Categories Setting

The coders were asked three questions: is the tweet misinformation? What kind of reasons did the tweet mentioned to support the pro- or anti- vaccine claims in the tweet? And what was the attitude towards vaccine in the tweet? The three questions were served as themes and respectively, there were 2, 14, and 3 categories under misinformation, reasoning and attitude themes.

Intercoder Reliability

In total, 4 coders coded 583 tweets collected from Apr 06, 2018 to May 09, 2018. We went through 3 runs of coder training before the 4 coders worked alone to finish the full dataset. During the training process, two set of random samples from the full dataset were used as training tests. Four coders worked independently on 45 (7.72%) tweets and the intercoder agreement among all coder is .45, which is much lower than the criterion of .70 for exploratory research. Based on this test, coding problems and disagreement were discussed and coding instrument was revised. Then, the coders recoded the same 45 tweets independently, and reached .70 criterion. This process was repeated. Another random 45 samples were generated, and were independently coded by the 4 coders with higher intercoder agreement of .60. Discrepancies were discussed and resolved after discussion. It was believed that the competent and trained set of coder following the coding instrument will result in a reliable coding decision. A pilot test of reliability was conducted after that. Each coder coded 100 (17.15%) tweets, while 2 coders worked on the same dataset. Both the intercoder agreement for the two set of 100 tweets reached .70, and to create the final dataset, the discrepancies were resolved among two coders. The remaining tweets were randomly divided into two groups and each group was assigned to two coders who worked the same dataset in the pilot test. With a .81 intercoder agreement, the coding decision were made based on the agreement and the discussion of discrepancies.

Twitter Dataset

We crawled and indexed 1024 tweets using the Twitter API in from Apr 06, 2018 to May 09, 2018, and from Jul 20, 2018 to Jul 25, 2018. All tweets are English. We first scraped tweets by using keywords. The keywords used for searching are: vaccine, vaccines, vaccination, vaccinate. In addition, to obtain a balanced dataset between factually accurate tweets and misinformation tweets, we target a specific number of users with prior history of posting misinformation tweets by getting all the tweets these users posted in a week before the date of webscraping. The dataset was manually annotated by 4 human annotators. The annotators were asked to label the tweets based on their content. First, tweets were labeled based on whether it was misinformation or not. Then, specific aspects were labeled to indicate what kind of reasons did the tweet mentioned to support the pro- or anti- vaccine claims in the tweet. Finally, the attitude towards vaccine in the tweet was labeled as well. Respectively, there were 2, 14, and 3 categories under misinformation, reasoning and attitude labels.

Annotators went through 3 runs of annotator training before the they worked alone to finish the full dataset. During the training process, two set of random samples from the full dataset were used as training tests. Four Annotators worked independently on 45 (4.39%) tweets and the inter-annotator agreement among all annotators is .45, which is much lower than the criterion of .70 for exploratory research. Based on this test, annotating problems and disagreement were discussed and annotating instrument was revised. Then, the annotators recoded the same 45 tweets independently, and they reached .70 criterion. This process was repeated. Another random 45 samples were generated, and were independently annotated by the annotators with higher inter-annotator agreement of .60. Discrepancies were discussed and resolved after discussion. It was believed that the competent and trained set of annotators following the annotating instrument will result in a reliable annotating decision. A pilot test of reliability was conducted after that. Each annotator coded 100 (17.15%) tweets, while 2 annotators worked on the same dataset. Both the inter-annotator agreement for the two set of 100 tweets reached .70, and to create the final dataset, the discrepancies were resolved among two annotators. The remaining tweets were randomly divided into two groups and each group was assigned to two annotators who worked the same dataset in the pilot test. With a .81 inter-annotator agreement, the annotating decision were made based on the agreement and the discussion of discrepancies.

Reasoning Template

We observed that for reasoning, Ingredient, Safety, Effectiveness, Autism, Conspiracy, Infant Mortality and N/A are the most frequent used reasons for anti-vaccine misinformation. We then worked to provide arguing templates for those frequent used reasons. For each reason, one to four arguing templates were generated, and all the templates refer to reliable sources. All the tweets labeled as those reasons were annotated for templates by four human annotators. Annotators were asked to assign a template to individual tweets according to their content. Annotators went through two runs of annotations; each annotator annotate 120 (30%) tweets independently with each tweet had two annotators annotated. For the first run, the inter-annotator agreement between two annotators was 67% which was below 70%. The discrepancies and annotating disagreements were discussed and annotating instrument was revised. Then the tweets were re-annotated by the same annotators and the inter-annotator agreement was 82%. The final result was generated based on the annotation and the discussion for discrepancies.