# Integrated Text Analysis Approach: COVID-19's Impact on Public Health Sentiments

Stephen Goosen - 10592099

Edith Cowan University

Supervisor: Dr Mary Hanson

**Integrated Text Analysis Approach: COVID-19's Impact on Public Health Sentiments**

The nearly four-year long COVID-19 pandemic has impacted public perceptions of vaccine risk, not just for COVID vaccines themselves, but for vaccines in general. Public health communicators need to adjust their risk communication strategies to address these concerns and combat vaccine hesitancy. In defining different types of risk communication, social media allows for a greater range of strategies than older forms of media. The strategies can generally be grouped into three main categories: educational posts, crisis updates, and community building posts (Malecki, Keating, & Safdar, 2020). Understanding the audience and treating the public as partners in risk communication during crises is important (Malecki, Keating, & Safdar, 2020).

Effective strategies need to acknowledge the public's fear of the disease, perceive likelihood of infection, and trust in the government institutions (Ibekwe et al., 2024). Malecki, Keating, and Safdar (2020) proposed that communication of disease information, which includes knowledge about symptoms, transmissibility, and safety, influence outrage factors. These outrage factors seemingly shape personal responsibility, fear, and institutional trust. Distrust in government institutions has often been linked to vaccine hesitancy, as seen in studies on voluntary COVID vaccinations (Ibekwe et al., 2024), vaccine uptake during the Ebola outbreak (Mesch & Schwirian, 2019), and parental confidence toward general childhood immunisations (Frew et al., 2019).

A common misconception in social and popular media discussions is that vaccine hesitancy stems from a lack of education. However, according to a study by Ibekwe et al. (2024), vaccine hesitancy (often driven by vaccine-injury risk perceptions) was highest among educated populations. They argue that vaccine hesitancy is not due to a lack of information but rather to the prevalence of misinformation. The prevalence of misinformation can sometimes be attributed to ineffective communication of changes to public health practices (Krause et al., 2023).

## Topic Modelling and Sentiment Analysis

Analysing social media responses is an important task for public health communicators to tailor their risk communication strategies, but analysis is often limited to short-text data. While short-text data is pervasive on social media platforms, its informal and unstructured nature poses challenges for text analysis algorithms.

One such text analysis method is topic modelling, a technique used to identify themes within a corpus of text. Latent Dirichlet Allocation (LDA) is a commonly used and well-founded topic modelling tool, developed by Blei, Ng, and Jordan (2003). LDA uses a 3-level Bayesian model to uncover an underlying set of topics in a collection of text documents. Following training on a selected range of topics, an LDA model can effectively predict the topic membership probability of individual documents or short sentences.

Sentiment analysis is another powerful method for text analysis. One widely adopted tool for short-text sentiment analysis is the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2018). The model was developed to understand sentiment by analysing text from both left to right and right to left (hence bidirectional), allowing it to understand word context more clearly. Sentiment analysis models can be trained to classify text into two or more emotions or sentiments, most commonly positive or negative.

The high dimensionality of text data features, and LDA's necessity to reduce noise when uncovering latent topics, means text analysis generally requires large datasets. For topic modelling, Ahammad (2024) has effectively trained an LDA model to predict 'fake news' headlines and "non-fake news" headlines concerning COVID-19, using 10'000 labelled observations. For sentiment analysis, a social media study on COVID-19 vaccines by Liu, Li, and Lui (2021) used only 5'000 labelled observations for transfer learning on the pretrained BERT-base-cased model.

There are a few models that have taken a combined approach to text analysis. One method combined Double-LDA with sentiment analysis, which yielded significantly better results than either method separately (Chen et al., 2024). Another method used a simple LDA model for feature extraction, a support vector machine classifier for feature evaluation, and a sentiment classifier to predict sentiment very effectively (F1 score = .923; Tao, 2020).

## Aims

This report aims to assess the efficacy of a combined method approach in analysing public sentiments around vaccine-related risk communication. This approach integrates topic modelling, sentiment analysis, and a custom feedforward neural network to analyse social media posts. The resulting model will be evaluated based on its ability to accurately predict pro-stance and anti-stance sentiments expressed in responses to health communicators social media posts on vaccination.

Another objective is to evaluate COVID-19's impact on vaccine hesitancy. A Chi-Squared test of independence will be conducted to compare overall public sentiment between the pre-COVID and post-COVID periods. The null hypothesis assumes that there is no significant difference in public sentiment regarding vaccine-related risk communication between post responses before December 2019 and post responses after January 2020.

## Methodology

Due to Facebook and Twitter’s attempts to monetize data analytics on their sites, data scraping has become expensive. Thus, the data collection was done manually. Initially the public health communication organisations was limited to the Australian Department of Health and Aged Care (DoHAC), but due to the lack of pre-COVID interactions on vaccine-related posts, the search included the Centers for Disease Control and Prevention (CDC) and the National Health Services (NHS). Table

1 shows the number of observations per organization as well as the number selected before December 2019 (pre-COVID) and after January 2020 (post-COVID).

**Table 1**

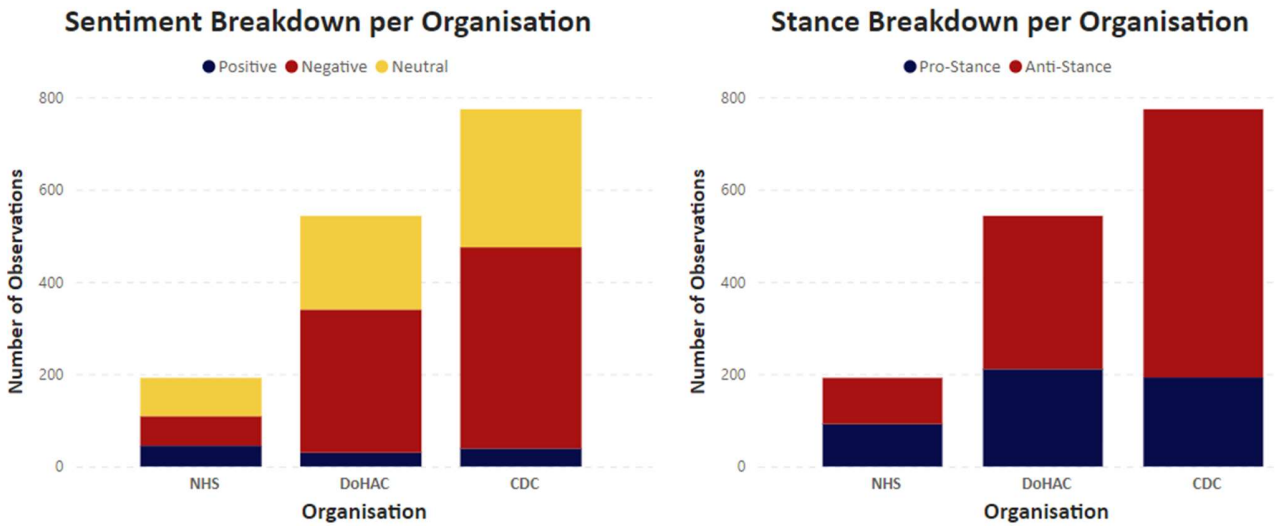
*Observation count before and after COVID-19 by organization*

| | Pre-COVID | Post-COVID | Total |
|---|---|---|---|
| NHS | 26 | 166 | 192 |
| DoHAC | 188 | 355 | 543 |
| CDC | 297 | 477 | 774 |
| | **511** | **998** | **1509** |

## Data Labelling and Pre-Processing

Labelling was done on all 1509 observations. The combined topic modelling and sentiment analysis approach complicated the labelling process, as there were positive and neutral sentiments that would be considered 'anti', while there were negative sentiments concerned with "anti-vaxxers" or vaccine availability that appear to be 'pro' overall. Therefore, two labels were created, one for sentiment (negative, neutral, and positive) and the other for stance (anti and pro). The labelled frequency distributions for sentiment and stance for each organisation are shown in Figure 1.

**Figure 1**




*Bar charts showing breakdown of stance and sentiment by organization.*

As can be seen in the second graph in Figure 1, the stance distribution for the original dataset had a balance issue (Pro=1013 and Anti=496), so bootstrapping was done to increase the number of pro-stance comments to equal that of the anti-stance comments. Many machine learning models performs better on balanced dataset. The new dataset will be referred to as the balanced dataset.

Pre-processing of the data is required to reduce noise and high dimensionality generally associated with text data. Capitalizations, special characters, hyperlinks, and hashtags were converted or removed. Using the NLTK English library, non-context stopwords such as “the,” “and,” and “was” were removed in a process called lemmatization. To reduce dimensionality further, stemming was used to reduce words to their root form. Although the PorterStemmer model from NLTK is rule-based and rather crude, making the words less legible (see examples in Figure 3), dimensionality reduction is important for model training and inference. Preprocessing was performed on both the original and balanced datasets.

**Tokenization, Vectorization, and the TF-IDF Matrix**

The comments from the pre-processed datasets were tokenized, converting each word into individual tokens. The tokenized comments are then vectorized and transformed it into the numerical format that is suitable for machine learning algorithms. Using the numerical representation of the balanced dataset’s comments, a Term Frequency-Inverse Document Frequency (TF-IDF) matrix was created which was limited to 1000 features to manage dimensionality further. The TF-IDF matrix is required to capture the importance of each word in the context of the entire dataset. The quality of topic modelling is dependent on the quality of the TF-IDF matrix (Blei, Ng, & Jordan, 2003).

**Topic Modelling**

Topic modelling was performed using LDA with 2 topics chosen. The architecture of the LDA model was provided by the Scikit-Learn decomposition package. To insure that the topics were

relevant to the study, seed words were incorporated to adjust the topic selection. Seed words were chosen based on their prevalence in each response type. The model was trained on the TF-IDF matrix and vectorized data. The purpose was to construct an LDA model that best distinguished between pro-response and anti-response, optimizing topic selection for the subsequent predictive models. Therefore, the binary prediction output for the LDA model will either be Topic 1 (Pro) and Topic 2 (Anti).

**Sentiment Analysis**

The second predictive model was the sentiment analysis model, which used the BERT-base-uncased architecture. Hyper-parameter tuning was done to optimise the model. A range of learning rates, batch sizes, and epochs were used to avoid overfitting. The output for the classification layer was modified to include 3 outputs: negative, neutral, and positive. Tokenized comments from the balanced dataset were used as training and validation input. Pre-processing was done to insure uniform input size to maintain computation efficiency.

**Custom Neural Network**

Based on the outputs from the LDA and sentiment analysis models, a feedforward neural network (FNN) was designed to predict the stance labels for each comment. The outputs from the LDA model (a 1x2 vector representing the topic distributions) and the sentiment analysis model (a 1x3 vector representing the sentiment probabilities). Each vector's values were standardized. The final combined output shape was a 5-dimensional feature vector for each comment. This was then used as the input to the FNN. The architecture of the model is as follows:

- **Input Layer:** The input 5x1 dimension which takes the combined outputs.
- **First Hidden Layer**: A dense layer with 16 neurons and ReLU activation function, followed by a dropout layer with a rate of 0.5 to prevent overfitting.
- **Second Hidden Layer**: A dense layer with 8 neurons and ReLU activation function.

- **Output Layer**: A dense layer with sigmoid activation function for the binary output.

The balanced dataset was split 80/20 into training and validations sets, respectively. The FNN was trained for 5 epochs with a batch size of 16. As an added measure to avoid the potential effects of data leaks, the model was evaluated against 500 newly collected test observations. This data was never seen by the IF-IDF matrix, the LDA model, the sentiment analysis model, or the FNN. Table 2 shows the frequency distributions for the test data. The same preprocessing method that was applied to the original and balanced datasets, was applied to the new dataset.

**Table 2**

*Test data sentiment and stance by organisation*

| | Sentiment | | | Stance | |
|---|---|---|---|---|---|
| | Negative | Neutral | Positive | Pro | Anti |
| NHS | 62 | 67 | 30 | 70 | 89 |
| DoHAC | 55 | 48 | 8 | 29 | 82 |
| CDC | 133 | 74 | 23 | 60 | 170 |
| | **250** | **189** | **61** | **149** | **341** |

## Chi-Squared Test

A Chi-Squared test was used to compare the frequency distributions for the labelled stances between the pre-COVID and post-COVID periods for the original dataset. The test assumptions of randomness, categorical data, and independence of observations were all met. A .05 α-level was selected. Table 3 shows the frequency distributions for the stances pre- and post-covid for the original dataset.

**Table 3**

*Frequency distributions for stance pre- and post-COVID*

| | Anti-Stance | Pro-Stance |
|---|---|---|
| Pre-COVID | 343.0 | 167.9* |
| Post-COVID | 670.0 | 328.1 |

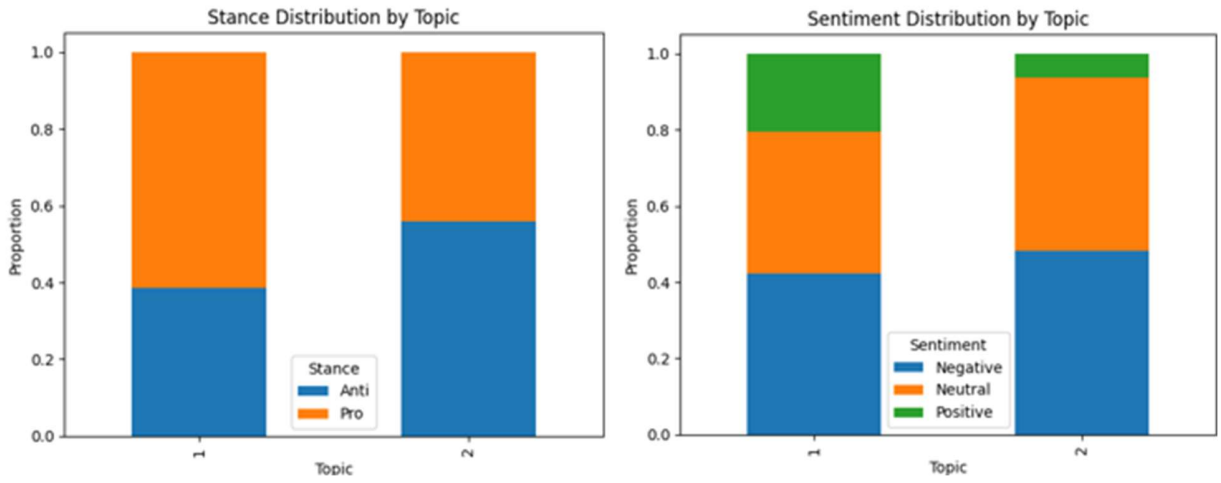
*Notes* Minimum frequency for chi-squared test exceeded.*

## Results

### Topic Modelling

The trained LDA model topics were very effective at distinguishing between response stances. The proportional stance distributions and sentiment distributions by predicted topic memberships are shown in Figure 2. According to the sentiment distribution in Figure 2, the sentiments were more balanced for Topic 1 than for Topic 2. For the stance distribution, a greater proportion of pro-stance comments were predicted to be in Topic 1, and a greater proportion of anti-stance comments were predicted to be in Topic 2. The frequency distributions for the predicted topic membership are shown in Table 4, indicating the model was better at predicting anti-stance responses. The word cloud in Figure 3 shows the common words associated with each Topic.

**Figure 2**




*Proportional bar charts for stance and sentiment distributions based on topics.*

**Table 4**

*Frequency distributions of stance in each topic*

| | Anti | Pro | **Total** |
|---|---|---|---|
| Topic 1 | 261 | 417 | **678** |
| Topic 2 | 752 | 596 | **1348** |

**Figure 3**


*Word clouds showing the most important words for each topic.*

A logistic regression model was used to evaluate the model's performance, where the overall accuracy was determined to be 57.9%. The scores for each topic are shown in Table 5. The LDA model performance metrics suggest Topic 1 has a better F1-score, as Topic 2 has very poor recall. The results are sufficient given the small dataset.

**Table 5**

*Performance metrics for LDA model classification*

| | Precision | Recall | F1-score |
|---|---|---|---|
| Topic 1 | 54.9% | 74.4% | 63.4 |
| Topic 2 | 62.4% | 42.4% | 50.3 |
| | **58.7%** | **58.3%** | **57.9** |

**Sentiment Analysis**

Based on the hyperparameter search for the sentiment analysis model, the model was trained using a batch size of 16 with the AdamW optimizer (lr = .0001) and Sparse Categorical Crossentropy loss function for 3 epochs. The balanced dataset was split 80/20 into training and validation sets, respectively. The pre-trained weights were not used, only the BERT-base-uncased architecture. The model performed very well, with a final training accuracy of 97.2% and validation accuracy of 82.8%.

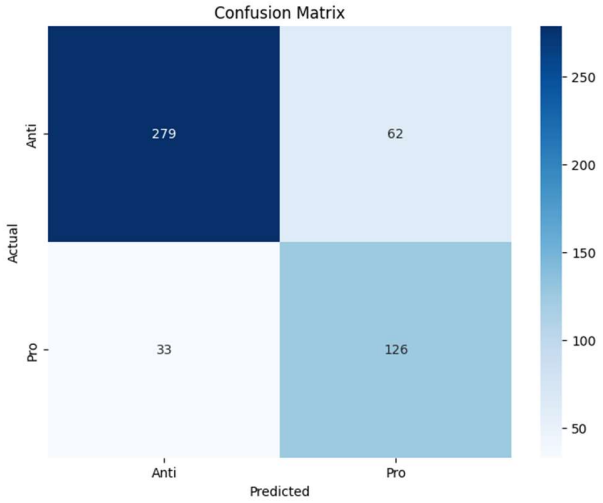
**Feedforward Neural Network**

The FNN was trained on the balanced dataset, which was split 80/20 into training and validation sets, respectively. The input for the model was prepared by making predictions on the balanced dataset

with the LDA and sentiment analysis models, and then combining the outputs into a 5x1 vector. The model was trained for 5 epochs, using the AdamW optimizer (lr=.001) and the Binary Crossentropy loss function. The model performed very well, with a training accuracy of 94.5% and a validation accuracy of 95.7%. This prompted the need to evaluate the performance with new, unseen data.

Understandably the model's performance on the unseen data was not as high, but it still performed very well, with an overall accuracy of 81.0%. This possibly suggests overfitting on the training data, or data leak resulting from the data's use in the creation of the models. The confusion matrix in Figure 4, shows the number of accurate and inaccurate predictions for each stance. The model seems to discern anti-stance comments better than pro-stance, which is consistent with the results from the LDA model (See Table 4).

**Figure 4**




*Confusion matrix showing FNN's performance on unseen data.*

**Chi-Squared Test**

Finally, the Chi-squared test to determine whether there is a difference in public sentiment. The test resulted in a p-value of 0.327 ($\chi^2$ = .960, df = 1) greater than the α-level of .05. Thus, we fail to reject the null hypotheses, and suggest there is no significant difference between stance pre-COVID and stance post-COVID.

## Discussion

The approach this paper has taken appears novel, albeit it similar to the study by Liu, Li, and Lui (2020). The integration of topic modelling using LDA and sentiment analysis with the BERT-base-uncased architecture, followed by a custom FNN, proved effective in predicting vaccine sentiment on social media posts. The LDA model was fine-tuned to differentiate between pro- and anti-vaccine sentiments and showed reasonable performance with an overall accuracy of 57.9%. The sentiment analysis model achieved a validation accuracy of 82.8%. Finally, the combined FNN demonstrated a high validation accuracy of 95.7% on the training dataset, and a slightly lower but still significant test accuracy of 81.0% on unseen data. These results indicate the effectiveness of the combined approach to classify vaccine stance. There still exists the issue of model generalizability to other public health domains, but the overall approach could be adapted to other areas of concern.

The use of seed words enhanced the topic selection for the LDA model but may also have led to the potential overfitting. Fine-tuning the topic selection further may require more expertise with the subject matter. Despite the slightly unbalanced performance seen in Table 5, the LDA model successfully captured the latent content that was crucial in predicting vaccine-related stance. The model’s ability to distinguish between topics representing pro- and anti-vaccine stances highlights its utility in identifying underlying themes in short-text social media discussions.

Sentiment analysis using the BERT model provided an additional understanding of public sentiments, capturing the positive, neutral, and negative tones in the responses. This step was beneficial, as it added depth to the analysis of the public sentiments, indicating a relationship between sentiments and topic membership.

The FNN's high accuracy in the training and validation phases suggests that combining outputs from topic modelling and sentiment analysis can significantly improve predictive performance.

However, the decline in accuracy when tested on new data indicates potential overfitting. This issue may be mitigated by increasing the size and diversity of the training dataset or incorporating cross-validation techniques. The accuracy improvements of the combined approach were supported by the results presented by other papers discussed in the literature review (Chen et al., 2024; Tao, 2020).

The Chi-Squared test comparing pre- and post-COVID sentiment distributions failed to reveal a significant shift in public sentiment regarding vaccine-related risk communication. This result does not support the findings by Ibekwe et al. (2024) or Mesch and Schwirian (2019), who observed changes in vaccine hesitancy trends. This could be attributed to the smaller dataset or potential data selection biases. Regardless, the findings from these papers emphasize the need for public health communicators to adapt their strategies in addressing vaccine hesitancy. Effective communication should focus on combating misinformation, building trust, and addressing fears about vaccine safety, as suggested by Krause et al. (2023).

The primary limitation of this study is the reliance on a relatively small and manually collected dataset. Expanding the dataset and automating data collection could improve the robustness of the models. Another concern is that the original posts, from which the responses were gathered, almost always contained hyperlinks to additional information, and neither Twitter nor Facebook track link interactions. So, the analysis of the comments may only reflect the response to the initial presentation of the information. Future research should investigate the efficacy of combined methods and explore the use of more holistic approaches to improve the performance of text analysis models.

## References

Ahammad, T. (2024). Identifying hidden patterns of fake COVID-19 news: An in-depth sentiment analysis and topic modeling approach. *Natural Language Processing Journal, 6*. doi:10.1016/j.nlp.2024.100053

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*. 993-1022.

Chen, X., Tang, W., Xu, H., & Hu, X. (2024). Double LDA: a sentiment analysis model based on topic model. *2014 10th International Conference on Semantics, Knowledge and Grids*. 49-56. doi: 10.1109/SKG.2014.20.

Devlin. J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019.* 4171-4186.

Frew, P. M., Murden, R., Mehta, C. C., Chamberlain, A. T., Hinman, A. R., Nowak, G., Mendel, J., Aikin, A., Randall, L. A., Hargreaves, A. L., Omer, S. B., Orenstein, W. A., & Bednarczyk, R. A. (2019). Development of a US trust measure to assess and monitor parental confidence in the vaccine system. *Vaccine, 37*(2), 325–332. doi:10.1016/j.vaccine.2018.09.043

Malecki, K. M. C., Keating, J. A., & Safdar, N. (2020). Crisis communication and public perception of COVID-18 risk in the era of social media. *Clinical Infectious Diseases, 72*(4). 699-704. doi:10.1093/cid/ciaa758.

Mesch, G. S. & Schwirian, K. P. (2019). Vaccination hesitancy: fear, trust, and exposure expectancy of an Ebola outbreak. *Heliyon, 5*(7). doi:10.1016/j.heliyon.2019.e02016

Krause, N. M., Beets, B., Howell, E. L., Tosteson, H., & Scheufele, D. A. (2023). Collateral damage from debunking mRNA vaccine misinformation. *Vaccine, 41*(4), 922–929. doi:10.1016/j.vaccine.2022.12.045

Liu, S., Li, J., & Lui, J. (2021). Leveraging transfer learning to analyze opinions, attitudes, and behavioral intentions toward COVID-19 vaccines: Social media content and temporal analysis. *Journal of Medical Internet Research, 23*(8). doi:10.2196/30251

Tao, W. (2020). Emotional classification model of review text based on topic and sentiment characteristics. *2020 2nd International Conference on Economic Management and Model Engineering (ICEMME).* doi:10.1109/icemme51517.2020.00213