Differences in Health News from Reliable and Unreliable Media

Sameer Dhoju¹, Md Main Uddin Rony², Muhammad Ashad Kabir², Naeemul Hassan¹
¹The University of Mississippi, ²Charles Sturt University

Motivation

- More than 50% of the top-20 Facebook stories containing “cancer” in headline were Fake.
- Fake news about vaccine caused measles outbreak in Europe.
- Deluge of misleading health news over social media
  - Continuously produced and propagated by unreliable outlets.
  - Reach to a broader audiences through social networking sites.

Health misinformation can be critical

- Erroneous health news can cause hazardous health condition.
- Spoil the credibility of the health-care providers & medicines.

Health misinformation is a relatively unexplored area

- Lack of reliable entities to debunk health misinformation.
- Very few computational approaches with limited success.

Problem Formulation

- Develop a health-oriented news corpus.
- Analyze the corpus to identify discriminating features.
- Build a classifier to distinguish unreliable media sourced health news from reliable news.

Data Collection

- Use of quotations & links indicates credibility of an article.

Structural Analysis

- Media outlets use longer headline to get more attention.
- Unreliable outlets (12.13 words/headline) use longer headlines than reliable outlets (8.56 words/headline).
- Unreliable outlets (40.03%) practice more clickbait than reliable outlets (27.29%).
- Unreliable outlets use demonstrative adjectives and numbers significantly more than the reliable outlets.

Topical Analysis

- Used Google Cloud NLP & Latent Dirichlet Allocation (k=3).
- Reliable & unreliable outlets cover different health topics.

Semantic Analysis

- Use of quotations & links indicates credibility of an article.

Feature Extraction

- Word (W): Took 5,000 most frequent n-grams (n = 1,2).
- Extracted Features (EF): 10 features extracted from the analysis.
- Four out of 10 extracted features make to the top-20 most important features including the top spot.

Classification

- Performed 5-fold cross-validation using several classical machine learning models.
- Linear Support Vector classifier outperformed others.
- Experimented with three different combinations of feature sets.
- Combination of both feature sets improves overall performance.

References