10-29-2022

COVID-19 and Health Misinformation: A Topology and Classification Model

Archana Shinde
University of Massachusetts Boston, archana.shinde001@umb.edu

Romilla Syed
University of Massachusetts Boston, romilla.syed@umb.edu

Follow this and additional works at: https://aisel.aisnet.org/neais2022

COVID-19 And Health Misinformation: A Topology And Classification Model

Research-in-Progress

Archana Shinde
University of Massachusetts Boston
Archana.Shinde001@umb.edu

Romilla Syed
University of Massachusetts Boston
Romilla.syed@umb.edu

ABSTRACT

Misinformation – content that lacks truth, but the motivation of falsehood is uncertain – on social media during a health pandemic presents a major concern for public health. Recently, the vast volume of news and information around COVID-19, which the World Health Organization refers to as "Infodemic," has led to an unprecedented increase of health misinformation woven into the online narrative about the pandemic. This misinformation on social media presents a significant threat to the public and society. Several fact-checking organizations worldwide are devoting their efforts to publishing an evidence-based analysis of online narratives to convince audiences of their inauthenticity. However, there is a lack of research on how this misinformation could potentially impact individuals' behavior. To that end, in this study, we develop a topology of health misinformation to understand individuals' health behaviors. We collected COVID-19 misinformation data from six fact-checking websites (snopes.com, politifact.com, factcheck.org, leadstories.com, factcheck.afp.com, and poynter.com) for a duration of eight months (January 2020 to August 2020). Building on the Health Belief Model, we analyzed over 5K fact-checked articles shared on social media platforms to identify different categories or topics of misinformation. We also analyzed the veracity of the misinformation topics. We wrote Python scripts using python libraries named Beautiful Soup for web crawling and Pandas for data analysis. Overall, thirteen topics emerged from our analysis, with most of the misinformation questioning the benefits of preventive actions such as masking or social distancing and undermining the severity of the pandemic. We also found a significant amount of misinformation related to sources such as the government, health agencies, and institutes communicating about the pandemic.

Furthermore, we utilized the thirteen misinformation topics for training a classifier that predicts the misinformation category for any new data. The aim of this multiclass classifier is to classify the misinformation topics and predict topic labels for any new data. We trained three classification models (Logistic regression, K-nearest Neighbor, and Multiclass Support Vector Machine) that were selected based on an intense literature review. All the classifiers were trained by train- test split method, which splits the dataset into three subsets 1] training dataset used to train or fit the classification model. 2] validation dataset used for unbiased model evaluation during the hyper-parameter tuning. 3] testing dataset used to evaluate the final model fitting the feature sets generated. We applied the four evaluation metrics - Accuracy, Precision, Recall, and F1-Score to assess the performance of the models. The evaluation results suggest that the Multiclass Support Vector Machine (MSVM)-based classifier achieved high performance for accuracy (88%), precision (85%), recall (83%), and F-measure (82%). We used the scikit-learn implementations of the best parameter settings and the Grid-search algorithm with k-fold cross-validation to ensure a less biased model. The findings have implications for social media and health research. Public health experts and policymakers might find insights helpful in designing better communication and intervention strategies to counter the false narrative about the pandemic. The study lays the ground to examine further individuals' health attitudes and behavior upon exposure to misinformation.