INFSCI 2160

Progress Report

Group: Procrastinator Club

Group Member:

Shangbin Tang (sht105@pitt.edu)

Chengchen Wang (chw183@pitt.edu)

Contents

1. Introduction

Over the past few years, people's access to information has gradually shifted from traditional paper-based media, radio, and television to digital media and social media applications. At the same time, the information explosion also makes the information itself more and more fragmented. It's become more and more critical for people to distinguish and filter out those false ones, or they can easily plant wrong views in the public's mind.

Hence, in this project, we aim to figure out a method to detect and classify false and true information—especially those on social media.

We used Python programming and a web scraping package, *Beautiful Soup*, to get all the fact-checked messages on a famous political fact-check website, *PolitiFact(https://www.politifact.com/)*, of this year. After pre-processing, with r programming and libraries of *tidyverse, tidytext, repr, stringr, wordcloud, reshape2*, and *tm*, we split the content into specific words to generate the document term matrix. Then we converted all content to lower case, punctuation, special characters, and stopwords. The following exploratory analysis, including nomogram, unigram and other statistics analysis, yielded some informative and interesting results, such as the differences of hot words, content length distributions among true, half-true, and false information, etc.

In the coming stage, we will use logistic regression, random forests, and the AdaBoost method to perform supervised classification then evaluate, compare the results.

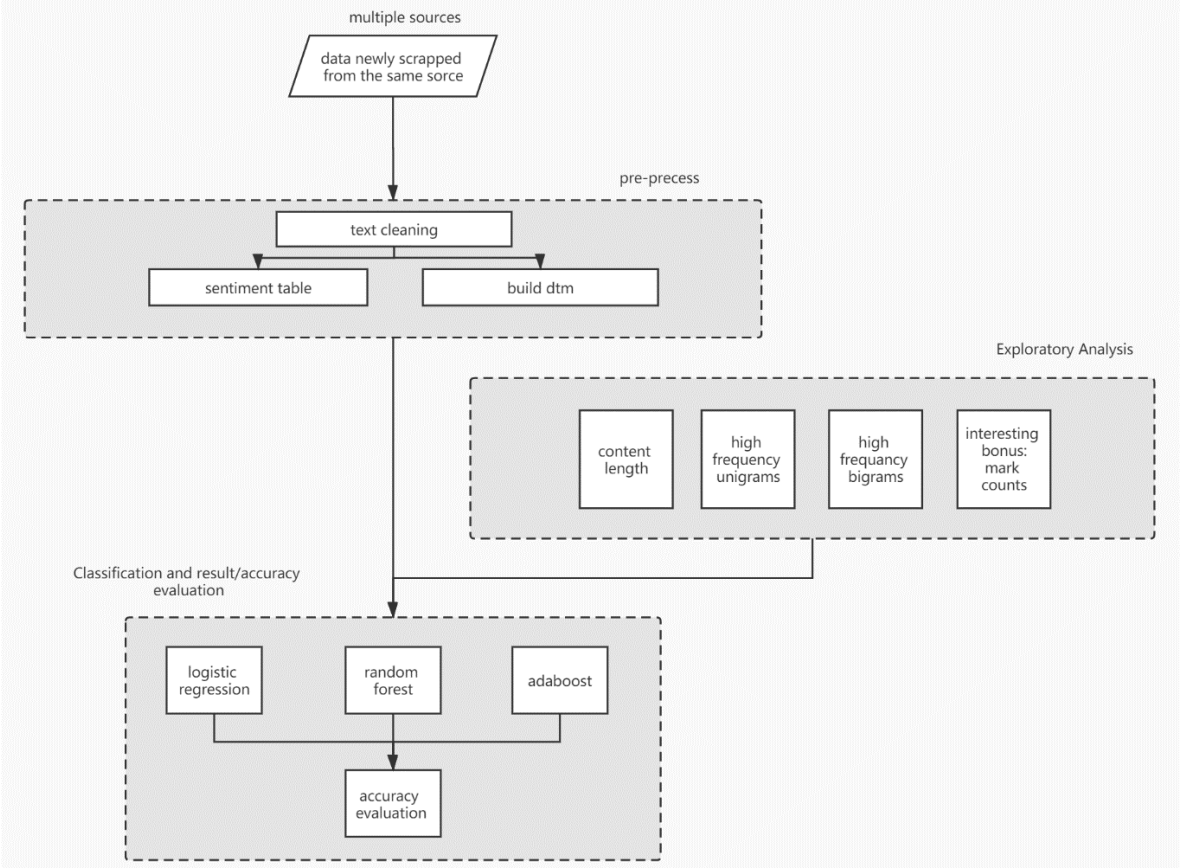

Fig.1. The workflow of this project.

2. Preliminary analysis results

After the preliminary analysis, we noticed that some social media platforms, like Facebook and web blogs, are the primary sources of false information, as well as the viral images, or memes, that have risen in these years(see Fig.2).

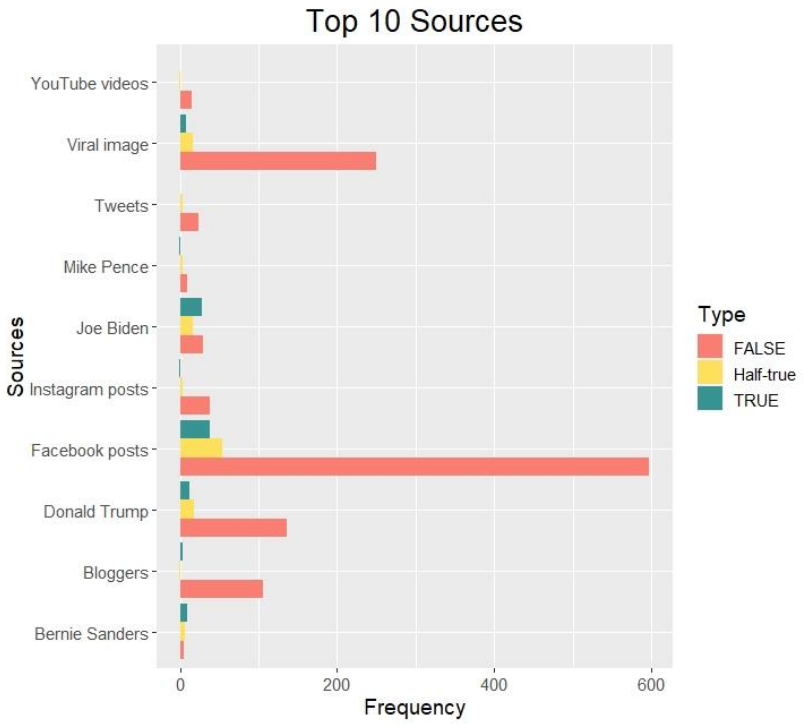

Fig.2 Top 10 major fact-checked information sources on politifact.com

The length of content, in this era of fragmented information, doesn't have significant differences among true, half-true, and false information(see Fig.3).

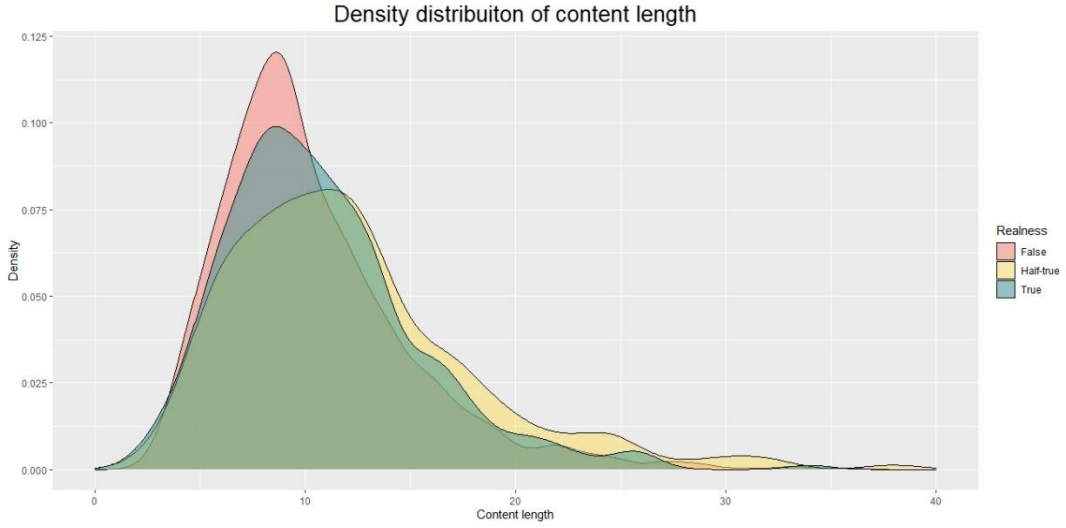

Fig.3 Density distribution of content length among true, half-true, and false information.

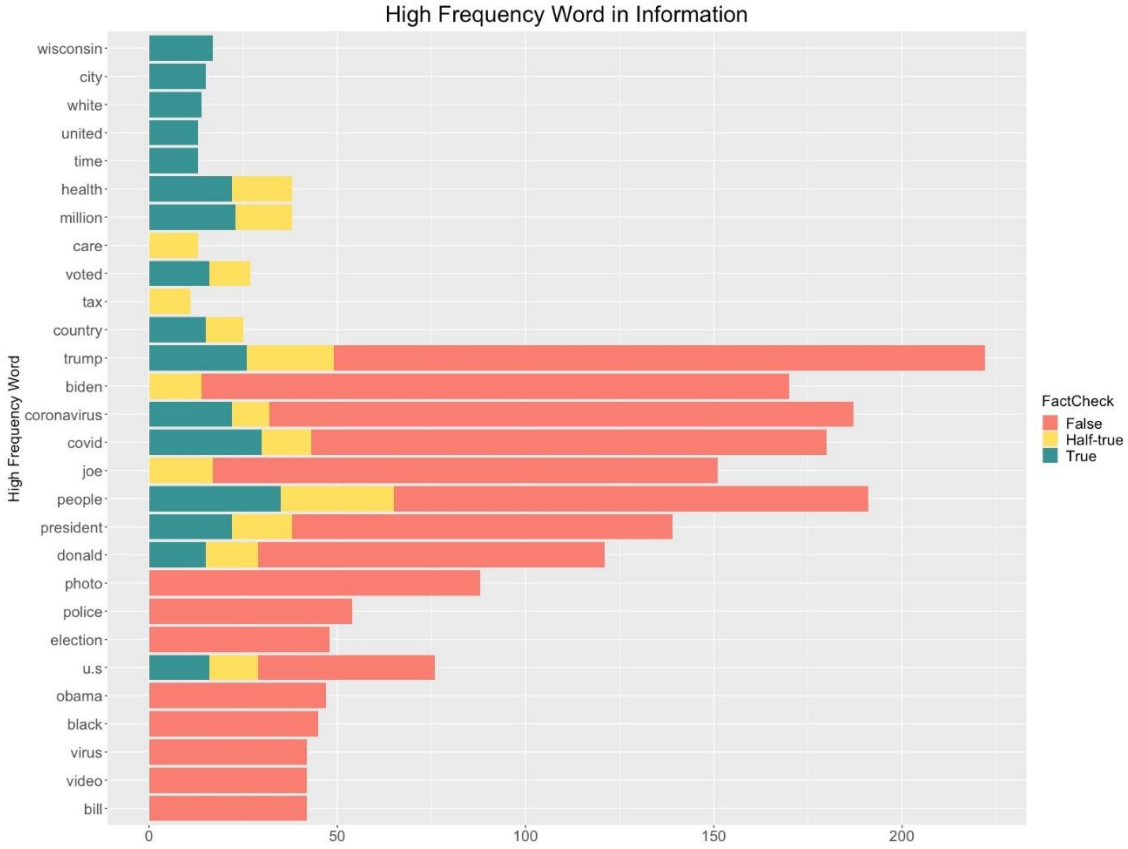

Fig.4 High-frequency words since the beginning of 2020 till this November

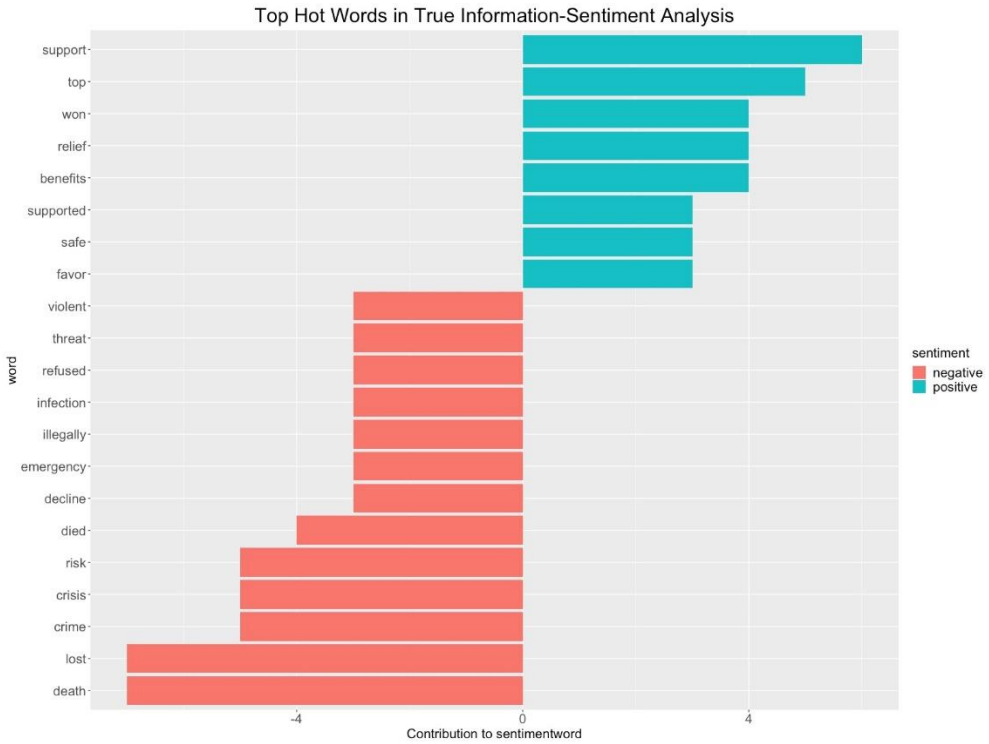

Fig.5 Hot words in sentiment analysis

The high-frequency words clearly show this year's themes: COVID-19, Black-lives-matter activity, and the presidential election(see Fig.4, 6-12). With these as the background, the hot words in the information sentiment analysis are also informative(see Fig.5).

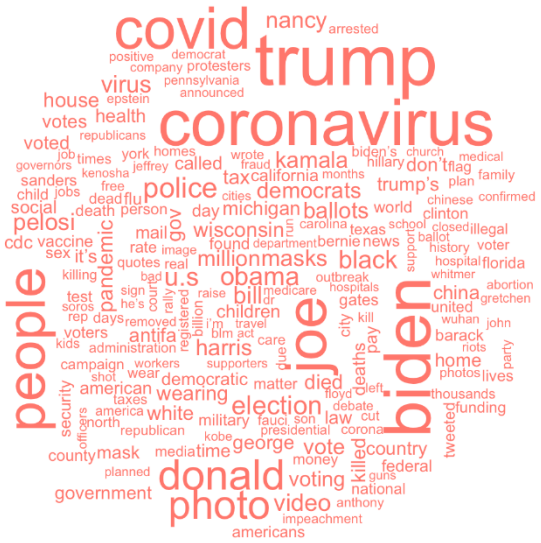

Fig.6 Word-cloud of the hot words in false information

Fig.7 Word-cloud of the hot words in half-true information

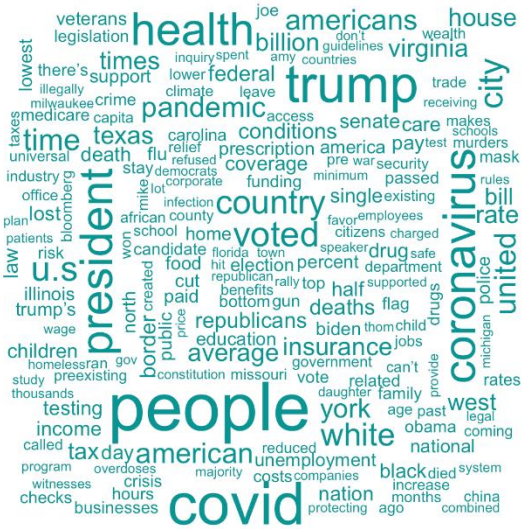

Fig.8 Word-cloud of the hot words in true information

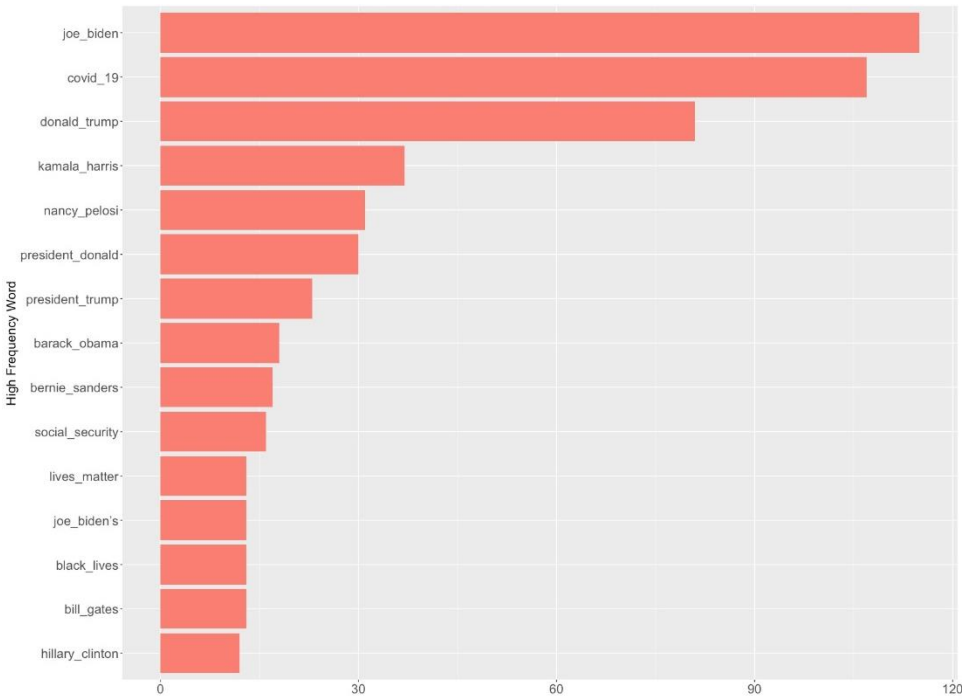

Fig.10 High-frequency bi-gram in false information

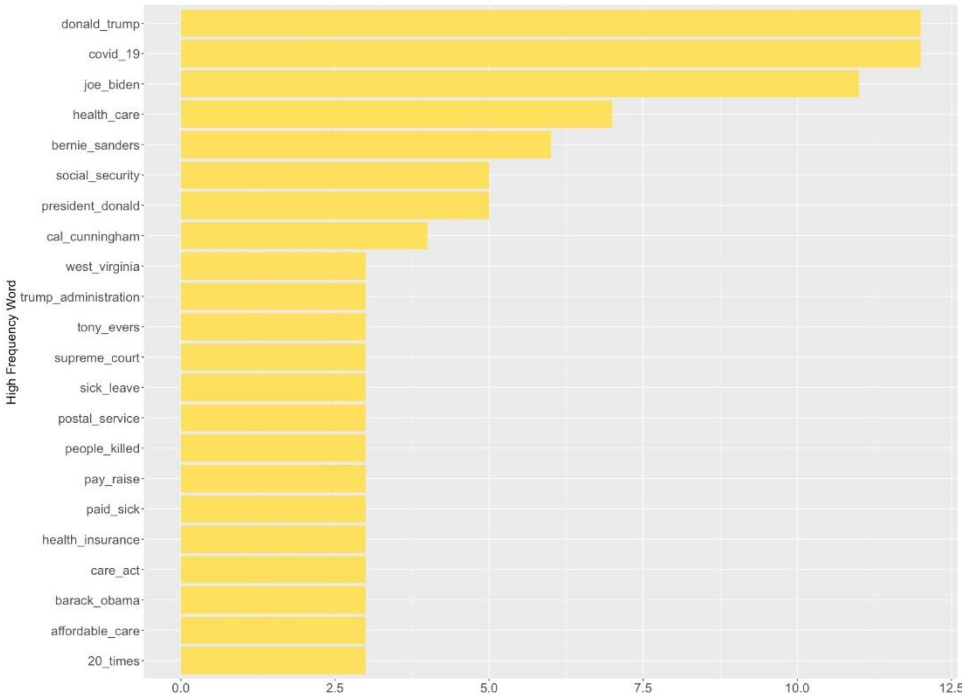

Fig.11 High-frequency bi-gram in half-true information

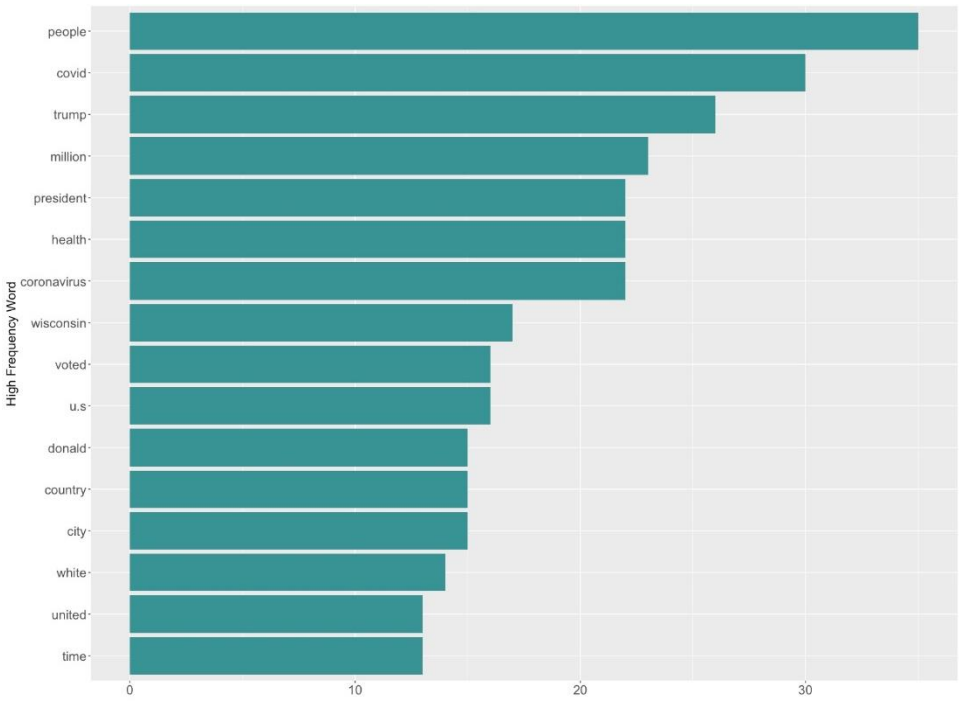

Fig.12 High-frequency bi-gram in true information

Besides these traditional parameters, we also noticed that people tend to use more questions and exclamation marks in false information(Fig.9). This could be an important variable in our following classification.

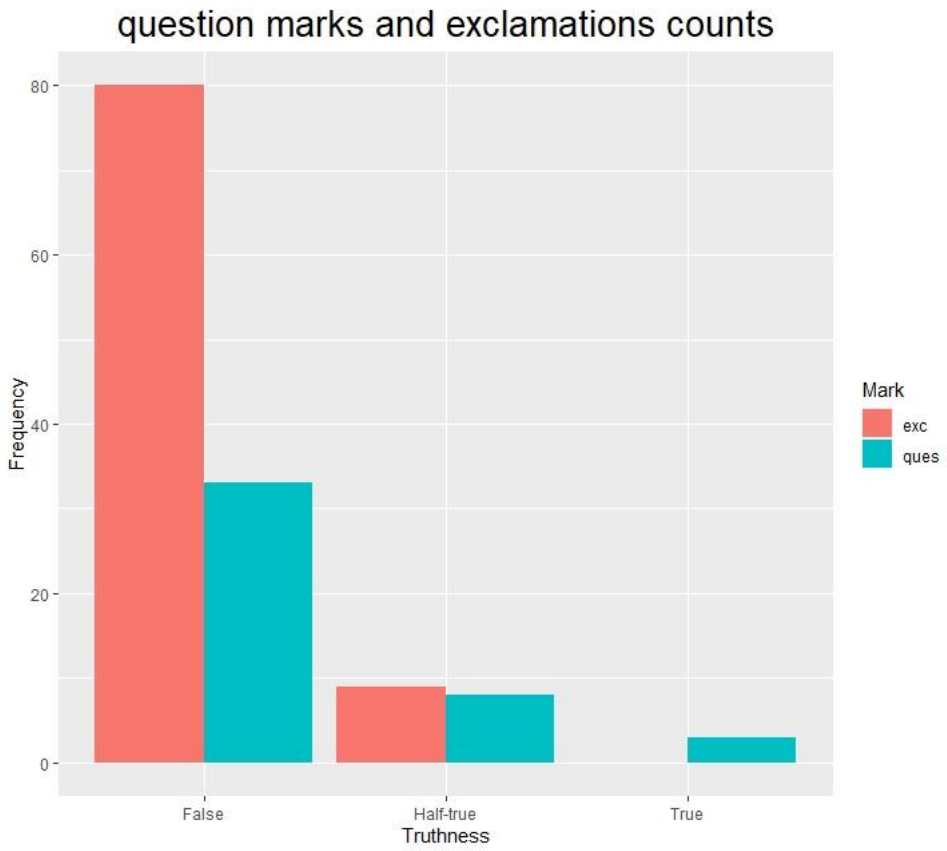

Fig.9 Question and exclamation mark's frequencies among different groups of information

3. Summary and the next steps

We aim to design a method to distinguish and classified true, half-true, and false information. Compared with other similar analyses, we focus on those short information. What's challenging is that we use the data scrapped by ourselves rather than those out-of-date datasets, and this will give us the most up-to-date analysis results, which is closely associated with the influential events that happened and are happening in this year. The topic and inspiration is the SBP-BRiMS 2020 Challenge 2 – Disinformation (http://sbp-brims.org/2020/challenge/challenge2_Disinformation.html); it's also one of the recommended topics.