DS-1011 Term Project Proposal: Textual Analysis of Communications in COVID-19 Infected Community on Social Media

Yuhan Liu, Yuhan Gao, Zhifan Nan, Long Chen
New York University
{yl7576,yg2417,zn2041,lc3424}@nyu.edu

1 Introduction

Starting in late 2019, the COVID-19 pandemic has rapidly impacted over 200 countries, areas, and territories. As of September 4, according to the World Health Organization (WHO), 26,121,999 COVID-19 cases were confirmed worldwide, with 864,618 confirmed deaths[1]. This disease has tremendous impacts on people's daily lives worldwide.

With the pandemic spreading in the United States, people who tested positive started sharing information about their physical condition, emotion and story with the virus. In addition, those who have not gotten infected are curious about the symptoms and nature of the virus, as well as procedures of testing services across the country. A community of those who want to share their own story and who want to know more about the virus emerged on Reddit, a platform for any user (older than 13 years) to discuss, connect, and share their experiences and opinions online. Under subreddit *r/COVID19positive*, people are sharing and discussing the virus, while seeking and giving emotional supports. An online community like this can have mixed emotions and splendid textual contents.

In this proposed study, we would like to investigate the linguistic features of contents in the subreddit. First, we classify the threads into different categories, including a) self-reporting of positive COVID-19 case, b) reporting of COVID-19 case of family and friends, c) question to those who tested positive, and d) general medical questions for COVID-19. Second, we aim to investigate linguistic characteristics of posts and subsequent comments in different contexts. Specifically, we try to detect differences in contents when people are posting for different reasons, or under different topics.

[1]https://www.who.int/docs/default-source/coronaviruse/situation-reports/wou-4-september-2020-approved.pdf?sfvrsn=91215c78_2

2 Related Work

A large number of studies were performed with LIWC, an API[2] for linguistic analysis of documents. Tumasjan et al. (2010) used LIWC to capture the political sentiment and predict elections with Twitter. The API was also used by Zhang et al. (2020) to provide insights into the sentiment of the descriptions of crowdfunding campaigns.

Previous studies have also attempted to make textual classification on social media data. Mouthami et al. (2013) implemented a classification model that approximately classifies the sentiment using Bag of words in Support Vector Machine (SVM) algorithm. Huang et al. (2014) applied SMOTE (Synthetic Minority Oversampling TEchnique) method to defecting online cyber-bullying behavior. In addition, a number of other studies performed textual classifications for various purposes using social media data (Chen et al., 2020; Chatzakou and Vakali, 2015; Lukasik et al., 2016).

3 Dataset

Data from subreddit *r/COVID19positive* between March 14, 2020[3] and October 14, 2020 is collected using Pushshift API [4]. In total, 17,285 submissions (contents that starts a Reddit thread) and 227,019 comments (contents that follows after submissions) were collected. As a medium-sized subreddit with 91.1K members[5], contents in this community should contain limited fake posts or misinformation, therefore leading to a relatively clean dataset. The details of the two types of data is discussed below.

[2]https://liwc.wpengine.com/
[3]The date when the subreddit was created.
[4]https://www.pushshift.io.
[5]As of October 14, 2020.

3.1 Submissions

Submission on Reddit starts a discussion with a title and an optional textual body. The title and the body are naturally good source for textual analysis. In addition, most submissions have *flair*, a hashtag-like label that describes the category of discussion under which the submission is about. The *flairs* serve as a perfect label for potential supervised classification tasks.

3.2 Comments

Comment on Reddit follows a submission with a textual body. Comments can be linked to a submission, serving as contents that are somehow related to the thread. Since the number of comments are significantly more than that of submissions, we can use comments as a means to understand reaction and community interaction towards different categories of submissions.

4 Experiment Design and Methods

We planned two natural language processing tasks in order to make classifications of different submissions, and to investigate sentiments and linguistic characteristics of Reddit threads, as more details and methodologies explained below.

4.1 Task 1: Classification

Thanks to the Reddit-exclusive *flair* system, we have self-reported labels for most submissions [6]. Most submissions have a flair of 10 most frequently used ones that consists of 98.6% of our dataset. The 10 flairs can roughly be classified into 4 categories: a) self-reporting of positive COVID-19 case, b) reporting of COVID-19 case of family and friends, c) question to those who tested positive, and d) general medical questions for COVID-19. Such flairs serve as perfect labels for classification models, motivating us to attempt a classification task with textual features.

A number of models were proposed to be used. First, we would like to build a baseline model with RNN with Word2Vec embeddings. Next, Bidirectional Long Short Term Memory (Bi-LSTM) model with Word2Vec is planned as a more advanced RNN model. In addition, we will attempt SOTA pre-trained language models with fine tuning, including BERT and XL-Net, aiming for better results. However, such models may struggle with limited corpus size of this study, as Ezen-Can points out in study (Ezen-Can, 2020).

4.2 Task 2: Analysis of Linguistic Characteristics

Linguistic Inquiry and Word Count (LIWC2015) is applied to extract the sentiment of submissions and comment of our corpus. LIWC2015 is a dictionary-based linguistic analysis tool that can count the percentage of words that reflect different emotions, thinking styles, social concerns, and capture people's psychological states[7]. We focus on 4 summary linguistic variables and 12 more detailed variables that reflect psychological states, cognition, drives, time orientation, and personal concerns of the Twitter users of both groups. We follow the similar methodology used by Yu et al. (2008) which concatenates textual contents of same category for a combined analysis. In addition, VADER Sentiment Analyzer will be used to investigate average sentiment score of submissions and comments. In this exploratory task, we aim to find discrepancies of texts with topics in different categories.

5 Work Distribution

The work distribution of this project is as the following:

- LC: Data retrieval and pre-processing, Task 2
- YL: Task 2
- YG: Task 1, RNN and Bi-LSTM classifiers
- ZN: Task 1, BERT and XL-Net

In addition, all members will work on write-ups and posters.

T

References

Despoina Chatzakou and Athena Vakali. 2015. Harvesting opinions and emotions from social media textual resources. *IEEE Internet Computing*, 19(4):46–50.

Long Chen, Hanjia Lyu, Tongyu Yang, Yu Wang, and Jiebo Luo. 2020. In the eyes of the beholder: Sentiment and topic analyses on social media use of neutral and controversial terms for covid-19. *arXiv preprint arXiv:2004.10225*.

Aysu Ezen-Can. 2020. A comparison of lstm and bert for small corpus. *arXiv preprint arXiv:2009.05451*.

[6] 15,548 out of 17,285 submissions have a flair apiece.

[7] https://liwc.wpengine.com/how-it-works/

Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. 2014. Cyber bullying detection using social and textual analysis. In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, pages 3–6.

Michal Lukasik, PK Srijith, Duy Vu, Kalina Bontcheva, Arkaitz Zubiaga, and Trevor Cohn. 2016. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 393–398.

K Mouthami, K Nirmala Devi, and V Murali Bhaskaran. 2013. Sentiment analysis and classification based on textual reviews. In *2013 international conference on Information communication and embedded systems (ICICES)*, pages 271–276. IEEE.

Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*.

Bei Yu, Stefan Kaufmann, and Daniel Diermeier. 2008. Exploring the characteristics of opinion expressions for political opinion classification. In *Proceedings of the 2008 international conference on Digital government research*, pages 82–91. Digital Government Society of North America.

Xupin Zhang, Hanjia Lyu, and Jiebo Luo. 2020. What contributes to a crowdfunding campaign's success? evidence and analyses from gofundme data. *arXiv preprint arXiv:2001.05446*.