Mapping of political events related to the COVID-19 pandemic on Twitter using topic modelling and keywords over time

Keywords: *topic modelling*, *keywords*, *political events*, *COVID-19*, *Twitter*

The COVID-19 pandemic outbreak paused societies worldwide. As cities were forced to go on lockdown, people turned to social media platforms like Twitter to discuss ongoing events. This research aims to study the relationship between actual, real-world events related to the COVID-19 pandemic and the impact these events produced on social media. To achieve this objective, we employ topic modelling and keyword extraction techniques. Topic modelling is a Natural Language Processing technique that attempts to identify topics automatically from a collection of documents (Vayansky and Kumar, 2020). This is similar to keyword extraction but, unlike this, topic modelling algorithms return clusters of words that make up the topic. Thus, a second objective is to compare the results of these two methods when it comes to identifying the salient topics in a corpus.

Several studies have looked into the social media dynamics in the context of COVID-19. For instance, Xue et al. (2020) examined COVID-19-related discussions, concerns, and sentiments on Twitter using the machine learning approach known as LDA. Jiang et al. (2020) linked Twitter users to locations within the United States to see local discussions about COVID-19. Boon-Itt and Skunkan (2020) studied the public's perception of COVID-19 by analysing keyword frequency, sentiment analysis, and topic modelling. In Spanish, Argüero-Torales, Villares, and López-Herrera (2021) applied topic modelling to study Twitter discussions at the beginning of the COVID-19 pandemic.

Methodologically, we have used the publicly available and multilingual COVID-19 Twitter dataset collected from January 21, 2020 (and still ongoing) available via the COVID-19-TweetsIDs GitHub repository (Chen, Lerman & Ferrara, 2020). The data is collected using Twitter's streaming application programming interface (API) and the Tweepy library to follow specific keywords and trending accounts. Each tweet is categorised as an original tweet, a retweet (with or without a comment), or a reply. For this study, we will focus on tweets written in English from 2020 and 2021. We limited our study to the years 2020 to 2021, which contains 1 billion tweets (31 billion tokens), and extracted a random, time-stratified sample of 0,1%, which resulted in a total of approximately 1 million tweets (31 million tokens). To our knowledge, there has not been a study which identifies events over a long period of time (2 years), by using topic modelling and keywords.

In terms of methods, we employed unsupervised machine learning methods for both tasks. For topic modelling we used BERT embeddings and the BERTopic library (Grootendorst, 2022). Our script generates a full list of topics and assigned terms, a coherence score, and several data visualisations, such as topics-over-time graphs, heatmaps, and topic hierarchies. For keyword extraction, we used *TextRank* (Mihalcea & Tarau, 2004), a language-independent, graph-based ranking model. We then compare results returned by both methods in terms of usefulness and, finally, provide an interpretation of results by relating the extracted topics to the situation of the global pandemic at different stages of the crisis.

Bibliography

Agüero-Torales, M. M., Vilares, D., & López-Herrera, A. G. (2021). Discovering topics in Twitter about the COVID-19 outbreak in Spain. *Procesamiento Del Lenguaje Natural*, *66*, 177–190.

Boon-Itt, S., & Skunkan, Y. (2020). Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study. *JMIR Public Health and Surveillance*, *6*(4), e21978. https://doi.org/10.2196/21978

Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health and Surveillance*, *6*(2), e19273. https://doi.org/10.2196/19273

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*.

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Texts. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 404–411.

Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582. https://doi.org/10.1016/j.is.2020.101582

Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y., & Zhu, T. (2020). Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach. *Journal of Medical Internet Research*, *22*(11), e20550. https://doi.org/10.2196/20550