Detecting health misinformation targeting aging populations using natural language processing

*Bozhao Wang[1], and Zirui Chen[2]**

[1] Department of Economics, University of Calgary, Alberta, Canada

[2] Haskayne School of Business, University of Calgary, Alberta, Canada

- Corresponding author

  E-mail: bozaowang24@gmail.com (B. Wang)

Word count: 1096

Funding

The authors received no specific funding for this work.

Competing Interests

The authors declare no competing interests.

Detecting health misinformation targeting aging populations using natural language processing

Keywords: COVID-19, health misinformation, natural language processing, BERT, aging population

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. [1] 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 and using more accurate labels to protect older adults from online health misinformation.

Introduction

Health misinformation has been massively spread across social media in the 21st century, especially during the COVID-19 pandemic. They are targeting aging populations who significantly lack the ability to classify fake news from real news. To better help elderly get authorized medical treatment and away from exaggerated claims on the internet, our research question concentrated on leveraging AI tools such as Machine Learning(ML) and Natural Language Processing(NLP) to classify COVID-19 health misinformation in online text. In this study, we introduced two primary variables:

- Textual Features: converting text to numerical features by using techniques such as TF-IDF, word embeddings, and tokenization.
- Platform domain: the specific platform where machine learning techniques are applied. In this case, they are Twitter vs. Reddit.

We will first train and test classifiers using the chosen dataset, and then check cross-platform robustness by applying the trained models to Reddit data.

Literature Review

Ajao et al. (2018) developed an end-to-end deep-learning approach where they fed raw Twitter texts into convolutional layers to learn phrase patterns, then used that to detect fake news which resulting in 82% accuracy[2]. While their method is useful and efficient, it was evaluated only on one platform instead of a comparison to simpler platforms. In our work, we added to their insight by using TF-IDF + logistic Regression model as a baseline model to compare with transformer-based classifier (BERT) for richer text representations, and testing both approaches on other platforms such as Reddit. Moreover, we followed closely with Yuehua Zhao, Jingwei Da, and Jiaqi Yan (2021)'s work, where they explored

health-misinformation detection by combining four different feature types: n-grams, topical, sentiment and behavioural within a supervised learning framework and achieved a high accuracy of 85%.[3] Their results show that adding user interaction data can booth text-only models' performance.

Methods

Data collection.

The dataset we used is the publicly available Constraint COVID-19 Twitter dataset, which has 6,420 twitter tweets. We also collected reddit posts from online sources such as (r/AskDoc andr/seniors) by using PRAW. After removing unnecessary content, our Reddit collected a total of 169 health-related sources.

Data preprocessing.

All tweets and Reddit posts were cleaned by stripping URLs, punctuation, and then lowercase. Twitter labels were mapped to fake = 0, real = 1. The 6,420 tweets were split into a 90 % training set and a 10 % development set.

Baseline model.

We trained the TF-IDF + logistic regression classifier with ℓ_2 regularization as our robust baseline model. Split the data into a 90% training and 10% development set.

Transformer model.

We fine-tuned a pretrained BERT-base-uncased model for binary classification. Tweets were tokenized to a fixed length of 128 tokens (padding shorter sequences), and the model was

trained for up to three epochs with AdamW, batch sizes of 16 (train) and 32 (dev), and early stopping based on development ROC-AUC.

Cross-domain pseudo-labelling.

We trained a multinomial Naïve Bayes classifier on the original training tweets and computed posterior probabilities for each Reddit post. Only posts with > 0.90 probability for either class were retained to maintain a high-confidence pseudo-labelled set.

Evaluation.

We compared models and evaluated their performance by using accuracy, precision, recall, F1-score and ROC-AUC curve. We also applied SHAP to better understand error patterns to ensure transparency in the classification pipeline.

Results

In-Domain Classification Performance

We first evaluated both the TF–IDF + Logistic Regression baseline and the fine-tuned BERT model on the Constraint COVID-19 tweet dataset.

Table 1: Development-set performance

Model	Accuracy	F1 score	ROC-AUC
TF–IDF + Logistic Regression	0.90	0.91	0.97
BERT-base fine-tuned	0.94	0.94	0.98

Table 2: Test-set performance

Model	Accuracy	F1- score	ROC-AUC
TF-IDF + Logistic Regression	0.91	0.91	0.96
BERT-based fine-tuned	0.54	0.48	0.54

Cross-Domain Evaluation on Health-Related Reddit Posts

After keyword filtering, we retained 169 Reddit posts with pseudo-labels from a high-confidence Naïve Bayes ensemble.

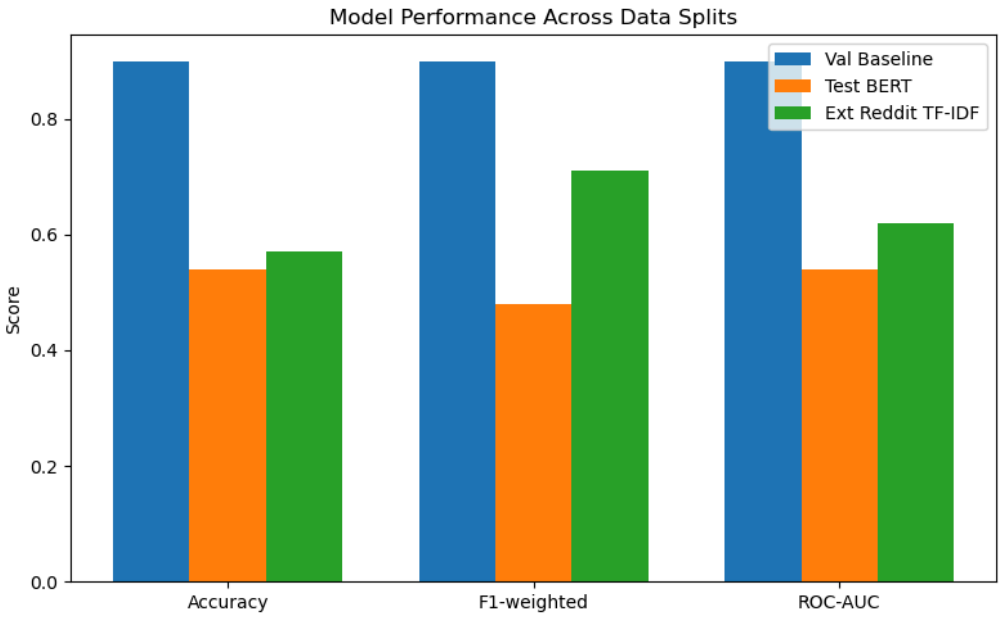

Figure 1: Model-comparison bar chart

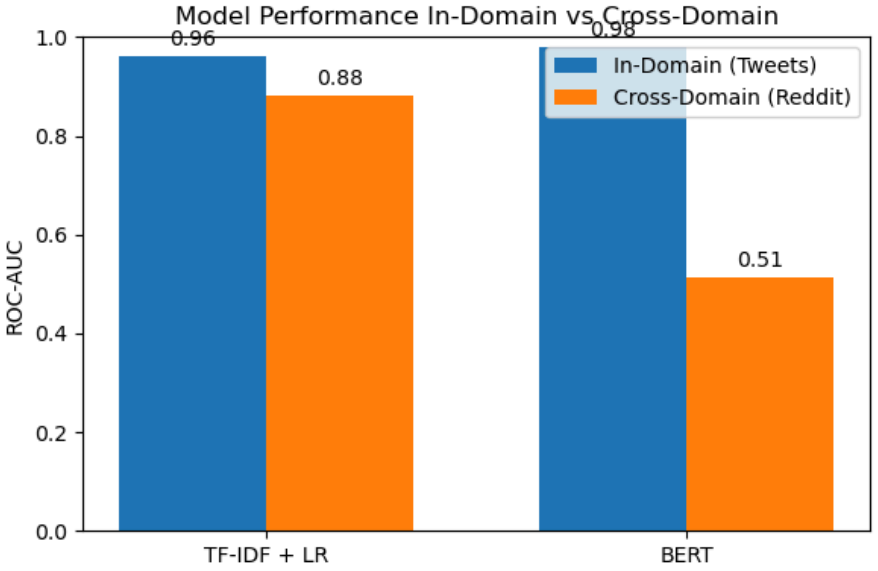

Figure 2: ROC-AUC: Tweets vs. Reddit (In-Domain vs. Cross-Domain)

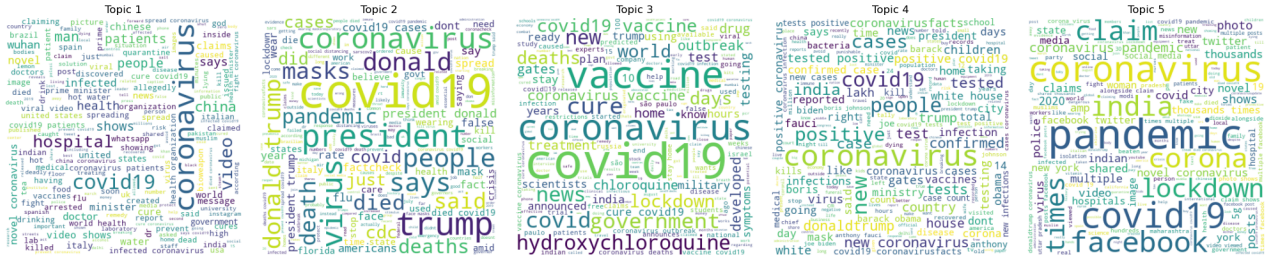

Figure 3: Topic Modelling Word Clouds.

```
Topic 1: coronavirus, covid19, video, hospital, china, man, patients, shows
Topic 2: covid19, trump, president, virus, coronavirus, says, people, donald
Topic 3: covid19, vaccine, coronavirus, cure, covid, government, news, new
Topic 4: coronavirus, new, people, covid19, positive, cases, india, case
Topic 5: pandemic, coronavirus, covid19, facebook, corona, india, claim, lockdown
```

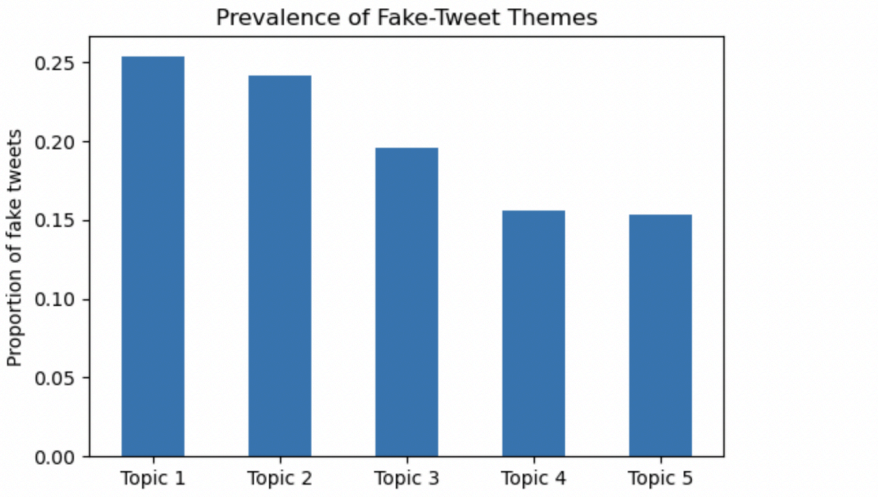

Figure 4: Fake-tweet themes

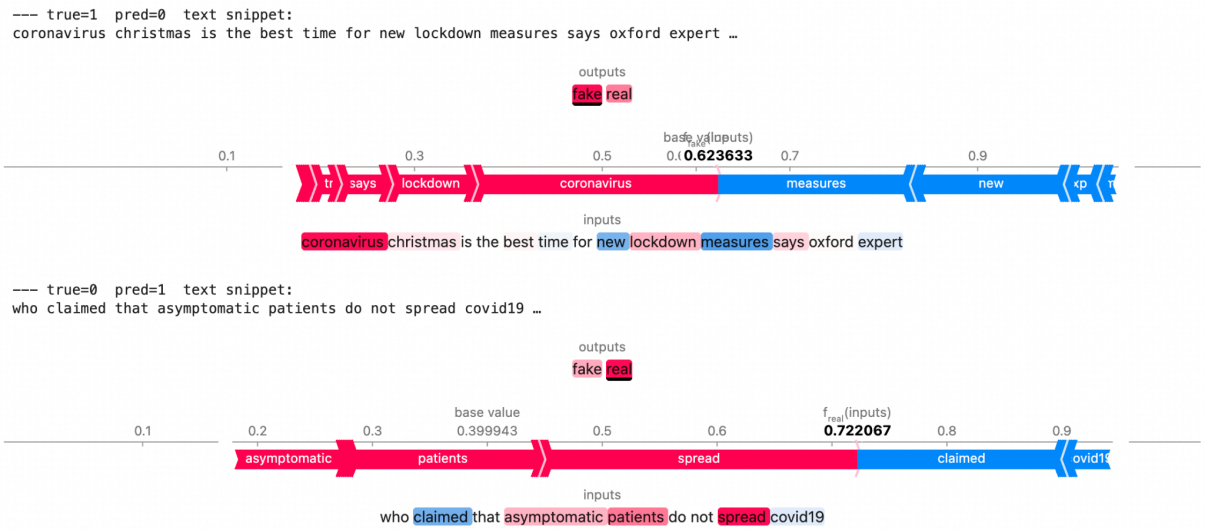

Figure 5: SHAP: Keywords driving predictions of fake or real COVID-19 content

Discussion

The results show that a simpler baseline model can achieve strong accuracy on labelled in-domain COVID-19 tweets. However, when transferring from Twitter to Reddit, the accuracy dropped significantly and the model became unstable. It highlights domain shift and label noise issues. Platforms such as Reddit often contain longer yet informal posts while Twitter's posts are relatively short. Switching domains made the model struggle on different writing patterns. Moreover, we used the high-confidence pseudo-labels from a Naïve Bayes rule ensemble which introduces biases and mistakes. To address these issues, further works are expected such as hand-labelling reddit samples and fine-tuning the model to correct errors.

Conclusion

In this research, we examined and evaluated two techniques for detecting COVID-19 health misinformation: the TF-IDF + Logistic Regression baseline and a fine-tuned BERT classifier. After training and testing the model, we observed high in-domain performance for both

models but low accuracy and stability when applying to Reddit posts. These results demonstrate that models trained on one platform on social media are not reliable cross-domains. To better protect elders, future works are expected to correct pseudo-label errors and make high accuracy classifications. After that, we can comfortably deploy it to real-world systems to flag fake health misinformation accordingly.

Acknowledgements

We thank the STEM Fellowship for dataset access and the support from the University of Toronto and University of Calgary.

Reference

- Statistics Canada. A portrait of Canada's growing population aged 85 and older from the 2021 Census [Internet]. Statistics Canada. 2022. Available from: https://www12.statcan.gc.ca/census-recensement/2021/as-sa/98-200-X/2021004/98-200-X2021004-eng.cfm
- Zhao Y, Da J, Yan J. Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches. Information Processing & Management. 2021 Jan;58(1):102390.
- Pérez-Rosas V, Kleinberg B, Lefevre A, Mihalcea R. Automatic Detection of Fake News [Internet]. ACLWeb. Santa Fe, New Mexico, USA: Association for Computational Linguistics; 2018. p. 3391–401. Available from: https://aclanthology.org/C18-1287/
- COVID19 Fake News Dataset NLP [Internet]. www.kaggle.com. Available from: https://www.kaggle.com/datasets/elvinagammed/covid19-fake-news-dataset-nlp
- Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Internet]. ArXiv. 2018. Available from: https://arxiv.org/abs/1810.04805
- 1.Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions [Internet]. Vol. 30, Neural Information Processing Systems. Curran Associates, Inc.; 2017. Available from: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html