Topic Modelling of Tweets on Omicron

Venkata Shiva Sai Mallikarjun Devasani
University of Florida
vdevasani@ufl.edu

Abstract – **In this project, the topic modeling analysis of Omicron's Twitter dataset taken from Kaggle is explained. One of the Covid-19 variants is Omicron. The dataset contains user tweets on this topic acquired from Kaggle. Three topic modeling techniques will be used to investigate the underlying distribution of topics across the tweets. Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-negative Matrix Factorization (NMF) are those algorithms. To begin, preprocessing of the data and exploratory analysis will be performed to better comprehend the tweets. After preprocessing of the tweets data, they are inputted to the algorithms, which generates the C-V coherence plots. The score from these plots will be compared amongst all the three techniques to evaluate their performance and also to arrive at the optimum number of topics for the respective technique. Also, various visualizations were used to analyze the results.**

Index Terms— **LDA, NMF, LSA, Omicron variant, Tweets, Kaggle, Topic modelling.**

I. INTRODUCTION

THE B-1-1-529 Covid-19 variant is popularly known by the name, “the Omicron variant." Initially, on the 24th of November ’21, it was first reported in South Africa. World Health Organization (WHO) has declared the Omicron to be a variant of concern on the 26th of November ’21. The organization also declared that they are not completely certain about the severity or the transmissibility of this variant. Also, they were not sure whether the existing vaccinations and safety precautionary countermeasures would be sufficient to fight with the Omicron. Preliminary research showed that it is a variant posing much greater reinfection risk than other previously reported ones. This indicates that persons who have already been infected with COVID-19 may get readily reinfected, which has caused anxiety among people all around the world. The utilization of virtual entertainment stage media to examine individuals' impression of infections has shown to be essentially valuable previously. Illnesses like H1N1 flu, Zika infection and Ebola infection attracted public discussions to figure out worldwide opinions on sicknesses. This is especially entirely important on account of the Omicron infection which is generally new as such data can be valuable. I have chosen to apply unaided subject displaying methods on tweets information on clients about this variation and comprehend the main themes examined.

Topic Modelling approaches have been utilized in many distributed articles on various points, for example, in sciences, programming, interpersonal organizations, etc. (Daud et al., 2010, p. 280-301) utilized point demonstrating with delicate grouping to characterize existing models in numerous classifications. Boundary assessment similar to Gibbs inspection and measures of execution assessment were utilized. Chen al., 2016 played out an overview on subject displaying with an emphasis on computer programming field. Top modelling was the subject of discussion across various articles from the December of 1999 to December of 2014.

From the inception of Covid-19 virus outbreak, numerous researchers have used social media data to acquire a deeper understanding of the disease. Here, Sentimental Analysis is the examination of the behavior of people towards the virus and handling the virus, their ideology regards the wearing of the facemasks. Heet al, 2021 is a researcher whose research is based on this sentimental analysis and filtering of the specific keywords. Ordun al, 2020 is another researcher whose work is based on one of the most popular topic modelling algorithms, LDA. He associated the tweets during the outbreak with certain events occurring in the world using this algorithm. Re-current neural networks in conjunction with the LDA algorithm is used by Jelodar al., 2020 to categorize the tweets into distinct topics and then perform sentimental analysis for each topic. Mackey al, 2020 employed Bi-term Topic Model for Covid19 tweets data by splitting them into topics consisting of certain words like topics clustering around symptoms, testing, and recovery.

DATA MODIFICATION

Data Description

The data collection for this project has been through the already available tweets dataset on Kaggle, as it is an online repository that has public datasets for free. The data collection comprises tweets regarding the new Covid19 version – Omicron – from people all around the world. As features or characteristics, the dataset includes all its critical information. The dataset is csv file with each tweet as rows and various information attributes regarding the tweet like the id of the tweet, location of the user, context of the tweet, username, date, hashtags, etc. All of this information is contributing to a total of 8073 tweet rows and 13 attribute columns for each row.

Preprocessing the data

As we are dealing with the textual data here, preprocessing it before inputting it to the algorithms is very important. The textual data in its original form might be untidy. The preprocessing operations are carried out to assist in the cleaning and transformation of data into formats that may be better understood, hence assisting in the achievement of better results. Data preparation not only removes noise but also lowers data, which speeds up calculations.

tweet_id	date	text	user_name	user_location	user_description	user_created	er_followe	user_friender	er_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 (áš‪á	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	2009-03-30 0	12232	3191	271	['GoldCoa	Zapier.con	FALSE
1.47E+18	2021-11-29	What Is	Ghost Of Gus H	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»ÎµÎ½Î± 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 H	Josh	Brisbane, Que	Oi Oi Oi ðŸ»ðŸ‡¦ð	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: Tweets data from Kaggle

Filtering/removing of punctuation, numbers, Stopwords pulled in from the NLTK library which does the natural language processing, unnecessary spaces, tokens that consist of only single characters, mathematical signs, links, hashtags, hyphens, underscores, slashes, underscores, periods, and index termed data is being lowercased. Lastly, using the WordNet lemmatize function, words were reduced to their most basic form.

Analysis of tweet data

From the preprocessed data corpus from the previous stage, we are now going to generate a word cloud consisting most of the frequently repeated words so that we can have a deeper insight over the tweets about the Omicron variant made by Twitter users. The analysis's outcome is displayed below. The words that appeared the most frequently have a higher font size here. The term omicron, as predicted, dominates the word cloud, which appears in multiple bigrams.

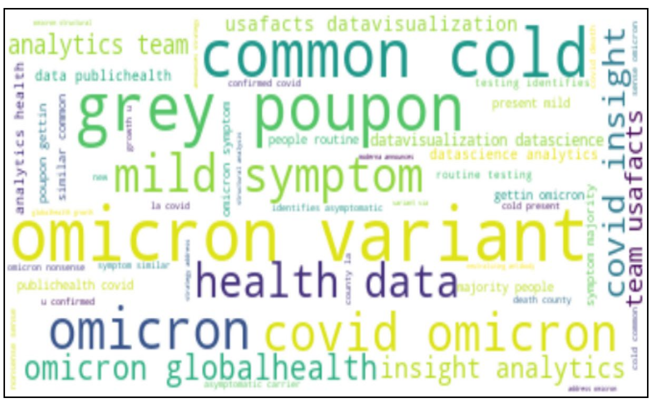

Figure 2 : Frequent words in the corpus

Another significant aspect to consider is the distribution of tweet durations in the dataset, which aids in understanding the length of user discussions. Figure 3 depicts the distributions, with the bulk of tweets ranging in length from 5 to 25 characters. Given Twitter's 280-character restriction, it was expected to fall somewhere in this region.

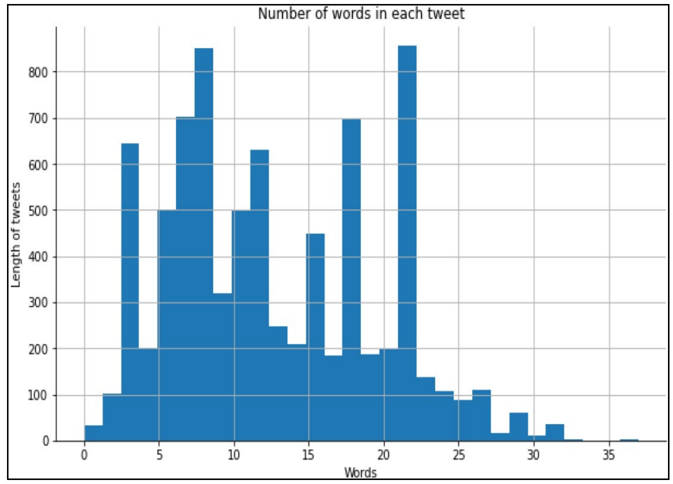

Figure 3: Tweet length vs words

Selection of features and vectorization

Right after preprocessing the data, we face another challenge in the form of unwanted words throughout the data. Each tweet has lot of the unwanted context due to conjunctions or grammar for forming or bonding the sentence. All these unwanted contexts need to be eliminated using the feature selection. This helps us with having only significant terms in our data that is going to be inputted to the algorithm. Also, it massively speeds up the training process. If a word is appearing at more than 3 instances and less than 60% of all the tweets as they might be domain-specific, then that particular term is deemed significant. Another filtration that needed to be done is for vocabulary repetitions. For this selection, Bigrams are being used. Initially we had 9834 unique token items which were reduced to 3263 after applying the feature selection. This data is then taken as an input for LDA, I constructed a dictionary and corpus.

Because computers cannot operate directly with words, they must first be converted into some numerical form. The corpus's TF-IDF scores were computed, and the data was now ready to be input into the topic modeling algorithms.

II. DESCRIPTION

To examine the dataset, I used three unsupervised topic modelling strategies. Topic modelling, also known as a probabilistic clustering technique, is a machine learning approach for categorizing massive collections of text corpora into abstract topics and themes.

Latent Dirichlet Allocation

LDA is one of the most popular Topic Modelling Algorithm which is widely implemented across various fields as it has extensively well-suited libraries which can make its implementation very simple. LDA leverages the Bayesian so that it could spot hidden or latent topics in the corpus of data. Using combination of different topics and vector space representation of words to count the terms a Probabilistic Model is generated. The TD-IDF matrix is used to represent the text corpus which requires the number of topics to be specified in prior as this is an unsupervised modeling.

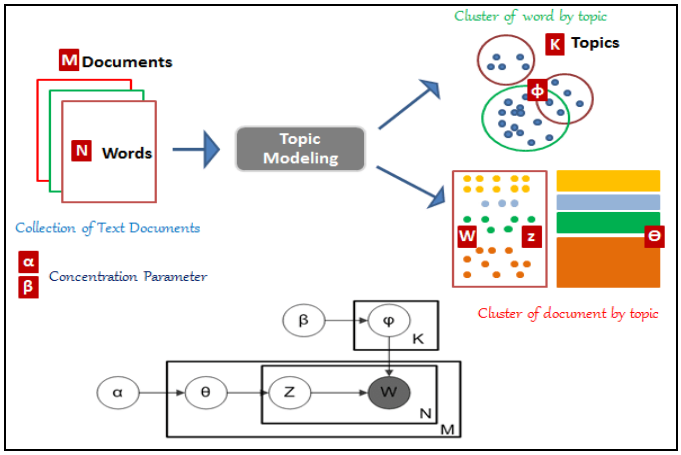

Figure 4: LDA Algorithm working model

Latent Semantic Analysis (LSA)

LSA uses the LSI function and is formerly known to be called Latent Semantic Indexing is a Natural Language Processing (NLP) approach for creating semantic content by vectorizing the text corpus. LSA identifies a group of words and concepts that are hidden of the property to frequently appear together, and which are expected to be relatively linked using the term co-occurrence. For this process, Singular-Value Decomposition is used where the whole term-document matrix is broken down to learn the hidden groups of topics by matrix decomposition. Thereby, giving us the scope to identify the patterns, concepts, and the relationships amongst the words in the corpus. Initially, the documents existing in the term-document matrix are converted as frequency matrices. Followed by the reduction of high-dimensional text documents vector-space representations into lesser dimensional representations even may be to a fixed figure of dimensions.

2.3 Non-Negative Matrix Factorization (NMF)

The NMF method is a linear algebra method for extracting important data information. Even though the NMF does not have any preceding knowledge or knows the data context, it is capable of extracting vital information from the input data matrix. Unlike the LSA or LDA, where the TD-IDF matrix is directly inputted to the algorithm, here, this input TD-IDF term-document matrix is further decomposed into topic matrix and topic-term matrix..

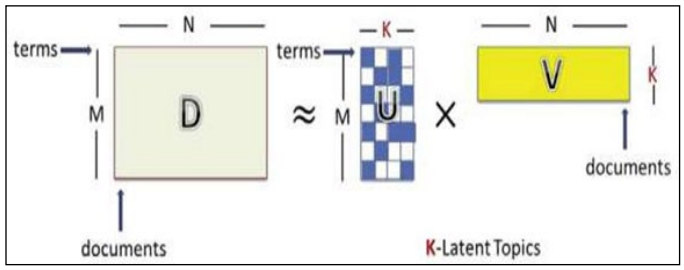

Figure 5: NMF Algorithm working model

Libraries used

NumPy, Pillow, Tabulate, Pandas, wordcloud, nltk, seaborn, genism, pyLDAvis, matplotlib.

III. Evaluation

Latent Dirichlet Allocation (LDA)

The Gensim package was used to create the LDA model, and the corpus was provided as an input in the form of its TF-IDF. It's critical to choose the right number of topics in unsupervised modeling while keeping the underlying topic distribution in mind. We are going to use the Coherence scores to determine a topic's quality. C-V, a metric that uses sliding cosine similarity, windows, and Normalized Pointwise Mutual Information (NPMI) to measure the quality of subjects, was chosen for this purpose.

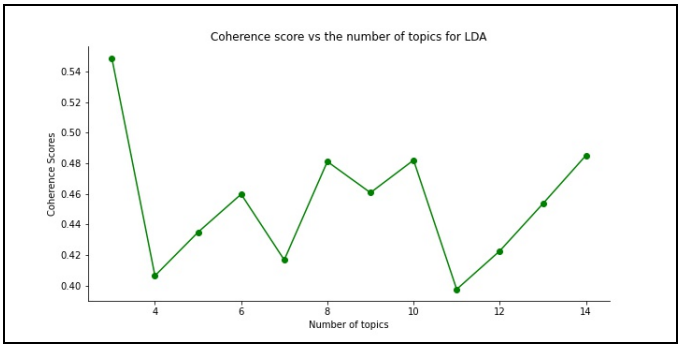

Figure 6: LDA coherence score

The coherence score was calculated on a scale of three to fifteen topics. Figure 6 shows the findings, as the topics are three, we find the coherence score to reach its maximum for this algorithm for the whole dataset.

Latent Semantic Analysis (LSA)

The LSA's main premise is that words will be semantically related as soon as they appear in alike texts. Here, we again use the exact same TF-IDF corpus as that of the one used for the LDA. The LSI function is then supplied with this, and the genism is employed with finding the data clusters. The best number of topics for this method was determined by measuring coherence scores from 3 to 15. Figure 7 depicts the coherence scores in relation to the number of subjects.

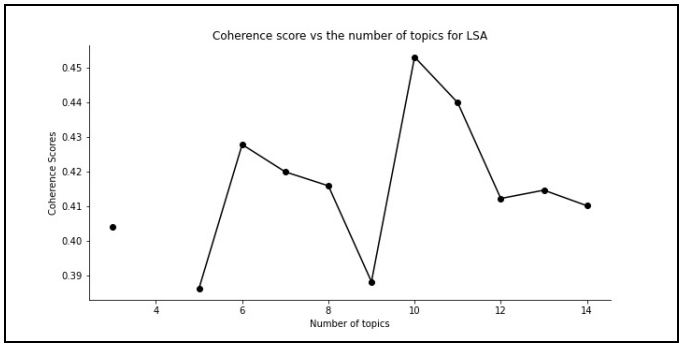

Figure 7: LSA coherence score

When the number of subjects was 10 and the coherence score was 0.45, the highest score was reached. In contrast to LDA, the coherence score here increased as the number of topics climbed until it reached ten, and then it fell.

Non-Negative Matrix Factorization (NMF)

The same corpus that was utilized in the preceding section is used for Gensim. The best number of topics was determined using c v coherence ratings, which were recorded from topic 3 to 15. The resulting outcomes are depicted in Figure 8.

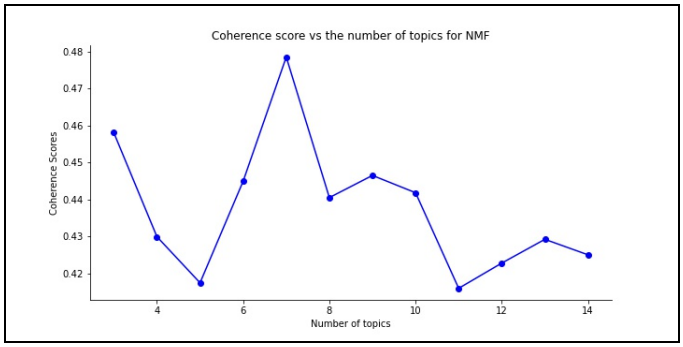

Figure 8: NMF coherence score

As seen in the graph above, the coherence scores had not been constant by achieving 0,47 as the highest score for 5 number of topics which further sees a decline to the least score of 0.41 for eleven. These findings indicate that the number of subjects under discussion among users is not very large.

IV. Related Work

Based on the coherence ratings acquired from all of the models, it can be concluded that LDA outperforms them all, with a coherence value of 0.54 being the greatest of all the methods utilized. The LDA method with three topics was chosen as the final model, and the results of this model were analyzed further to better understand how consumers perceive sickness. Pyldavis was built specifically for this purpose, as illustrated in the diagram below.

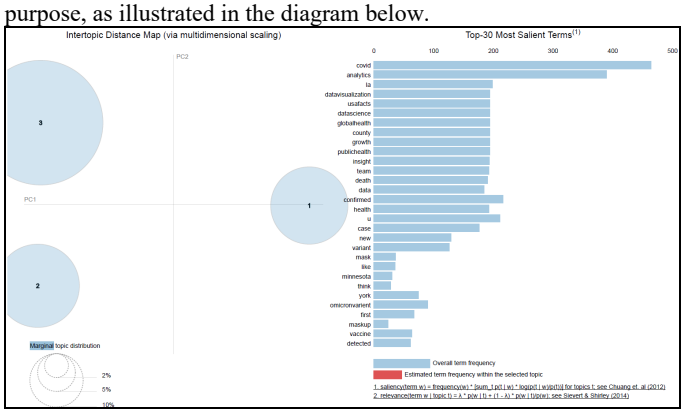

Figure 9: Pyldavis graph for the LDA

The cluster on the left side of the accompanying diagram exists in a two-dimensional space, and they are far away, indicating a greater semantic distance between them. The most essential terms in the corpus are displayed on the right side, and the length of their bars is determined by their frequency.

The words that are often used under each topic were taken from word clouds created from instances of each topic. It's also worth noting that there were 6233 instances of 1st topic, 1738 instances of 2nd topic, and 95 instances of 3rd topic. The word clouds are depicted in the illustration below.

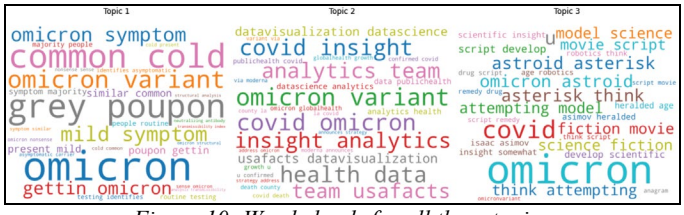

Figure 10: Word clouds for all three topics

In addition to examining the content of these clusters, the themes were examined using some of the other attributes contained in the original dataset. Figure 11 is one such visualization, which attempts to investigate any association between the themes and the source used to tweet. The pattern is nearly identical across all sources, implying that the type of device used to tweet had little bearing on the topic.

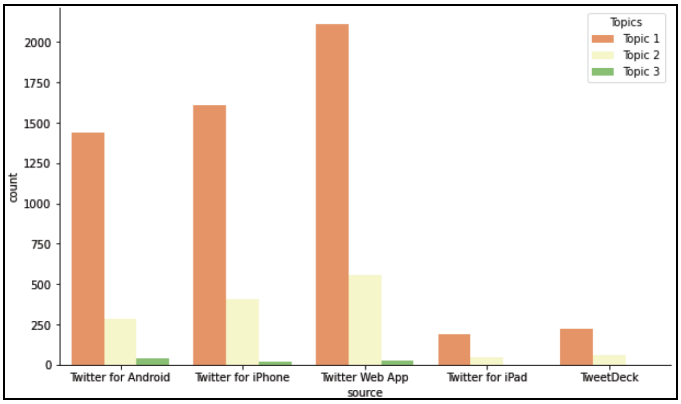

Figure 11: Top source topics

Another graph was created to investigate the link between the subjects and the number of followers a person has. In the dataset, the predicted number of users was supplied in integers and was classified into several classes based on the number of followers. Figure 12 depicts this visualization, and it can be seen that the pattern is consistent across all categories. After viewing these visualizations, one may conclude that the themes are evenly divided across all sorts of consumers.

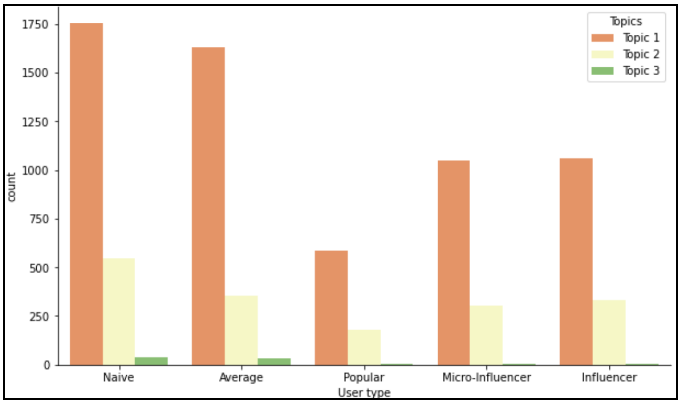

Figure 12: User topics

V. Conclusion

Three unsupervised Topic Modelling techniques, LDA, LSA and NMF are successfully implemented for the tweets dataset that is obtained from the Kaggle pertaining to the omicron variant. Amongst all the three algorithms used, I have concluded that LDA has performed much better than the others with an achievement coherence score of 0.54. The LDA model is further comprehended through various alternative graphical representations.

The coherence score we have achieved is acceptable, but far from exceptional. The main reason for this is a lack of data. Because the Omicron variation was in its early stages, the number of individuals discussing it was small, limiting our capacity to achieve better findings. It can be proven that considerable effort was used in assessing the available data and doing the research behind the issues. This project can also be used for various other datasets which might even reach much better results as there might be much more data that can be comprehended for the same issues.

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 Surveillance (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] 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)

[5] 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)

[6] 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)

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