COSC 519 Project Proposal

Problem statement

1.What problem are you trying to solve? A description of the problem

The Research Questions that we intend to solve:

1. How do social media posts correlate with vaccination rates?
2. How different is the sentiment about vaccination by region?
3. How accurately can the vaccination rate be predicted from social media data of a particular region?

The problem we are trying to solve is how posts on social media may impact or affect public health initiatives by influencing those who interact with these posts.

Motivation

2.Why is it a problem? Why is it important to solve it? Why is it novel?

This problem is relevant due to widespread reports of misinformation posted regarding the current and ongoing COVID-19 Pandemic. It is important to know whether or not there is a correlation between these online posts and public health metrics like vaccination rates in order to understand how drastic the impact of social media is on initiatives like public health. This is a novel problem because Covid-19 is still a fairly new disease and there is lots of new unverified information coming out about it every day. Furthermore, The effect of social media on vaccination rates will help future researchers to gain a general idea of the effect of social media in a public health crisis.

3. Solving the Problem & Desired Information

We plan to solve this problem by mapping post sentiments with time stamps. This would be implemented with sentiment analysis and creating a geographical network between posts that discuss the COVID-19 Vaccine. We hope to ascertain public sentiments regarding the COVID-19 Vaccine by geographical region from this data with which we can compare vaccination rates by geographical location. This will allow us to cross-examine the effect of social media sentiments in regards to the vaccine versus actual vaccination rates by location.

Data Collection

4.What type of data will you require? Where will you get data from?

We will require the COVID vaccination rates by country or region, the country/region of the social media poster, and the overall sentiment of the social media posts. We will access the COVID vaccination rates from the World Health Organization website. For the other two data pieces, we will access those from Reddit using an API already built. For further data, we intend to analyze other social media websites from where we can extract data(such as Facebook, Instagram, Twitter, Youtube).

5.What type of analysis will you carry out?

Data Analysis Strategy

We intend to 3 different types of analysis

1. Sentiment analysis and grouping regions by sentiment
2. Social network analysis (Visualization based on centrality measures and finding different communities)
3. Predicting the vaccination status based on Social Media data

For Social network analysis, we plan on using Network tools such as Gephi, NodeXL

For extracting features we plan to use Natural Language Processing Techniques such as Tokenization, Word Embeddings. We will also extract additional features depending on the data we extract.

We primarily intend deep neural networks(CNN, LSTM, Transformer model, GRU) to perform sentiment analysis and prediction [1]. Furthermore, for better accuracy, we can use pre-trained models for sentiment analysis such as BERT (Bidirectional Encoder Representations from Transformers).

We will also use traditional Machine learning classifiers we would have SVM, Random Forest, Naive Bayes, Decision Tree. We plan on using Rapid Miner and also Sklearn libraries.

Results and Model Evaluation

6.How do you plan to evaluate your results or get feedback from others? What methods do you plan to use (Interviews, usability testing, on-site observation, observing yourself, listening to people) Who will evaluate your application? How many people?

For quantitative performance [2] we will measure:

1. F1 score, 2. Accuracy, 3. Precision, 4. Recall, 5. Mean Squared Error

We will compare these metrics among the models used in our research. Finally, we will compare our best model with researchers who have done similar work.

Our plan of attack to evaluate our results is to build a network with edges and nodes that connect certain countries/regions vaccination rates with social media posters. We plan on looking into the relationship between where the social media poster is from and the vaccination rates in that region. This network will give us access to other things learned in class such as the betweenness, closeness, etc… Which will advance our results section even further.

Based on these results we will design an interview or questionnaire to get people's feedback on their thoughts on how social media, in general, can affect things in their day to day lives. We won't tell them the results we have seen through our study but just get an overall opinion on how it affects them.

If we end up building a questionnaire we will simply post it to Amazon Mechanical Turk which allows hundreds of participants to participate in it. If we design an interview we will have real-life interviews either in person or over Zoom with some participants. As for the number of people, if we do end up posting to Amazon Mechanical Turk we can expect anywhere from 100-1000 participants. For the real-life interviews, we will be aiming to get 5-10 participants.

7. Group Roles

Each group member has a duty to complete their fair share of work. Thus, we have decided to divide group roles according to our individual strengths. The roles are as follows:

1. Conceptualization: Amanat , Karanmeet
2. Methodology: Amanat , Zach
3. Data processing: Karanmeet, Amanat, Will, Zach
4. Implementation: Will, Karanmeet, Zach, Amanat
5. Validation: Will, Karanmeet
6. Writing and editing: Will, Karanmeet
7. Background study: Zach, Will

Each person is responsible for maintaining their above role through the entirety of the project unless special circumstances arise and the rest of the group approves of the recommended changes.

8. Proposed work plan with dates

Week	To Do	Deadline
Oct 11	Meet as a group and finalize the proposal and split up work on it	Project proposal (October 14)
Oct 18	Gathering data on what we are going to analyze on social media sites Meet as a group	
Oct 25	Visualize the dataset (charts, graphs etc…) Meet as a group	
Nov 1	Background study (review what others have done) Meet as a group	
Nov 8	Write the mid-project milestone report Finalize our working plan Meet as a group	
Nov 15	Start on our working plan Meet as a group	Mid-project milestone report (Nov 15)
Nov 22	Start writing the report and get to work on the project presentation Finish our working plan Meet as a group	
Nov 29	Finalize presentation and finish off the report Meet as a group	Project presentation / report (Nov 30 & Dec 2)
Dec 6	Finalize report	Project report (Dec 7)

Reference:

1. Yadav, A., Vishwakarma, D.K. Sentiment analysis using deep learning architectures: a review. *Artif Intell Rev* 53, 4335−4385 (2020). https://doi.org/10.1007/s10462-019-09794-5
2. Odiete, O., Jain, T., Adaji, I., Vassileva, J., & Deters, R. (2017, July). Recommending programming languages by identifying skill gaps using analysis of experts. a study of stack overflow. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization* (pp. 159-164). https://doi.org/10.1145/3099023.3099040