Combating Misinformation[*] Invited Talk - Extended Abstract

Leonard Grokop

Facebook, Menlo Park, CA, USA
lgrokop@fb.com

1 Introduction

Misinformation can occur across the spectrum of our apps, and be produced in a number of different formats such as memes, videos, comments, messages, links to external URLs, etc. We have a responsibility to combat it and this means addressing several challenges such as fake accounts, deceptive behavior, and misleading and harmful content. This talk describes how we tackle the latter. I start by giving an overview of our high-level approach.

Several strategies are employed. Our starting point is connecting people with reliable information from trusted experts. This is done by attaching labels to certain posts that direct users to information hubs like our COVID-19 Information Center, Climate Science Information Center and US 2020 Voting Information Center. Further, we notify users when they post misinformation and reduce the distribution of Pages and websites that repeatedly share it.

2 Our Approach

The main strategy involves identifying misinformation and taking mitigating actions on the posted content, to limit its harm. A machine learning model identifies potential misinformation using a variety of signals. The most serious kinds of misinformation such as false claims about COVID-19 and vaccines, and content intended to suppress voting, are removed. The remaining content is eligible to be reviewed and rated by one of our third-party fact-checking partners. These fact-checkers are independent and certified through the non-partisan International Fact-Checking Network. When they rate something as false, we reduce its distribution in News Feed and apply a visual treatment that displays the rating and provides a link to their fact-check article.

To help fact-checkers address false content faster, we employ a team of Community Reviewers: contractors hired through one of our partners. They are not making final decisions themselves but instead are sent content identified by our

[*] Copyright ©2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). Presented at the MISINFO 2021 Workshop, held in conjunction with the 30th ACM The Web Conference, 2021, in Ljubljana, Slovenia.

machine learning models, asked to identify the main claim being made and to conduct research to find other sources that either support or refute that claim. Fact-checking partners are then shown the collective assessment of community reviewers as a signal in selecting which stories to review and rate.

3 Why Misinformation is Hard

Having laid out the high-level approach I next delve into why misinformation is a hard problem. There are several reasons. The primary one is that identifying what is actually misinformation is often a difficult task for humans, let alone algorithms. I walk through the steps that Community Reviewers took for a real-world example, illustrate the ambiguity of claim identification, and show how humans themselves can struggle to reach a unanimous conclusion on the veracity of a post or article. I then highlight other difficulties such as the sudden growth in viewership that misinforming posts can experience if left unchecked, and adversarial behavior that attempts to hinder our detection capabilities.

From here I discuss how our detection models work, primarily relying on engagement signals such as user reports and comments that express disbelief. The problem with relying on engagement signals though is latency. The signals don't become reliable until enough comments, user reports, etc., have come in, but by that point in time many views of the content have occurred. This is a chicken and egg problem.

Thorough reporting by our fact-checking partners can also take time. Content that is restating an existing misinformation claim requires fact-checkers to identify the fact-check article associated with the existing hoax. Content that is stating a new misinformation claim requires investigation and writing of a new fact-check article. That's why we're always looking for ways to help our partners focus their time and journalistic expertise on the clearest misinformation and fake news that can harm and mislead.

4 Reducing latency

The remainder of the talk describes our approach to solving the latency problem, which is to rely on matching content to existing rated content, without first waiting for user engagement. Matching takes two forms. Some matches are near-exact duplicates of already debunked misinformation. We refer to these matches as "copies". Other matches make the same claim as previously debunked content, but may either not immediately resemble the original, or may be a completely new and independent version of the same hoax. We refer to these as "semantic" matches. I illustrate examples of each of these forms.

For both forms we detect the matches by computing pre-trained embeddings of misinforming content already rated as false/misleading by fact-checkers and add these embeddings to an index. When new content is uploaded, we compute its embedding and query the index to retrieve rated content with similar embeddings. We then perform an alignment and verification step on each retrieved candidate to make a final match determination.

The difference in our approach to handling copies versus claim matches is in the embedding used, and in the mitigating action taken. Copies use an embedding that captures only image similarity. The two images need not be pixel exact but need to be sufficiently similar that we can make a high-confidence determination that the content is the same. Detected copies have a visual treatment applied to them informing the user that the content contains the same information found to be false by fact-checkers in another piece of content. We also provide a link to the associated fact-check article. In contrast, the embedding in semantic matches uses text and OCR information, in addition to the image similarity of key objects detected. Semantic matches are sent to fact-checkers for review.

By matching posts to already debunked content at creation time, we remove the detection latency component that depends on user engagement, making it possible to act on misinformation prior to many views occurring. This is an overview of the multi-pronged approach Facebook takes to stop the spread of misinformation.