text within has open ended but no structure.

Topic modeling:

Topic modeling is a unsupervised machine learning technique, it is the process of diving the document into groups. Several sets of documents are grouped by topics, topic modeling is used to understand and organize large collection of unstructured data. The topic modeling is useful for document clustering, information extraction, feature selection and information retrieval from unstructured data.it does not require training data, topic modeling can be used to identify the topics by detecting patterns and reoccurring words. Topic modeling involves grouping similar words, counting of words within the unstructured dataset. Topic modeling require less manual input than supervised because it does not require training data or trained by humans.

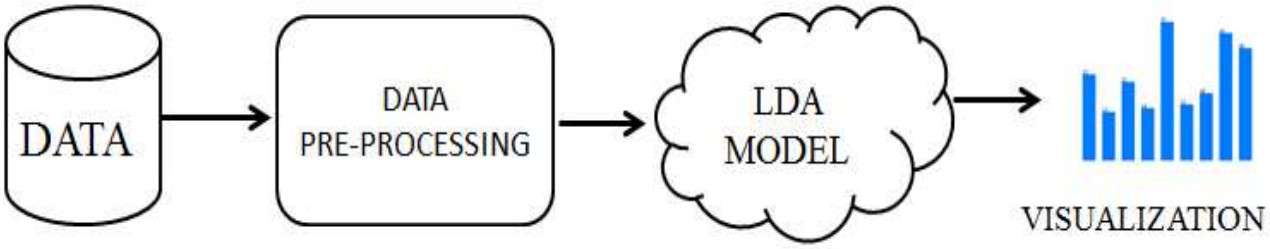
Latent Dirichlet Allocation (LDA):

LDA is used for mapping each document into list of topics, it determines the number of words in the document, it is used for dimensionality reduction. By using LDA algorithm one can extract the relevant information by minimizing the noise and reducing the redundancy. LDA is an unsupervised probabilistic model that is used to discover latent themes in a document.

LDA makes the following two assumptions for topic modeling

1. Each article is made up of some topic distribution.
2. Each topic is made of some word distribution.

FLOW DIAGRAM



IV. IMPLEMENTATION

Environmental Setup

- **IDE** - Jupyter Notebook
- **Language** - python
- **Dataset used** - covid19_articles.csv/from kaggle

Importing data: covid-19 public media dataset is used which consist of more than 52 thousand articles of different domains such as sports , government, science, political etc… pandas library is used to import data, the dataset used are manually classified and almost all the articles are named as general category, this general category data has to be grouped into different domains. For grouping the data, topic modeling has to be done.

Pre-processing

- Removing blank rows in data
- Changing all text to lower case
- Remove stop words
- Tokenization

LDA algorithm for topic modeling: LDA algorithm is applied for pre-processed covid-19 articles dataset, unsupervised learning does not required training data, as the dataset contained more than 52 thousand articles, all the data was pre-processed and used for topic extraction

Visualization: The visual representation of data will give the better understanding and a clear picture of the experimental result, where we can easily find which of the non- medical domain has affected much comparatively with other domain

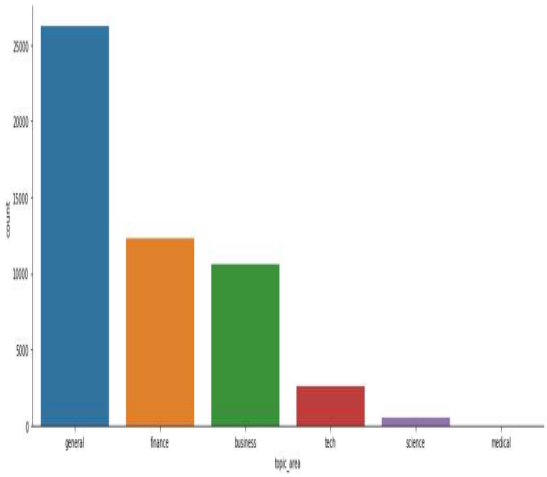
Experimental Result

The table depicts the topics obtained by applying LDA algorithm

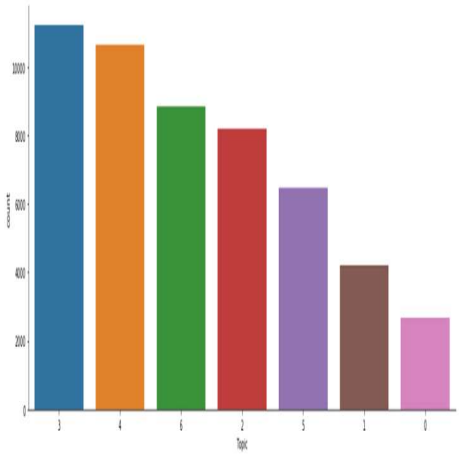
Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
Prime	Election	Customer	Novel	Company	Future	Event
Country	Governor	Food	Public	Demand	Cost	Video
Minister	Government	Home	Confirm	Cut	Market	Team
Power	Administration	Worker	Spread	Rate	Business	Club
Historic	President	Employee	Covid	Price	Income	Player
order	State	Service	Infection	Market	Product	League
Government	federal	pay	Hospital	Rise	financial	Season

Visual Representation:

Before Applying LDA



After Applying LDA



Initially most of the articles were classified as general category. After applying LDA algorithm we have got different domains namely medical which is blue in color and here in this research we are focusing on non-medical domains the most affected domain is Finance which is orange in color, and next is the Sports which is green in color, and next is the Employment and Business which is red and purple in color, Politics and Government are less affected compared to other domains which are brown and pink in color as shown in the above visual representation.

V. CONCLUSION

Text mining is one of the best techniques used to study the pattern or to extract information. In this paper covid-19 public media dataset was used which contained news articles of initial covid-19 period from January 2020 to May 2020 dataset has been grouped into different topics to study how this initial covid-19 pandemic had effected on non-medical domains. In the future study different supervised machine learning algorithms will be applied to classify the articles and see which classifier will give the accurate result.

REFERENCES

1. Kanika, Sangeeta (2019) Applying Machine Learning Algorithms for News Articles Categorization: Using SVM and kNN with TF-IDF Approach. In: Luhach A., Hawari K., Mihai I., Hsiung PA., Mishra R. (eds) Smart Computational Strategies: Theoretical and Practical Aspects. Springer, Singapore.
2. A. Salloum, Mostafa Al-Emran, Azza Abdel Monem and Khaled Shaalan(2018) Using Text Mining Techniques for Extracting Information from Research Articles, at: https://www.researchgate.net/publication/321150349
3. Tong, Zhou, Zhang, Haiyi(2016/05/21) A Text Mining Research Based on LDA Topic Modelling,volume-6, Computer Science & Information Technology
4. Springer International Publishing AG 2018 K. Shaalan et al. (eds.), Intelligent Natural Language Processing: Trends and Applications, Studies in Computational Intelligence 740
5. A. Fesseha, S. Xiong, E. D. Emiru and A. Dahou, "Text Classification of News Articles Using Machine Learning on Low-resourced Language: Tigrigna," *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, Chengdu, China, 2020, pp. 34-38, doi: 10.1109/ICAIBD49809.2020.9137443.

6. Gui Y., Gao Z., Li R., Yang X. (2012) Hierarchical Text Classification for News Articles Based-on Named Entities. In: Zhou S., Zhang S., Karypis G. (eds) Advanced Data Mining and Applications. ADMA 2012. Lecture Notes in Computer Science, vol 7713. Springer, Berlin, Heidelberg.
7. A. Fesseha, S. Xiong, E. D. Emiru and A. Dahou, "Text Classification of News Articles Using Machine Learning on Low-resourced Language: Tigrigna," *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, Chengdu, China, 2020, pp. 34-38, doi: 10.1109/ICAIBD49809.2020.9137443.