Fake News Lab

Dang Pham and Kairuo Yan

Abstract— **Covid-19 myths are widely discussed. In this research, we would like to explore the most discussed topics and keywords using LDA (Latent Dirichlet Allocation) and text summarization. We also plotted a heatmap and bar graph to help us better understand the most popular topics discussed during Covid. As a result, we found the most popular topics are individuals, government, conspiracies, spread, and lockdown. However, our study is also limited, and we could improve our analytical methods to understand topic popularity better.**

I. INTRODUCTION

Since the 2020 Covid-19 pandemic, there are some public discussions of Covid-19 myths. In this lab, we look at how the frequencies of Covid myths topic is discussed. In our expectations, people are more likely to discuss topics involving medical care, prevention methods, and personal experiences. To find the answer, we took three significant steps. In the beginning, we plotted a heatmap to explore the relationships among topics, and we compared the topics' relationships for true and false statements. Then we performed Latent Dirichlet Allocation (LDA) to modeling the most discussed topics and keywords. Finally, we summarized the text we got using the deep learning method to find the top discussed topics. As a result, we discovered that LDA is a better method to find the most discussed topics than deep learning. Second, we also found the most discussed topics are individuals, government, conspiracies, spread, and lockdown.

II. DATA

There are 2526 observations in our dataset and there are twenty-six variables, include index, summary, label, topic_info, aid, animals, conspiracies, detection, food, governments, hospitals, individuals, laws, lockdown, medical equipment, medicine, origins, other diseases, predictions, religion, risk factors, spread, symptoms, travel, vaccines and text.

Variables	Description
Time	The transaction duration
Amount	Transaction amount
Class	A label that 1 is fraudulent transactions and 0 is non-fraudulent case
V1-V28	28 different features about users and being concealed due to privacy issues

III. RESULTS

There are several findings in our study. First, we found there are twenty-one topics, and they are aid, animals, conspiracies, detection, food, governments, hospitals, individuals, laws, lockdown, medical equipment, medicine, origins, other diseases, predictions, religion, risk factors, spread, symptoms, travel, and vaccines. According to the waffle chart by topic categorization, the top 5 discussed topics that have the highest discussion are individuals, government, conspiracies, spread, and lockdown, which are 742, 664, 591, 407, and 277. The topics that are least discussed are animals, travel, and symptoms. Secondly, In the heat map that explores the relationship between topics, we found that topics mentioned together for most of the times are individuals and government, government and conspiracies, conspiracies and individuals, government and lockdown, and individuals spread. We found that the most mentioned topics are spread, individuals, medicine, risk factors, and predictions for the true statement. The least mentioned topics are religion, laws, animals, travel and aid. We found that the most mentioned topics are individuals, governments, conspiracies, spread and lockdown for the false statement. And the least mentioned topics are symptoms, risk factors, and animals. For the true statement, we found that topics that mentioned together for most of the times are risk factors and individuals, spread and individuals, and spread and detection. For the false statement, we found that topics that mentioned together for most of the times are individuals and governments, governments and conspiracies, and individuals and conspiracies. We found that there are more false statements than true statements, and true statements often related to the reality and facts, while the false discussion are involved with discussion for individuals and government, and conspiracies. Thirdly, we found the top five topics with highest count are 'conspiracies', 'governments', 'individuals', 'lockdown' and 'spread'. According to the word cloud and bar graph, the top five most searched words are "coronavirus", "covid", "19", "facebook" and "äu". Thus, the results for LDA analysis are successful. We believe that we should still need to rely on manual labeling to some extent since there are overlapping for the topics found in LDA analysis. Last but not least, we observed the summary of the post and the text. We found in our histograms, the word count for the text message is around 500 words, and most are less than 1000 words. For the summary, the word counts are mostly less than 25 words, and only a few summaries are more than 50 words. The accuracy for train and test datasets remained low as data sizes get bigger. As a result, we found text summarization is not very reliable. LDA(Latent Dirichlet Allocation) provides greater insight.

One possible way to improve the analysis is to use both the classification model and the text summarization to find the most discussed topics. One causal inference problem could be simultaneity, a particular keyword got popular when it got most discussed, but people also refer to the word more often because of its popularity.

IV. VISUALIZATION

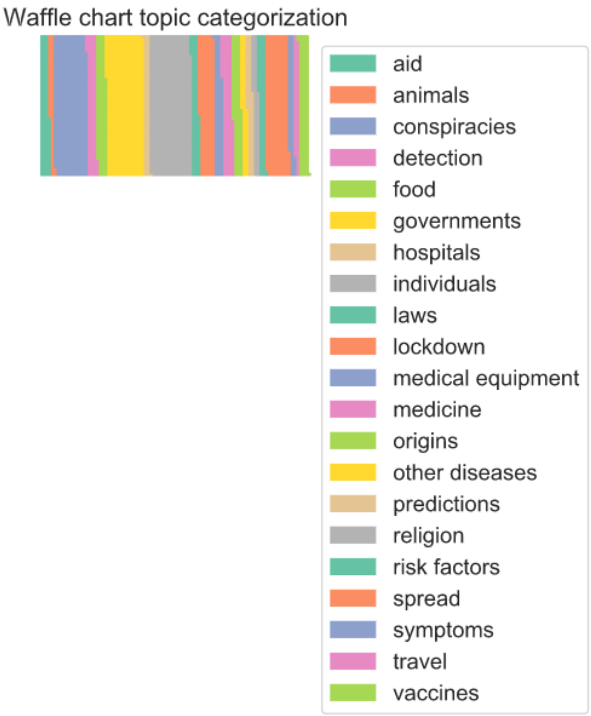

Fig. 1. Histogram of Amount distribution

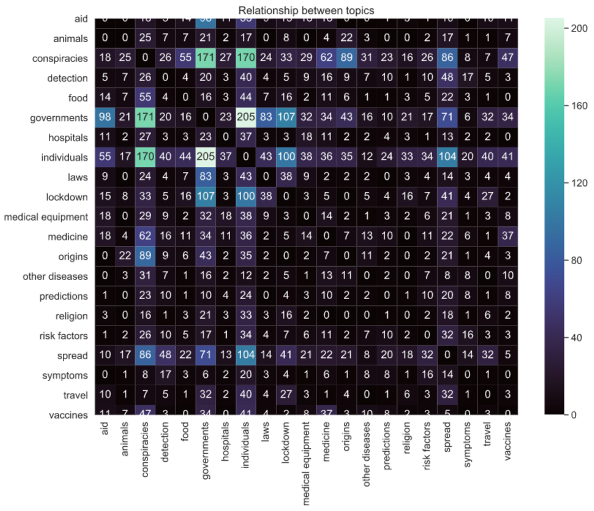

Fig. 2. Histogram of Time distribution

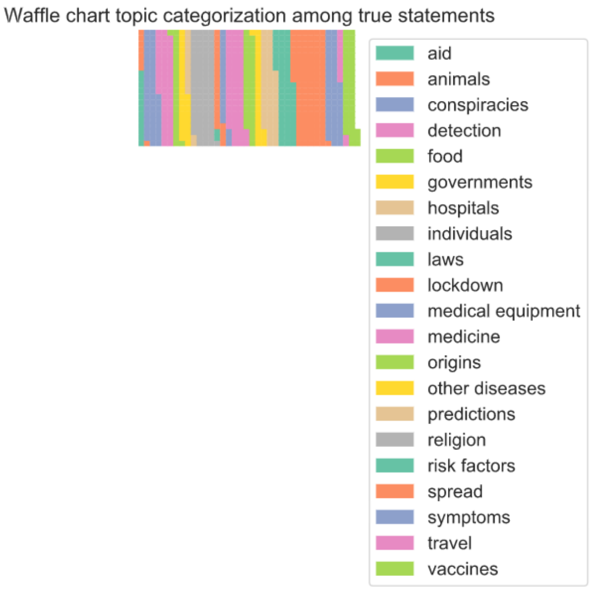

Fig. 3. Pairplot among five Vs variable by Amount and Time

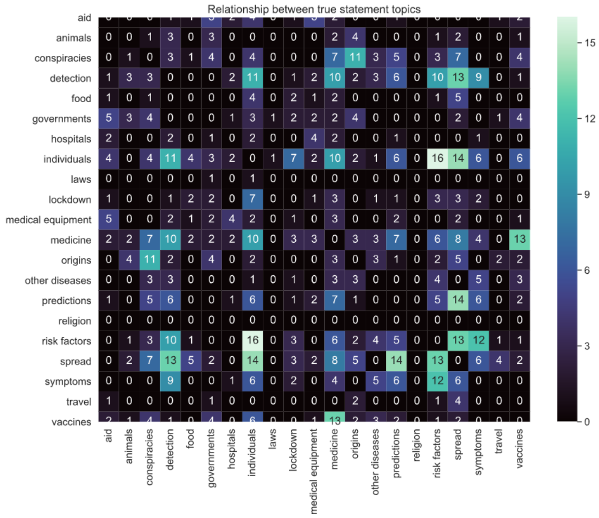

Fig. 4. Pairplot among five Vs variable by Amount and Time

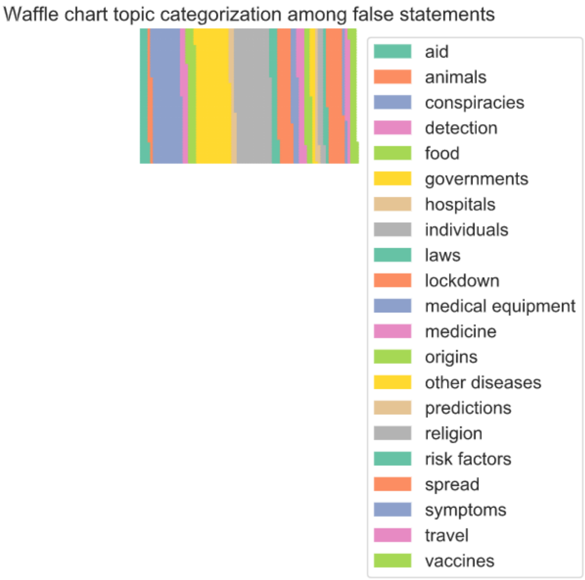

Fig. 5. Pairplot among five Vs variable by Amount and Time

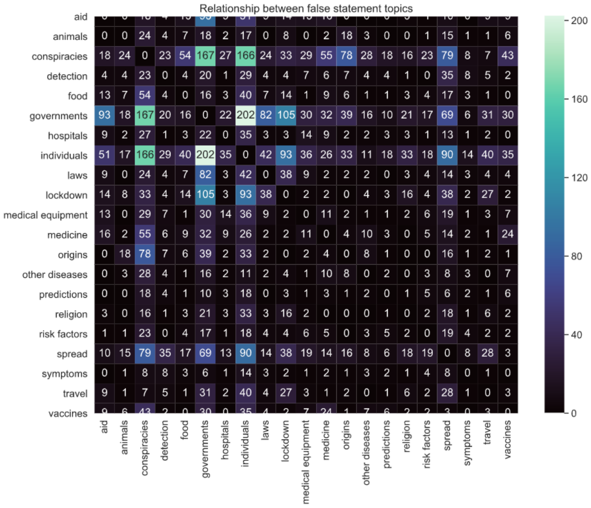

Fig. 6. Pairplot among five Vs variable by Amount and Time

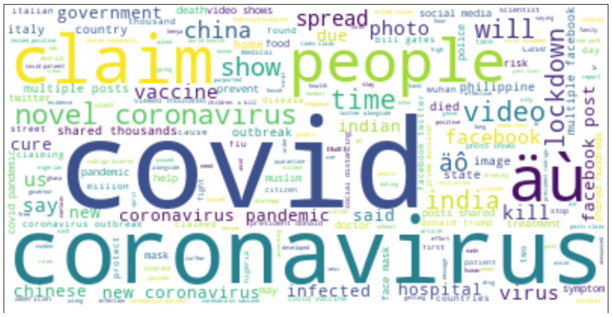

Fig. 7. Pairplot among five Vs variable by Amount and Time

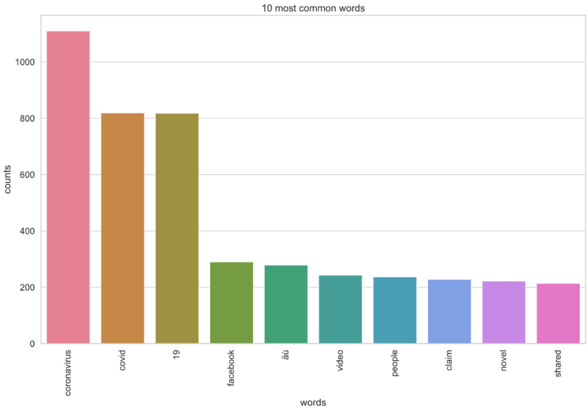

Fig. 8. Pairplot among five Vs variable by Amount and Time