



The Clinical Data
Science Conference

EU 2024

10-13 November
Strasbourg Convention Centre



Topic and Sentiment Analysis of Anti-Vaccine activism using Natural Language Processing and Large Language Models

Agenda

01

Overview of vaccine hesitancy

02

Analysis objectives

03

Review of similar studies

04

Methodology and process steps

05

Data

06

Results

07

Conclusions and future work

Vaccine hesitancy overview

- Criticism around vaccines was always present globally
- But much more prevalent since COVID-19
- Vaccine hesitancy is considered one of the top 10 global health threats by the WHO
- Influenced by a variety of factors:
 - Complacency
 - Convenience
 - Confidence
- In the age of the internet forums and social media, public opinion can be influenced
 - By campaigns (of companies, NGOs, governments)
 - By content produced and made viral via individuals and their networks
 - Bad actors are also playing a role here
- Understanding and profiling public opinion can help fight negative opinions

Analysis objectives

- Can we use NLP to detect and profile public opinion (on Social Media)?
- Can we use more traditional tools as well as LLMs?
- Use topic modelling techniques in uncovering opinions
- Analyse sentiment
- Try to categorise stance to see anti- or pro-vaccines
- Can we compare how the public opinion have changed since COVID-19?

Review of related studies

- Supervised topic modelling on depression-related language on Twitter
- VaxxHesitancy: human annotated dataset and study
- Understanding Side Effects of Antidepressants on Social Media Data
- PHUSE paper: Leverage Social Media Analysis and NLP to Study the effectiveness of antidepressants
- Measuring the impact of COVID-19 vaccine misinformation on vaccination intent
- Multilingual dataset of COVID-19 vaccination studies

Methodology and process 1

- Data collection using the x.com APIs – query and other parameters
- Preprocessing tweets
 - Removing retweets
 - Detect/check language
 - Special characters, emojis, links and usernames' removal
- Natural Language Processing steps
 - Tokenisation
 - Removing stop words
 - Stemming and lemmatisation
- Features from text approaches
 - Bag of words
 - Unigram-bigram
 - Term frequency – inverse document frequency



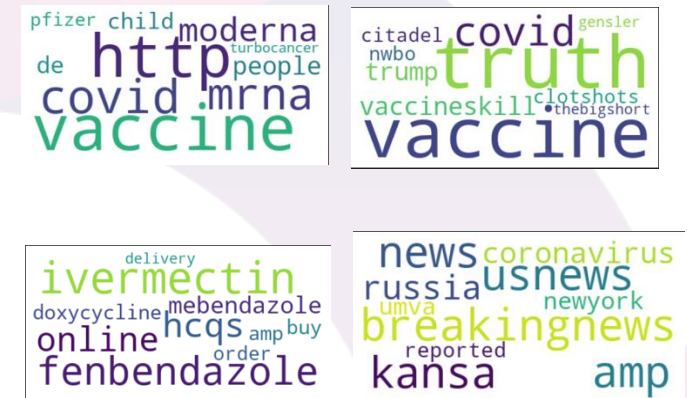
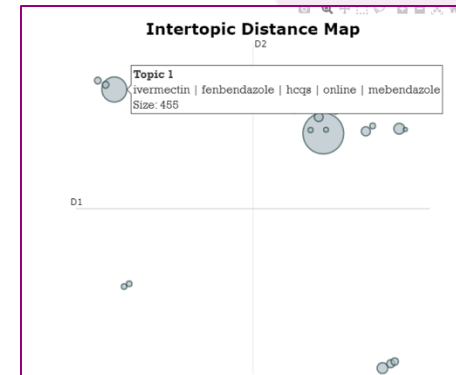
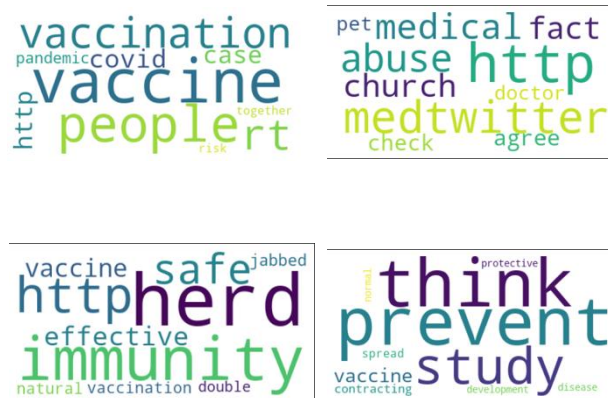
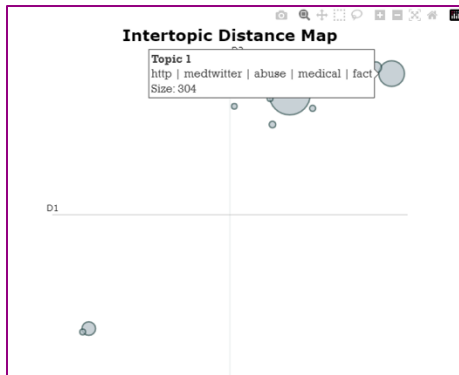
Methodology and process 2

- Topic modelling
 - LDA (Latent Dirichlet Allocation)
 - NMF (Non-Negative Matrix Factorisation)
 - BERTopic
 - Using generic LLMs
 - Profiling of topics
- Sentiment analysis
 - Vader sentiment – negative/neutral/positive
 - BERT-based sentiment analysis – negative/positive with a score
 - TextBlob sentiment analysis – polarity and subjectivity score
- Stance detection
 - Using generic LLMs
 - Categorise topics to stances

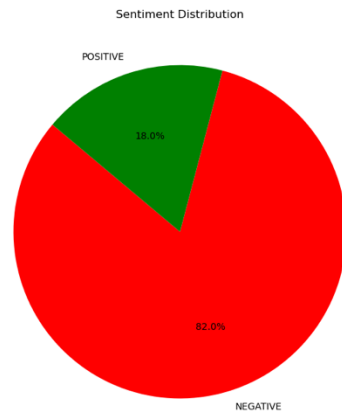
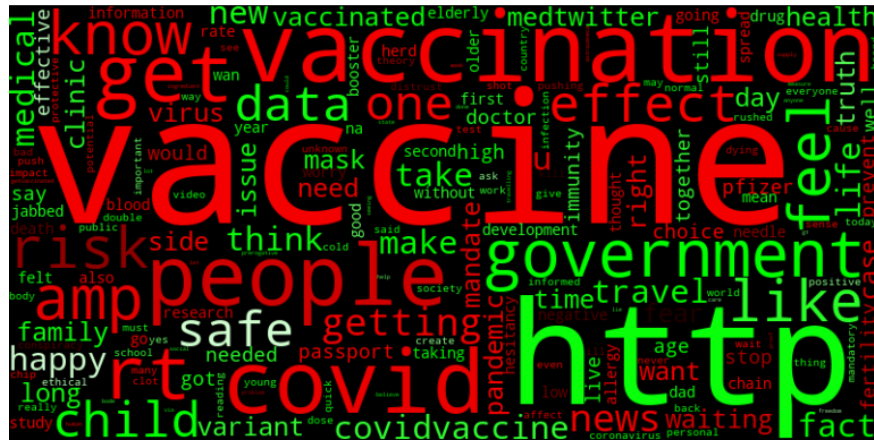
Data

- 2020-2022 based on existing dehydrated dataset
 - Get content using the x.com API
 - ~ 2,700 tweets after cleaning
- 2024 queries on the X.com API
 - similar query expressions as the 2020-2022 dataset
 - ~ 2,700 tweets after cleaning
- English language

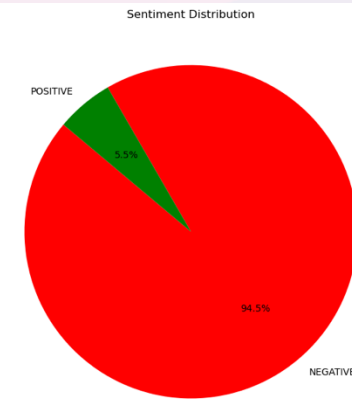
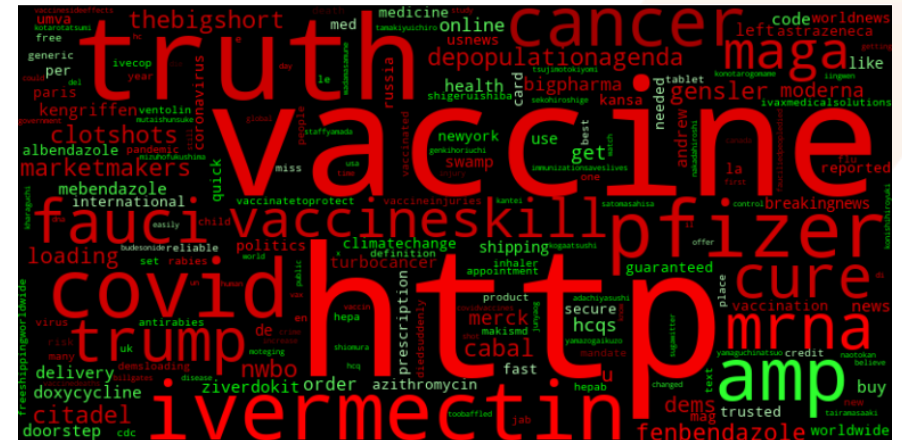
Topic modelling results



Sentiment analysis results



2020-2022



2024

Conclusions

- Topic modelling shows interesting results. BERTopic proves to be more useful than LDA or NMF approaches
- Various themes and beliefs can be uncovered analysing tweets
- Themes include information sharing about vaccines, general hesitancy, misbeliefs and conspiracy theories
- Comparing 2020-2022 to 2024, conspiracy theories seem to get more frequent and polarisation of sentiments are higher
- It is useful to understand the general sentiment of tweets
- However more detailed stance analysis can uncover more on pro- or anti-vaccine opinions

Future work

- Focus on non-English languages using appropriate tools
- Use human annotated datasets to train and evaluate pro/anti vaccine classifiers
- Combine human annotation with LLM (GPT) annotation
- Apply certain feature extraction techniques to uncover more on the nature of the public opinion
- Research more into the likely increase in polarisation of opinions on vaccines



Do you have any questions?

Contact information:

Tamas Bosznay

Katalyze Data

The Old School Hall OX28 6ZJ Witney

Oxfordshire United Kingdom

tamas.bosznay@katalyzedata.com