Topic Modelling of Tweets on Omicron

Venkata Shiva Sai Mallikarjun Devasani
Electrical and Computer Engineering
University of Florida
Gainesville, FL, USA
vdevasani@ufl.edu

***Abstract*— In this project , Topic Modelling study of the twitter dataset of Omicron obtained from Kaggle is discussed. Omicron is one of the Covid19 variant. The dataset includes user tweets on this topic which was obtained from Kaggle. Three different topic modelling algorithms - Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and Latent Semantic Analysis (LSA), will be applied to study the underlying topic distribution among tweets. Firstly, data preprocessing and data exploratory analysis will be carried out to understand the data. These pre-processed tweets will be then fed to the algorithms and C_V coherence score will be used to compare the performance of the models as well as to choose the best number of topics for each model. The results were examined using different visualizations.**

***Index Terms*— Topic modelling, Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), Latent Semantic Analysis (LSA), Omicron, Twitter dataset.**

I. Introduction

THE B.1.1.529 variant of Covid19 also known as “the Omicron variant” was first reported by South Africa on 24th November 2021. The World Health Organization classified the new variant as a variant of concern, named Omicron on 26th November 2021. According to WHO, not a lot is known about the transmissibility, severity of the disease, the effectiveness of existing countermeasures and vaccines on the new variant and the effectiveness of currently used PCR tests in detecting the new variant. Preliminary evidence suggests that this new variant has a higher reinfection rate compared to other earlier discovered variants. This means that people who have previously had COVID-19 could become easily reinfected and this has been a cause for concern amongst people all over the world. The use of social media platform media to analyze people’s perception of diseases has proven to be significantly useful in the past. Diseases such as H1N1 influenza (Chew & Eysenbach, 2010), Zika virus (Stefanidis et al., 2017) and Ebola virus (Odlum & Yoon, 2015) have used public conversations as an indicator to understand global sentiments on diseases. This is particularly very valuable in the case of the Omicron virus which is relatively new as such information can be useful. I have decided to apply unsupervised topic modeling techniques on tweets data on users about this variant and understand the most important topics discussed.

Topic modelling approaches have been used in many published articles on numerous topics such as in sciences, software engineering, social networks and so on. (Daud et al., 2010, p. 280-301) used topic modelling with soft clustering to classify existing models in many categories. Parameter estimation like as Gibbs Sampling and performance evaluation measures was used. (Chen et al., 2016, p. 1843-1919) performed a survey on topic modelling with a focus on software engineering field. The focus was on the application of topic models on articles from December 1999 to December 2014. A total of 167 articles was discovered to have used topic modelling in this field and that most of the tasks performed focused on basic topic models.

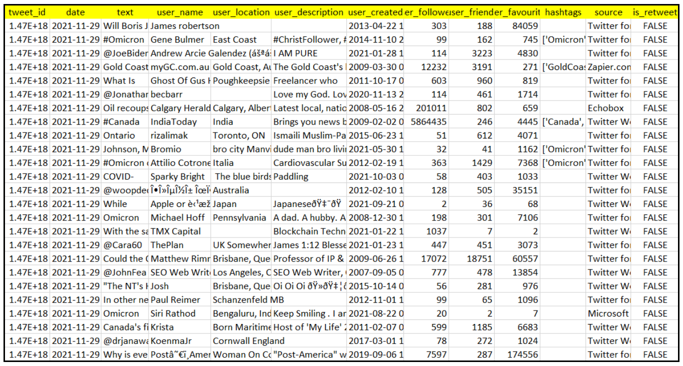
Since the initial outbreak of the Covid19 virus, many researchers have utilized social media data to gain better insights to the disease. One of those researchers is (He et al., 2021) who used the process of keywords filtering and sentiment analysis to analyze people’s behavior and sentiments towards the wearing of facemask during the Covid19 outbreak. More particularly, (Ordun et al., 2020) used topic modelling, particularly LDA topic modelling method to track the outbreak of the Covid19 virus over a period and correlated these tweets with certain news events overtime. Another similar research was by (Jelodar et al., 2020) who used LDA topic modelling method together with Recurrent Neural Networks to classify tweets into different topics and carried out sentiment analysis of tweets within each topic. (Abd-Alrazaq et al., 2020) also used English twitter data from users to identify 12 topics using Latent Dirichlet Allocation (LDA). These topics were further grouped into themes which were the virus origin or source, countries and economy, solutions to mitigating risk and its impact on the population. (Mackey et al., 2020) also used BTM (Bi-term Topic Model) on some tweet data related to Covid19 symptoms to separate tweets containing words with related topics about testing, symptoms, and recovery into same topic cluster.

1. DATA OVERVIEW AND PROCESSING

1.1 Data Description

The data set used for this project was obtained from an online repository called Kaggle from where public datasets can be downloaded for free. The dataset used in this project can be accessed using this link: https://www.kaggle.com/shivamb/omicron-covid19-variant-tweets.

The data set contains tweets from different people across the world about the new Covid19 variant – Omicron. The dataset contains all of its vital information as features or attributes. Altogether, the dataset contains 8073 rows or tweet instances and 13 columns. The included attributes are Tweet_id, date, text, username, user location, user description, user created, user followers, user friends, user favorites, hashtags, source, Is_retweet.



tweet_id	date	text	user_name	user_location	user_description	user_created	er_followe	user_friender_favourit		hashtags	source	is_retweet
1.47E+18	2021-11-29	Will Boris J	James robertson			2013-04-22 1	303	188	84059		Twitter fo	FALSE
1.47E+18	2021-11-29	#Omicron	Gene Bulmer	East Coast	#ChristFollower, #	2014-11-10 2	99	162	745	['Omicron'	Twitter fo	FALSE
1.47E+18	2021-11-29	@JoeBider	Andrew Arcie	Galendez (áš@á	I AM PURE	2021-01-28 1	114	3223	4830		Twitter fo	FALSE
1.47E+18	2021-11-29	Gold Coast	myGC.com.au	Gold Coast, Au	The Gold Coast's I	2009-03-30 0	12232	3191	271	['GoldCoa:	Zapier.con	FALSE
1.47E+18	2021-11-29	What Is	Ghost Of Gus I	Poughkeepsie	Freelancer who	2011-10-17 0	603	960	819		Twitter fo	FALSE
1.47E+18	2021-11-29	@Jonathar	becbarr		Love my God. Lov	2020-11-13 2	114	461	1714		Twitter fo	FALSE
1.47E+18	2021-11-29	Oil recoups	Calgary Herald	Calgary, Alber	Latest local, natic	2008-05-16 2	201011	802	659		Echobox	FALSE
1.47E+18	2021-11-29	#Canada	IndiaToday	India	Brings you news b	2009-02-02 0	5864435	246	4445	['Canada',	Twitter W	FALSE
1.47E+18	2021-11-29	Ontario	rizalimak	Toronto, ON	Ismaili Muslim-Pa	2015-06-23 1	51	612	4071		Twitter fo	FALSE
1.47E+18	2021-11-29	Johnson, M	Bromio	bro city Manvi	dude man bro livi	2021-05-30 1	32	41	1162	['Omicron'	Twitter fo	FALSE
1.47E+18	2021-11-29	#Omicron (	Attilio Cotrone	Italia	Cardiovascular Su	2012-02-19 1	363	1429	7368	['Omicron'	Twitter fo	FALSE
1.47E+18	2021-11-29	COVID-	Sparky Bright	The blue birds	Paddling	2021-10-03 0	58	403	1033		Twitter W	FALSE
1.47E+18	2021-11-29	@woopdec	Î•Î»Î»Î½Î± ÎœÏ	Australia		2012-02-10 1	128	505	35151		Twitter fo	FALSE
1.47E+18	2021-11-29	While	Apple or è‹¹æž	Japan	JapaneseðŸ‡¯ðŸ	2021-09-21 0	2	36	68		Twitter W	FALSE
1.47E+18	2021-11-29	Omicron	Michael Hoff	Pennsylvania	A dad. A hubby. A	2008-12-30 1	198	301	7106		Twitter fo	FALSE
1.47E+18	2021-11-29	With the sa	TMX Capital		Blockchain Techn	2021-01-22 1	1037	7	2		Twitter W	FALSE
1.47E+18	2021-11-29	@Cara60	ThePlan	UK Somewher	James 1:12 Blesse	2021-01-23 1	447	451	3073		Twitter fo	FALSE
1.47E+18	2021-11-29	Could the (	Matthew Rimr	Brisbane, Que	Professor of IP &	2009-06-26 1	17072	18751	60557		Twitter fo	FALSE
1.47E+18	2021-11-29	@JohnFea	SEO Web Writ	Los Angeles, C	SEO Web Writer, (	2007-09-05 0	777	478	13854		Twitter W	FALSE
1.47E+18	2021-11-29	"The NT's I	Josh	Brisbane, Que	Oi Oi Oi ðŸ»ðŸ‡¦	2015-10-14 0	56	281	976		Twitter W	FALSE
1.47E+18	2021-11-29	In other ne	Paul Reimer	Schanzenfeld MB		2012-11-01 1	99	65	1096		Twitter fo	FALSE
1.47E+18	2021-11-29	Omicron	Siri Rathod	Bengaluru, Ind	Keep Smiling . I ar	2021-08-22 0	20	2	7		Microsoft	FALSE
1.47E+18	2021-11-29	Canada's fi	Krista	Born Maritime	Host of 'My Life' 2	2011-02-07 0	599	1185	6683		Twitter W	FALSE
1.47E+18	2021-11-29	@drjanawa	KoenmaJr	Cornwall England		2017-03-01 1	78	272	1024		Twitter W	FALSE
1.47E+18	2021-11-29	Why is eve	Postâ˜€ï¸Amer	Woman On Co	"Post-America" w	2019-09-06 1	7597	287	174556		Twitter fo	FALSE

Figure 1: shows a partial overview of the Omicron variant tweets corpus

1.2 Data Pre-processing

As far as textual data is concerned, preprocessing of the data before feeding it to different algorithms is always a critical step. These processes are necessary because data in its raw form can be messy. These processes are carried out to help clean and transform data into formats that can be better understood thereby, helping to achieve better results. Data preprocessing not only results in the removal of noise but also reduces the data which speeds up calculations.

The preprocessing task carried out for this project included the filtering/removing of punctuation, numbers, Stopwords, unnecessary spaces, tokens that consist of only one character, links, slashes, underscores, hashtags, equal signs, periods, dashes, symbols, and data was made sure to be lowercased for all the index terms. The list of stop words for English language was imported from NLTK library and every word was checked to see if it is present in the stop word list or not. Finally, words were converted to their most basic form using the WordNet lemmatize function.

1.3 Exploratory Data Analysis

A word cloud of frequently used or appearing words in our pre-processed corpus was constructed to provide deeper insights to tweets related to the Omicron variant as posted by Twitter users. The result of the analysis is shown below. Here the words that appeared more frequently have a larger font-size. As expected, omicron word dominates the word cloud which appears in several bigrams.

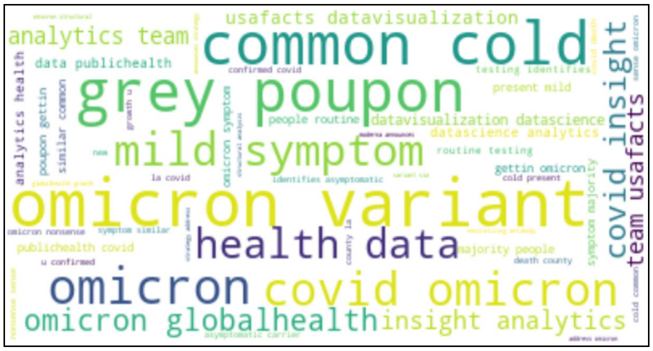


Figure 2 : Shows some of the most popular words related to the Omicron variant in our corpus

Another important thing to note is the distribution of the lengths of tweets in the dataset, it helps to understand the length of the conversations by the users. The distributions are shown in the figure 3 and majority of the tweets are between the lengths of 5 and 25. Considering the 280-character limit of twitter, it was expected to lie between this range.

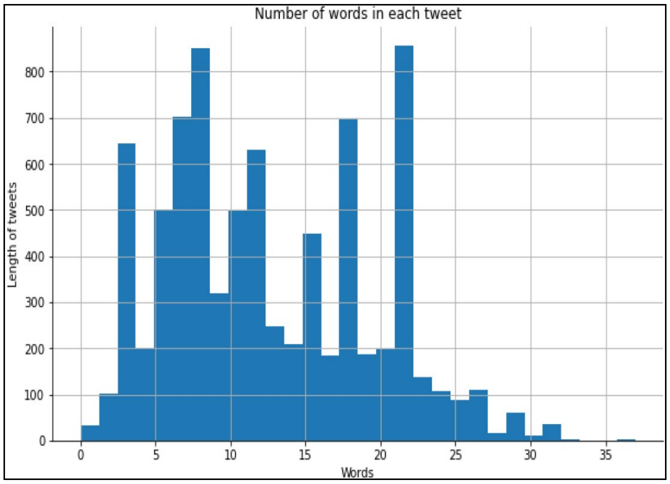


Figure 3: Distribution of the lengths of tweets after pre-processing

1.4 Feature Selection and Vectorization

Feature selection was applied on the training data to include only the important words. This is usually carried out to reduce the data available for training as it helps to speed up the training process. Only words that were present in more than 3 documents and in less than 60% of the documents were deemed important. Here it is assumed that the words that emerge in more than 60% of the documents can be rendered as domain specific and thus can be ignored and the words that appear in less than 3 documents can be considered to have appeared randomly. Bigrams were also added as features in vocabularies which appeared at least twice in the whole of the corpus. The unique number of tokens before this selection step was 9834 and this was reduced to 3263. The processed corpus was initially split into tokens and a bigram model was trained to include those bigrams as tokens who occurred at least 6 times in the whole dataset. Then I created a dictionary and corpus as input for LDA.

As computers can't work directly with words, it is required to first convert these words into some numerical representation. TF-IDF scores were calculated for the corpus and at this stage data was ready to be fed to the topic modeling algorithms.

II. DESCRIPTION

I have applied three topic modelling methods to analyze the dataset. Topic modelling also known as a probabilistic clustering algorithm is an unsupervised machine learning method used to organize large collection of text corpus into abstract topics and themes.

2.1 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a popularly used Topic Modeling method mostly because it comes with well suited libraries which makes implementation relatively straightforward. Latent Dirichlet Allocation works by leveraging Bayesian to spot hidden topics (latent topics) in a corpus of documents (Blei et al., 2003). It does this by assuming a generative probabilistic model where each document contains a combination of different topics, and each topic is a mixture of words that follows a bag-of-words model of documents, and each document is a vector space representation of words and is considered a count of terms. The TD-IDF matrix is used to represent the text corpus as LDA uses a bag-of-words model of documents. The number of topics need to be specified ahead of time.

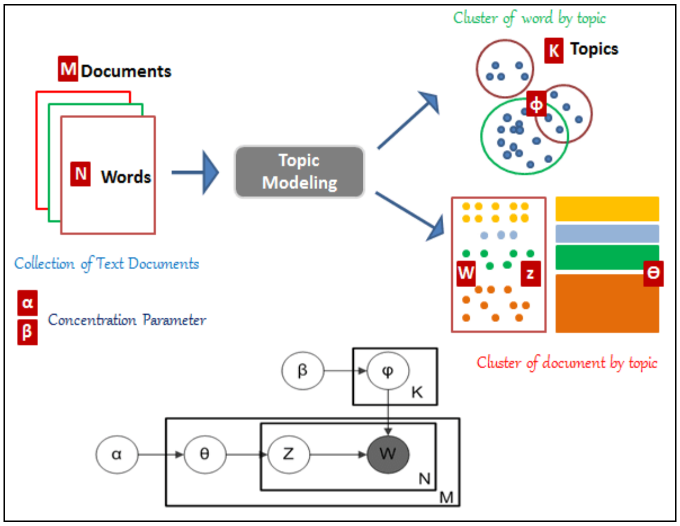


Figure 4: LDA-based topic modeling process (Amara et al., 2021)

The above figure provides a visual representation of the LDA process where M is the number of documents and Nd is the number of words in each document with d ∈ {1, …, M}. To model the document collections, the following LDA generative process is executed. [Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, Zhao L (2019) Latent Dirichlet Allocation (LDA) and topic modeling: Models, applications, a survey. Multimedia Tools Appl 78:15169-15211]:

for every topic t (t ∈{1, ..., T}) a multinomial distribution ϕ t is selected from a Dirichlet distribution with hyper-parameter β

for every document d (d ∈{1, ..., M}) a multinomial distribution θ d is elected from a Dirichlet distribution with hyper-parameter α.

For every word w n (n ∈{1, ..., N d }) in document d,

A topic z n is selected from θ d.

A word w n is selected from ϕ zn.

The words in documents are considered observed variables while other components are considered hidden or latent variables (ϕ and θ). The hyper parameters are α and β. The probability of the corpus is expressed as:

$$p(D \mid \alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \times \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta) \right) d\theta_d$$

2.2 Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) formerly known as Latent Semantic Indexing (LSI) is a concept derived from Natural Language processing (NLP) used to make semantic content by representing a text corpus in vector form. LSA uses term cooccurrence to select a set of hidden concepts and words that occur frequently together are said to be somewhat associated. Latent Semantic Analysis utilizes term-document matrix by firstly converting documents into this frequency matrices, then reduces this high dimensional vector space representations of the text documents to a fixed number or fewer dimensional representation. LSA uses singular value decomposition (SVD) for this decomposition and learns hidden topics by carrying out a matrix decomposition on the term-doc matrix. Concepts in the text, patterns and relationships between terms is determined using the singular value decomposition (SVD).

2.3 Non-Negative Matrix Factorization (NMF)

Non-Negative Matrix factorization (NMF) is linear algebra method used to extract vital information from data. It does this without having preceding knowledge or context of the data. Mathematically, NMF decomposes an input matrix into two and the product of both matrices is equal or almost equal to the initial input. In the case of topic modelling, the document-term matrix represents the input matrix where the matrix is decomposed into a document-topic matrix and a topic-term matrix. The figure below shows the decomposition of the TF–IDF term–document matrix, D with a dimension of "M * N" into two matrices - U and V in such a way that D ≈ UV where U is the term-topic matrix and V is the topic-document matrix with K coordinate axes and N points.

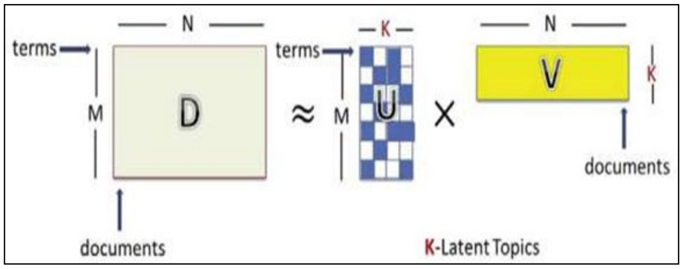


Figure 5: NMF topic modelling process (Vayansky & Kumar, 2020)

2.4 Used Technologies for Implementation

Libraries and packages imported for this project are: Gensim for topic modeling algorithms, feature selection and vectorization, NLTK (Natural Language Processing) for preprocessing, Matplotlib, Seaborn and pyLDAvis for Visualization and these are supported by Pandas and NumPy. Genism is an open-source library used mainly for natural language processing and topic modelling.

III. Evaluation

3.1 Latent Dirichlet Allocation (LDA)

The LDA model was implemented using the Gensim library and the corpus in the form of their TF-IDF was provided as an input. In unsupervised modeling, it is very important to choose the correct number of topics by keeping underlying topic distribution in mind. Coherence scores are often used to determine the correct number of topics as they measure the quality of topics learned by determining the semantic similarity between top words under each topic. The metric which was chosen for this purpose was C_V which employs sliding window, cosine similarity and Normalized Pointwise Mutual Information (NPMI) to determine the quality of topics.

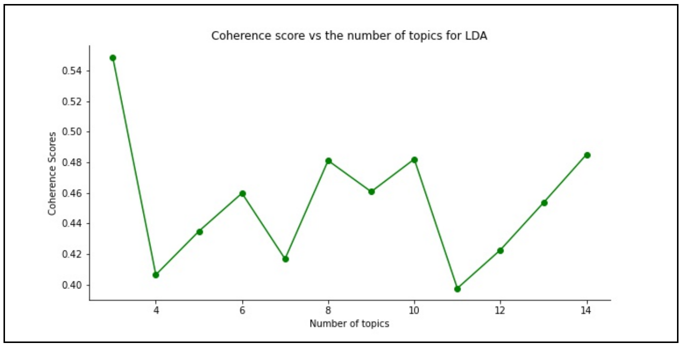


Figure 6: Coherence scores against number of topics for LDA

The coherence score was measured from 3 number of topics to 15 number of topics. The results are shown in the figure 6 where the score was highest at 0.54 when the number of topics was 3.

3.2 Latent Semantic Analysis (LSA)

The major idea behind the LSA is that words will be semantically like each other if they occur in similar texts. The TF-IDF corpus which was used for the LDA model was used here as well. This was fed to the LSI function provided by the genism to find clusters within the data. Then coherence scores were measured from 3 to 15 to find the best number of topics for this algorithm. The coherence scores against the number of topics are shown in the figure 7 below.

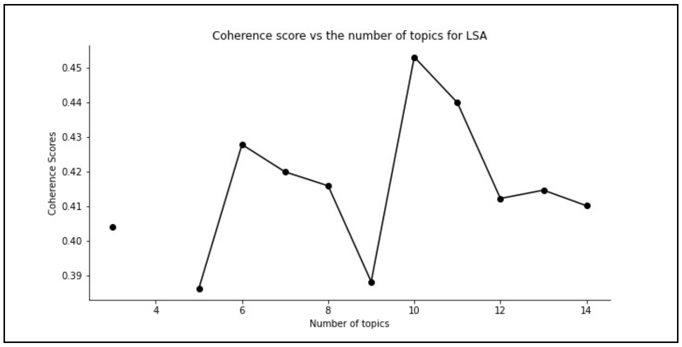


Figure 7: Coherence scores against number of topics for LSA

The highest coherence score was achieved when the number of topics were 10 and it was 0.45. Contrary to LDA, coherence score here increased by increasing the number of topics till 10 and then it decreased afterwards.

3.3 Non-Negative Matrix Factorization (NMF)

Gensim is applied on the same corpus which was used in the previous section. To find the best number of topics, coherence scores of c_v were used, and it was recorded from the topic number 3 to 15. The obtained results are shown in the figure 8.

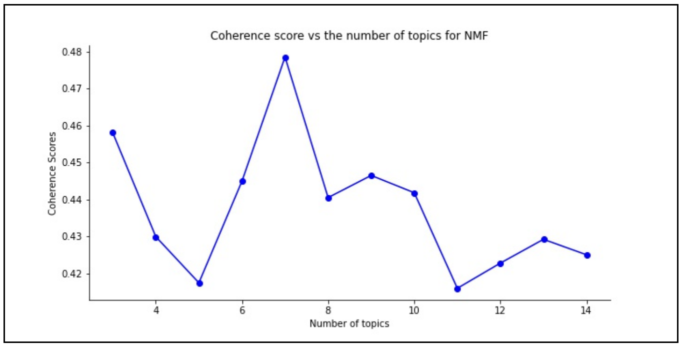


Figure 8: Coherence scores against number of topics for NMF

It is evident from the above figure that the best score of 0.47 was obtained when number of topics were 5 and it roughly decreases with the increase in the number of topics. The lowest score was 0.41 when the number of topics was 8. It can be deduced from these results that the range of topics under discussion among users are not very high.

IV. Related Work

From the coherence scores obtained from all the models, it can be said that LDA performs better than all as it's coherence score of 0.54 was highest among all used algorithms. LDA algorithm with 3 number of topics was chosen as the final model, so further analysis was carried out on the results of this model to better understand the disease perception among the users. Pyldavis was constructed for this purpose, and it is shown in the figure below. In the figure, left hand side represents where the cluster exist in a two-dimensional space, and they are far apart from each other which shows a greater semantic distance between them. The right hand shows the most important terms in the corpus and the length of their bar is influenced by their frequency.

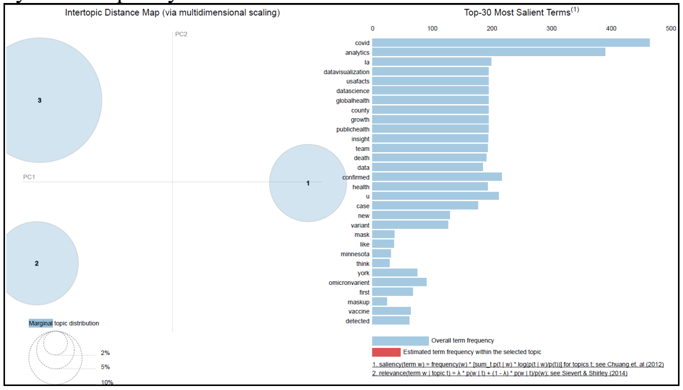


Figure 9: Pyldavis for LDA model

Despite being familiar with the semantic distance between the topics, it is also very important to know what each topic represents. The best way to do so is by looking at the top 10 most dominant words under each topic and these are shown in the table below.

	Topic 1	Topic 2	Topic 3
1	Covid	mask	Case
2	Analytics	like	New
3	La	Omicron variant	Variant
4	USA facts	Minnesota	York
5	Data science	Omicron variant	First
6	data visualization	think	Vaccine
7	GlobalHealth	mask up	State
8	Country	anyone	Detected
9	Growth	auspol	coronavirus
10	Public health	Biden	Omicron variant

Table 1: Top words under each topic

From the table above, one can understand, how the model categorizes the documents into different clusters of topics. When provided with the text corpus, model outputs the probabilities against each topic cluster to which it can belong. So, based on these probabilities, dominant topics has been extracted for each tweet to further explore the dataset.

Moreover, word clouds were extracted from instances of each topic to know the words that are being used frequently under each topic. It is also very important to mention that there were 6233 instances for topic 1, 1738 for topic 2 and 95 for topic 3. The word clouds are shown in the figure below.

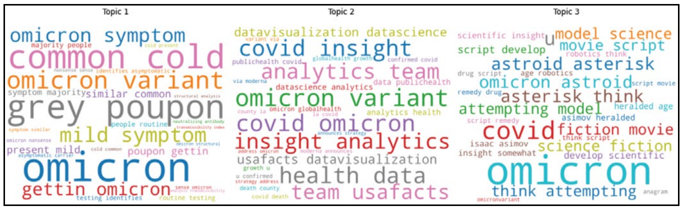


Figure 10: Word clouds for each topic

Besides analyzing the content of these clusters, the topics were also analyzed with some other features available in the original dataset. One of such visualization is shown in the figure 11 where it tries to study any relationship between the topics and the source which has been used to tweet. The trend is almost familiar with each source, so it can be interpreted that the type of device which the users uses to tweet didn't influence the topic.

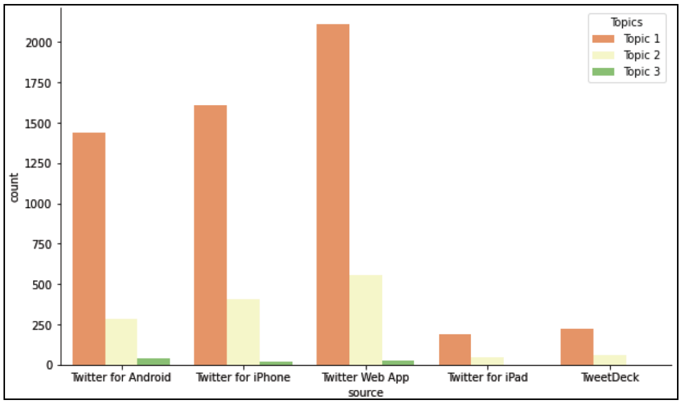


Figure 11: Topics discussed by top sources

Another visualization was plotted to study the relationship between the topics and the number of user's followers. In the dataset, expectedly number of users were given in integers and were categorized into different classes based on the number of followers. This visualization is shown in the figure 12 and it can be noted that trend remains similar across the categories. After studying these visualization, one can deduce that the topics are distributed evenly across all types of users.

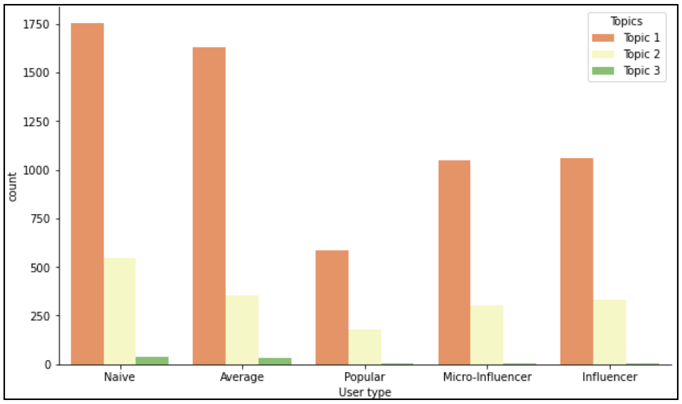


Figure 12: Topics discussed by the different users

V. CONCLUSION

Three unsupervised topic modeling algorithms were implemented on the twitter dataset concerning the omicron variant which was obtained from Kaggle. These models included Latent Dirichlet Allocation, Non-Negative Matrix Factorization, and latent semantic analysis. The best result was obtained with LDA where the coherence score of 0.54 was achieved. Several other visualizations were also used to understand the results of the LDA model.

The achieved coherence score can be considered good, but it is far from being awesome. The major reason for this can be attributed to the lack of data as the Omicron variant was in its early stages, so the people talking about this were not in a great number thus it limits our ability to obtain better results. It can be established that extensive work was carried out in analyzing the available data and study underlying the topics. This work can be considered as a base to extend it further to different datasets or better results can be obtained by using more data of same topics.

VI. REFERENCES

[1] Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: content analysis of tweets during the 2009 H1N1 outbreak. PloS one. (URL: https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0014118)

[2] Stefanidis, A., Vraga, E., Lamprianidis, G., Radzikowski, J., Delamater, P. L., & Jacobsen, K. H. (2017a). Zika in Twitter: Temporal Variations of Locations, Actors, and Concepts JMIR Public Health and Surveillaince (URL: https://publichealth.jmir.org/2017/2/e22/)

[3] Odlum, M., & Yoon, S. (2015a). What can we learn about the Ebola outbreak from tweets? *American Journal of Infection Control*. (URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4591071/)

[4] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research,* 3, 993–1022. (URL: https://dl.acm.org/doi/10.5555/944919.944937)

[5] Chen, T. H., Thomas, S. W., & Hassan, A. E. (2016) A survey on the use of topic models when mining software repositories. *Empirical Software Engineering*. 21(5). p.1843-1919

[6] Daud, A., et al., (2010). Knowledge discovery through directed probabilistic topic model. *A survey. Frontiers of computer science in China*. 4(2). p. 280-301

[7] He, L., He, C., Reynolds, T. L., Bai, Q., Huang, Y., Li, C., Zheng, K., & Chen, Y. (2021). Why do people oppose mask wearing? A comprehensive analysis of U.S. tweets during the COVID-19 pandemic. *Journal of the American Medical Informatics Association. (URL:* https://doi.org/10.1093/jamia/ocab047)

[8] Ordun, C., Purushotham, S., & Raff, E. (2020). Exploratory Analysis of Covid-19 Tweets using Topic Modeling, UMAP, and DiGraphs. (URL: http://arxiv.org/abs/2005.03082)

[9] Jelodar, H., Wang, Y., Orji, R., & Huang, S. (2020). Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: *NLP Using LSTM Recurrent Neural Network Approach. IEEE Journal of Biomedical and Health Informatics*. 24(10). p. 2733–2742. (URL: https://doi.org/10.1109/JBHI.2020.3001216)

[10] Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M. & Shah Z. (2020) Top concerns of tweeters during the COVID-19 pandemic. *Infoveillance study, J. Med.Internet Res*. 22 (4) (URL: http://dx.doi.org/10.2196/19016)

[11] Mackey, V. Purushothaman, J. Li, N. Shah, M. Nali, C. Bardier, B.Liang, M. Cai, R. Cuomo (2020). Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with COVID-19 on Twitter: Retrospective big data invigilance study, JMIR Publ. Health Surveillance 6 (2) (URL: http://dx.doi.org/10.2196/19509)

[12] Vayansky, I. & Kumar, A. P. S. (2020) A review of topic modeling methods, Information Systems. 94. (URL: https://doi.org/10.1016/j.is.2020.101582)

[13] https://www.sciencedirect.com/science/article/pii/S0306