Cecilia Bell 5.25.2021

MVP

Detecting Non-Factual Content on Facebook

Problem

At Facebook, a fake news detector has been developed to fight misinformation, which flags posts with fake news articles. However, posts with non-factual articles, including satirical, comic, and opinion articles, are also being flagged as fake news. The problem, then, is: how can non-factual articles be identified and distinguished from fake news articles (an identification problem)?

Impact

The desired impact of this project is to offer Facebook users more accurate truth-rating content warnings.

Solution

My first proposed solution path is to build a natural language processing classification model, which classifies articles as mostly true, mostly false, or non-factual (these categories are subject to change). The solution requires a large dataset of articles posted on facebook, including their titles and truth ratings. The truth ratings will be determined and cross-checked by a group of fact-checkers given a specific set of guidelines.

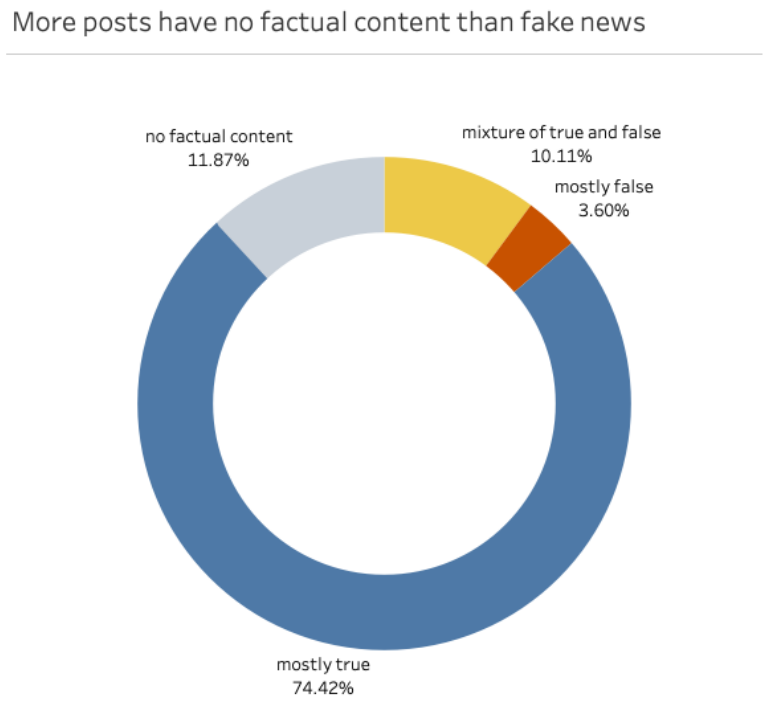
Other potential solution paths include:

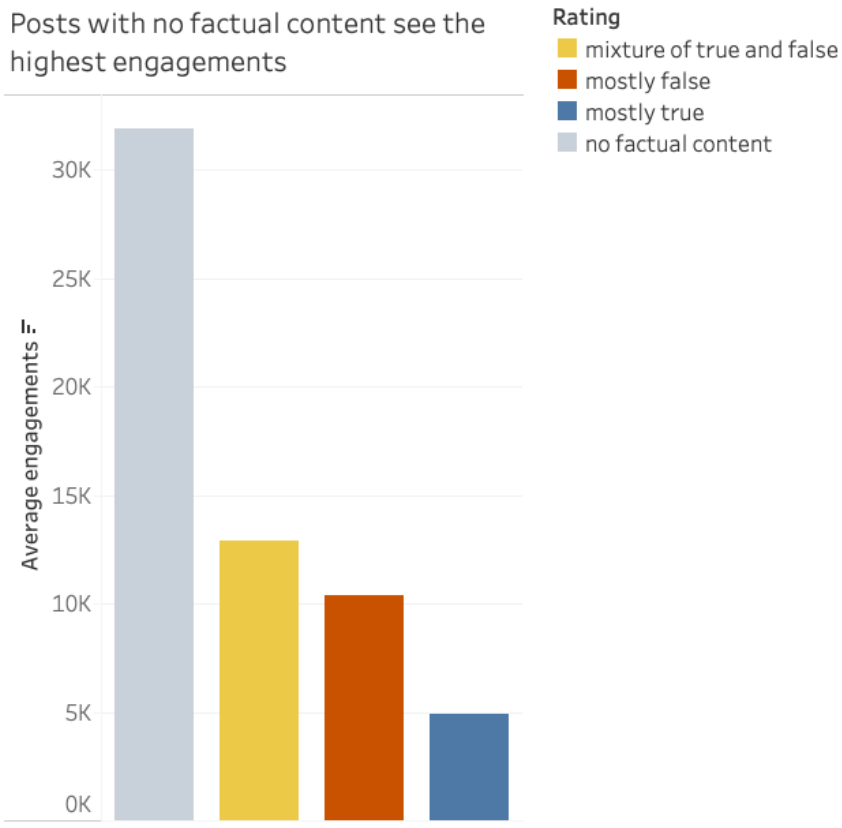
- Using third-party (human) fact-checkers
- Allowing users to flag misinformation

Impact Hypothesis

Using an NLP classification model to classify posts into more granular truth categories will give Facebook users more accurate truth-rating content warnings.

Preliminary Supporting Visualizations


More posts have no factual content than fake news
no factual content
11.87%
mixture of true and false
10.11%
mostly false
3.60%
mostly true
74.42%


Posts with no factual content see the highest engagements
Rating
mixture of true and false
mostly false
mostly true
no factual content
Average engagements
30K
25K
20K
15K
10K
5K
0K