Misinfo Workshop 2019

Differences in Health News from Reliable and Unreliable Media

Sameer Dhoju[1], **Md Main Uddin Rony**[1], Muhammad Ashad Kabir[2], Naeemul Hassan[1]

[1]The University of Mississippi, [2]Charles Sturt University

THE UNIVERSITY of MISSISSIPPI 1848

dear.lab

Charles Sturt University

Motivation

- Deluge of misleading health news over social media
  - More than 50% of the top-20 Facebook stories containing “cancer” in headline were False. [Katie Forster 2017]
  - “Pricking someone’s fingers and ears during a stroke can save their life” - went viral. [Daniel Funke 2019]
- Click-through-rate (CTR) -based pay policies intensify the phenomenon
  - Bots in social networks significantly promote unsubstantiated health-related claims.
- Alarming for general people
  - 35% of U.S. adults have gone online to self-diagnose a medical condition.[Michelle Castillo 2013]

THE UNIVERSITY of MISSISSIPPI

dear.lab

Charles Sturt University

Motivation

- Health misinformation can be critical
  - Fake news about vaccine caused measles outbreak in Europe [Muiris Houston 2018].
  - Can damage the credibility of the health-care providers and create a lack of trust in taking medicine, food, and vaccines.
- Health misinformation is a relatively unexplored area
  - Lack of reliable entities to debunk health misinformation.
  - Very few computational approaches with limited success.

Doctor Blows Whistle on Flu Shot: 'It's Designed to Spread Cancer'

April 26, 2018 by Edward Morgan

Share Tweet Newsletter Signup

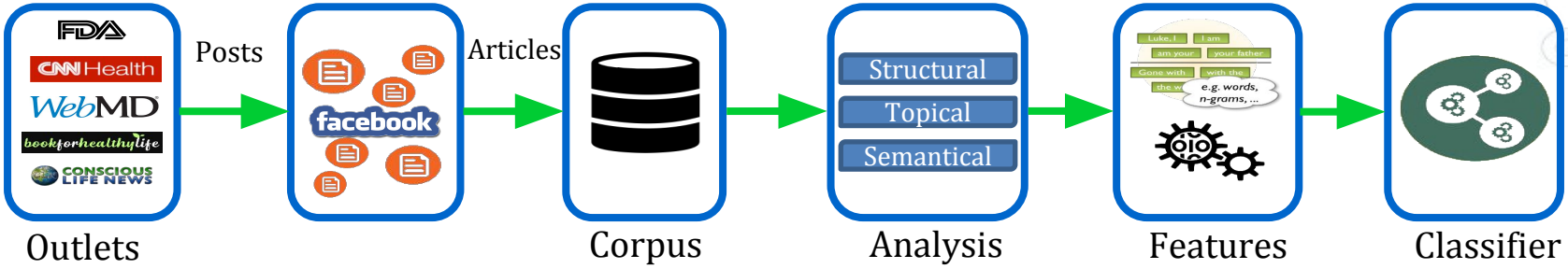
Dr. John Bergman issues warning to the public of 'flu panic'. Dr. John Bergman says the flu vaccine is laced with cancer-causing ingredients. A top doctor has gone on the record to blow the whistle in a video statement and reveal that flu vaccines have been laced with "cancer-causing ingredients."

THE UNIVERSITY of MISSISSIPPI 1848

dear.lab

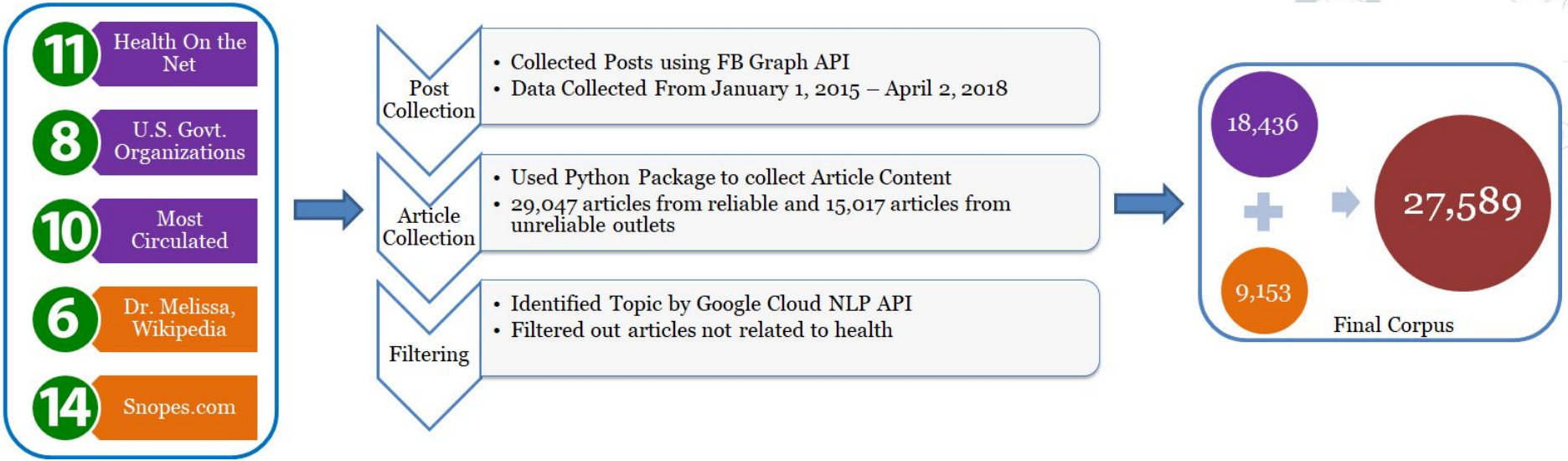
Charles Sturt University

Workflow


FDA
CNN Health
WebMD
bookforhealthylife
CONSCIOUS LIFE NEWS
Posts
facebook
Articles
Structural
Topical
Semantical
e.g. words, n-grams, ...
Outlets
Corpus
Analysis
Features
Classifier

Data Preparation

Media Outlet Selection & Data Collection



THE UNIVERSITY of MISSISSIPPI

dear.lab

Charles Sturt University

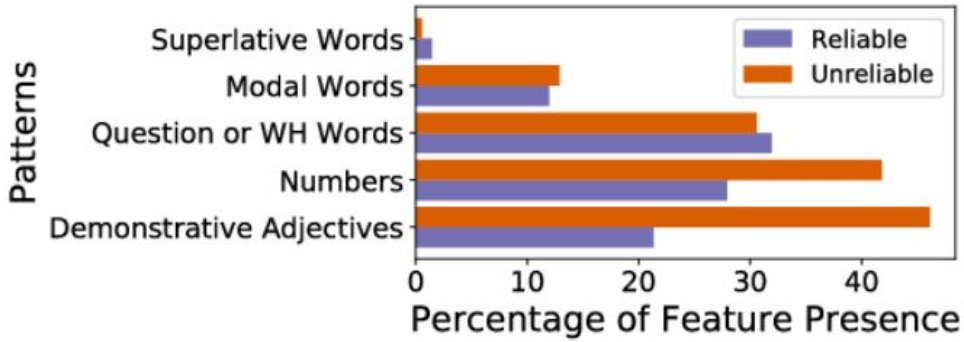
Analysis

Structural Analysis

- Headline is the most important part of a news article.
  - Only 4 out of 10 Americans read beyond the headline. [Breaux, C. (2015)]
- A longer headline receives more click than a short line does. [Breaux, C. (2015)]
- Unreliable outlets (**12.13 words/headline**) use longer headlines than reliable outlets. (**8.56 words/headline**)
- An unreliable outlet's headline has a higher chance of receiving more clicks or attention than a reliable outlet's headline.

Structural Analysis

- Examined the clickbaityness of the headlines.
  - Used two supervised clickbait detection model (Cohen's $\kappa = 0.44$)
  - Considered headline as a clickbait if both models labeled it as clickbait.
- Unreliable outlets (**40.03%**) practice more clickbait than reliable outlets (**27.29%**).
- Unreliable outlets use demonstrative adjective and numbers significantly more than the reliable outlets.



THE UNIVERSITY of MISSISSIPPI 1848

dear.lab

Charles Sturt University

Topical Analysis

- Used Latent Dirichlet Allocation (k=3).
- Representations are different for the common topic, e.g. “Cancer”
  - In reliable outlets, the topic is associated with research studies, facts, and references.
  - In unreliable outlets, the discussions are on an unsubstantiated claim - how vaccines put people under autism and cancer risk.

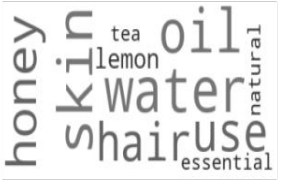
(a) RT1

(b) RT2

(c) RT3

(d) UT1

(e) UT2



(f) UT3

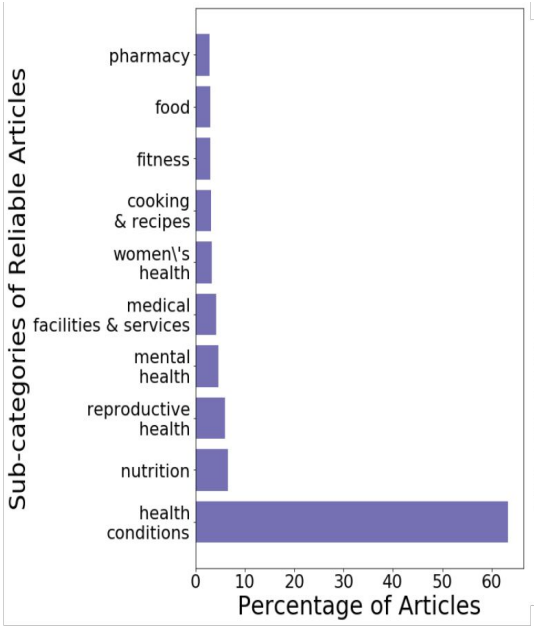
THE UNIVERSITY of MISSISSIPPI

dear.lab

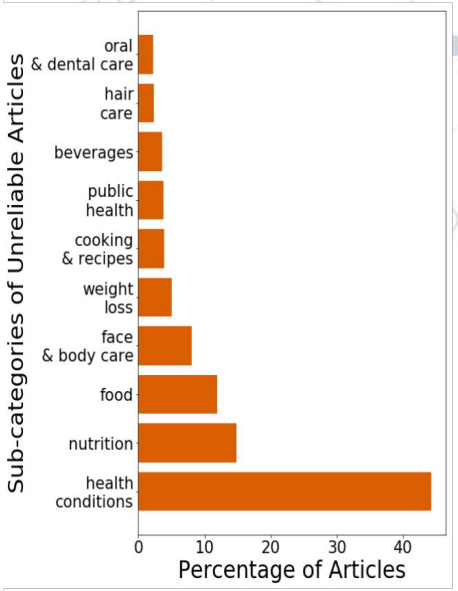
Charles Sturt University

Topical Analysis

- Identified topic by Google Cloud NLP API.
- For reliable, the distribution is significantly dominated by **health condition.**
- Percentages of **nutrition** and **food** are noticeable for unreliable outlets.
- Reliable and Unreliable outlets cover different topics.
  - Only 4 of the 10 categories are common.



(a) Reliable



(b) Unreliable

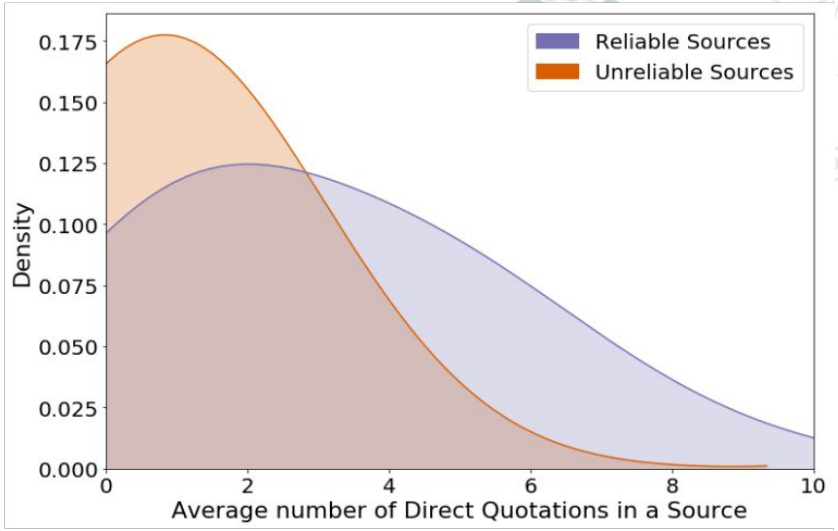
THE UNIVERSITY of MISSISSIPPI

dear.lab

Charles Sturt University

Semantic Analysis

- Use of quotations and links indicates credibility of an article [Sundar, S. S. (1998), De Maeyer, J. (2012)].
- Used the **Stanford QuoteAnnotator** to identify the quotations from a news article.
- Reliable outlets **(3/article)** use more number of quotes than unreliable outlets **(1/article)**.



(a) Quotation

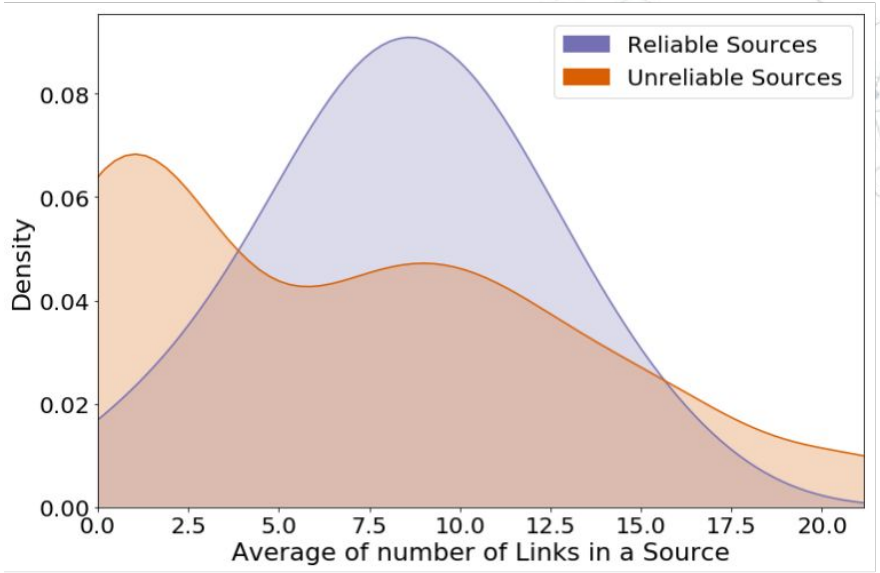
THE UNIVERSITY of MISSISSIPPI 1848

dear.lab

Charles Sturt University

Semantic Analysis

- On average, a reliable outlet sourced article contains **8.4** hyperlinks and an unreliable outlet sourced article contains **6.8** hyperlinks.
- Articles from reliable outlets (**median 8**)contain more hyperlinks than the articles from unreliable outlets (**median 2**).



(b) Link

THE UNIVERSITY of MISSISSIPPI

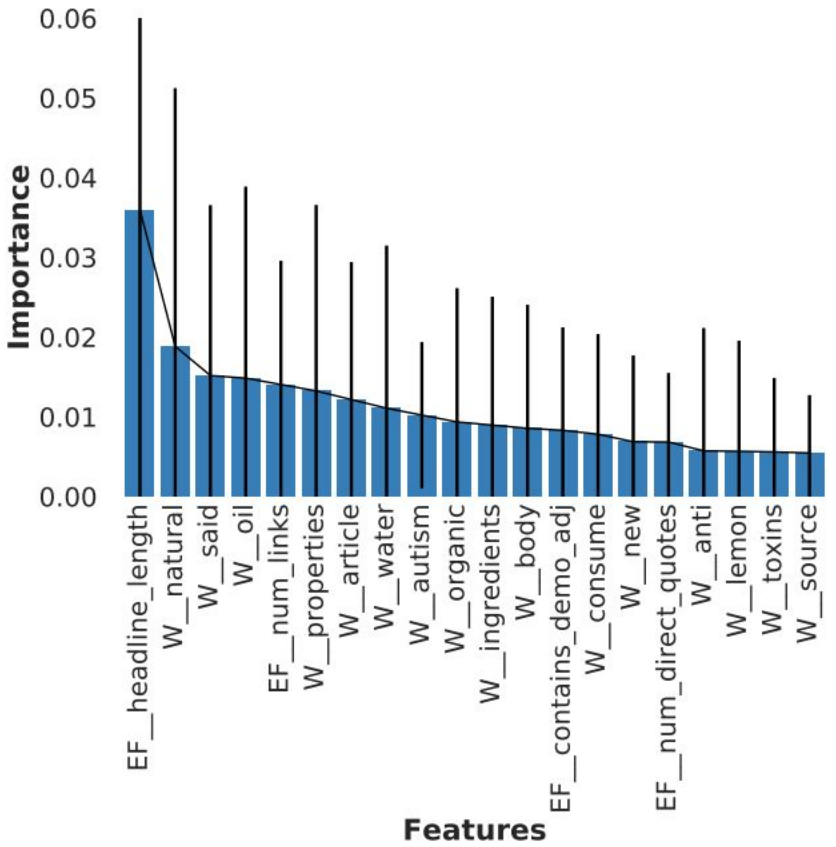
dear.lab

Charles Sturt University

Source Classification

Feature Importance

- Word (W): Took 5, 000 most frequent n-gram (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.



THE UNIVERSITY of MISSISSIPPI

dear.lab

Charles Sturt University

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.

Features	Labels	Precision	Recall	F-1
Word (W)	Unreliable	0.94	0.92	0.93
	Reliable	0.96	**0.97**	**0.97**
	Macro-Avg	0.95	**0.95**	0.95
Extracted Features (EF)	Unreliable	0.76	0.47	0.58
	Reliable	0.78	0.93	0.85
	Macro-Avg	0.77	0.70	0.72
W + EF	Unreliable	**0.95**	**0.93**	**0.94**
	Reliable	**0.97**	**0.97**	**0.97**
	Macro-Avg	**0.96**	**0.95**	**0.96**

Conclusion and Future Work

- Analyzed structural, topical, and semantic differences between articles from reliable and unreliable outlets.
- Identified some patterns that can potentially help classify articles of reliable outlets from unreliable outlets.
- Our classification model showed better performance with the inclusion of these patterns.
- In future, we want to incorporate the videos, cited experts, users' reaction and other metadata in combating health disinformation.

References

- De Maeyer, J. (2012). The journalistic hyperlink: Prescriptive discourses about linking in online news. Journalism Practice, 6(5-6), 692-701.
- Sundar, S. S. (1998). Effect of source attribution on perception of online news stories. Journalism & Mass Communication Quarterly, 75(1), 55-68.
- Chris Breaux. (accessed September 28, 2018). "You'll Never Guess How Chartbeat's Data Scientists Came Up With the Single Greatest Headline". https://tinyurl.com/nleq7ph
- Muiris Houston. (accessed October 31, 2018). Measles back with a vengeance due to fake health news. https://tinyurl.com/y4a7bbak
- Katie Forster. (accessed October 30, 2018). Revealed: How dangerous fake health news conquered Facebook. https://tinyurl.com/y2plehsu
- Daniel Funke. (accessed May 12, 2019). On Facebook, health misinformation is king. And it's a global problem. https://tinyurl.com/y2kujtar
- Michelle Castillo. (accessed May 2012, 2019). More than one-third of U.S. adults use Internet to diagnose medical condition. https://tinyurl.com/y2655995

Feel free to contact at:

nhassan@olemiss.edu, mrony@go.olemiss.edu

thank you!

THE UNIVERSITY of MISSISSIPPI

dear.lab

Charles Sturt University