Overview | Task 1

> "We don't have a choice on whether we DO social media, the question is how well we DO it." - Eric Qualman.

At the best of times, humanity is struggling to cope with the impact of social media on our world and environment. When a global pandemic such as the current Coronavirus (SARS-CoV-2, or COVID-19) is mixed into the equation, social media fuels the surge of numerous rumours, hoaxes and misinformation, thus amplifying the impact and severity of it exponentially (Tasnim et al., 2020). The latest Coronavirus variant, Omicron (B.1.1.529), was discovered by a team of infectious disease specialists at the University of KwaZulu-Natal in Durban, South Africa, and first reported to the World Health Organization (WHO) on 24 November 2021. The release of the news statement by WHO on 26 November 2021 has caused a major knee-jerk reaction throughout the world (World Health Organization, 2021). Multiple countries have imposed immediate travel bans on international travellers from possible exposure sites, even though the actual severity or impact of Omicron has not been fully realised.

There is a clear disparity between the perceived impact or severity of such a variant through social media platforms, and the reality as presented by the WHO and other infectious disease specialist teams around the globe. Tasnim et al identified this same issue, highlighting that fake news is spread more rapidly on social media than news from reliable sources. This overload of information can lead to media fatigue, causing individuals to discontinue healthy behaviours that are essential to personal protection. The social media hype can lead to erratic behaviour, with potential dire consequences for the individuals involved in the process.

This assignment sets out to clarify how the hype around a new strain or variant of Coronavirus is fuelled by social media, more specifically, to identify some of the key issues social media commentators are Tweeting around the Omicron variant, and to determine the sentiment behind those Tweets (whether it is positive or negative). This is not a fact verification service; no attempt has been made to verify the validity or content of any of the Tweets. Rather, this Twitter web crawler scrapes Tweets for the specified time period and extracts the Tweeted text in its raw format. The Tweeted text is cleaned by removing hyperlinks and punctuation marks, removing stop words and lemmatisation of the text to create a cleaned corpus. This corpus is then fed into the first Natural Language Processing (NLP) task, using latent dirichlet alloction (LDA) to identify key topics from the corpus that social commentators are Tweeting about. The second NLP task, Sentiment Analysis, takes each of the corpus entries to determine the sentiment of the Tweet, whether positive, neutral or negative. These prototype NLP tasks can be used by any development team as basis to identify key topics in a corpus, and to measure the sentiment of entries within the corpus.

As Quelman eluded to, this assignment is a modest contribution towards improving how well we do social media, with the hope that these prototypes may help transform how we respond to new coronavirus variants.

References

Tasnim, S., Hossain, M. M., & Mazumder, H. (2020). Impact of Rumors and Misinformation on COVID-19 in Social Media. J Prev Med Public Health, 53(3), 171-174. https://doi.org/10.3961/jpmph.20.094

World Health Organization (WHO). (2021, 26 November 2021). Classification of Omicron (B.1.1.529): SARS-CoV-2 Variant of Concern [Statement]. https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern