Topic Modelling of Tweets on Omicron

Venkata Shiva Sai Mallikarjun Devasani
Gainesville, FL, USA
vdevasani@ufl.edu

***Abstract*— In this proposal, the project discussed is Topic Modelling study of the twitter dataset of Omicron obtained from Kaggle. 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.**

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

I. Project Plan

THE data containing tweets across the world about the Omicron variant is collected from the following Kaggle link (https://www.kaggle.com/shivamb/omicron-covid19-variant-tweets).

Loading & pre-processing the data: This involves loading the data and the removal of noise from the dataset. This also include the removal of symbols, whitespaces, duplicate rows, hashtags, converting to lower cases, tokenization, normalization, removal of stop words, lemmatizing and stemming. Expecting to complete it by 28th March 2022.

Model Selection: For this project we will compare topic modelling with Latent Dirichlet Allocation (LDA) method with two other algorithms - Latent Semantic Analysis (LSA) and non-negative factorization. LDA is an unsupervised algorithm used to dictate semantic relationship between words in a group using certain indicators. Latent Semantic Analysis uses Term Frequency-Inverse Document Frequency (TF-IDF) to analyze documents. TF-IDF is a ranking method in statistic that shows the relevance of a word to a document inside a corpus. Non-Negative Factorization is a statistical method used to reduce the dimension of the input corpora. It can also be used for topic modelling where the input is the term-document matrix. It uses the factor analysis method to give comparatively less weightage to the words that are having less coherence. Expecting to complete it by 10th April 2022.

Outcome: The three methods will be compared. As I have been looking into this project for a while and gathered literature required for this project. I expect to complete a proper final report by 17th April 2022.

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] 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