PROJECT PROPOSAL

Group 7

Research Question:

Can Europeans sentiment towards vaccination be quantified by analyzing threads scraped within Reddit subforums?

03.01.2021

Introduction

As the majority of populations in countries around the world turn to online networks as a medium for information, individuals have the power to efficiently create and proliferate ideas at scale. Thus, sentiment analysis is now a vital tool for capturing public opinion about political trends, socio-economic and societal changes. News corporations monitor political trends and now governments and the healthcare industry need to monitor attitudes towards medical innovation. Our project aims to investigate the lively discussions on the popular online network Reddit, concerning vaccine sentiment specifically. In the fightback against the COVID-19 pandemic, vaccination programs worldwide are beginning. The UK Government, which is well underway in its vaccination program, will define a successful vaccination campaign if uptake reaches 75% of the population. In care homes and over 75's, the uptake has been at 90%. However, as we move through the population, they see BAME communities and healthcare workers experience lower uptakes (Department of Health & Social Care, 2021). A study by the BMJ outlined that an NHS trust in England noted Black Caribbean/African and Bangladeshi/Pakistani healthcare workers were half as likely to take a vaccine as their other ethnic counterparts (Razai, Osama, McKechnie and Majeed, 2021). In Europe, the picture is more worrisome, with online anti vaccination sentiment promoting the idea that the vaccine is not safe, and that immunization is part of a broader business strategy (Porreca, Scozzari and Di Nicola, 2020).

There is precedent for studying public perception of vaccines using sentiment analysis. In a paper published by Reghupathi, et al. in 2020, natural language processing was used to explore sentiment trends on vaccinations on the social media platform Twitter (US National Library of Medicine National Institutes of Health, 2021). Over four months in 2019, as the vaccination debate took place in the United States, Reghupathi et al. scraped over 9581 tweets with the keyword "vaccine" globally from the site (US National Library of Medicine National Institutes of Health, 2021). The NLTK and the sentiment reasoner VADER analyzed each sentence and determined a sentiment score based on whether it conveys a positive, negative or neutral sentiment (US National Library of Medicine National Institutes of Health, 2021). The results highlighted that over half the tweets analyzed in the sample expressed negative sentiment towards vaccination, highlighting the scale of opinion online (US National Library of Medicine National Institutes of Health, 2021).

However, unlike Twitter and Facebook, Reddit posts and subforums are less regulated than other social media platforms. For this project's purposes, we can see more representative insight into opinion on vaccines. Furthermore, previous studies have used Reddit posts as training data for well-known NLP models (Kerrigan, Slack and Tuyls, 2020). Work by Kerrigan, G.et.al 2020 demonstrated how OpenAI's pre-trained GPT-2 was fine-tuned on a Reddit components data set of 10,000 comments. (Kerrigan, Slack and Tuyls, 2020).

Methodology

a. Data Gathering

Data will be obtained using Python's Wrapper Reddit API: PRAW.[1] An account in Reddit to gain access to the API will be created. Moreover, potential subreddit forums for data extraction include:

- r/CovidVaccine
- r/Information about vaccine safety and combating anti-science rhetoric
- r/$100% True Stories from the Anti-vaxx Crowd
- r/Novel Coronavirus (COVID-19)

b. Data Cleaning

Data will be cleaned by removing:

- Leading and trailing white spaces
- Punctuation and numbers
- Stop words (aided by NLTK)

Words will be transformed into lower case and stemming, and tokenization will be carried out (using CountVectorizer). These transformations will enable to construct vectors which will be fed to an algorithm.

c. Creating Bag of Words

Word's importance will be measured using measures such as *td_idf(t,d) = wc(t,d)/wc(d) / dc(t)/dc() (Ismiguzel, 2020).*

d. Labelling

Each team member is expected to label 100-150 sentence labels regarding people's sentiment to receive a vaccine as -1 (negative), 0 (neutral) or 1 (positive). Labelling will be done using Excel (Lee, 2020).

e. The Model

We consider TextBlob as a starting point. TextBlob measures polarity (how positive or negative it is) and subjectivity (if it's closer to opinion or ground truth) on the labels. TextBlob will assign a score to each sentence ranging from -1 to 1 (Tran, 2020). Data will be split between train, test and validation (part of training for hyperparameter tuning). Then, GridSearch will be used to reduce error metrics to arrive at an optimal model possibly using Logistic Regression (Selvaraj, 2020).

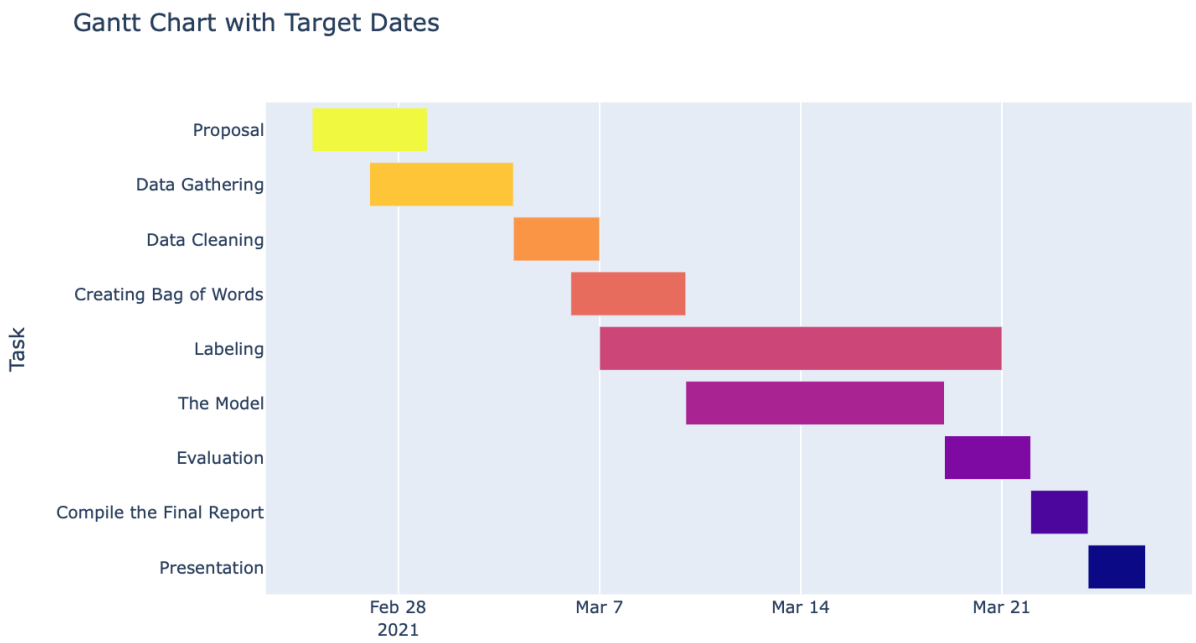
f. Evaluation

Based on the model's performance, the project scope might be expanded to evaluate a longer time span regarding vaccine sentiment.

[1] PRAW: The Python Reddit API Wrapper: https://praw.readthedocs.io/en/latest/

Project Plan

We have outlined the tasks required to successfully complete this project above. Below, using Python and Plotly we have also visualized the tasks as well as the timeline that we expect to follow as we track our progress.



4. Key resources (datasets, models, sample code, and pre-trained weights)

As mentioned in the Methodology, we will use The Python's Reddit API Wrapper (PRAW) to extract data from subreddits. It is a python package that allows for simple access to Reddit's API. With PRAW, not only can we obtain the titles of the posts, but we can also get the body, number of comments, as well as the score of the posts. After collecting the data, we will then perform data cleaning and labelling in order to train the model.

To demonstrate how we can obtain the data through PRAW, we have created the sample code (see Appendix). The code extracts the newest ten posts created in the subreddit "r/CovidVaccine". The PRAW model returns a list of posts fetched from the subreddit. We then create a Pandas DataFrame to store the posts for further executions.

5. Conclusion

The sentiment analysis that our project will utilize, aims to be a litmus test for vaccine opinion throughout multiple Reddit subforums. We aim to help governments, NGOs, and marketing firms understand and contextualize the type of hesitancy being promoted and quantify the ratio of positive to negative sentiment around the COVID-19 vaccines online.

References

1. Department of Health & Social Care, 2021. *https://www.gov.uk/government/publications/covid-19-vaccination-uptake-plan/uk-covid-19-vaccine-uptake-plan*.
2. Ismiguzel, I., 2020. *Applying Text Classification using Logistic Regression: A comparison between BoW and Tf-Idf*. [online] Medium. Available at: <https://medium.com/analytics-vidhya/applying-text-classification-using-logistic-regression-a-comparison-between-bow-and-tf-idf-1f1ed1b83640> [Accessed 28 February 2021].
3. Kerrigan, G., Slack, D. and Tuyls, J., 2020. *Differentially Private Language Models Benefit from Public Pre-training*. [online] Arxiv.org. Available at: <https://arxiv.org/pdf/2009.05886.pdf> [Accessed 28 February 2021].
4. Lee, I., 2020. *Data Labeling For Natural Language Processing*. [online] TOPBOTIvan LeeS. Available at: <https://www.topbots.com/data-labeling-for-natural-language-processing/> [Accessed 28 February 2021].
5. Porreca, A., Scozzari, F. and Di Nicola, M., 2020. Using text mining and sentiment analysis to analyse YouTube Italian videos concerning vaccination. *BMC Public Health*, 20(1).
6. Razai, M., Osama, T., McKechnie, D. and Majeed, A., 2021. Covid-19 vaccine hesitancy among ethnic minority groups. *BMJ*, p.n513.
7. Selvaraj, N., 2020. *A Beginner's Guide to Sentiment Analysis with Python*. [online] Medium. Available at: <https://towardsdatascience.com/a-beginners-guide-to-sentiment-analysis-in-python-95e354ea84f6> [Accessed 28 February 2021].
8. Tran, J., 2020. *NLP Sentiment Analysis for beginners.*. [online] Medium. Available at: <https://towardsdatascience.com/nlp-sentiment-analysis-for-beginners-e7897f976897> [Accessed 28 February 2021].
9. US National Library of Medicine National Institutes of Health, 2021. *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7277574/*.

Appendix

Sample Code of using PRAW:

```
import praw
import pandas as pd
from datetime import datetime
```

```
reddit = praw.Reddit(client_id=client_id, client_secret=client_secret, user_agent=user_agent)

wsb_sub = reddit.subreddit('CovidVaccine')

posts = []

for post in wsb_sub.new(limit=10):
    posts.append([post.title, post.score, post.id, post.subreddit, post.url, post.num_comments, \
                  post.selftext, post.created])

posts = pd.DataFrame(posts,columns=['title', 'score', 'id', 'subreddit', 'url', 'num_comments', 'body', 'created'])
posts['created'] = posts['created'].apply(lambda x: datetime.fromtimestamp(x).strftime('%Y-%m-%d'))
posts.head()
```

	title	score	id	subreddit	url	num_comments	body	created
0	Given vaccine by mistake	1	lufqrr	CovidVaccine	https://www.reddit.com/r/CovidVaccine/comments...	2	I am considered a government employee and my s...	2021-02-28
1	Nervous about side effects because of chronic ...	2	luarls	CovidVaccine	https://www.reddit.com/r/CovidVaccine/comments...	5	I finally got me and my family (all high risk,...	2021-02-28
2	odd vascular symptoms second day after vaccine	4	lu7ysu	CovidVaccine	https://i.redd.it/imgjxwkao5k61.jpg	7		2021-02-28
3	2nd Shot New State ? AZ / TX	3	lu6h7s	CovidVaccine	https://www.reddit.com/r/CovidVaccine/comments...	4	Hi All, what a pickle I'm in now. Have been i...	2021-02-28
4	Chest pain (like heartburn?) sustained after c...	3	lu6h3j	CovidVaccine	https://www.reddit.com/r/CovidVaccine/comments...	4	Hi! I got the first dose of the Pfizer covid v...	2021-02-28