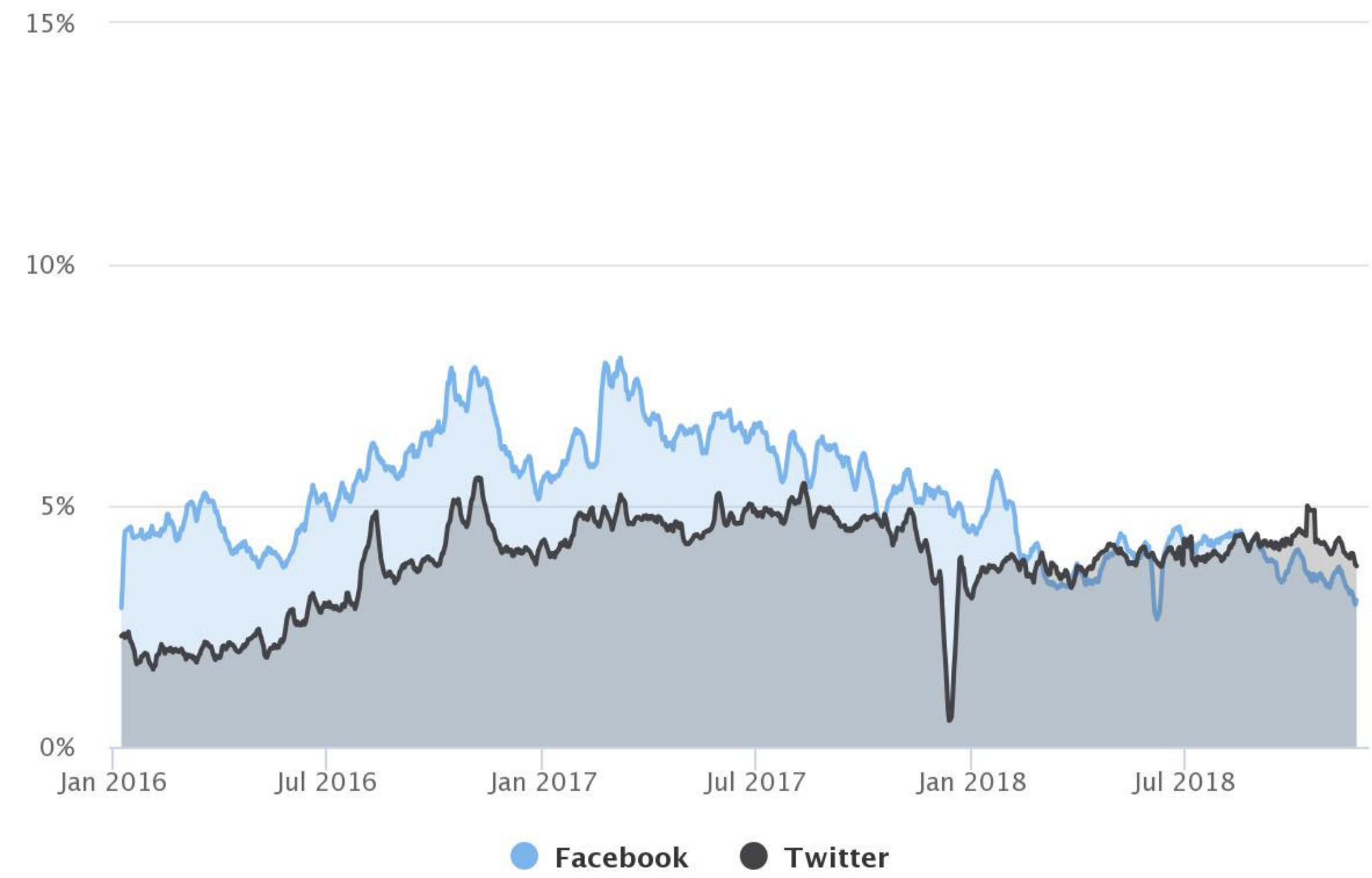
Platform Health Metrics

Paul Resnick
Michael D. Cohen Collegiate Professor
Associate Dean for Research and Faculty Affairs
November 29, 2018


15%
10%
5%
0%
Jan 2016
Jul 2016
Jan 2017
Jul 2017
Jan 2018
Jul 2018
Facebook
Twitter

Outline

- Platform Health Metrics
  - Motivation
  - Desiderata
- The Iffy Quotient
  - Current Status
  - Future Improvements
- Other Metrics Under Development
- Brainstorming

PLATFORM HEALTH METRICS

Problems

- Viral misinformation
- Toxic public conversations
- Filter bubbles
- Polarization
- Popularity manipulation (with bots)
- Troll accounts influencing media
- Harassment silencing minority voices
- …

What: Prevalence Metrics

- Collect (Sample)
- Classify
- Summarize

Why

- Assess Importance of Problems
- Maintain Accountability for Progress

Desiderata

- Understandable
- Credible
- Robust
- Comparable
  - Between sites
  - Over time

THE IFFY QUOTIENT

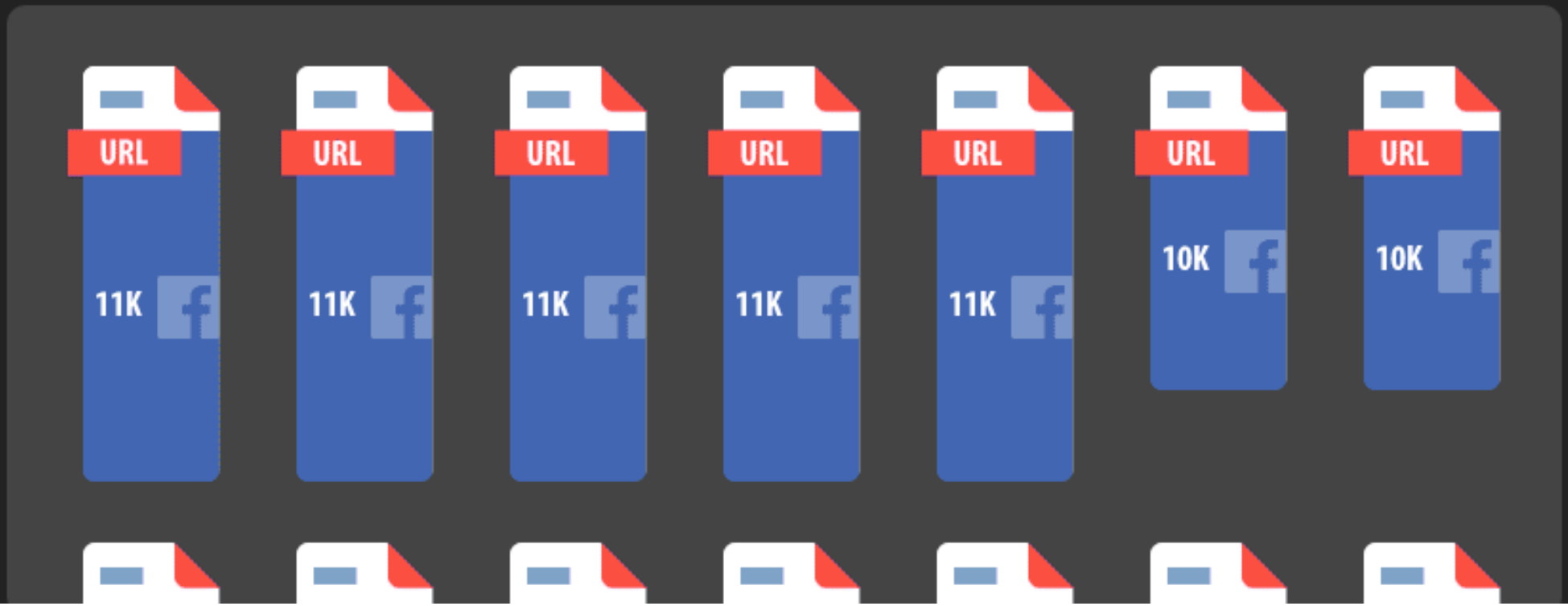
STEP 1 NewsWhip tracks the creation of URLs on more than 400K sites.

STEP 2 For each news URL, NewsWhip gathers engagement data on Facebook and Twitter.

STEP 3 We query NewsWhip for each day's top 5,000 URLs from each site.

STEP 4 We classify each URL based on whether its domain name appears on one of Media Bias/Fact Check's lists.

CSMR



SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

MBFC Criteria

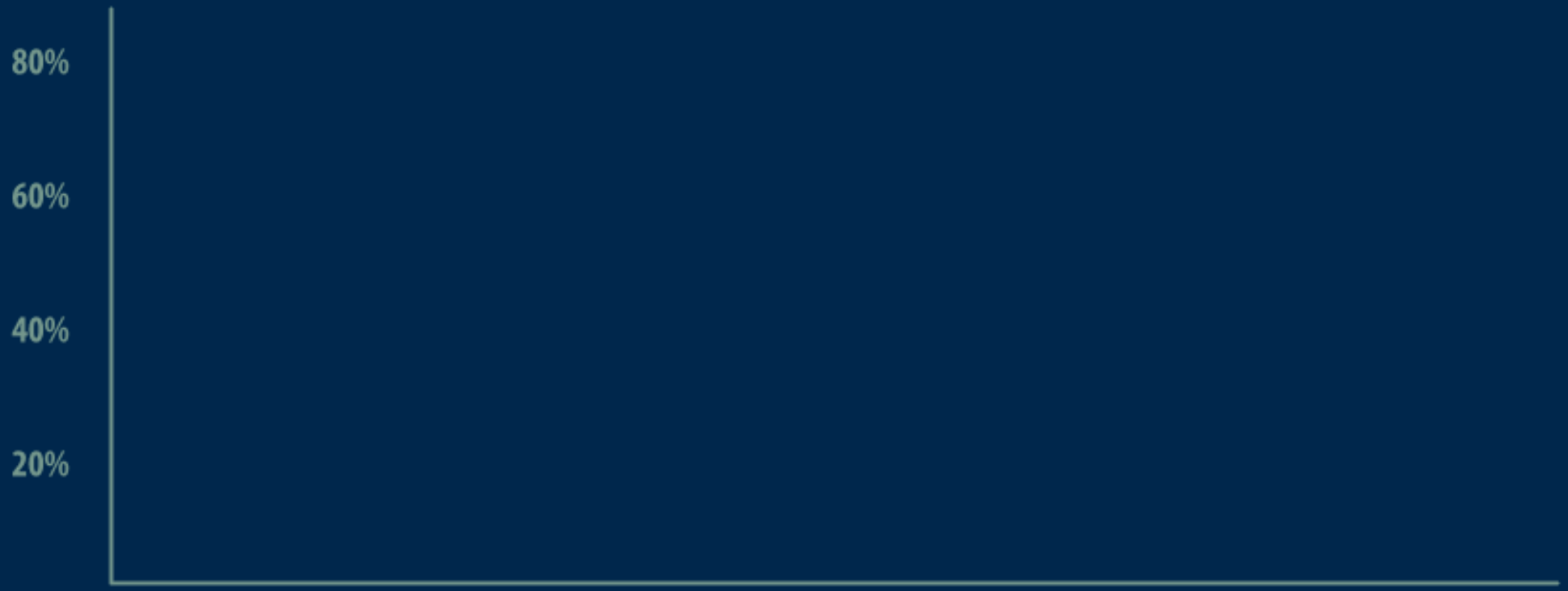
- Questionable Source

  A questionable source exhibits *one or more* of the following: extreme bias, overt or no sourcing to credible information and/or is fake news. Fake News is the *deliberate attempt* to publish hoaxes and/or disinformation for the purpose of profit or influence. Sources listed in the Questionable Category *may* be very untrustworthy and should be fact checked on a per article basis.

- Conspiracy/Pseudoscience

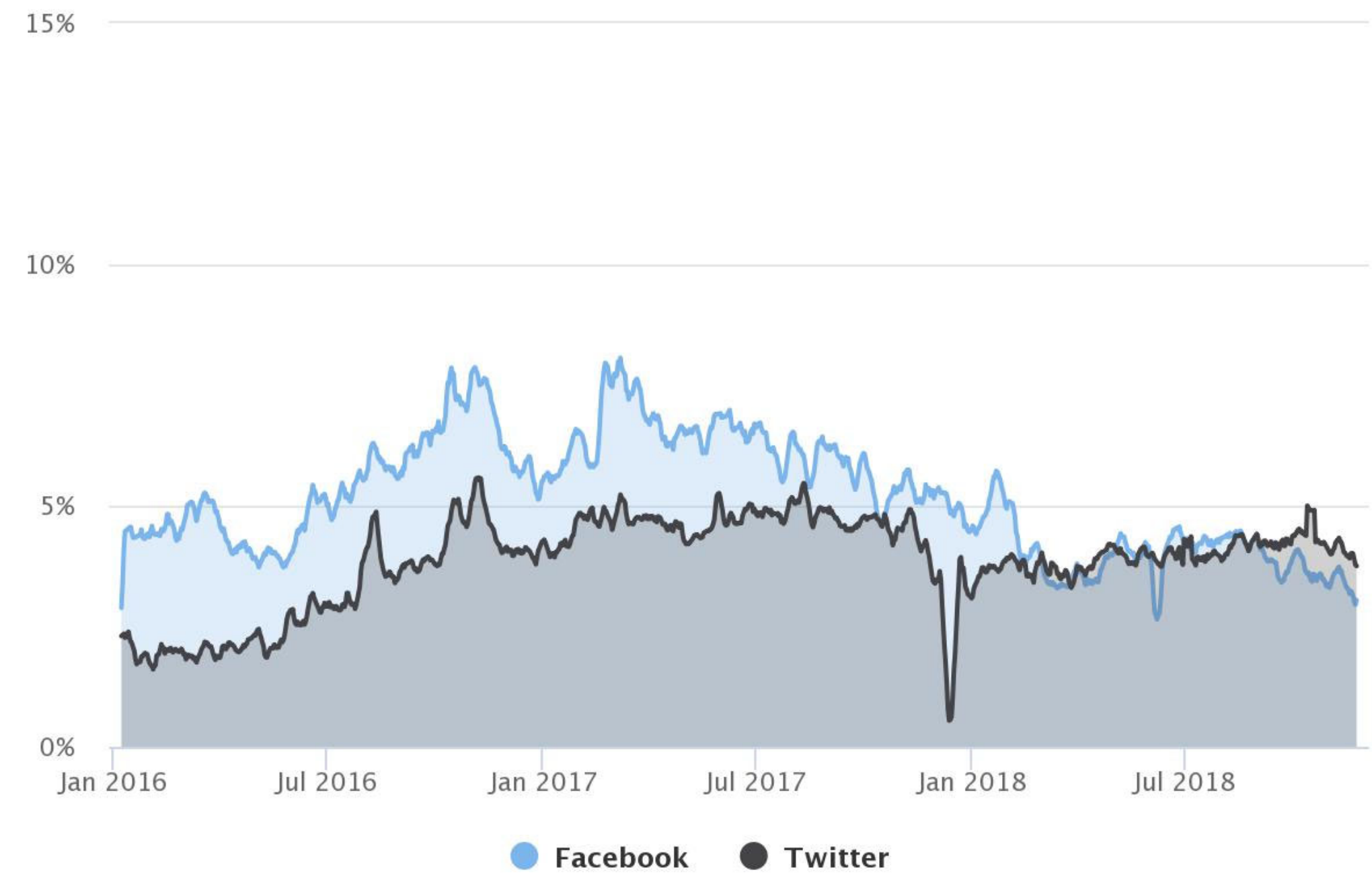
  Sources in the Conspiracy-Pseudoscience category *may* publish unverifiable information that is *not always* supported by evidence. These sources *may* be untrustworthy for credible/verifiable information, therefore fact checking and further investigation is recommended on a per article basis when obtaining information from these sources.

STEP 5 We show the percentage of URLs from iffy sites, the Iffy Quotient.

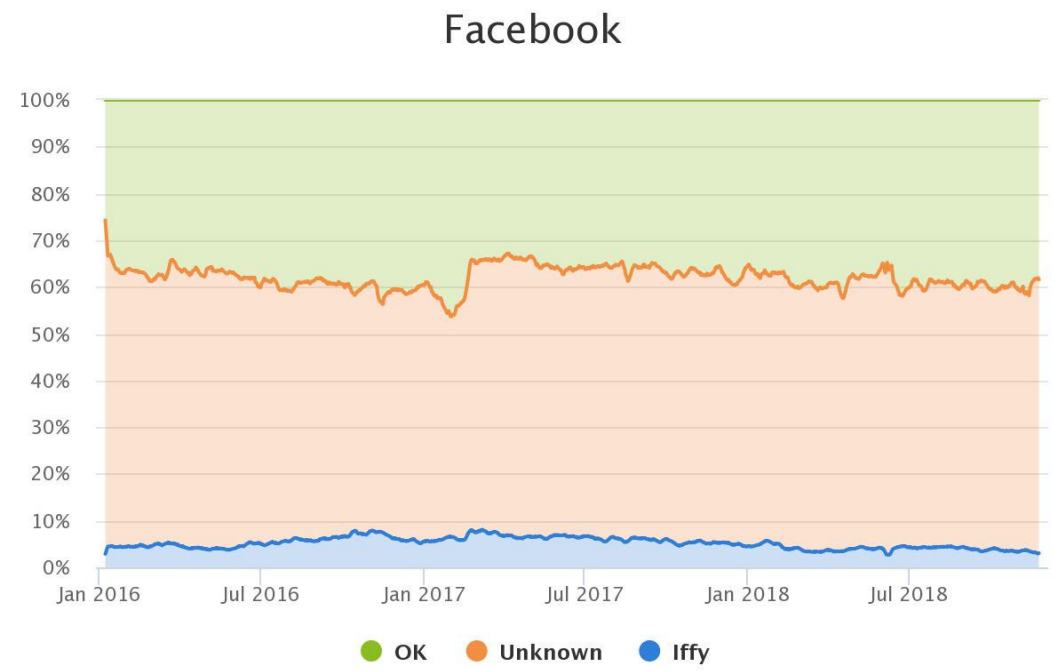


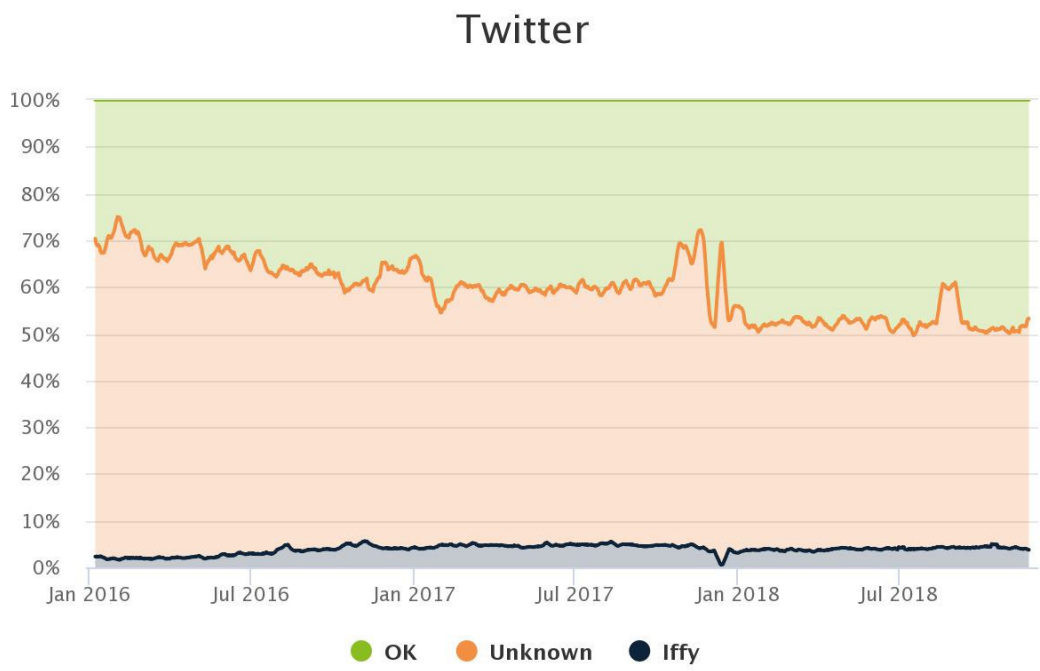
Summary

- Collector
  - NewsWhip, top 5K URLs daily, by “engagement”
- Classifier
  - MBFC
    - Questionable Source or Conspiracy/Pseudoscience → Iffy
    - Other labels → OK
    - Unlabeled → Unknown

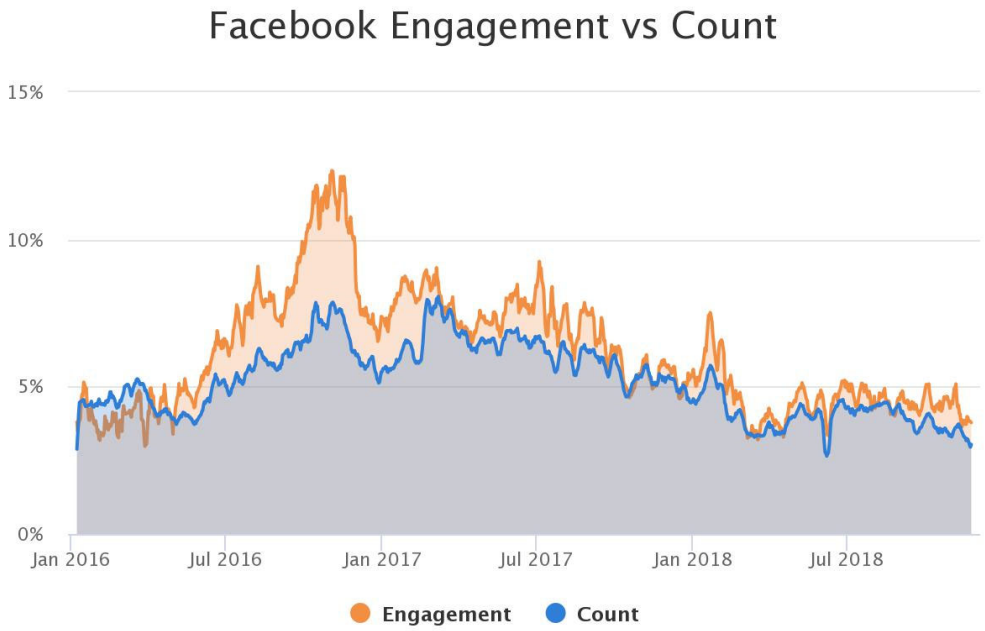

15%
10%
5%
0%
Jan 2016
Jul 2016
Jan 2017
Jul 2017
Jan 2018
Jul 2018
Facebook
Twitter

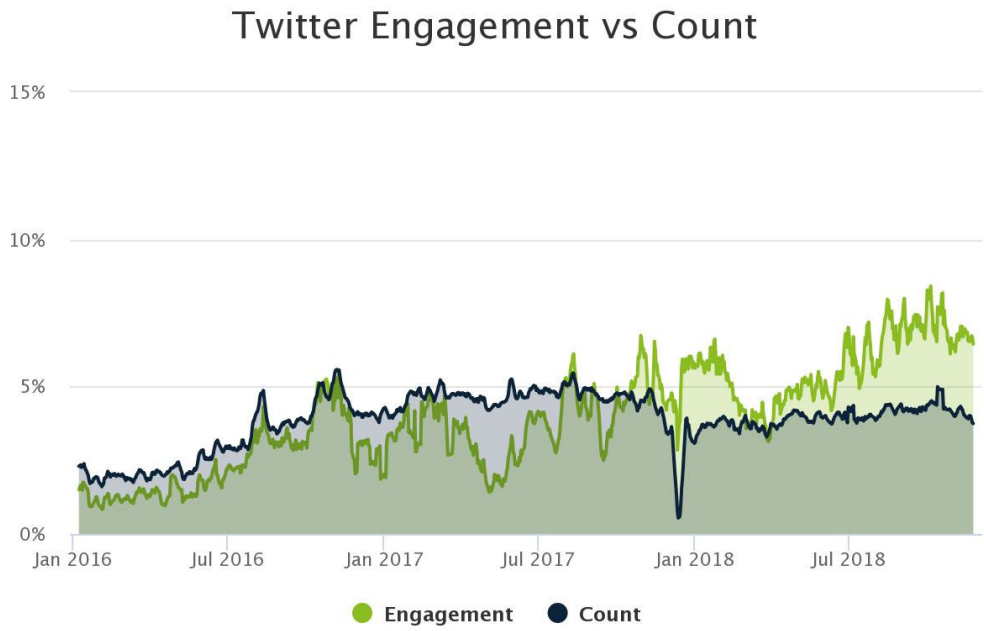
Classifier Decay?



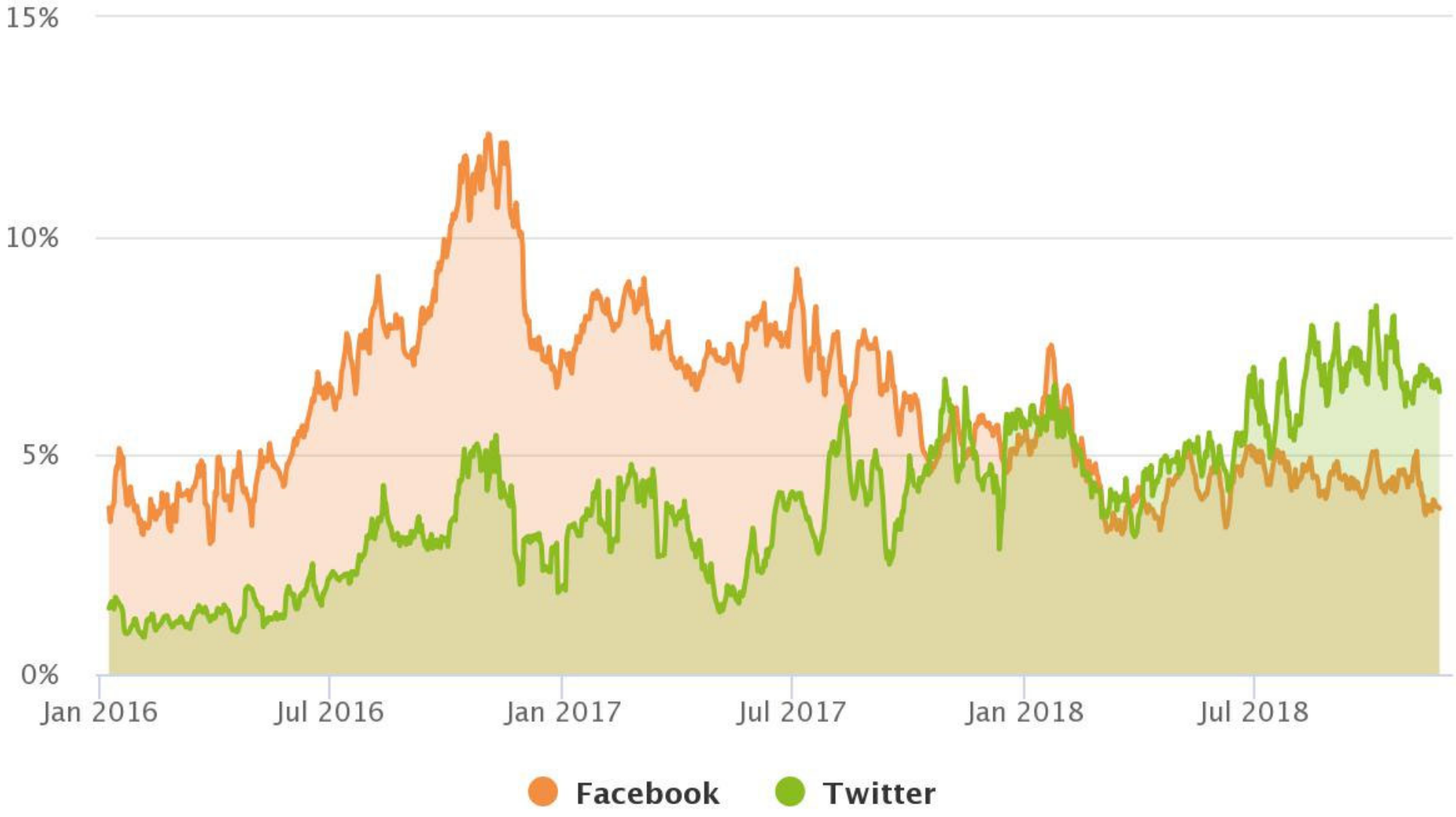


Engagement Weighted

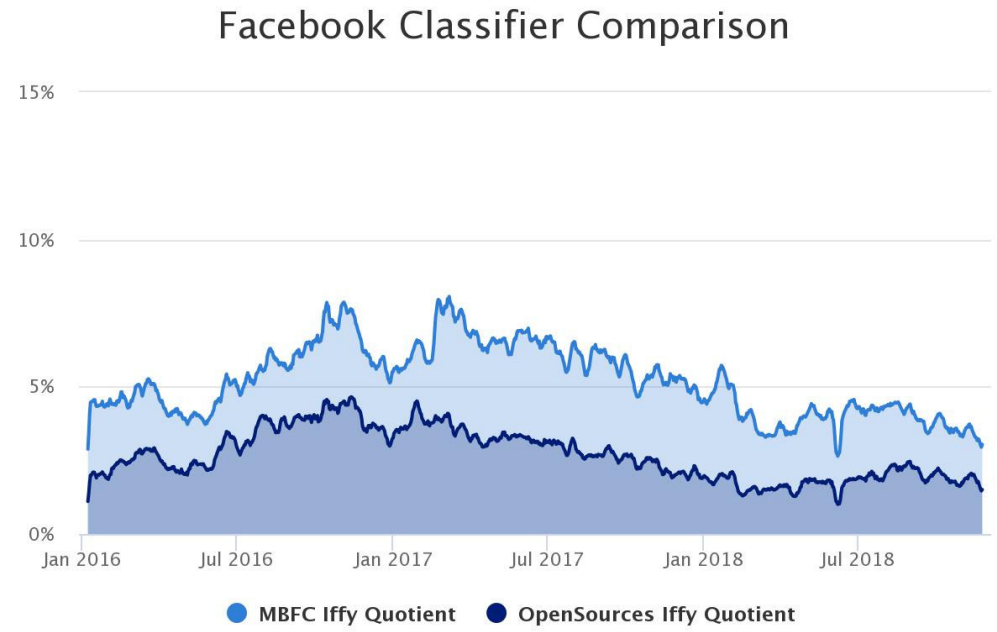


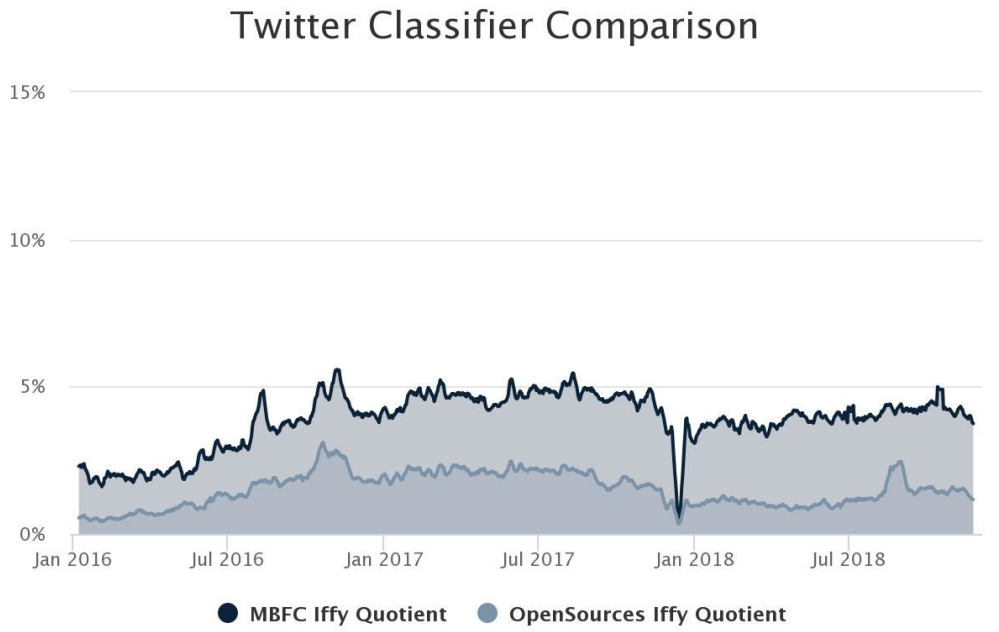


Engagement-weighted Iffy Quotient



Alternative Classifier





Future Improvements

- Collector
  - Filter URLs for “newsiness”
  - Requires a classifier…
- Classifier
  - NewsGuard site labels by journalists
  - URL-level classification?

METRICS UNDER DEVELOPMENT

Conversation Quality

- Collector
  - Seed: news and politics articles from mainstream sites
  - Collect comments from:
    - Publisher's comment section
    - Publisher's Facebook page
    - Twitter
    - SubReddits
- Classifier
  - Jigsaw Perspective API personal attacks classifier

YouTube Recommender Polarization

- Collector
  - Seed: search on popular political topics
  - Crawl
    - From each video, get next recommend one, 20 times
- Classifier: (-1, +1) liberal to conservative
  - 1: Based on text of comments
  - 2: Based on audience (inferred from ads API?)
- Rollup
  - Each video: polarizer score = difference in classifier score from start video to 20th recommendation
  - Average across videos

Desiderata

- Understandable
- Credible
- Robust
- Comparable
  - Between sites
  - Over time

Brainstorming

- What other metrics would be valuable?
- What collectors are available/possible?
- What classifiers are available/possible?