Detecting COVID-19 Health Misinformation Targeting Older Adults Using NLP and Transformer Models

Bozhao Wang, University of Calgary • Zirui Chen, University of Calgary

Abstract

Canada's aging population has been growing at a rapid rate since the 1960s, with over 0.86 million people counted as aged 85 or older in the 2021 Census. This number is a 12% increase from 2016 and currently occupies 2.3% of the entire Canadian population. As seniors live longer, they have been facing challenges accessing verified medical expertise and consistently suffering from misleading online health claims, especially during the COVID-19 pandemic crisis. To address such urgency, this study will develop and compare two automated detection approaches:

- TF-IDF vectorization + Logistic Regression
- Bidirectional Encoder Representations from Transformers(BERT)

By using a publicly available Constraint COVID Twitter dataset from Kaggle, we applied standard text preprocessing before training the model. The result shows a strong performance for both the baseline and the fine-tuned BERT model, but it collapsed when applied to Reddit posts targeting elders. The finding highlights the challenges of domain shift and label noise, suggesting training on diverse sources, adapting models to different platforms, and using more accurate labels to protect older adults from online health misinformation.

Research Question

How can Natural Language Processing detect COVID-19 health misinformation in online text?

We test how well NLP models can detect health misinformation across multiple social media platforms.

Hypothesis

Models trained on tweeter's data will achieve high performance in-domain but will fail to keep the consistency to Reddit posts due to domain shift and different formatting styles.

TF-IDF + Logistic Regression will show better performance than BERT because it is less sensitive to training noise.

Background

- Health misinformation occupied on online platforms.
- Older adults are the majority target and are suffered to the misleading health information online.
- Previous literaure focuses on single platforms like Twitter, we want to test on multiple social media platforms.

STEM FELLOWSHIP

Methods

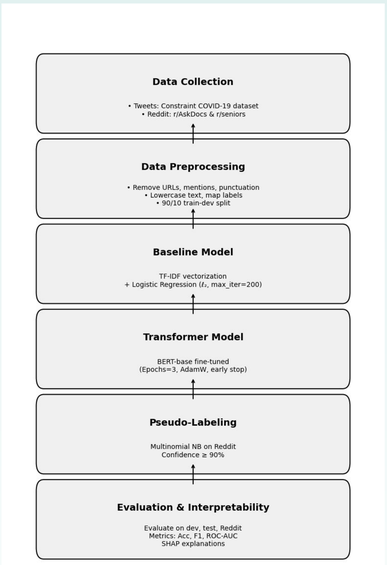

Figure 1: Method Pipeline

- Data: COVID-19 tweet dataset (6,420); Reddit posts from r/AskDocs, r/seniors
- Preprocessing: Remove URLs, mentions, punctuation, lowercase text
- Models: TF-IDF + Logistic Regression vs. fine-tuned BERT
- Pseudo-label Reddit posts with Naive Bayes
- Evaluate via F1, ROC-AUC, SHAP

Model Interpretability

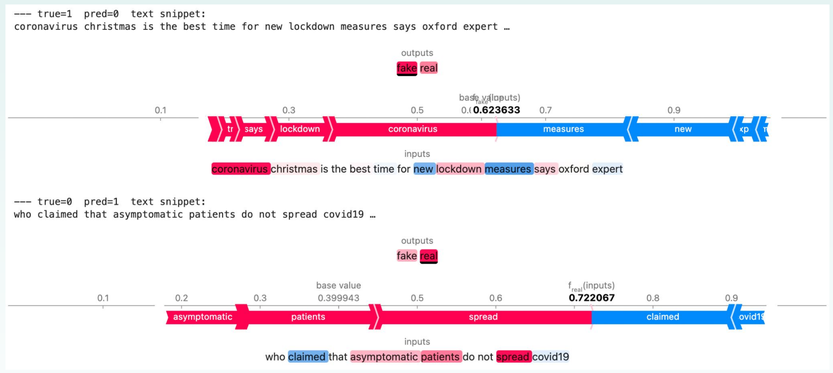

Figure 2: SHAP Interpretability: Key tokens driving predictions of fake or real COVID-19 content

Model Performance

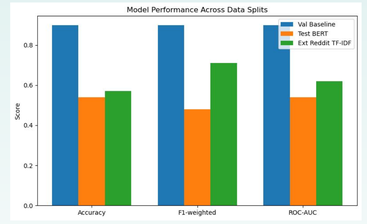

Figure 3: Model performance on in-domain, tweet test, and senior-focused Reddit data.

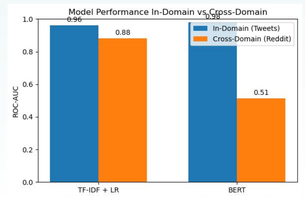

Figure 4: ROC-AUC: Tweets vs. Reddit (In-Domain vs. Cross-Domain)

Key Results & Discussion

- In-domain (Tweets): TF-IDF + LR and BERT both got F1 = 94%, AUC = 0.98
- Cross-domain (158 high-confidence Reddit posts):
  TF-IDF + LR: accuracy = 60%, AUC = 0.88
  BERT: accuracy = 37%, AUC = 0.51
- Severe performance drop highlights domain shift & noisy labels

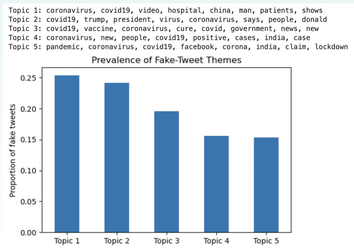

Figure 5: prevalence of fake-tweet themes

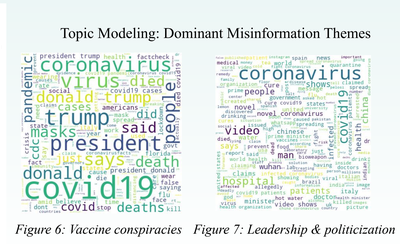

Figure 6: Vaccine conspiracies

Figure 7: Leadership & politicization

Conclusion

Key Findings

- **In-domain success:** TF-IDF + Logistic Regression and fine-tuned BERT both achieve F1 = 0.94 on the Constraint COVID-19 tweet data.
- **Cross-domain drop:** On senior-focused Reddit posts, BERT's F1 falls to 0.47 while TF-IDF retains 57% accuracy, highlighting the robustness of simpler models.

Limitations

- **Domain shift:** Differences in length, style, and vocabulary between Twitter and Reddit severely degrade performance.
- **Label noise:** Reliance on pseudo-labels introduces errors that hurt fine-tuned model quality.
- **Context complexity:** Informal, multi-sentence Reddit posts challenge short-text classifiers.

Future Directions

- **Targeted fine-tuning:** Hand-label a small Reddit subset and apply continued BERT training (domain adaptation).
- **Feature enrichment:** Incorporate behavioral and metadata signals (e.g., user activity, thread structure).

References

- Devlin J et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018).
- Zhao Y, Da J, Yan J. Detecting health misinformation in online health communities. Inf Process Manage.
- Pérez-Rosas V et al. Automatic Detection of Fake News. Proc. ACL 2018:3391–3401.