Jing Li (S5374553)

A Report on Extract Keywords and Word Embedding for COVID-19 Related Tweets

This report presents the findings of an analysis of a dataset of tweets in several languages about COVID-19. The dataset has 46 languages, with English being the most common with 14849 messages. Other common languages in the dataset include Spanish, which has 5538 messages, and French, which has 1193 messages. Additional examination of the dataset reveals that the top five most frequently used hashtags in messages are "COVID19", "coronavirus", "Coronavirus", "Covid19", and "COVID19".

Extract Keywords

To gain a deeper understanding of messages in the most commonly used language, English, I used tokenization and TF-IDF analysis to extract the top five keywords. I started by removing URLs, punctuation, and stop words from the text data. The TfidfVectorizer was then used to compute the TF-IDF values for the pre-processed texts.

The resulting TF-IDF matrix was converted into a dictionary, and the top five words with the highest TF-IDF scores were extracted. The analysis revealed that the most frequent keyword in the English messages was "coronavirus" with a TF-IDF score of 0.878, followed by "covid19" with a score of 0.279. Other keywords included "case", "covid", and "trump". These results supported the widespread concern and discussion on social media about the COVID-19 pandemic.

This analysis provides a snapshot of the content and themes present in the dataset. These insights might inform further research and analysis on social media trends and public discourse.

Word Embedding

Word embeddings are a well-known technique for capturing semantic similarity between words and their context inside a corpus. In this section, I used Word2Vec to train word embeddings on four different samples of COVID-19-related tweets and examine the top 10 words that are comparable to the words "covid", "coronavirus", and “virus".

I first preprocessed the English-only tweets. I removed URLs and usernames with placeholders, and reduced words to their most basic form, also stopwords were removed. And I trained four different Word2Vec models with varying sample sizes of 1000, 1500,

2000, and 2500 tweets respectively. For each of the four samples, I computed the top 10 similar words for the words "covid", "coronavirus", and "virus".

	covid	coronavirus	virus
Sample1	mid, ecoin, register, slated, world, please, way, corona, fools, results	url, cases, user, england, th, said, country, update, year, outbreak	wuhan, called, name, hubei, chinese, market, bed, make, hoax, focus
Sample2	mid, fools, fighting, pandemic, may, via, third, lt, ecoin, two	breaking, first, update, via, pm, url, today, reported, uk, people	wuhan, chinese, called, epicenter, corona, hubei, racist, make, funky, came
Sample3	two, due, usa, impact, cases, spread, effect, update, deadly, help	outbreak, update, reported, confirmed, health, breaking, case, death, total, via	china, wuhan, outbreak, spread, epicenter, called, cases, country, face, deadly
Sample4	world, pandemic, may, virus, people, via, life, help, time, crisis	outbreak, update, confirmed, via, news, case, death, reported, pandemic, quarantine	china, wuhan, outbreak, cases, epicenter, spread, deadly, mask, face, corona

Table 1 Top 10 similar words with “covid” “coronavirus” and “virus”

The result (Table 1) shows that related words for "covid," "coronavirus," and "virus" are generally consistent across sample sizes, with some variation in the exact words discovered. For example, in all samples, "wuhan" is consistently selected as a similar word for "virus," while other related words vary significantly depending on sample size.

And for the last step, I used the pre-trained GloVe Twitter model of 50 dimensions to find the most similar words to the words “covid”, “coronavirus” and “virus”. Comparing the outputs, we can see that words listed in the top 10 most similar words are different for each model. For example, in the Word2Vec output, "viral" was listed as the most similar word to "coronavirus", while in the GloVe output, "wikileaks" was listed as the most similar word. The word “covid” is even not present in the vocabulary.