 doi:10.20944/preprints202011.0056.v1

Article

Exploring the Non-Medical impacts of Covid-19 using Natural Language Processing

Amol Agade [1] * and Samta Balpande [2]

[1] Department of Information Technology; University of the Cumberlands; 6178 College Station Drive; Williamsburg; KY 40769; aagade8666@ucumberlands.edu
[2] School of Business Administration; Oakland University; Elliott Hall; 275 Varner Drive; Rochester, MI 48309; samta2012@gmail.com
* Correspondence: aagade8666@ucumberlands.edu

Abstract: Ongoing COVID-19 Pandemic has resulted into massive damage to various platforms of global economy which has caused disruption to human livelihood. Natural Language Processing has been extensively used in different organizations to categorize sentiments, perform recommendation, summarizing information and topic modelling. This research aims to understand the non-medical impact of COVID-19 on global economy by leveraging the natural language processing methodology. This methodology comprises of text classification which includes topic modelling on unstructured COVID-19 media articles dataset provided by Anacode. Like other Natural Language Processing algorithms, Latent Dirichlet allocation (LDA) and Non-negative matrix factorization (NMF) has been proposed to classify the media articles dataset in order to analyze COVID-19 pandemic impacts in the different sectors of global economy. Model Accuracy was examined based on the coherence and perplexity score which came out to be 0.51 and -10.90 using LDA algorithm. Both the LDA and NMF algorithm identified similar prevalent topics that was impacted by COVID-19 pandemic in multiple sectors of economy. Through intertopic distance map visualization produced by LDA algorithm, it can be reciprocated that general industries which includes children schooling, parental care, and family gatherings had the major impact followed by business sector and the financial industry.

Keywords: COVID-19; Deep Learning; Natural Language Processing; Topic Modelling; Text Classification; Latent Dirichlet allocation (LDA); Non-negative matrix factorization (NMF)

1. Introduction

Natural Language Processing (NLP) is the deep learning methodology that has evolved greatly to obtain meaningful insights from human language in the form of tweets, documents, articles or texts. This has enabled computer to learn human emotions and sentiments from the wide variety of data available. Due to availability of abundance data and large computational power, NLP is booming in large number of industries to derive meaningful insights from texts and drive organizational decisions. Whatever we express either verbally or written, has some or the other information in it. Natural Language Processing algorithm uses this information to predict human behavior.

NLP is getting attention very much as it has capability to learn from unlabeled data to classify relevant topics [1-3]. NLP is booming in mass media industry and social media platform due to its capability of deriving meaningful insights from the news headlines, tweets, sentiments shared by the people in the content of news article and understanding the impact of this analysis on human lives [4]. Furthermore, Natural language processing has shown its caliber in almost all the disciplines such as healthcare, finance, food industry, technology and political science etc. to understand the mankind and its behavior to optimize the experience [5-12]. People relies more on newspaper and social media to seek updates from every corner of the world. Thus, it has become necessary to extract this meaningful information from social media and newspapers which is stored

 doi:10.20944/preprints202011.0056.v1

in the form of texts, documents and article. Topic modeling is the method that provides us with the concept of extracting prevalent texts and information from the set of documents and collection of texts [13].

Due to the outbreak of the Covid-19 pandemic, world has been suffering from huge loss in various domains and not just limited to healthcare facilities. The only source to rely on updates regarding COVID-19 impacts, necessary measures and treatments were newspaper article, social media and research articles. Thanks to the internet, through which we could access real time updates and take necessary measures against this deadly virus to ensure safety of people around us. However, healthcare industry was largely affected with the daily rise in Corona positive cases, it created adverse impact on other domains as well hampering the global economy. With the availability of mass media dataset, it has become wide possible to understand the impact and affected areas of global economy other than just health care industry.

The media dataset is considered to be an unstructured data with large set of texts and information. This needs to be structured into topic or categories by the use of dictionaries or corpus or supervised learning methods. The aim of this research is to leverage the capability of Latent Dirichlet allocation (LDA) and Non-negative matrix factorization (NMF) to correctly identify the non-medical impact of COVID- 19 pandemic. LDA was considered as the base model for this research, where each news article can be described by distribution of topics and each topic can be described as distribution of words. Fang et al. suggested Contrastive Opinion Modelling based on LDA whose purpose was to find out the opinions from multiple perspective on the given topic and compare those opinions with the individuals. That model was able to generate a process of how opinions would occur in different collections of documents [14]. Few authors have also presented the study where they collected the topics in software engineering that are evolving rapidly and the topics that are fading out or are not catching much attention from the public as it requires more testing due to data duplication in the source code repositories. They used diff model to detect the distinct topic in software engineering to separate out the duplication from the source code repositories [15].

LDA was considered to be the best choice for this research because it works similar to dimensionality reduction methodology of the original dataset. The primary goal of LDA is to determine the mixture of topics that a news article contains [16]. This facilitates the distribution of the words in topics as more Dirichlet distribution, hence creating more transparent vector representation of topics in news articles. Similarly, NMF works as same as that of dimension reduction to cluster the news article to determine most relevant topics by considering only non-negative elements. NMF performs well with text or document clustering and topic modeling. Also, NMF is more efficient and faster than LDA calculations to produce the results from bag of words. Over the last few decades, NMF has gained enormous amount of attention in the field of text mining, spectral data analysis, Bioinformatics, image processing, hyperspectral data analysis, computational biology, clustering and many others due to its computational capability and efficiency [17-19]. In the past, NMF has been also used to detect the protein protein interactions between HIV-1 and human proteins where multiple datasets were collected to form biological network and were then utilized to predict the accuracy for the model [20].

With the Covid-19 media dataset and resources, topic modeling techniques like LDA and NMF successfully detects the prevalent topic in the dimension that was impacted by Covid-19 worldwide. This study also includes the comparative analysis of LDA and NMF algorithm to determine the accuracy of topics in both the algorithms. This can help public to understand the effects of Covid-19 on global economy and take necessary strategic measures against it to bring back everything on track.

Literature Review are described in Section 2, proposed and pretrained models are presented in Section 3, Results and Analysis are depicted in Section 4 and lastly Section 5 concludes the finding of paper.

 doi:10.20944/preprints202011.0056.v1

2. Literature Review

There has been extensive workaround on the topic covid-19 on how it is impacting everybody's livelihood. We are trying to do a workaround from non-medical perspective on how Covid-19 is affecting people's capabilities. Few researchers have directly considered distinct topic area and have defined their research-based work on how covid-19 is altering us globally.

Author Hee tried to find out the impact of COVID- 19 in financial world, where he tried to assess the sentiments of US stock market using Daily News Sentiment Index and Google Trends searches on the topic relating to covid-19 for the time span of Jan-May 2020. According to Hee, strategic investment decision is needed by considering the time lag perspectives by visualizing the changes in the correlation level by time lag differences [21]. Bollen et al. proposed a solution where they tried to analyze whether a public emotions or mood from Twitter Feed are directly correlated to Dow Jones Industrial Average (DJIA). They used Fuzzy Neural Network to predict the public mood from the Twitter feed and tried to corelate that with DJIA. There model found an accuracy of 87.6 % to predict daily changes in DJIA by inclusion of public mood [22]. Similarly, Bharati et al. proposed a study where they have combined Indian stock Market Sensex data points, Twitter data and Really Simple Syndication (RSS) feeds to predict the daily changes in the stock Market. Their study proved there is correlation between stock market index, RSS and Twitter Feeds [23]. In addition to this, Pereira et al. used the non-parametric models like ANN, SVMs with polynomial kernels, and RBF kernels to predict the movement of Korean stock Market (KOSPI 200) where Google Trend was found to be inadequate input factor to predict the price of KOSPI 200 Index [24].

Owing to capabilities of technologies, Chamola et al. have explored the use of technologies like Internet of Things, Artificial Intelligence, Drone technology, Autonomous Vehicles and wearable devices to mitigate the risk posed by covid-19 [25]. These technologies are used for spraying disinfectants, delivering medical equipment's, monitoring health of the patient from remote location, screening masses and do 24 X 7 crowd surveillance to ensure strict social distancing protocols are in place. Because telehealth is growing rapidly, Hedge et al. have demonstrated that many government organizations around the world are developing the digital contact tracing application which could help Health Officials to gather the information on patient with covid-19 symptoms in order to isolate them [26]. There is also a greater concern in the community, if such tools developed by authorities should undergo regulatory process before it is rolled out for general public usage.

Businesses like Food Industries are getting impacted to maintain proper supply chain as many of the workers are falling sick to SARS-CoV-2. In order to create safe food environment, this companies are applying antimicrobial coatings to high touch surfaces like doors, handles, touch screen to inactivate the virus [27]. Similarly, Employees working in the retail sectors like shopping malls, Grocery stores who has direct customer exposure are 5 times more likely to have tested positive for SARS-CoV-2 [28].

In the healthcare domain, According to Sethi et al. Nurses are experiencing an extreme workload in managing healthcare facilities at their workplace. Most of them are feeling anxious, distressed and depressed due to Covid-19 Pandemic [29]. On the budget aspect, Dauner et al. suggested the steps on how hospital management system can prepare themselves so that they provide clear documentation and create a dashboard which tracks all the metrics related to financial expenses due to covid-19 impact on weekly basis [30].

Automobile sectors have also seen a hit from the ongoing COVID-19 pandemic where their sales have plummeted to the numbers not seen before. Due to strict social distancing guidelines from March 2020, General Motors is prioritizing new vehicle redesigned model over already existing

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0056.v1

freshened model. They have already lost 2 months of new production for a new car to comply with government policies to protect their workforce from covid-19 Pandemic [31]. According to Mead et al. Covid-19 pandemic caused short term disruption in the economy where many researchers saw a shift where people were found to be consuming home cooked food rather than consuming restaurant food. Prices for the meat products were increased for the US consumers and there was price volatility in all the BLS price index [32].

Despite the myriad of research in the field of Covid-19 and its impacts it has caused in everybody's livelihood, at the time of this writing, there is no specific research based study that provides how the world economy was shaping in different sectors due to COVID-19 outbreak. Furthermore, no work in existing literature attempts to review the role of Natural Language Processing Algorithm such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) in mapping out COVID-19 pandemic in different sectors. This presents an urgency to detail non-medical impacts of COVID-19 on the global economy considering various sectors such as Finance, Business, Technology, healthcare, Automobile, International relations and general Industries. In this work, using natural language processing statistical method called topic modelling, we present a comprehensive review of how COVID-19 is not only affecting the medical domain but also it had massive impact on overall sectors of global ecosystem. Nevertheless, Topic Modelling model is constructed and trained with the dataset that is available from Jan-May 2020 provided by Anacode. With this base model it can be scaled to later dataset as well. Before divulging into a thorough analysis of the COVID-19 pandemic, we take a brief look at some of the past pandemics in the section below.

3. Proposed Topic Modelling on Covid-19 Media Article Dataset

The whole system of topic modelling from COVID-19 media dataset comprises of key steps that was followed during the complete analysis – collection of data, cleaning and pre-processing of data, creating bag of words, training the model and evaluation and analysis of the model. The architectural flow of the topic modelling model is portrayed in Figure 1. Firstly, the dataset on COVID-19 media article was collected from Jan 2020 to May 2020 timeframe. The dataset was then studied to understand the key features and elements that will be required to perform topic modelling. After this step, data cleaning and preprocessing took place using the Regex library to get rid of emails, distracting characters and new line characters. Then, models are trained based on the cleaned and preprocessed data using python libraries. Lastly, the models were evaluated and based on the key metrics like coherence score and perplexity. Following part of this section discusses how models were constructed using topic modelling and its performance.

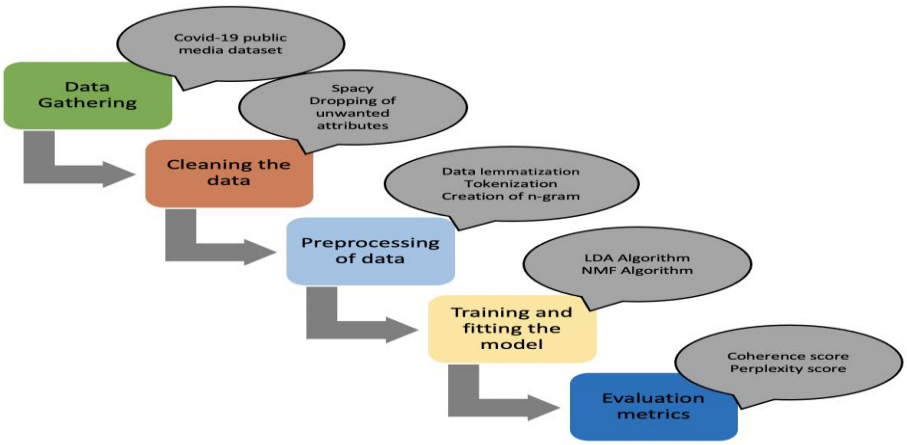

Figure 1. System architecture of topic modelling on COVID-19 media dataset.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0056.v1

3.1. Dataset Collection and Modeling

For constructing the topic modelling model, COVID-19 media article dataset was grabbed from Kaggle dataset naming "Covid-19 Public Media Dataset by Anacode" [33]. This dataset contains over 200,000 news articles with full content that was filtered out using web scraping from online media since January 2020 timeframe. The focus of this dataset is to analyze the non-medical impact of COVID-19 pandemic in various aspects of society. The heart of this dataset are the online articles from news websites and blogs which are in text form and was shared to help the community to explore non-medical impacts of COVID-19 pandemic in various dimensions such as economic, political, social and technological aspects. Thus, this dataset is considered to be best for NLP and text mining to retrieve information regarding the fake news or rumors and compare it with the real time situation that COVID-19 pandemic has resulted into. January 2020 to May 2020 timeframe dataset consists of total 76471 rows and 9 columns namely headlines title, content of the articles, domains, publish date of the article, author, crawled time, topic area and URL from where the article was extracted. Exploratory data analysis on the COVID-19 public media dataset is represented in below graphical representations in Figure 2. which clearly depicts the impacted categories of the global economy due to Covid-19 pandemic

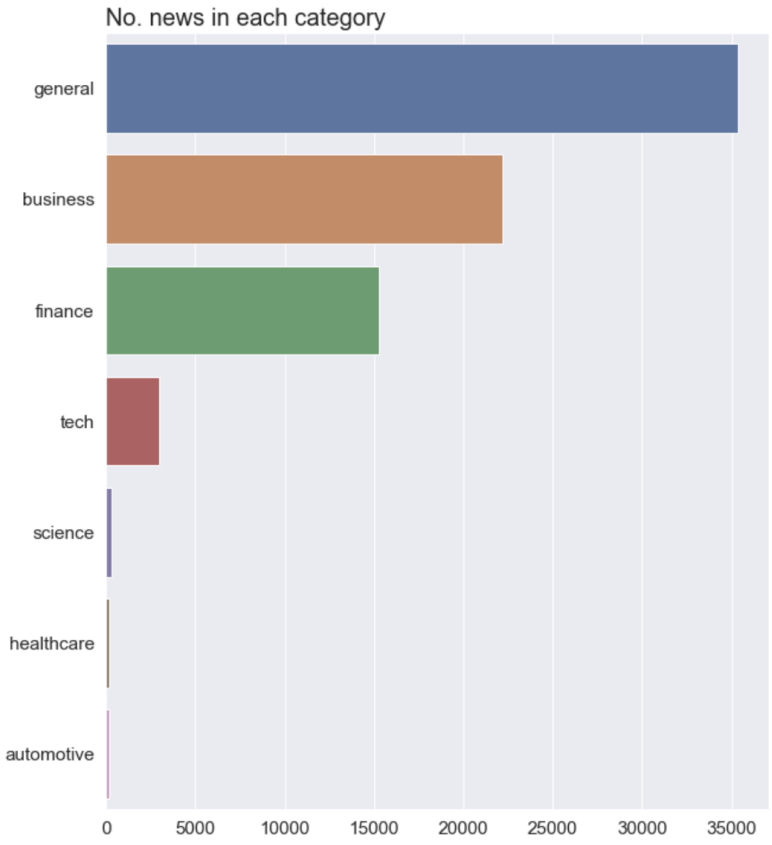

Figure 2. The impact of the COVID-19 pandemic in each dimension.

Figure 3. and Figure 4. elaborates the headlines collected in this dataset are usually 10 to 170 character long and generally it is between 70 to 120 characters and if we look at the data exploration at word level, it clearly shows that words in news title appears to be ranging from 5 to 27 words long. However, most of the news article title with number of words falls within the range of 10 to 20 words long.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0056.v1

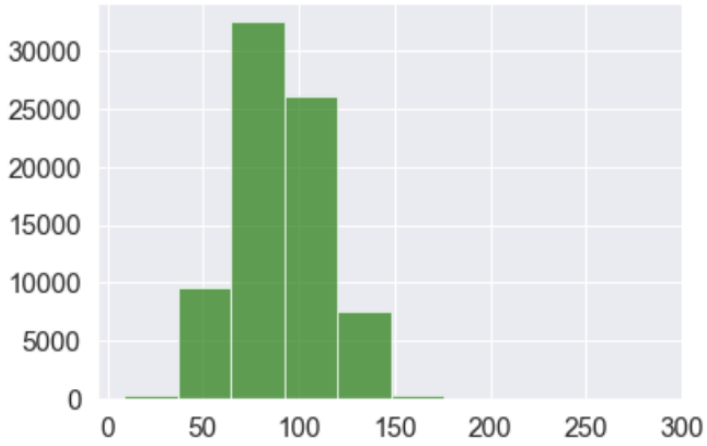

Figure 3. Number of characters in media article title.

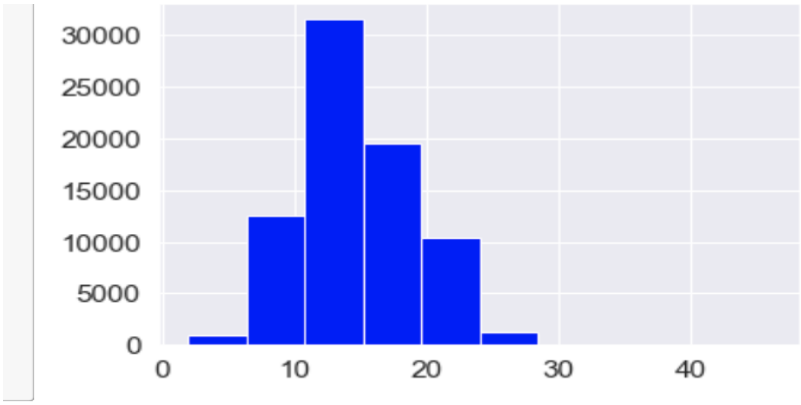

Figure 4. Number of words in media article title.

3.2. Topic Modelling

In Deep learning and natural language processing, topic modelling technique plays an important role in discovering abstract topics that occur in the collection of documents and set of texts. This approach is basically very useful for search engines, to automate the customer service and other areas where knowing the topics from texts is crucial. There are several algorithms available that be trained for topic modelling such as LDA, LSA, NMF and Clustering [34-44]. These algorithms are unsupervised methods, which means, the relationship among document is not revealed prior to the model being executed. This research aims at using LDA and NMF algorithm to group public media articles in fixed number of topics and to estimate the optimal number of topics. In the following section, we provided a taxonomy of LDA algorithms that is used in this research.

3.2.1. Latent Dirichlet allocation (LDA)

LDA falls under unsupervised technique which considers documents, article and texts as the bag of words where the order of document or texts does not matter. This algorithm works by considering the set of articles as it was generated by picking a set of topics and then each topic by considering the set of words. To perform this calculation and execution, LDA uses reverse engineering concept where each document can be represented as a probabilistic distribution over latent topic . LDA can also be used to represent the topics by word probabilities in the media article dataset that shares a common Dirichlet. Therefore, LDA has the caliber to identify subtopics for various dimensions such as technology, business, finance, automobile, international relationship

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0056.v1

composed of many articles and represent each article in an array of topic distribution [45]. The basic functioning of LDA is to take M documents and N number of words, it generates k number of topics, word distribution for each topic (psi) and topic distribution for each document(phi). LDA workflow model has been depicted in Figure 5. It first makes an assumption that there are k topics across M documents and then it distributes these k topics by assigning each word a topic within each document. Its samples θ from the Dirichlet distribution of α. Then for each of the N word, a topic k is sampled from a θ distribution. Here, α is the concentration parameter that represents document topic density. Higher the α is, document is made of more topics resulting into more specific topic distribution per document. β is also considered as the concentration parameter which represents topic word density resulting into more specific word distribution per topic.

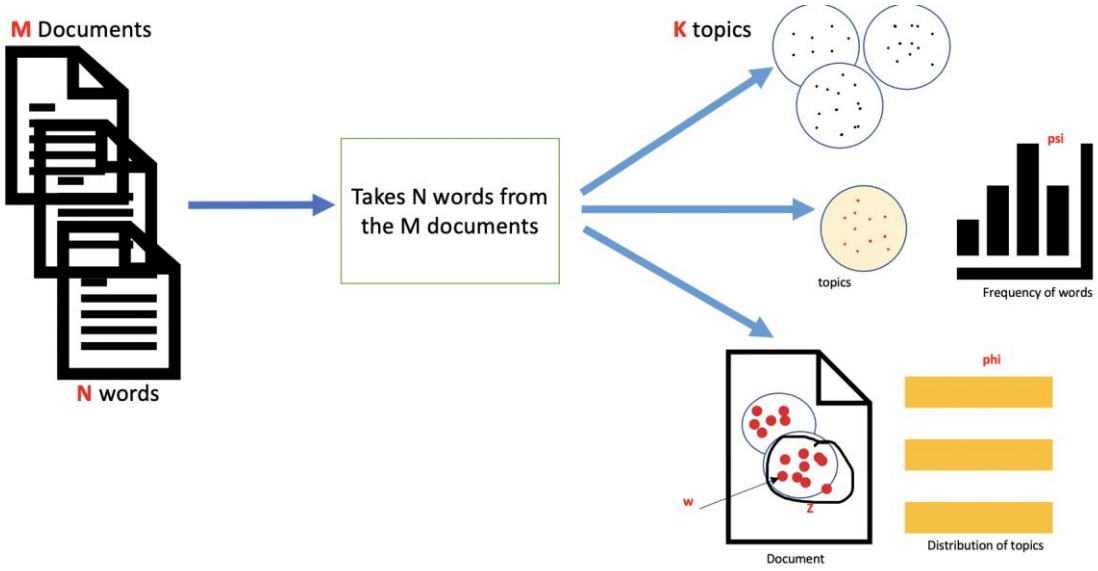

Figure 5. Latent Dirichlet allocation(LDA) workflow model**.**

Topic is considerably the collection of keywords which clearly shows what the topic is all about. Once the number of topics is provided to LDA algorithm, all it does is rearrange the distribution of topics within the set of documents and keywords with the set of topics. To have a good segregation topic, what matters the most are number of topics that are fed to an algorithm, quality of text processing and the choice of topic modelling algorithm.

With the help of LDA algorithm, COVID- 19 public media dataset was used to extract the naturally discussed topics in the public domain. During this research, LDA was used from the Gensim package which was extracted and imported in python. As a prerequisite, stop words from NLTK library were downloaded for text preprocessing. Core packages in python that are used in this research are for regular expressions to wipe out the emails addresses and special characters, genism– for building the LDA model and evaluation of its metrics, spacy- to lemmatize the data and PyLDAvis to visualize the intertopic distance map. Data lemmatization is needed to convert the words in an article to its root words without changing the meaning of the words. As the stop's words are already downloaded from NLTK library, they were imported to use for further analysis and model processing.

The COVID-19 public media dataset is available in csv format from January 2020 to May 2020 timeframe and this dataset was imported in jupyter notebook using panda's library. As a part of data cleaning and pre-processing, we got rid of the extra spaces, emails and single quotes to avoid any distraction during the modelling process. However, this process is not enough to feed the data

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0056.v1

into LDA algorithm as the data is still messy and not tokenized. To overcome this, we broke each sentence from the Content of dataset into list of words by tokenizing it and cleaned up the messy text to be utilized for further processing. Tokenization was basically carried out to remove punctuations and unnecessary characters from the Content columns in dataset. Then we created the sequence of N- words as N-gram to understand the sequence of words occurring consecutively in the sentences of articles. Specifically, we created Bigrams to figure out two words that are occurring together in the article and trigram to list out three words occurring together in the set of sentences of an article. Additionally, Dictionary and corpus was created to feed as a parameter in the LDA model as they are the key parameters for LDA modelling. Dictionary is created to bias corresponding words towards similar topic in the texts and corpus is about selecting the relevant text in the set of articles. Gensim library creates a unique id for each single word in the article along with its frequency. With all these as inputs, LDA model can be constructed where the key parameters are given along with the number of topics that we need to extract from an algorithm. We created LDA model with 20 different topics as a combination of some keywords and how each keyword contributes to that topic. As the results of LDA are hard to interpret just by looking at the output generated by LDA algorithm, we used PyLDAvis package to visualize the results which is elaborated in result and analysis section of this research article.

3.2.2. Non-negative matrix factorization (NMF)

Similar to LDA, Non-negative matrix factorization techniques also contributes to unsupervised learning where there are no labelling of the topics that a deep learning model will be trained on [46]. NMF factorizes high dimensional vectors into lower dimension representation having non-negative coefficients. Due to availability of enormous amount of mass media data, text mining has gained huge popularity and document clustering is one of the method involved into it. Document clustering is the technique of organizing set of documents or articles into several semantic clusters to help users to derive meaningful insights out of it. Topic modelling in this case, deals with the semantic meaning of each topic from the document and models it as a weighted combination of keywords.

NMF algorithm can also be categorized as dimension reduction that works by teasing out the key topics that the body of the text is about. Figure 6. shows the graphical representation of NMF model,

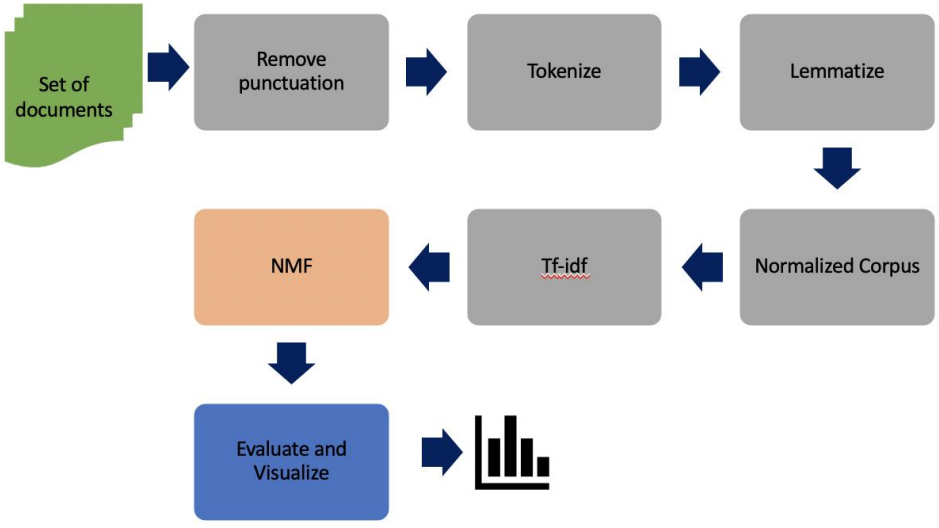

Figure 6. Non-negative matrix factorization (NMF) workflow model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0056.v1

Cleaned Covid-19 data corpus was feed into the NMF algorithm to obtain the design matrix. For results improvisation, tf-idf transformation was applied to the counts. To get count design matrix, CountVectorizer module from Sklearn python library was used. This will give the matrix of article and features where the value of each cell will be the frequency of each word in that article. Tf-idf was applied to transform the count with the model and it was then normalized to unit length for each row. NMF algorithm was then applied to these normalized tf-idf values to iterate over to each topic in the cluster and list down the important scoring words in each cluster. To ease out the computation and faster processing, we generated 20 relevant topics using NMF algorithm.

4. Results and Analysis

In this section results are divided into two parts i.e. LDA Learnings and NMF Learnings to give in-depth analysis from the study.

4.1. LDA Learnings

LDA is typically evaluated by measuring performance on how the documents are classifies or by information retrieval. It can also be evaluated by training the model first and then test it with some unseen documents to investigate how model performs on unseen data.LDA models are also evaluated on the measure of coherence score and Perplexity. Coherence score is used to measure a single topic by examining the degree of similarity between high scoring words in that topic. It measure the relative distance between the word and the topic. There are multiple measure of topic coherence through which coherence score can be calculated. We used c_v as our choice of metric to calculate the coherence score that used sliding window.The coherence score of the LDA model came out to be 0.5128. Perplexity defines on how well the topic modelling technique predicted sample and for LDA model it came out to be -10.90. To visualize LDA algorithm output, PyLDAvis library is used to from the Gensim package of python.

```
[(0,
  '0.020*"rise" + 0.020*"fall" + 0.018*"week" + 0.018*"market" + 0.017*"year" '
  '+ 0.016*"say" + 0.016*"high" + 0.015*"month" + 0.014*"low" + '
  '0.014*"economy"'),
 (1,
  '0.048*"food" + 0.034*"restaurant" + 0.020*"water" + 0.020*"shop" + '
  '0.019*"mental" + 0.015*"plant" + 0.015*"exercise" + 0.013*"eat" + '
  '0.013*"relax" + 0.013*"shopper"'),
 (2,
  '0.077*"patient" + 0.040*"drug" + 0.035*"symptom" + 0.035*"hospital" + '
  '0.029*"doctor" + 0.025*"treatment" + 0.023*"study" + 0.019*"treat" + '
  '0.017*"medical" + 0.016*"health"'),
 (3,
  '0.127*"child" + 0.093*"school" + 0.080*"family" + 0.053*"parent" + '
  '0.043*"student" + 0.030*"year" + 0.026*"old" + 0.024*"young" + '
  '0.019*"friend" + 0.019*"mother"'),
 (4,
  '0.044*"pay" + 0.028*"job" + 0.022*"money" + 0.021*"worker" + 0.020*"people" '
  '+ 0.019*"work" + 0.015*"state" + 0.015*"cost" + 0.014*"benefit" + '
  '0.014*"may"'),
 (5,
  '0.060*"stock" + 0.047*"market" + 0.036*"company" + 0.029*"investor" + '
  '0.022*"buy" + 0.021*"price" + 0.020*"share" + 0.015*"investment" + '
  '0.013*"sell" + 0.013*"covid"'),
 (6,
  '0.045*"people" + 0.037*"home" + 0.029*"say" + 0.025*"work" + 0.020*"worker" '
  '+ 0.020*"care" + 0.018*"mask" + 0.017*"health" + 0.014*"public" + '
  '0.014*"staff"'),
 (7,
  '0.019*"use" + 0.017*"new" + 0.016*"app" + 0.015*"online" + 0.015*"help" + '
  '0.014*"service" + 0.014*"user" + 0.013*"launch" + 0.012*"datum" + '
  '0.012*"customer"'),
 (8,
  '0.067*"oil" + 0.056*"production" + 0.049*"price" + 0.039*"demand" + '
  '0.029*"supply" + 0.027*"cut" + 0.027*"trade" + 0.022*"energy" + '
  '0.021*"producer" + 0.021*"global"'),
```

Figure 7. Topics generated via LDA Algorithm.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0056.v1

PyLDAvis produces the visualization that consist of intertopic distance map which is interactive map that showcases each topic as the bubble on left hand side consisting of keywords. Each bubble on the left of the intertopic distance map represents a topic. Larger the size of the bubble, prevalent the topic is. Closeness of the topic can be measured how closer the topic is to each other. A good topic model resembles some dominant bubble with smaller one disseminated on the plane. If intertopic distance map consists of more overlapping bubbles, it means there are more topics and model did not perform well. The right hand side is showcasing the most relevant bigrams of the topic. This is the interactSive chart which represents the bubble when we highlight or select any word from the right handside list.

The performance of the model can be evaluated how scattered the bubbles are on the plane of intertopic distance map. As this can be seen in the Figure 8. The intertopic distance map is created with the mmds (via multidimensional scaling) and there are some bubble overlapped on each other in first and 4th quadrant. "mmds" (multi-dimensional scaling)parameter which takes topic_term distance as input and outputs number of topics by 2 distance matrix. As there are still bubbling overlapping with the mmds as parameter, its time to investigate the result with other parameter like tsne.

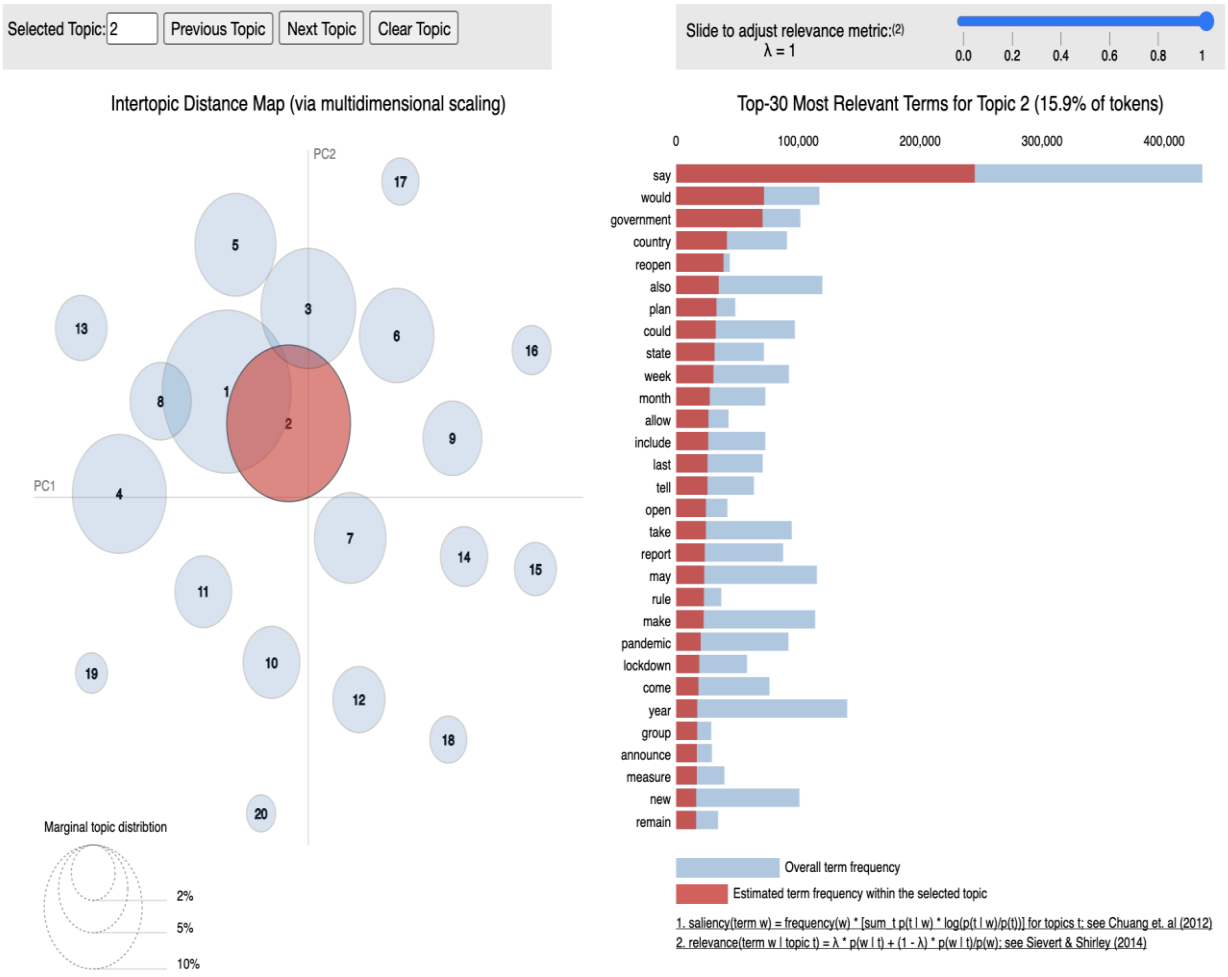

Figure 8. Intertopic Distance map using mmds parameter.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0056.v1

Figure 9. is created with tsne as the parameter of pyldavis. This is another dimessionality reduction technique which means t-distributed stochastic neighbour embedding and is used to visualize high dimensional data. This seems to give us the better result as compare to mmds as there are hardly any overlapping of the bubble on the left hand side. This confirms that the LDA model performed well on the COVID-19 public media article dataset.

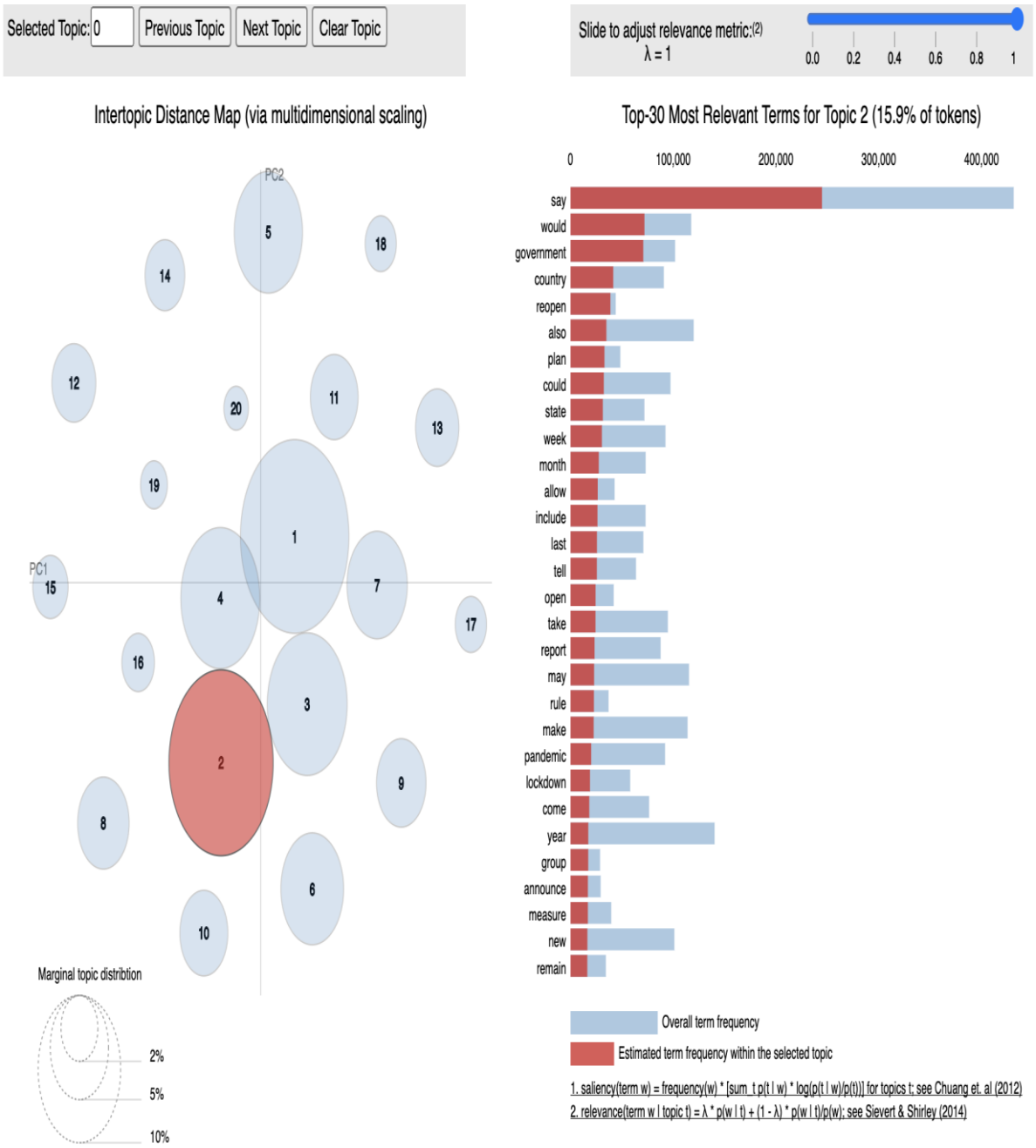

Figure 9. Intertopic Distance map using tsne parameter

Word cloud is the best way to represent the data. The size of the word in the topic resembles to its frequency and importance. As it is seen in below graph, some of the words are highlighted in larger font. This clearly shows the importance and frequency of the word with respect to other word around it.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0056.v1

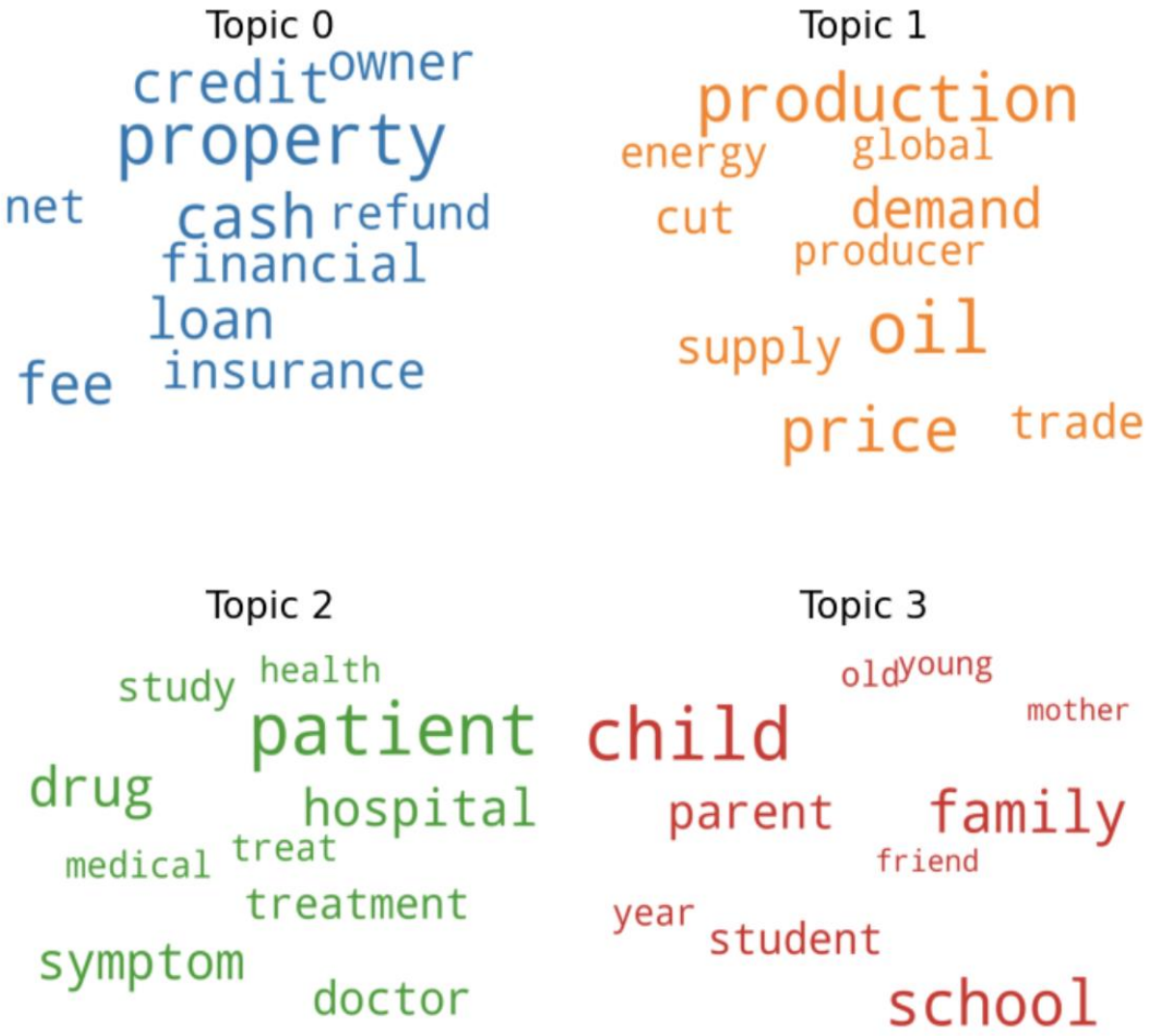

Figure 10. word cloud for each topic.

Above Figure 10. entails the keywords generated in each topic from each article. For example, keywords generated in topic 0 resembles to Finance. Likewise, topic modelling algorithm helped us to segregate the topic for each domain as shown in Figure 11.

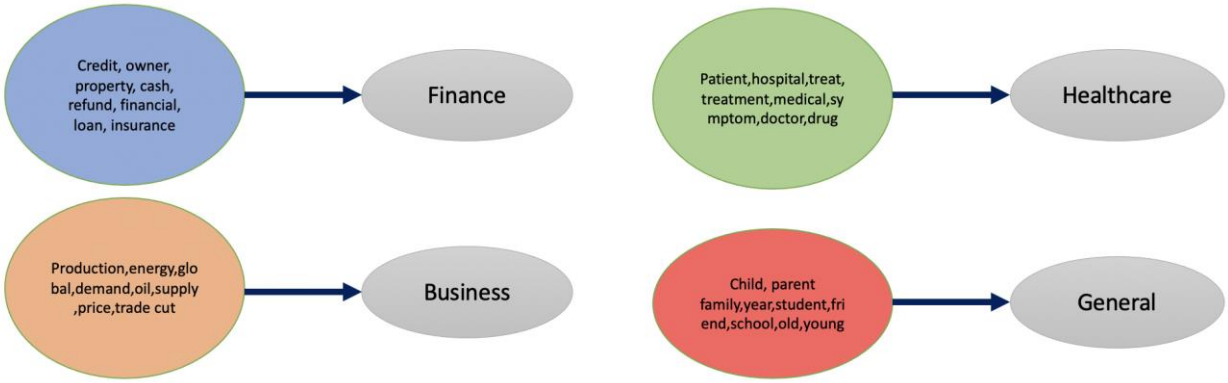

Figure 11. Inferring the topics from keyword.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 November 2020 doi:10.20944/preprints202011.0056.v1

4.1. NMF Learnings

TfIdf transformation was applied to NMF to improve the results fr covid-19 public media dataset. While evaluating NMF model, we need to consider the meaning of each topic, how prevalent the topic is in the overall corpus and the how the topics are interrelated. NMF is a deterministic model for us to modify the probabilities of key terms and determine how they vary within each topic. To achieve better topic coherence, LDA is the best choice.

```
The top 25 words for topic #0
['develop', 'case', 'dr', 'nhs', 'infection', 'trial', 'coronavirus', 'treatment', 'doctor', 'study', 'medical', 'dru
g', 'people', 'testing', 'vaccine', 'disease', 'care', 'virus', '19', 'covid', 'symptom', 'hospital', 'health', 'pati
ent', 'test']
The top 25 words for topic #1
['analyst', 'high', 'bond', 'point', 'trade', 'barrel', 'gain', 'global', 'rise', 'crude', 'year', 'economic', 'dolla
r', 'cut', 'rate', 'investor', 'economy', 'index', 'low', 'fall', 'bank', 'price', 'stock', 'oil', 'market']
The top 25 words for topic #2
['match', 'solskjaer', 'sancho', 'star', 'old', 'newspaper', 'fan', 'man', 'utd', 'team', 'arsenal', 'summer', 'manch
ester', 'game', 'chelsea', 'united', 'play', 'football', 'transfer', 'liverpool', 'premier', 'season', 'player', 'clu
b', 'league']
The top 25 words for topic #3
['authority', 'world', 'flight', 'government', 'official', 'italy', 'infection', 'city', 'reuters', 'number', 'sprea
d', 'wuhan', 'confirm', 'travel', 'people', 'health', 'chinese', 'outbreak', 'coronavirus', 'report', 'virus', 'deat
h', 'country', 'case', 'china']
The top 25 words for topic #4
['day', 'queen', 'page', 'prime', 'think', 'daily', 'issue', 'child', 'royal', 'coronavirus', 'express', 'family', 't
oday', 'work', 'nhs', 'time', 'minister', 'home', 'johnson', 'government', 'lockdown', 'mr', 'newspaper', 'people',
'uk']
The top 25 words for topic #5
['worker', 'loan', 'impact', 'provide', 'industry', 'information', 'look', 'pay', 'result', 'financial', 'work', 'ser
vice', 'product', 'share', 'include', 'revenue', 'year', 'quarter', 'store', 'sale', 'employee', 'statement', 'custom
er', 'business', 'company']
The top 25 words for topic #6
['business', 'economy', 'cuomo', 'reopen', 'pandemic', 'democrats', 'vote', 'coronavirus', 'federal', 'senate', 'repu
blican', 'democratic', 'washington', 'americans', 'election', 'york', 'governor', 'administration', 'donald', 'bide
n', 'white', 'house', 'state', 'president', 'trump']
```

Figure 12. Topics generated via NMF Algorithm.

The two figures above Figure 7. And Figure 12., in each section, show the results from LDA and NMF on the datasets. Of course there is consistency between the words in each clustering. In LDA, topic#2 shows the words associated with healthcare industries as evident with words such as "drug","treatment", "patient" and so on. In NMF, we can see that in topic#0 there are many names clustered into the same category. These type of subjective headlines are very common in the articles during pandemic and outbreak.

5. Conclusions and Future Directions

Topic Modelling is an evolving area in natural language processing and deep learning. With this, it has become helpful to understand the underlying semantic structure of documents and article and classify them accordingly. Using LDA and NMF methodology , we can apply topic modelling to set of text to correctly classify them based on their underlying structure. LDA performed better and allowed us to better learn the relationships among words, topics and article. It provided us the clear picture to visualize which are the areas impacted by COVID-19 outbreak. Through this research, we have presented how Covid-19 pandemic caused financial fragility on businesses and other industries like technology, automobile and general industries like child day care, schools, colleges etc. In our research-based study, this has not only impacted small businesses but also international relationship and trade deals among two countries for instance, oil supply, energy, production of merchandise were also severely impacted. None the less, restaurants and hospitality industries were at the peak that got impacted by Covid-19 pandemic. Corresponding regions, cities and states with

the exposure to these industries were impacted as well and turned into bankruptcy. The proposed topic modelling identified the topics associated with the articles which in given to the classifier and generates the appropriate topics for the articles. The LDA model with number of topics as 20 achieved better performance and identified relevant topics. Also, the results provided by the system are easy to understand and infer and can be helpful in further strategic decision-making process. This approach can help government and policy makers to take decisions based on the current scenario. Different Governments are trying to stimulate the economy across the world and the findings from the paper can be used to bring back the economy on track. Thus, for any end user application that involves human interaction or that carries human intelligence, flexibility and coherence advantage of LDA warrants strong consideration. The excellence of LDA and NMF in topic modelling presents exciting research directions. Topic modelling is not an easy task as it requires lot of domain expertise and good knowledge about the underlying algorithm it is using. LDA is difficult to train due to the time-consuming calculations and its results need human interpretation. Our research shows that the words of the learned topic are not properly similar however they are relevant to the topic predicted.

Future directions of this research consist of optimizing hyperparameter using Grid Search. Also, finding dominant topic and determining the best number of topics for best modelling performance can be best future efforts to deep dive into topic modelling area.

Author Contributions: Conceptualization, A.A. and S.B.; methodology, A.A. and S.B.; software, S.B.; validation, A.A. and S.B.; formal analysis, A.A.; investigation, S.B.; data curation, S.B.; writing—original draft preparation, A.A.; writing—review and editing, A.A. and S.B.; visualization, S.B.; supervision, S.B.; project administration, A.A. All authors have read and agreed to the published version of the manuscript."

Funding: "This research received no external funding"

Acknowledgments: In this section you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Conflicts of Interest: "The authors declare no conflict of interest."

References

1. Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X.; Jiang, X.; Li, Y.; & Zhao, L. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications* **2018**, 78, 15169–15211, doi:10.1007/s11042-018-6894-4
2. Toubia, O.; Iyengar, G.; Bunnell, R.; & Lemaire, A. Extracting Features of Entertainment Products: A Guided Latent Dirichlet Allocation Approach Informed by the Psychology of Media Consumption. *Journal of Marketing Research* **2018**, 56, 18–36, doi:10.1177/0022243718820559
3. Feuerriegel, S.; & Pröllochs, N. Investor Reaction to Financial Disclosures across Topics: An Application of Latent Dirichlet Allocation: Investor Reaction to Financial Disclosures across Topics. *Decision Sciences* **2018**, doi:10.1111/deci.12346
4. Xu, Z.; Liu, Y.; Xuan, J.; Chen, H.; & Mei, L. Crowdsourcing based social media data analysis of urban emergency events. *Multimedia Tools and Applications* **2017**, 76, 11567–11584, doi:10.1007/s11042-015-2731-1
5. Chew, C.; & Eysenbach, G. Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. *PloS One* **2010**, 5, e14118–e14118, doi:10.1371/journal.pone.0014118
6. Zhao, W.; Zhang, G.; Yuan, G.; Liu, J.; Shan, H.; & Zhang, S. The Study on the Text Classification for Financial News Based on Partial Information. *IEEE Access* **2020**, 8, 100426–100437, doi:10.1109/ACCESS.2020.2997969
7. El-Haj, M.; Rayson, P.; Walker, M.; Young, S.; & Simaki, V. In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse. *Journal of Business Finance & Accounting* **2019**, 46, 265–306, doi:10.1111/jbfa.12378
8. Batra, R.; & Daudpota, S. Integrating Stock Tweets with sentiment analysis for better prediction of stock price movement. **2018**, 1–5, doi:10.1109/ICOMET.2018.8346382

9. Li, Q.; Deleger, L.; Lingren, T.; Zhai, H.; Kaiser, M.; Stoutenborough, L.; Jegga, A.; Cohen, K.; & Solti, I. Mining FDA drug labels for medical conditions. *BMC Medical Informatics and Decision Making* **2013**, 13, 53–53, doi:10.1186/1472-6947-13-53
10. Frost & Sullivan Honors Linguamatics for Developing a Best-in-Class NLP-based Data Mining Platform for the Healthcare Industry: I2E makes natural language processing-based text mining intuitive and interactive. (2017, July 20). PR Newswire.
11. Handoyo, E.; Arfan, M.; Soetrisno, Y.; Somantri, M.; Sofwan, A.; & Sinuraya, E. Ticketing Chatbot Service using Serverless NLP Technology, **2018**, 325–330, doi:10.1109/ICITACEE.2018.8576921
12. Greene, D.; & Cross, J. Unveiling the Political Agenda of the European Parliament Plenary: A Topical Analysis. **2010**,1–10, doi:10.1145/2786451.2786464
13. Fatemi, M.; & Safayani, M. Joint sentiment/topic modeling on text data using a boosted restricted Boltzmann Machine. *Multimedia Tools and Applications* **2019**,78, 20637–20653, doi:10.1007/s11042-019-7427-5
14. Fang, Y.; Si, L.; Somasundaram, N.; & Yu, Z. Mining contrastive opinions on political texts using cross-perspective topic model, **2012**, 63–72, doi:10.1145/2124295.2124306
15. Thomas, S.; Adams, B.; Hassan, A.; & Blostein, D. Modeling the evolution of topics in source code histories, **2011**, 173–182, doi:10.1145/1985441.1985467
16. Tran, B.; Nghiem, S.; Sahin, O.; Vu, T.; Ha, G.; Vu, G.; Pham, H.; Do, H.; Latkin, C.; Tam, W.; Ho, C.; & Ho, R. Modeling Research Topics for Artificial Intelligence Applications in Medicine: Latent Dirichlet Allocation Application Study. *Journal of Medical Internet Research* **2019**, 21, e15511–e15511, doi:10.2196/15511
17. Kuang, D.; and Park, H. Fast rank-2 nonnegative matrix factorization for hierarchical document clustering. In Proc. the 19th ACM *International Conference on Knowledge Discovery and Data Mining* **2013**, 739–747.
18. Kim, J.; He, Y.; and Park, H. Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *Journal of Global Optimization* **2013**.
19. Gonzales, E.F.; and Zhang, Y. Accelerating the Lee-Seung algorithm for non-negative matrix factorization. Technical Report.
20. Ray, S.; & Bandyopadhyay, S. An NMF based approach for integrating multiple data sources to predict HIV-1–human PPIs. *BMC Bioinformatics* **2016**, 17, 121–121, doi:10.1186/s12859-016-0952-6
21. Hee, S. Exploring the Initial Impact of COVID-19 Sentiment on US Stock Market Using Big Data. *Sustainability* **2020**, 12, 6648, doi:10.3390/su12166648
22. Bollen, J.; Mao, H.; & Zeng, X. Twitter mood predicts the stock market. **2010**, doi:10.1016/j.jocs.2010.12.007
23. Bharathi, S.; Geetha, A.; & Sathyanarayana, R. Sentiment Analysis of Twitter and RSS News Feeds and Its Impact on Stock Market Prediction. *International Journal of Intelligent Engineering & Systems* **2017**, 10, 68.
24. Pereira, D.; Junior, N.; and Caloba, L. "Financial Time Series Forecasting Using Non-Linear Methods and Stacked Autoencoders," *2018 International Joint Conference on Neural Networks (IJCNN*) **2018**, Rio de Janeiro,1-8, doi: 10.1109/IJCNN.2018.8489425.
25. Chamola, V.; Hassija, V.; Gupta, V.; and Guizani, M. A Comprehensive Review of the COVID-19 Pandemic and the Role of IoT, Drones, AI, Blockchain, and 5G in Managing its Impact, *IEEE Access* **2020**, vol. 8, 90225-90265, doi: 10.1109/ACCESS.2020.2992341.
26. Hegde, A.; & Masthi, R. Digital Contact tracing in the COVID-19 Pandemic: A tool far from reality. *Digital Health* **2020**, 6, 205520762094619–2055207620946193, doi:10.1177/2055207620946193
27. Zuber, S.; & Brüssow, H. COVID 19: challenges for virologists in the food industry. *Microbial Biotechnology* **2020**, 13, 1689–1701, doi:10.1111/1751-7915.13638
28. Lan, F.; Suharlim, C.; Kales, S.N.; and Yang, J. Association between SARS-CoV-2 infection, exposure risk and mental health among a cohort of essential retail workers in the United States.medRxiv,2020.2006.2008.20125120.**2020**,URL:https://www.medrxiv.org/content/medrxiv/early/2020/06/09/2020.06.08.20125120.full.pdf
29. Sethi, B.; Sethi, A.; Ali, S.; & Aamir, H. Impact of Coronavirus disease (COVID-19) pandemic on health professionals. *Pakistan Journal of Medical Sciences* **2020**, 36, S6–S11, doi:10.12669/pjms.36.COVID19-S4.2779
30. Dauner, C.; Perlman, J.; & Dougherty, T. 5 ways hospitals should prepare to access COVID-19 disaster funding. *Healthcare Financial Management* **2020**, 74, 36–39.
31. Lutz, H. GM freshenings pushed back by pandemic. *Automotive News* **2020**, 94, 6.

32. Mead, D.; Ransom, R.; Reed, S.; and Sager, S. The impact of the COVID19 pandemic on food price indexes and data collection," Monthly Labor Review, *U.S. Bureau of Labor Statistics*, August **2020**, doi:10.21916/mlr.2020.18.
33. Lipenkova, J. Covid-19 Public Media Dataset by Anacode. Available online: https://www.kaggle.com/jannalipenkova/covid19-public-media-dataset (accessed on 10 October 2020).
34. Asmussen, C.B.; Møller, C. Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data* **2019**, 6, 93, doi:10.1186/s40537-019-0255-7
35. Rana, T.; Cheah, Y.; & Letchmunan, S. Topic Modeling in Sentiment Analysis: A Systematic Review. *Journal of ICT Research and Applications* **2016**, 10, 76–93, Doi:10.5614/itbj.ict.res.appl.2016.10.1.6
36. Suri, P.; & Roy, N. Comparison between LDA & NMF for event-detection from large text stream data. **2017,** 1–5, doi:10.1109/CIACT.2017.7977281
37. Wang, Y.; Zhao, X.; Sun, Z.; Yan, H.; Wang, L.; Jin, Z.; Wang, L.; Gao, Y.; Law, C.; & Zeng, J. Peacock: Learning Long-Tail Topic Features for Industrial Applications. *ACM Transactions on Intelligent Systems and Technology* **2015**, 6, 1–23, doi:10.1145/2700497
38. Kalepalli, Y.; Tasneem, S.; Phani , P.; & Manne, S. Effective Comparison of LDA with LSA for Topic Modelling. **2020**, 1245–1250, doi:10.1109/ICICCS48265.2020.9120888
39. George, L.; & Birla, L. A Study of Topic Modeling Methods. **2018**, 109–113, doi:10.1109/ICCONS.2018.8663152
40. Barde, B.; & Bainwad, A. An overview of topic modeling methods and tools. **2017,** 745–750, doi:10.1109/iccons.2017.8250563
41. Korshunov, A.; & Gomzin, A. Topic modeling in natural language texts. Proceedings of the *Institute for System Programming of RAS* **2012**, 23, 215–244,doi:10.15514/ISPRAS-2012-23-13
42. Sapul, M.; Htike, A.; & Jiamthapthaksin, R. Trending topic discovery of Twitter Tweets using clustering and topic modeling algorithms. **2017**, 1–6, doi:10.1109/JCSSE.2017.8025911
43. Alhawarat, M.; & Hegazi, M. Revisiting K-Means and Topic Modeling, a Comparison Study to Cluster Arabic Documents. IEEE Access **2018**, 6, 42740–42749,doi:10.1109/access.2018.2852648
44. Rashid, J.; Shah, S.; Irtaza, A.; Mahmood, T.; Nisar, M.; Shafiq, M.; & Gardezi, A. Topic Modeling Technique for Text Mining Over Biomedical Text Corpora Through Hybrid Inverse Documents Frequency and Fuzzy K-Means Clustering. *IEEE Access* **2019**, 7, 146070–146080,doi:10.1109/access.2019.2944973
45. Gao, Y.; Xu, Y.; & Li, Y. Pattern-based Topics for Document Modelling in Information Filtering. *IEEE Transactions on Knowledge and Data Engineering* **2015**, 27, 1629–1642, doi:10.1109/TKDE.2014.2384497
46. Alostad, J. Reducing Dimensionality Using NMF Based Cholesky Decomposition, **2017**, 49–55, doi:10.1145/3129676.3129697