Team QCRI-MIT at SemEval-2019 Task 4: Propaganda Analysis Meets Hyperpartisan News Detection

Abdelrhman Saleh[1], Ramy Baly[2], Alberto Barron-Cedeno[3], Giovanni Da San Martino[3], Mitra Mohtarami[2], Preslav Nakov[3], and James Glass[2]

[1]Harvard University; [2]MIT CSAIL; [3]Qatar Computing Research Institute, HBKU

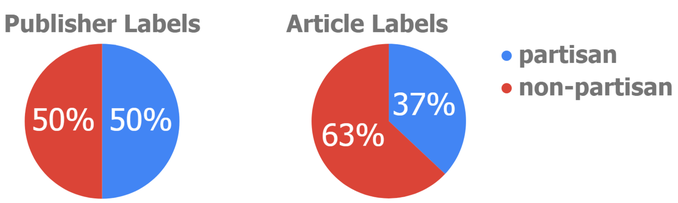
Motivation & Data

- **Hyperpartisan news are inherently propagandistic:**
  - Promote a particular non-objective point of view
  - Primarily used to influence public opinion or promote an agenda

- **Thus, we explore features known to work well for propaganda from A. Barron-Cedeno et al. (2019)**

- **Previous work:**
  - Determine factuality and bias of news sources rather than articles
  - Claim verification and stance detection

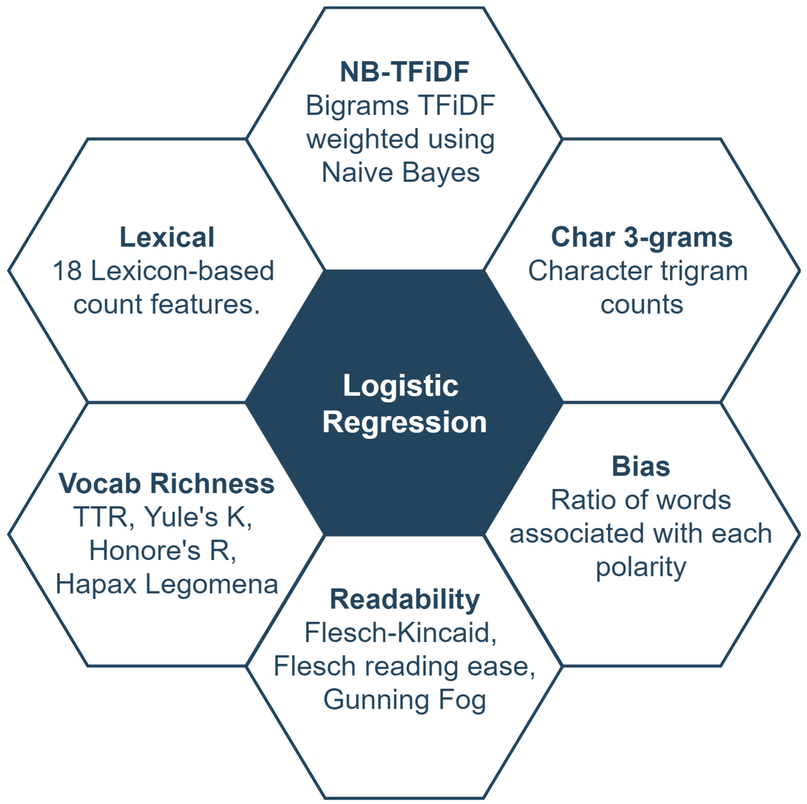
Dataset



- **Publisher labels:**
  - 750k articles labeled using distant supervision.
  - Labels inferred from the publisher's bias
- **Article labels:**
  - 645 articles labeled through crowdsourcing
  - More relevant for this task, more fine-grained labels

System Description

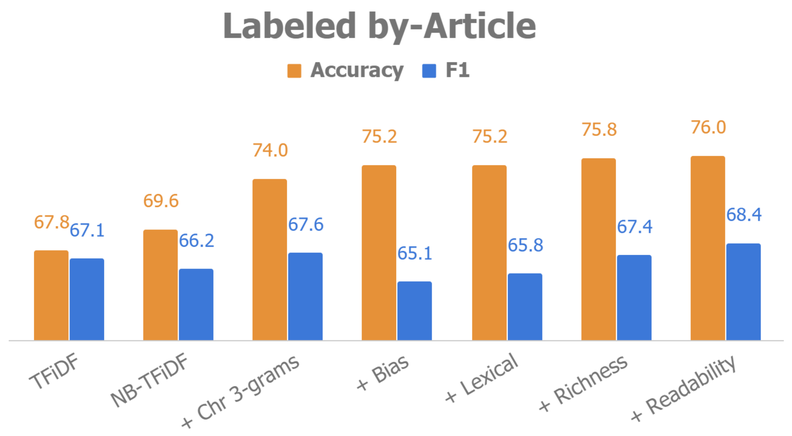
Features Used

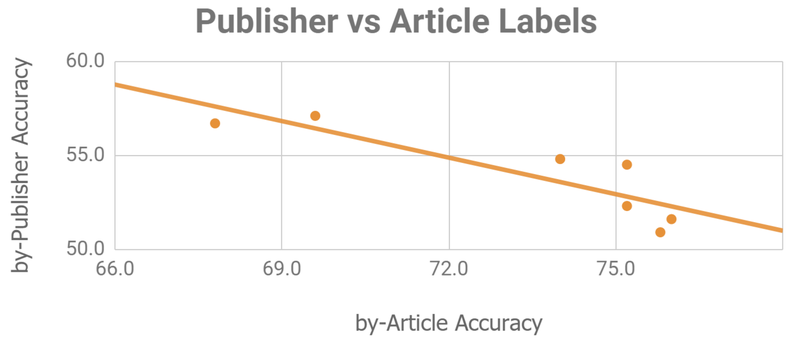


- **Training details:**
  - Train on the by-publisher data
  - Use the by-article data for validation
  - *Thus, avoid overfitting on the by-article data*

- **Model:**
  - logistic regression instead of more complex models
  - *Helps generalization since the training data is quite different from the test data*

Results





- **Propaganda features classify hyperpartisan news:**
  - Significant improvements in accuracy for the by-article-labeled data
- **Modest improvements in F1 score:**
  - Adding propaganda features improves precision, but it also hurts recall.
  - However, recall is more important here.
- **Char trigrams are important:**
  - Largest improvement for a single added feature.
- **Publisher labels not ideal:**
  - By-Article accuracy negatively correlates with by-Publisher accuracy (r = -0.87)
  - Difficult to generalize from by-publisher data

This research was carried out in collaboration between the MIT CSAIL and the Qatar Computing Research Institute, HBKU.