ISSN: 2366-1313

Document Clustering and Topic Classification Using LDA

[1]Mr.P.V.Ram Gopal Rao, [2]Pitta Shivani, [3]Sairam Shettigari, [4]Puliga SaiKumar, [5]Sai Vamshi

[1]Assistant Professor, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,

ramgopal.cse@tkrec.ac.in

[2]BTech student, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,
shivanireddy821@gmail.com@gmail.com

[3]BTech student, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,
sairamshettigari123@gmail.com

[4]BTech student, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,
lakshmisai91544@gmail.com

[4]BTech student, Dept.of CSE, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad,
saivamshi35122@gmail.com

Abstract: *With the world becoming more data-driven, enormous amounts of data are being added to databases, documents, libraries, and papers. The whole data would be useless if it is not utilized properly to get useful information out of it. And this useful information is never directly available to us as the data is mostly unstructured, confusing, and open-ended. The data can be well utilized if it is properly classified and could be categorized to have separate studies on each category of data. In the era of technology, computer science, and mathematics is continuously working to resolve problems like these. With a huge amount of data being available it is not an easy task to analyze the data. A number of studies have been conducted to extract features from the data. Text Mining is a sub-field of machine learning that deals with unauthorized and unorganized text to get meaningful data that can be analyzed and used. It is intensely being used to analyze unstructured text, get structured features from the text and find patterns. So, unsupervised learning is aiding the supervised algorithms to work on such kinds of text. One of the growing text mining algorithms is topic modelling, which distributes documents into topics and topics into words. This algorithm is working better than other algorithms like tf-idf as it and considers the actual scene where the topic can be related to multiple words, so LDA is working closer to the real-world topics and their distributions. Latent Dirichlet Allocation, sometimes known as LDA, is one of the most popular topic modelling algorithms. More importantly, it provides straightforward and understandable subjects that are similar to what the human mind assigns when reading a*

ISSN: 2366-1313

document. It is scalable, computationally quick, and scalable. LDA can be used as a feature extraction technique for supervised tasks like text classification, even though its primary applications are unsupervised tasks like topic modelling and document clustering.

***Keywords**: Document Clustering, Latent Dirichlet Allocation, Text mining, Machine learning.*

I. INTRODUCTION

In the text domain, document clustering (Aggarwal and Zhai, 2012; Cai et al., 2011; Lu et al., 2011; Ng et al., 2002; Xu and Gong, 2004; Xu et al., 2003) and topic modeling (Blei et al., 2003; Hofmann, 2001) are two widely studied problems which have many applications. Document clustering aims to organize similar documents into groups, which is crucial for document organization, browsing, summarization, classification and retrieval. Topic modeling develops probabilistic generative models to discover the latent semantics embedded in document collection and has demonstrated vast success in modeling and analyzing texts.

Document clustering and topic modeling are highly correlated and can mutually benefit each other. On one hand, topic models can discover the latent semantics embedded in document corpus and the semantic information can be much more useful to identify document groups than raw term features. In classic document clustering approaches, documents are usually represented with a bag-of-words (BOW) model which is purely based on raw terms and is insufficient to capture all semantics. Topic models are able to put words with similar semantics into the same group called topic where synonymous words are treated as the same. Under topic models, document corpus is projected into a topic space which reduces the noise of similarity measure and the grouping structure of the corpus can be identified more effectively. On the other hand, document clustering can facilitate topic modeling. Specifically, document clustering enables us to extract local topics specific to each document cluster and global topics shared across clusters. In a collection, documents usually belong to several groups. For instance, in scientific paper archive such as Google Scholar, papers are from multiple disciplines, such as math, biology, computer science, economics. Each group has its own set of topics. For instance, computer science papers cover topics like operating system, network, machine learning while economics papers contain topics like entrepreneurial economics, financial economics, mathematical economics. Besides group-specific topics, a common set of global topics are shared by all

ISSN: 2366-1313

groups. In paper archive, papers from all groups share topics like reviewing related work, reporting experimental results and acknowledging financial supports. Clustering can help us to identify the latent groups in a document collection and subsequently we can identify local topics specific to each group and global topics shared by all groups by exploiting the grouping structure of documents. These fine-grained topics can facilitate a lot of utilities. For instance, we can use the group-specific local topics to summarize and browser a group of documents. Global topics can be used to remove background words and describe the general contents of the whole collection. Standard topic models (Blei et al., 2003; Hofmann, 2001) lack the mechanism to model the grouping behavior among documents, thereby they can only extract a single set of flat topics where local topics and global topics are mixed and cannot be distinguished.

Naively, we can perform these two tasks separately. To make topic modeling facilitates clustering, we can first use topic models to project documents into a topic space, then perform clustering algorithms such as K-means in the topic space to obtain clusters. To make clustering promotes topic modelling, we can first obtain clusters using standard clustering algorithms, then build topic models to extract cluster-specific local topics and cluster-independent global topics by incorporating cluster labels into model design. However, this naive strategy ignores the fact that document clustering and topic modelling are highly correlated and follow a chicken-and-egg relationship. Better clustering results produce better topic models and better topic models in turn contribute to better clustering results. Performing them separately fails to make them mutually promote each other to achieve the overall best performance. In this paper, we propose a generative model which integrates document clustering and topic modelling together. Given a corpus, we assume there exist several latent groups and each document belongs to one latent group. Each group possesses a set of local topics that capture the specific semantics of documents in this group and a Dirichlet prior expressing preferences over local topics. Besides, we assume there exist a set of global topics shared by all groups to capture the common semantics of the whole collection and a common Dirichlet prior governing the sampling of proportion vectors over global topics for all documents. Each document is a mixture of local topics and global topics. Words in a document can be either generated from a global topic or a local topic of the group to which the document belongs. In our model, the latent variables of cluster membership,

ISSN: 2366-1313

document-topic distribution and topics are jointly inferred. Clustering and modelling are seamlessly coupled and mutually promoted.

II. LITERATURE SURVEY

Indonesia is a developing country and supports the program of Sustainable Development Goals (SDGs) which consist of 17 goals. SDGs are not only the government's duty but a shared duty from any elements. Online media has a crucial role in implementing the goals of Indonesia ' s SDG. Latent Dirichlet allocation (LDA) is one of the methods of topic modelling to find out the trend of topics of SDGs news. Development is a continual process that occurs in various dimensions,

including economic, social, and environmental dimensions. The goal is to promote the welfare of the community. Even though the development is not yet perfect, it forfeits a lot of natural resource exploitation that is carried out arbitrarily, without paying attention to aspects of the environment. As a result, damage to the environment which can disrupt life occurs more frequently.

The greatly accelerated development of information technology has conveniently provided adoption for risk stratification, which means more benefits for both patients and clinicians. Risk stratification offers accurate individualized prevention and therapeutic decision making etc. Hospital discharge records (HDRs) routinely include accurate conclusions of diagnoses of the patients. For this reason, in this paper, we propose an improved model for risk stratification in a supervised fashion by exploring HDRs in coronary heart disease (CHD). Cerebrovascular accidents (CVA), coronary heart disease (CHD) and other cardiovascular diseases (CVD) are the leading causes of death and serious family burden in China nowadays. According to the World Health Organization (WHO), risk factors can increase the chances that a person suffers from a disease (WHO, 2014). Risk stratification incorporating these risk factors can be used by physicians to assess the risk of atherosclerosis in the individua patient, such as taking treatment with drugs to lower blood pressure and blood cholesterol based on an individual ' s absolute cardiovascular risk.

Feature extraction is one of the challenging works in the Machine Learning (ML) arena. The more features one is able to extract correctly, the more accurate knowledge one can exploit from data. Latent Dirichlet Allocation (LDA) is a

ISSN: 2366-1313

form of topic modeling used to extract features from text data. But finding the optimal number of topics (on which the success of LDA depends) is tremendously challenging.

In-text mining and feature extraction algorithms provide a base for supervised, unsupervised or reinforcement learning, etc. For measuring distances among text documents, the Information Retrieval (IR) system needs to exploit semantic information. LDA is of great use in the scenario where exploiting semantic information is the first priority. But according to, the efficacy of LDA largely depends on the value of a vital parameter named "number of topics", which actually requires prior knowledge about the contents of the dataset

Topic models, such as Latent Dirichlet Allocation (LDA), allow us to categorize each document based on topics. It builds a document as a mixture of topics and a topic is modelled as a probability distribution over words. However, the key drawback of the traditional topic model is that it cannot handle the semantic knowledge hidden in the documents. Therefore, semantically related, coherent and meaningful topics cannot be obtained. However, semantic inference plays a significant role in topic modelling as well as in other text-mining tasks. In this paper, in order to tackle this problem, a novel NET-LDA model is proposed. In NET-LDA, semantically similar documents are merged to bring all semantically related words together and the obtained semantic similarity knowledge is incorporated into 10 the model with a new adaptive semantic parameter.

Introduction: Nowadays, customers share what they like and what is not and what is in their mind by clicking on a website from where they are. These online review websites have revolutionized daily economic life by changing the way of companies by which they can boost their sales and marketing. Based on the survey conducted by BrightLocal in 2017, 85% of customers trust online reviews as much as personal recommendations. On this basis, the huge amount of electronic word-of-mouth data provides a goldmine to the data scientists for the evaluation of others' opinions about what they have purchased or benefited from. Scientists mine web media for sentiment analysis and evaluate customers' choices and can also help others in the decision-making process.

Children are the future of the nation. All treatment and learning they get would affect their future. Nowadays, there are various kinds of social problems related to children. To ensure the right solution to their problem, social workers usually refer

ISSN: 2366-1313

to the social-child-case (SCC) documents to find similar cases in the past and adapt the solution of the cases. Nevertheless, reading a bunch of documents to find similar cases is a tedious task and needs much time. Hence, this work aims to categorize those documents into several groups according to the case type. We use topic modeling with Latent Dirichlet Allocation (LDA) approach to extract topics from the documents and classify them based on their similarities. The Coherence Score and Perplexity graph are used in determining the best model. The result obtains a model with 5 topics that match the targeted case types.

III. SYSTEM ANALYSIS

The experimental model diagram illustrated as in Figure represents the steps of analysis in this research study from the input (Raw bible data) followed by pre-processing to remove the noise in the data and further removal of stop words and to find the root word of the given word by Lemmatization process to further assess the text to vector conversion and comparison of two topic modelling methods (LSA and LDA) in identifying the document similarity within-corpus and with the unseen document to categorize the word associations and coherence score as a measure for topic comparison and goodness of the topic model.

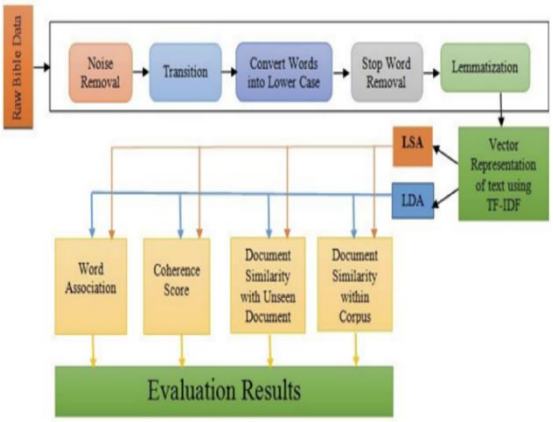

Fig.1 Comparative analysis of LSA and LDA

Unlike LSA, LDA does not directly output document similarities. Instead, LDA outputs a matrix, whose rows represent all the documents in the dataset, and columns represent all the topics. Each value represents a particular topic's weight in a document. The user specifies the total number of topics that the words are sorted into, and columns in the matrix range between 0 and the user-defined number of topics. LDA was run with different numbers of topics until a good topic range was found for the dataset. The final step is to compare the document similarity matrices output by LDA and LSA. If only minor differences can be found between them, it can be inferred that LSA and LDA are more or less equal in their ability to sort the mechanics of the data. Nevertheless, if the two results differ significantly, the more efficient algorithm is determined by comparing one document with the other documents.

ISSN: 2366-1313

IV. PROPOSED METHODOLOGY

TOPIC MODELING

Topic Modelling in NLP seeks to find hidden semantic structure in documents. They are probabilistic models that can help you comb through massive amounts of raw text and cluster similar groups of documents together in an unsupervised way.

This post specifically focuses on Latent Dirichlet Allocation (LDA), which was a technique proposed in 2000 for population genetics and re-discovered independently by ML-hero Andrew Ng et al. in 2003. LDA states that each document in a corpus is a combination of a fixed number of topics. A topic has a probability of generating various words, where the words are all the observed words in the corpus. These 'hidden' topics are then surfaced based on the likelihood of word co-occurrence. Formally, this is Bayesian Inference problem

.

SYSTEM DESIGN

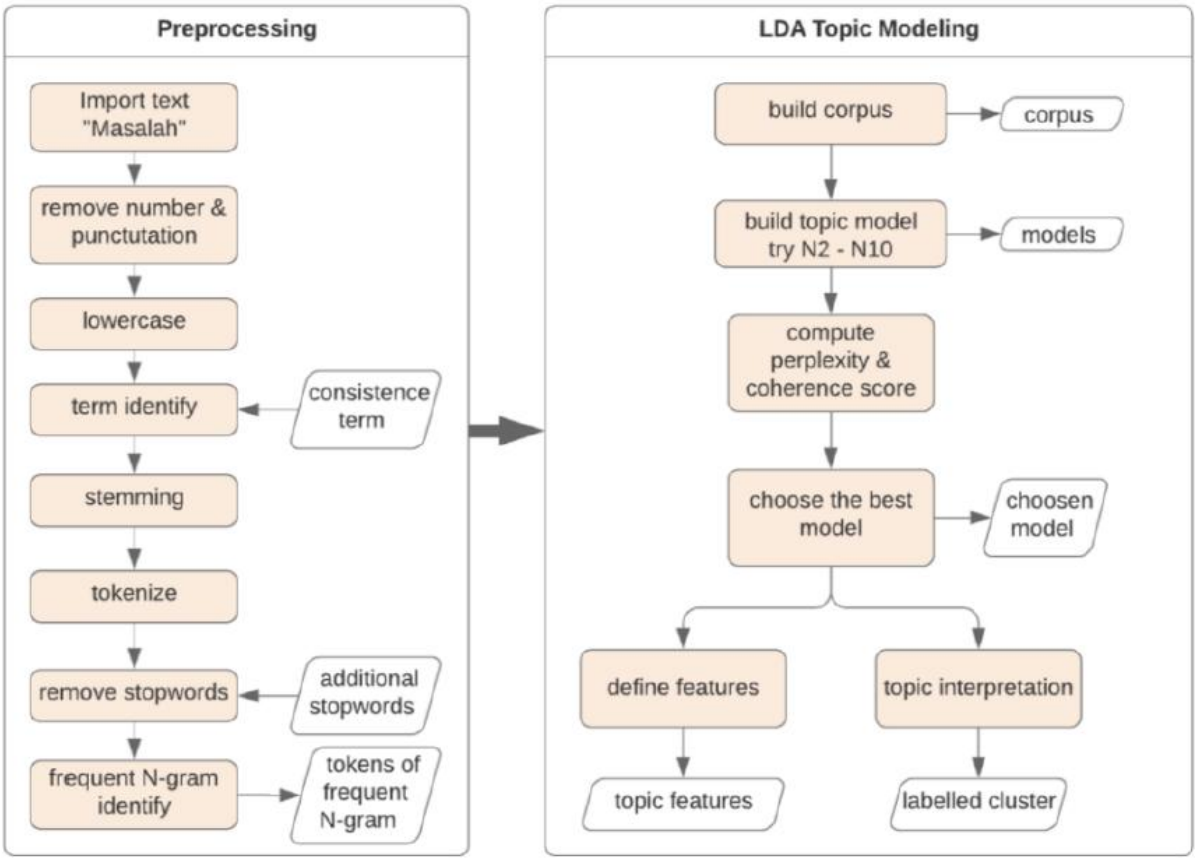

Fig.2 Experiment Design

ISSN: 2366-1313

The experiment design was divided into two significant parts. The first part is data pre-processing, and the second part is the topic modelling of pre-processed data.

Data Pre-processing: According to and preprocessing is considered an essential step in text processing, as it provides a standard and consistent form of data that affecting the whole experiment result. As shown in Figure, we use a series of preprocessing steps consist of 1) number and punctuation removal; 2) case-folding (turn all the text into lowercase); 3) term identification; 4) stemming; 5) stopword removal and 6) frequent n-gram identify. Among the preprocessing, there are default steps (data cleaning and case-folding), the additional step (term identification), and the optional steps (stemming, stopword removal, and frequent n-gram identify).

The term identification usage is performed to handle too many words found in the text, which have the same meaning. So, we consider them as SCC terms with inconsistent writing, as shown in Table 1. Those inconsistencies are the results of differences in abbreviations, the use of words, letters, and space. Inconsistent writing can lead to errors in recognizing a term or word. which affects the whole experiment.

Table.1 Inconsistence term

ISSN: 2366-1313

NO	Inconsistence Term	FixTerm
1	satuan bhakti, satuan bakti	sakti
2	petugas, pekerja sosial, satuan peksos, sakti peksos	peksos
3	satuan polisi pamong praja, polisi pamong praja	satpol pp
4	dinas sosial	dissos
5	rumah perlindungan sosial	rps
6	yayasan sayap ibu, sayap ibu, yayasan ysi	ysi
7	assessment, assesment, asessment, asesment, assessmen, assesmen, asessmen	asesmen
8	camp asesmen	camp
9	balai rehabilitasi sosial dan pengasuhan anak, balai rspa	brspa
10	balai perlindungan dan rehabilitasi sosial remaja, balai prsr	bprsr
11	balai perlindungan dan rehabilitasi sosial wanita, balai prsw	bprsw
12	balai rehabilitasi terpadu penyandang disabilitas, balai rtpd	brtpd
13	balai rehabilitasis sosial bina karya dan laras, bina karya, bina laras, balai rsbkl	brsbkl
14	case conference	cc
15	orang tua	orangtua
16	rumah sakit	rs
17	sekolah dasar	sd
18	taman kanak-kanak	tk
19	tracing	penelusuran
20	activity daily living	adl
21	penyerahan kembali, diserahkan kembali, menyerahkan kembali	reunifikasi
22	support	dukungan
23	family	keluarga

The next steps that are quite significant contain stemming and stopword removal. Actually, in and, Schofield found that in some instances, the use of stemming and stopword removal does not affect the topic model. Since it can even reduce its stability, the use of those steps requires some consideration. In this study, we try experimenting using both steps with the consideration that the text addresses a fairly specific domain. For example, without stemming, some words with the same context (e.g. ―pengasuhan‖, ―diasuh‖, ―mengasuh‖ and ―asuhan‖) are recognized as different tokens. And without some additional stopword, some words which in this context are general terms (e.g. ―klien‖, ―kondisi‖ and ―mengalami") frequently appear in many documents, even though they do not mean anything.

The frequent n-grams identification is made based on the assumption that a series of words frequently appear in the documents, have specific meanings that can become the text features,

as done in. As shown in Table 2, the results written in italics are named-entity. We get them by experimentally re-run this step while changing the threshold and min_count values. The higher the threshold values, the lesser the n-gram produced. Min_count is a minimum

ISSN: 2366-1313

number of n-word's occurrences in sequence. In this experiment, we use bigram and trigram, where both use threshold = 75 and min_count = 3

Table.2 N-gram

NO	WORDS	NO	WORD	NO	WORD	NO	WORD
1	dissos_diy	11	kena_razia_satpol_pp	21	media_sosial	31	senjata_tajam
2	satpol_pp	12	keras_fisik	22	habis_uang	32	retardasi_mental
3	lingkung_sosial	13	uang_saku	23	lampu_merah	33	obat_larang
4	*psbk_bekasi*	14	ojek_online	24	kelompok_punk		
5	rehabilitasi_sosial	15	interaksi_sosial	25	pondok_pesantren		
6	kelas_sd	16	penuh_butuh_hidup	26	biaya_salin		
7	kelas_smp	17	tonton_konser_musik	27	perangkat_desa		
8	ganggu_jiwa	18	tumpang_kendara	28	*gotong_royong*		
9	habis_bekal	19	konsumsi_obat	29	*pondok_sadar*		
10	kena_jangkau_satpol_pp	20	mantan_suami	30	penyalahgunaan_napza		

LDA Topic Modeling:

The final process output from the preprocessing becomes a corpus consisting of n-gram tokens. From this corpus, the topic model is built using the LDA algorithm. As a generative probabilistic model of the corpus, LDA assumes that each document is represented as a probabilistic distribution over latent topics, and each topic is characterized by a distribution over words. M represents the number of documents, while N represents the number of words in the document. The first level is the corpus level parameter (α and ß), which considered as samples in the corpus production process.

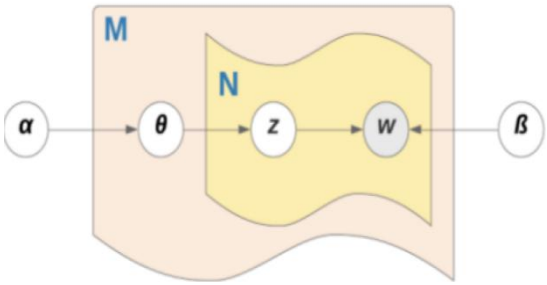

Fig.3 Graphical model representation of LDA

We first modeled the topic from the corpus and replicated this process several times, as was done in. In generating a model, the LDA algorithm requires an input parameter (n) to determine the number of generated topics. Because there was no absolute knowledge about the topic number of SCC documents, we determined the value of n based on the expert‘s (social worker) assumption on the range of the topic‘s number. Based on the experts' assumption, we experimented using n = 2 to n = 10.

Determination of the best model (n topic) was carried out with two measurements, which are the Perplexity value and the Coherence score. The value of perplexity showed the confusion metrics or ways to capture the level of 'uncertainty' of a model's prediction result. In contrast, the coherence score indicated the level of

ISSN: 2366-1313

semantic similarity between words on a topic.

IMPLEMENTATION OF MODULES

Scikit-learn

Defining scikit learn, it is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. Scikit-learn was initially developed by David Cournapeau as a Google summer of code project in 2007.Later Matthieu Brucher joined the project and started to use it as a part of his thesis work. In 2010 INRIA got involved and the first public release (v0.1 beta) was published in late January 2010.The project now has more than 30 active contributors and has had paid sponsorship from INRIA, Google, Tinyclues and the Python Software Foundation. In general, a learning problem considers a set of n samples of data and then tries to predict properties of unknown data. If each sample is more than a single number and, for instance, a multi-dimensional entry (aka multivariate data), it is said to have several attributes or features. Learning problems fall into a few categories:

supervised learning, in which the data comes with additional attributes that we want to predict (Click here to go to the scikit-learn supervised learning page).This problem can be either:

classification: samples belong to two or more classes and we want to learn from already labeled data how to predict the class of unlabeled data. An example of a classification problem would be handwritten digit recognition, in which the aim is to assign each input vector to one of a finite number of discrete categories. Another way to think of classification is as a discrete (as opposed to continuous) form of supervised learning where one has a limited number of categories and for each of the n samples provided, one is to try to label them with the correct category or class.

regression: if the desired output consists of one or more continuous variables, then the task is called regression. An example of a regression problem would be the prediction of the length of a salmon as a function of its age and weight.

unsupervised learning, in which the training data consists of a set of input vectors x without any corresponding target values. The goal in such problems may be

ISSN: 2366-1313

to discover groups of similar examples within the data, where it is called clustering, or to determine the distribution of data within the input space, known as density estimation, or to project the data from a high-dimensional space down to two or three dimensions for the purpose of visualization (Click here to go to the Scikit-Learn unsupervised learning page)

.

V. RESULTS

```
# Let's have a look at the class balance.
sns.countplot(data.Sentiment_Type)
plt.xlabel('review score')
```

```
Text(0.5, 0, 'review score')
```

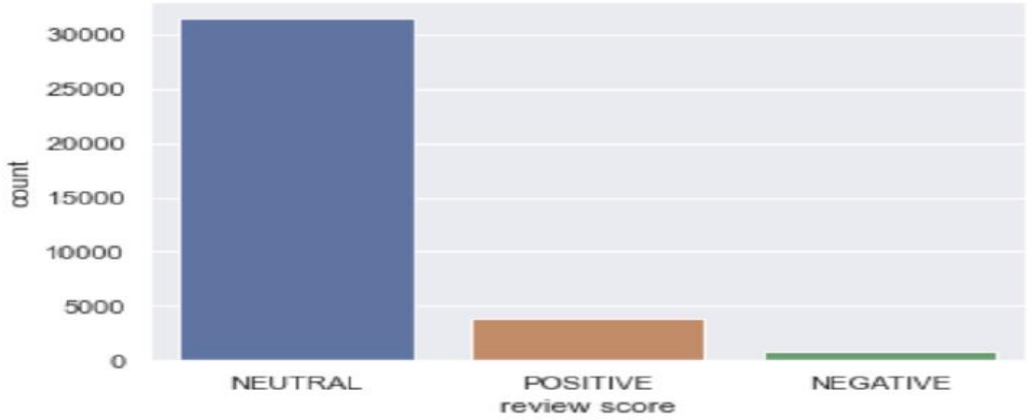

Fig.4 Review score

```
# Plot the distribution
class_names = ['negative', 'neutral', 'positive']
ax = sns.countplot(data.Sentiment_Type)
plt.xlabel('review sentiment')
ax.set_xticklabels(class_names)
```

```
[Text(0, 0, 'negative'), Text(1, 0, 'neutral'), Text(2, 0, 'positive')]
```

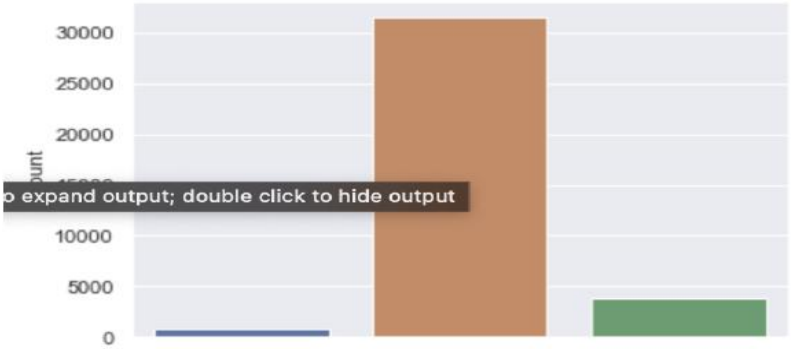

Fig.5 Review Sentiment Type

ISSN: 2366-1313

```
plt.figure(figsize=(20,10))
sns.countplot(x = 'all_hashtags',
              data = all_hashtags_df,
              order = all_hashtags_df['all_hashtags'].value_counts().iloc[:15].index)

plt.xlabel('Hashtags')
plt.ylabel('Count')
```

```
Text(0, 0.5, 'Count')
```

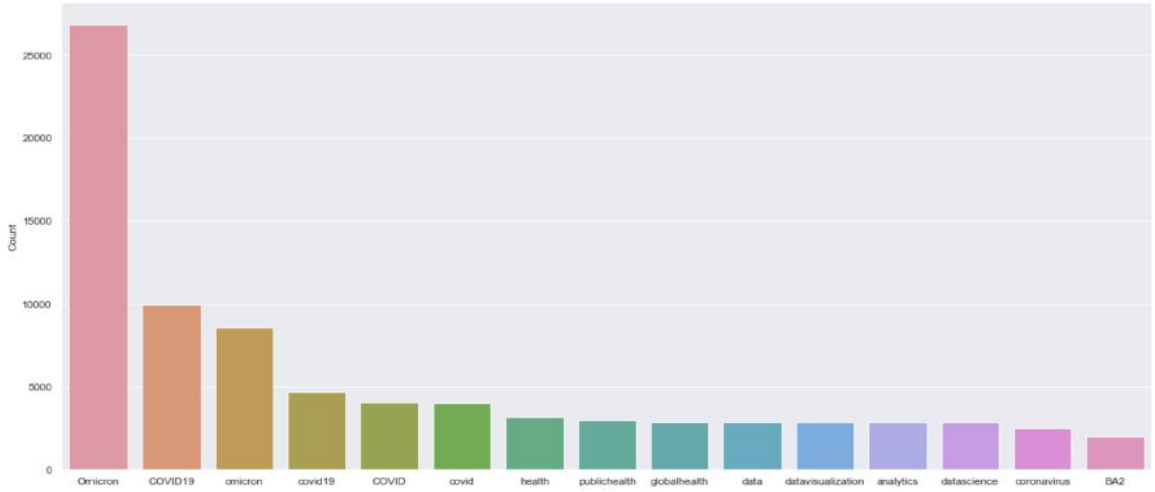

Fig.6 All Hashtags Count

```
# Word Cloud is a data visualization technique used for representing text data
# in which the size of each word indicates its frequency or importance.

wordcloud = WordCloud(stopwords=stopwords.words('english'),max_words=50).generate(str(data.text))
plt.figure(figsize=(10,6))
plt.imshow(wordcloud)
plt.axis('off')
plt.show()
```

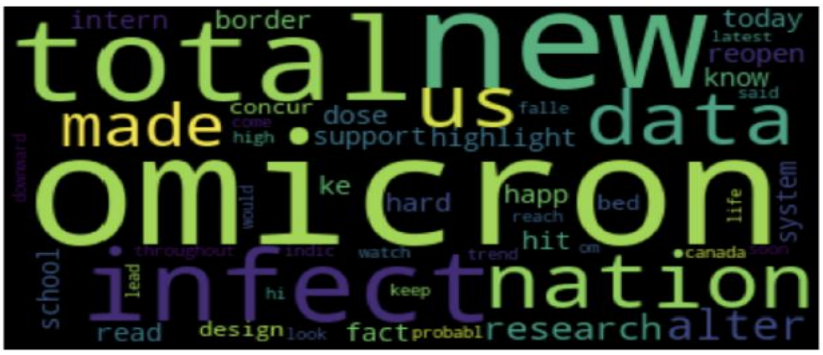

Fig.7 Word Frequency

ISSN: 2366-1313

```
Document 0:
Topic  0 :  3.290148329215143 %
Topic  1 :  3.2901306854810763 %
Topic  2 :  70.38740652114046 %
Topic  3 :  3.2902108671922248 %
Topic  4 :  3.2903727772130758 %
Topic  5 :  3.290137415748867 %
Topic  6 :  3.290695222330146 %
Topic  7 :  3.2903475710533274 %
Topic  8 :  3.290314392786178 %
Topic  9 :  3.2902362178395013 %
Text(0, 0.5, '% in Document')
```

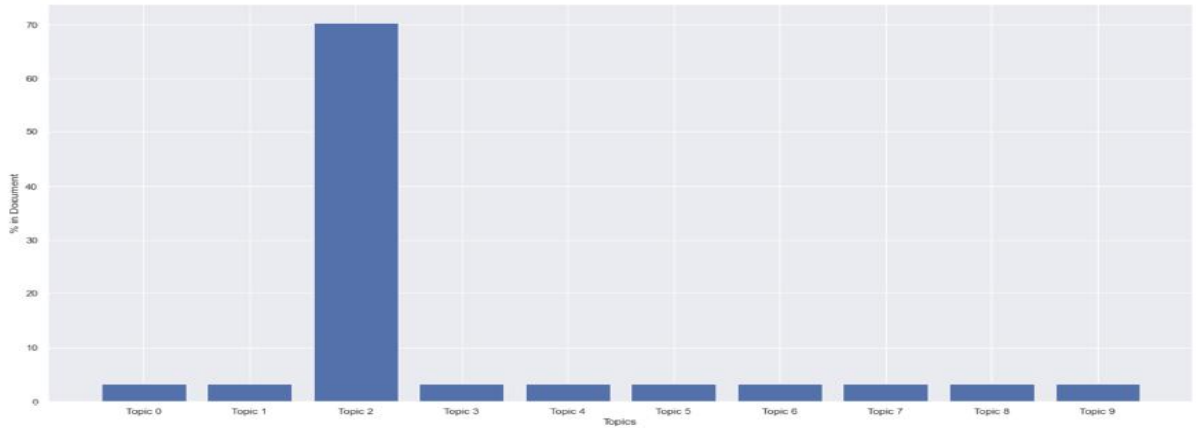

Fig.8 Data Visualization Of Document

Fig.9 Word Frequency of Paragraph

VI. CONCLUSION

Once LDA topic modeling is applied to a set of documents, you're able to see the words that make up each hidden topic. In my case, I took 40,000 rows of texts, hashtags and sentiment types from twitter during the period of pandemic. Based on the results of topic modeling using 10 topics, topic 2 had the highest coherence score of 0.5405. And all the other topics that had least scores. In the bar graph, the document is divided into topics and in the second figure, all the 10 topics inside the document are divided into individual

ISSN: 2366-1313

topics and in each topic based on their priority the word is highlighted. The modeling topic can be applied in finding topics that are often discussed or often appear in a document. In this study, the concept of LDA with TF-IDF was used to get the best coherence score. In the long run, it is hoped that other studies can apply other methods and combine them with LDA to get even more coherence scores and more datasets.

REFERENCES

1. Mahesh Korlapati, Tejaswi Ravipati, AbhilashKumar Jha and KollaBhanu Prakash, *Categorizing Research Papers ByTopics Using Latent Dirichlet Allocation Model*, ISSN

2277-8616.

2. Z. Tong and H. Zhang, "A text mining research based on LDA topic modelling", *International Conference on Computer Science Engineering and Information Technology*, pp. 201-210, 2016, May.

3. H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, et al., "Latent Dirichlet allocation (LDA) and topic modeling: models applications a survey", *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15169-15211, 2019.

4. C. B. Asmussen and C. Møller, "Smart literature review: practical topic modelling approach to exploratory literature review", *JournalofBigData*, vol. 6, no. 1, pp. 1-18, 2019.

5. A. Panichella, "A Systematic Comparison of search-Based approaches for LDA hyperparameter tuning", *Information and Software Technology*, vol. 130, no. 106411, 2021.

6. H. B. Yalamanchili, S. J. Kho and M. L. Raymer, "Latent dirichlet allocation for classification using gene expression data", *2017 IEEE 17th international conference on bioinformatics and bioengineering (BIBE)*, pp. 39-44, 2017, October.

7. C. Jacobi, W. Van Atteveldt and K. Welbers, "Quantitative analysis of large amounts of journalistic texts using topic modelling", *Digital journalism*, vol. 4, no. 1, pp. 89-106, 2016.

8. J. W. Uys, N. D. Du Preez and E. W. Uys, "Leveraging Unstructured information using topic modelling", *PICMET '08 – 2008 Portland International Conference on Management of Engineering Technology*, pp. 955-961, 2008.

9. D. J. Hu, Latent dirichlet allocation for text images and music., San Diego:University of California, vol. 26, 2013.

ISSN: 2366-1313

10. I. B'lro', J. Szabo' and A.A. Benczu'r, "Later dirichlet allocation in web spam filtering", *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pp. 29-32, 2008, April.

11 Prasadu Peddi (2019), "Data Pull out and facts unearthing in biological Databases", International Journal of Techno-Engineering, Vol. 11, issue 1, pp: 25-32.