Topic Modelling

of Tweets on Omicron

Presented by – Venkata Devasani

Contents

- Introduction – The topic under study, existing approaches
- Methodology
- Latent Dirichlet Allocation (LDA)
- Latent Semantic Analysis (LSA)
- Non-Negative Matrix Factorization (NMF)
- Performance Evaluation
- Conclusion

Introduction

- In the times of Covid outbreak, a deadly variant of the virus named as Omicron which is known for its severe transmissibility and higher reinfection rate has been discovered.

- Tweet data from Kaggle is obtained and is topic modelled to better understand the perspective of users about this variant.

- Mackey et al., 2020 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. Here, I am using 3 topic modelling algorithms namely LDA, LSA and NMF.

Methodology

- Using the NLTK library and WordNet lemmatize function for pre-processing of the tweet dataset.

- The data is explored in ways to analyze the most frequent words appearing and also the word length of the tweets.

- Feature selection was applied on the training data to include only the important words.

- Words are converted into numerical representations and TF-IDF scores for the corpus are calculated.

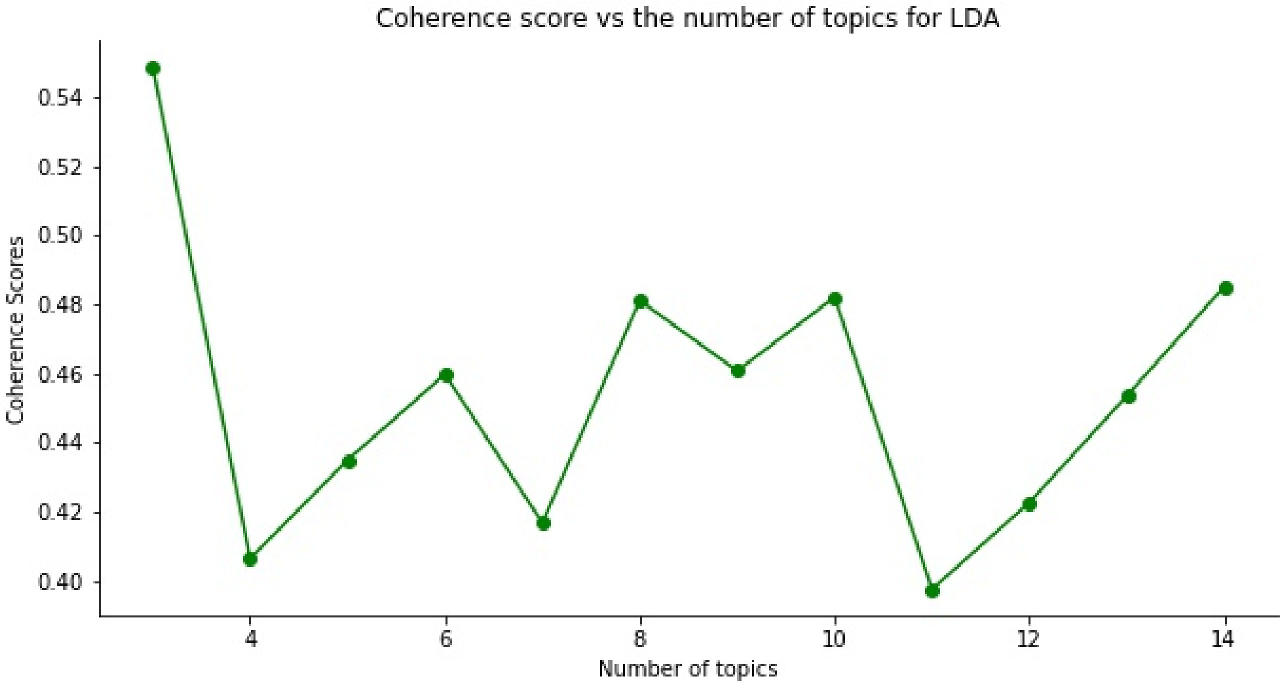
Latent Dirichlet Allocation (LDA)

- LDA works by leveraging Bayesian to spot hidden topics (latent topics) in a corpus of documents.

- Probabilistic model is generated here.

- The TF-IDF matrix is used here.

Results of LDA



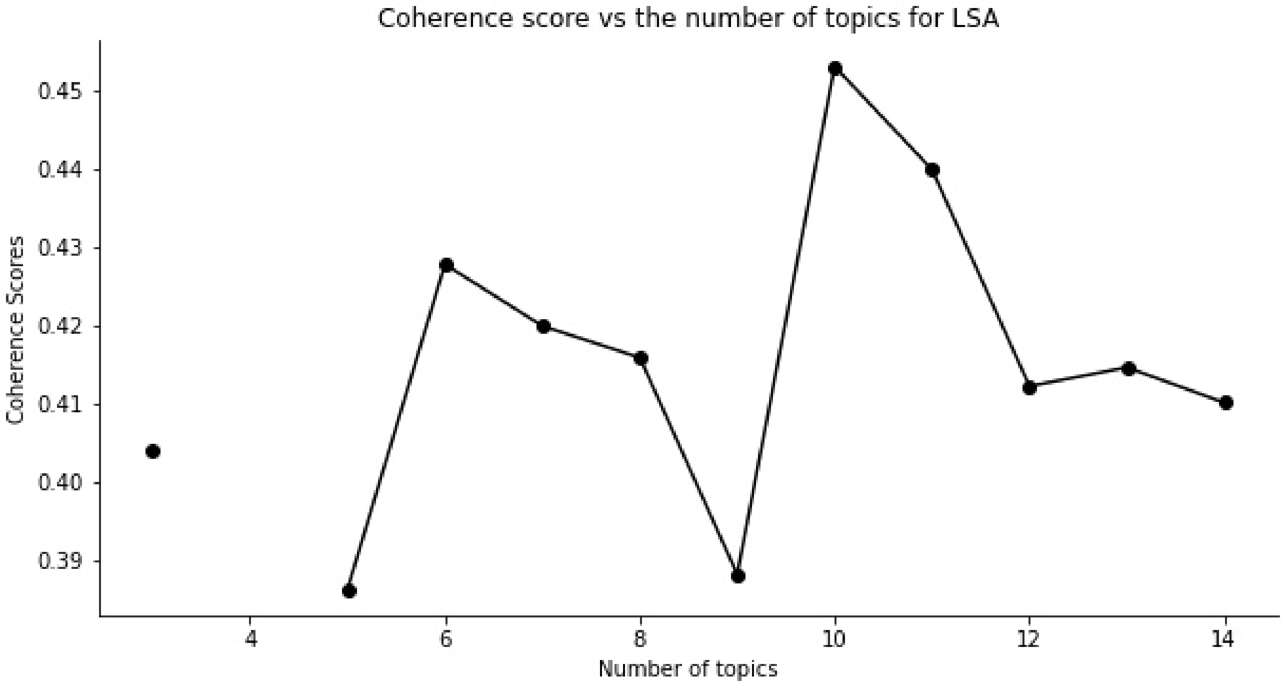
Latent Semantic Analysis (LSA)

- LSA is a concept derived from Natural Language processing (NLP) used to make semantic content by representing a text corpus in vector form.

- 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.

Results of LSA



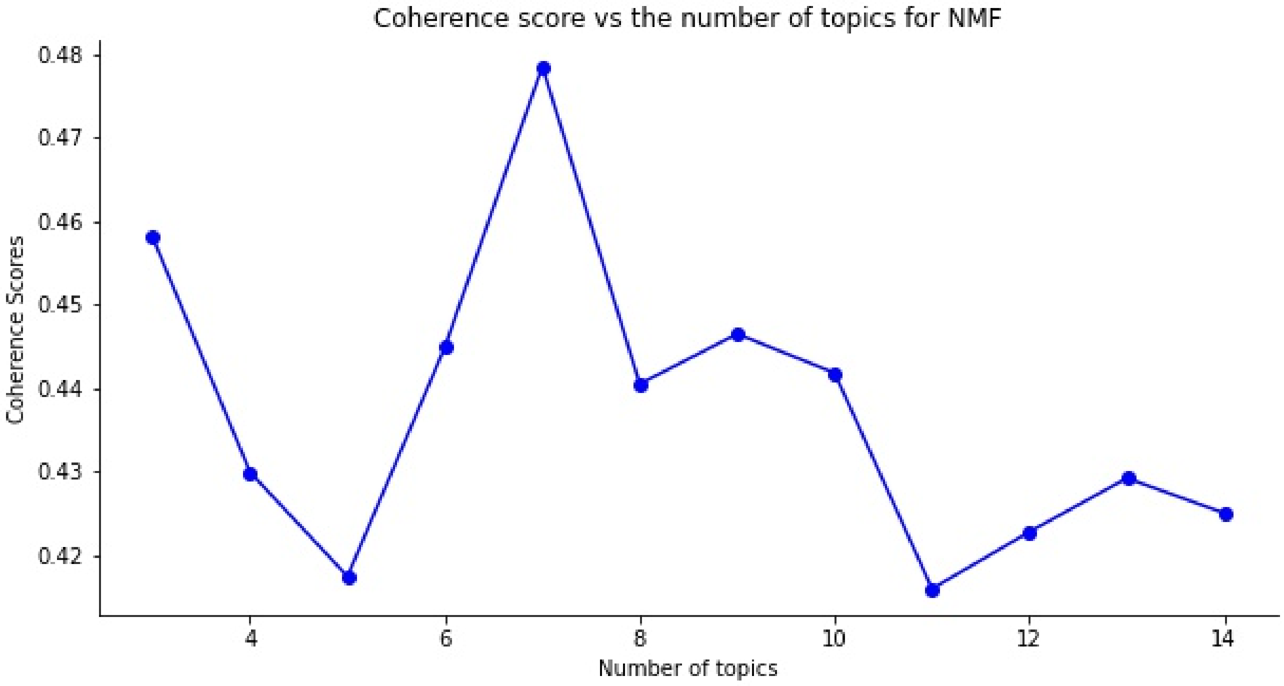
Non-Negative Matrix Factorization (NMF)

- NMF is linear algebra method used to extract vital information from data.

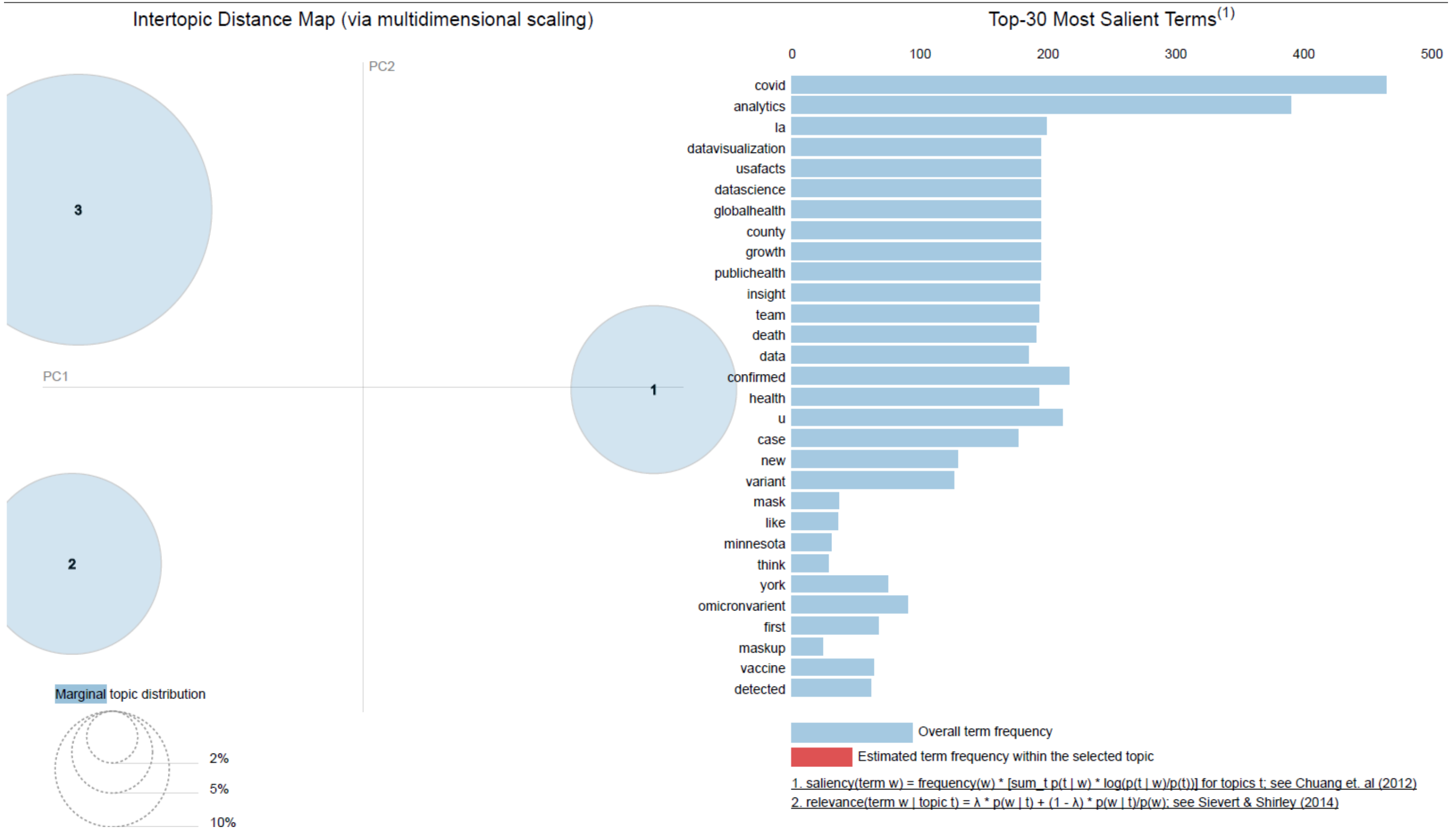
- Input matrix into two matrices and the product of these matrices is equal or almost equal to the initial input.

- The TF-IDF corpus which was used for the LDA model was used here as well. This was fed to the NMF function provided by the genism to find clusters within the data.

Results of NMF



Performance Evaluation – Pyldavis Model



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Conclusion

- Based on the coherence graphs for all the 3 algorithms used here, I conclude that LDA has the best performance among all the 3 algorithms with highest coherence score of 0.54.

- Limited data availability on Omicron limited us to getting much more better results. This work can be considered as a base to extend it further to other datasets as well.

THANK YOU