Labeling Fact-Check Relevance for Crowd Workers: Do Detailed Instructions Help?

In response to fears about the spread of online misinformation, there has been a rapid growth and investment in fact-checking. However, only a small percentage of people who are exposed to problematic online content are presented with corrective information like fact-checks. In the era of big data and artificial intelligence, a key agenda item for the fact-checking movement has been to develop automated fact-checking systems that leverage techniques like machine learning and natural language processing to limit peoples' exposure to misinformation by either having platforms present relevant fact-checks alongside problematic content.

In November 2017, as part of an expanded effort to provide users with context about news publishers, Google released Reviewed Claims a component of the Knowledge Panel that appears on many search engine results pages (SERP). The Reviewed Claims feature displayed authoritative, third-party fact-checks about content produced by certain news publishers. This feature was publicized as a meaningful aid to users[1] in the fight against misinformation. However, without informing fact-checking organizations, news publishers, or information seekers, Google's Reviewed Claims "claim matched" over half of the fact-checks that appeared in the component. Claim matching is a process in which fact-check articles that do not list the original source of a fact-checked statement (i.e. the claimant) are algorithmically assigned to a piece of the online content produced by a certain publisher, which I refer to as the document). Systems like Reviewed Claims that claim match existing fact-checks to online content have the ability to increase the number of stories that have been fact-checked and simplify the process of displaying fact-checks alongside problematic online content.

Crowd workers are the individuals that train these machine learning claim matching models. Crowd workers are people who do crowdsourced tasks. Crowdsourced tasks often require a large pool of workers to perform a small task. Platforms like Amazon Mechanical Turk facilitate the interactions between those with tasks (like researchers) and crowd workers.

Quality training data is essential for quality models. Previous pilot study experiments I have run have revealed that Mechanical Turk workers struggle to correctly label <news article, fact-check article> pairs as "relevant" or not "relevant." However, these surveys often provide little direction what it means for a <news article, fact-check> pairs as relevant.

---

[1] https://www.blog.google/products/search/learn-more-about-publishers-google/

I propose asking the following research question: **Does providing classification examples in the instructions increase agreement rates of Mechanical Turk workers when assessing the relevance of claim matches?**

In summary, two surveys would be created, one with instructions similar to those of past claim matching experiments, and the other with a more expansive instructions section. Participants will be randomly assigned to either control (basic instructions) or treatment (more elaborate instruction) and then perform the same claim matching task of matching <news article, fact-check article> pairs. There will be blocking by partisan affiliation, and no clustering will be necessary. The objective is to test whether having more detailed instructions increases the consistency of MTurker ratings.

The <news article, fact-check article> pairs would be drawn from the Reviewed Claims dataset which I scraped in January 2018. The news articles come from news organizations like *The Daily Caller* and *The Babylon Bee*, which are sources often described as hyperpartisan news outlets. To minimize people making snap judgements of relevance based on the news source name, I propose stripping all identifying website style and references to the outlet. In fact, all news sources would be referred to as *The Nevada Register* and all fact-check articles would be from *Fact-checking.com*, both of which are not existing publishers.

Studies examining the intersection of crowd workers and misinformation have often blocked by partisan affiliation[2] as it has been observed that Democrats and Republicans have different notions of fact vs. rumor. While this survey does not ask participants to assess the validity of the news article or fact-check, but rather the relevance of the news article to the fact-check article, it seems reasonable to think that partisan effects will remain. Therefore, I propose to do block in this survey. In fact, an interesting secondary research question is how partisanship affects crowd worker labeling.

There is also the question of how to best ensure crowd worker quality. I propose following the practices of similar crowd worker studies where a 98% HIT (task) approval rate, US residency, and >1000 tasks completed are prerequisites. Because labeling fact-checks requires a high level of understanding of the political landscape, we will also require a 3-question political knowledge test, again modeled on previous research.

Survey takers would be asked to label 5 <news article, fact-check article> pairs in a binary classification as "relevant" or "not relevant." While a Likert scale is an alternative scale option,

---

[2] Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, *116*(7), 2521–2526. https://doi.org/10.1073/pnas.1806781116

the purpose of this task is to train a binary classification model, so it is desirable to rely on a binary labeling. However, to understand the difficulty users had with labeling each pair, there will be follow up question about the difficulty of the task where users are asked to label each determination on a 5-point Likert scale of difficulty.

The five pairs of articles will cover different challenges frequently seen in automated fact-checking on five different topics. To draw out these challenges, I will use a simple example. Let's say we are asking crowd workers to determine whether the fact-check article "Vaccines Don't Cause Autism" applies to these five different articles, which I have labeled with the corresponding "type" of claim:

- **Diluted Claim:** "Vaccines May Cause Autism:" note that this article doesn't say that vaccines cause autism. The headline qualifies the link between vaccines and autism with the word "may."
- **Broad Claim:** "We Don't Know Everything We Should Know About the Effects of Vaccines:" this headline is a more extreme version of the first. In fact, his article doesn't mention the word autism anywhere in the article, but it does recommend that parents wait as long as possible to vaccinate their children because of potential risks.
- **Reporting on Third Party Opinion:** "Jessica Biel says Vaccines May Cause Autism:" This article doesn't take a stance about whether vaccines cause autism, and this news publisher is seemingly only reporting on what a celebrity said.
- **Topical Shift:** "Vaccines May Cause Breast Cancer:" This article doesn't mention autism but does say that there is a direct link between the MMR vaccine and breast cancer.
- **Satire:** "Vaccines Cause Autism and I'm the Queen of England": This article says multiple times that vaccines cause autism, but it was intended to be satire.

In this case, it seems reasonable that the first two types should be labeled as relevant, and the last three should be labeled as not relevant. However, the point of this experiment is not to definitively determine the criteria of "relevance," which is a notoriously difficult concept to define, but rather to see if giving the examples in the bulleted list above helps improve agreement rates of crowd workers. Ideally, higher agreement rates will result in more predictable and accurate models.

Conversations with team members are necessary to build consensus on the five examples that will actually be included in the survey. This study would first be run in a pilot, and then expanded to 100 observations. This is simple to implement on Mechanical Turk.

In terms of the statistical test to be used, we are interested in whether the distribution of answers changes with the addition of instructions. A Kolmogorov-Smirnov two sample test may be an effective way of testing that. We could also ask a few journalism graduate students if they would

be willing to label these pairs and use their evaluations as a "ground truth" and then perform t-tests.

The survey will conclude with a few demographic questions that include gender, educational attainment, age, and partisan affiliation. The target mean completion time of this survey would be ~20-30 minutes as reading the news article and fact-check articles require close reading.