Investigating the Trustworthiness of Wikipedia and the Media in the Scope of COVID-19

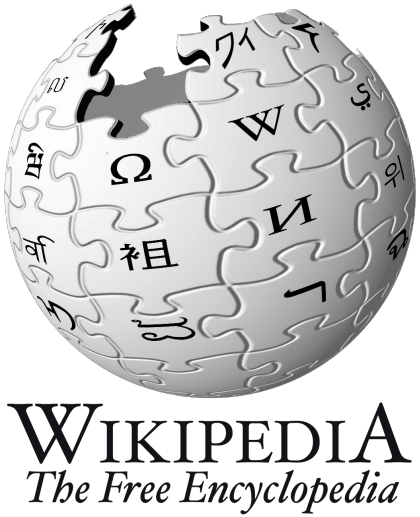
Scott Huang, Calvin Tam, Leena Elamrawy

Overview: Wikipedia vs Media?

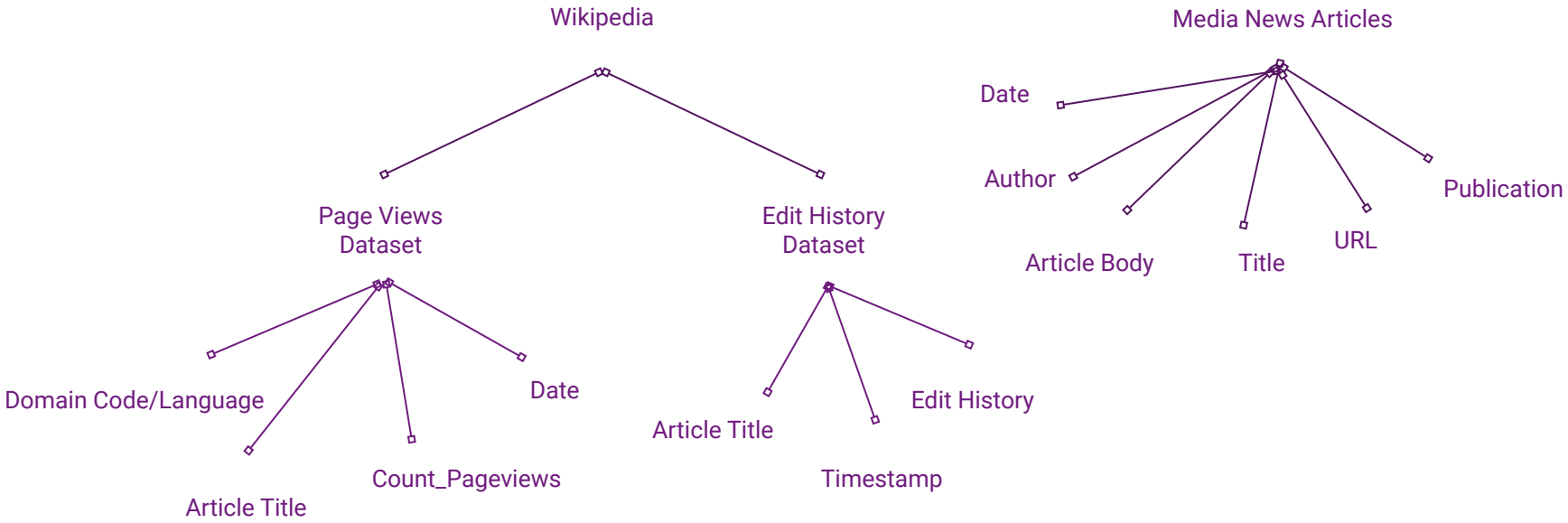
Question: How biased are certain popular news outlets in how they cover the effects and the spread of COVID-19, and how does the language used in these news articles compare to the language used in Wikipedia articles?

Hypothesis: Wikipedia vs Media?

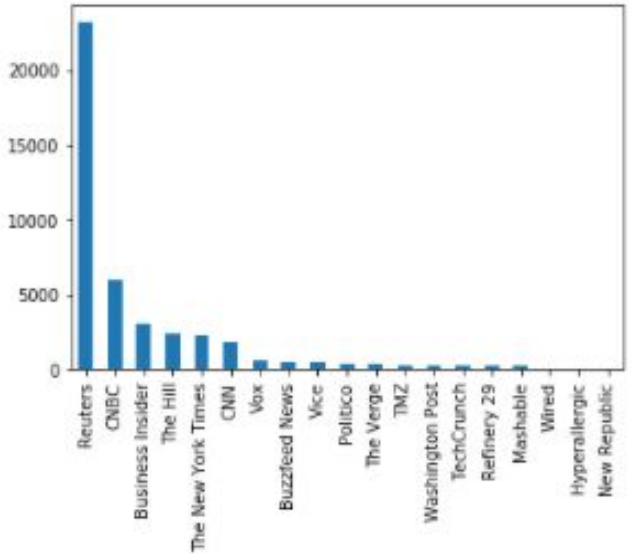
We hypothesize that certain historically unbiased, trustworthy journalism brands such as the New York Times will be the least biased, and that Wikipedia will be utilizing more biased language than that of the respected, reputable news sources.

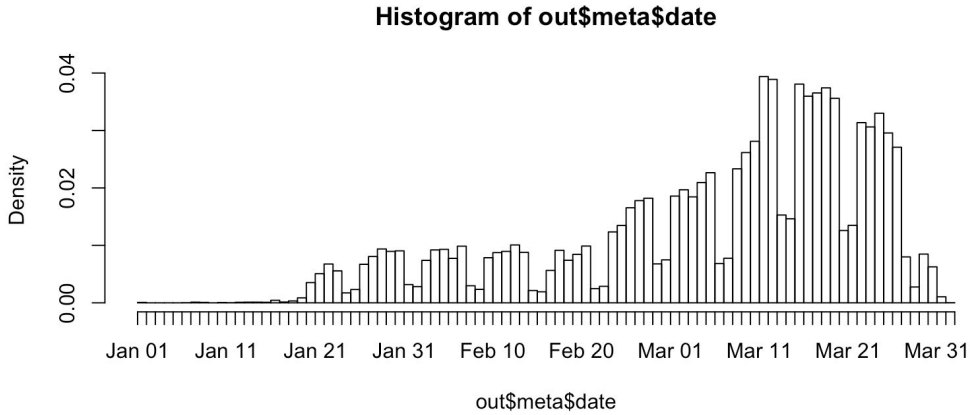


Data Ingestion Process

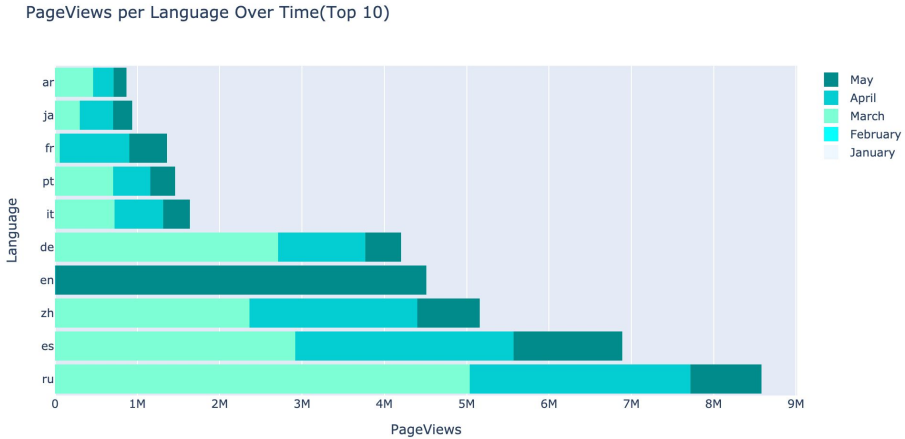


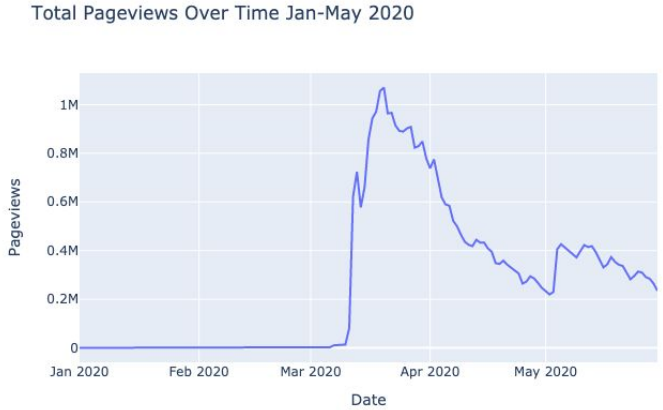
News Articles Dataset

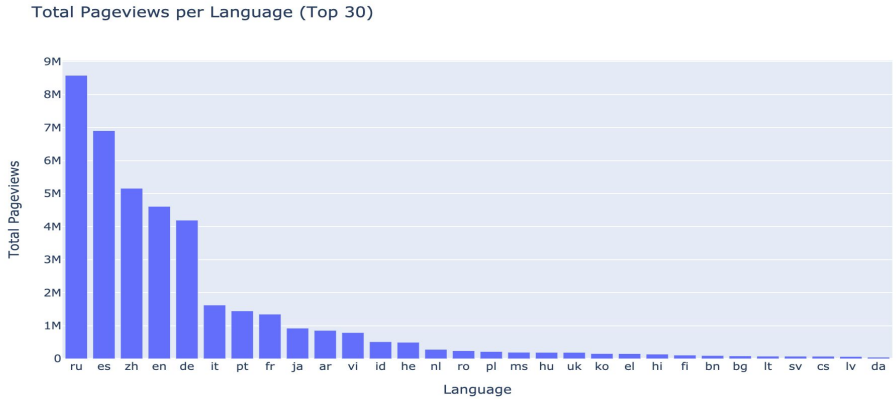


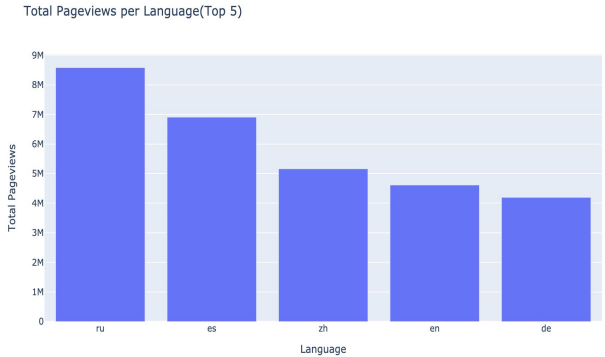


Wikipedia Dataset





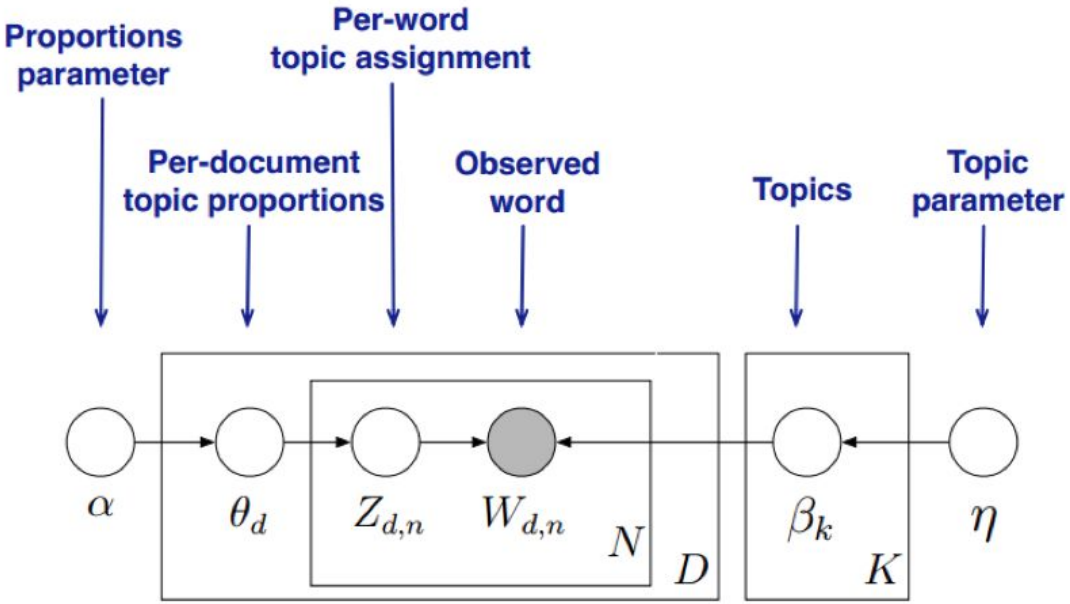




Methods

- Topic Modeling
  - Latent Dirichlet allocation(LDA)
  - Structural Topic Model(STM)
- Topic Distribution
- Distance Functions
  - Wasserstein Distance
  - Frobenius Norm

Latent Dirichlet allocation - LDA



Choose $\theta \sim \mathrm{Dir}(\alpha)$.

a = Probability on per document topic distribution

$\mathcal{B}$ = Probability on per topic word distribution

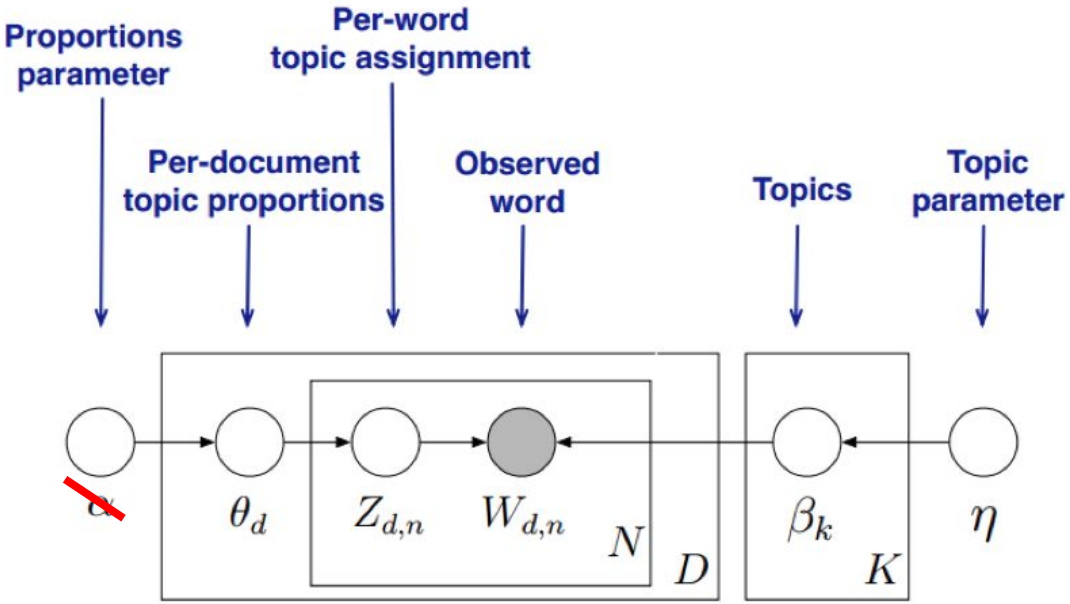
Θ_m = The topic distribution for document M

Φ_k = The word distribution for Topic K

Z_{mn} = The topic for the n-th word in document M

W_{mn} = The specific word

Structural Topic Model - STM



a = Probability on per document topic distribution

$\mathcal{B}$ = Probability on per topic word distribution

Θ_m = The topic distribution for document M

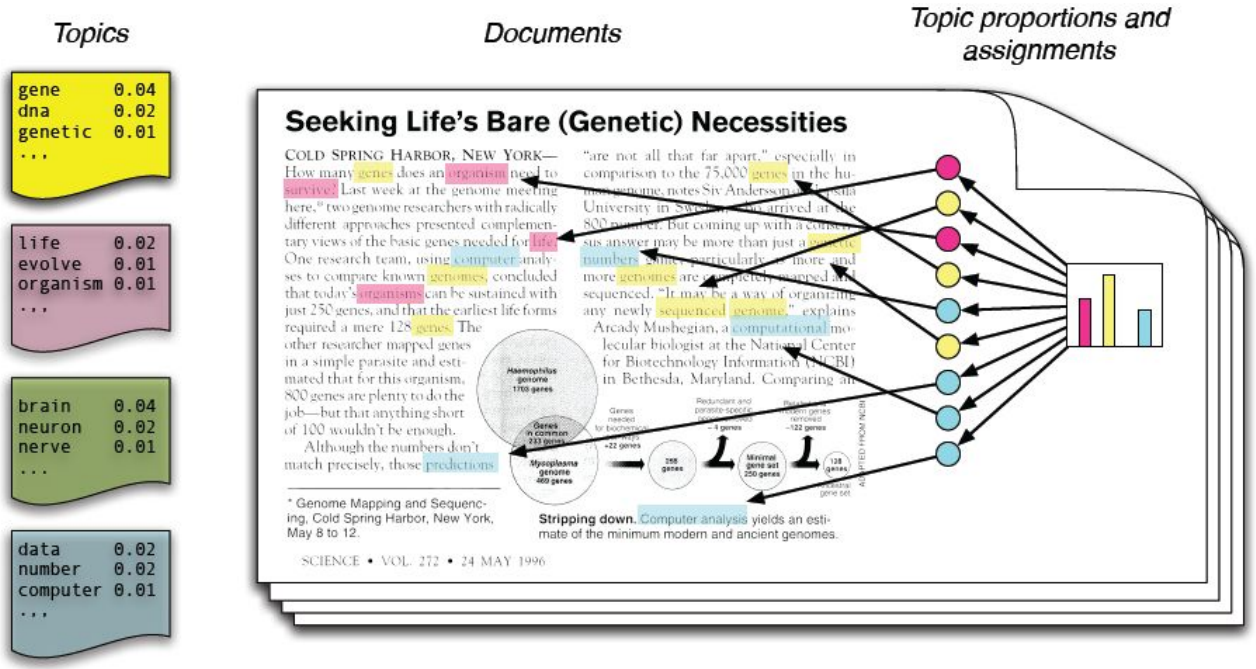
Φ_k = The word distribution for Topic K

Z_{mn} = The topic for the n-th word in document M

W_{mn} = The specific word

$$\vec{\theta}_d | X_d\gamma, \Sigma \sim \text{LogisticNormal}(\mu = X_d\gamma, \Sigma)$$

Topic Distribution

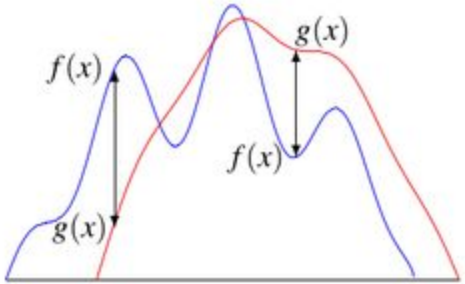


Topic Distribution

$$A_{n \times m} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix}_{n \times m}$$

Where n is the number of days of the time interval,
m is the number of the topics,
Aij = is the proportion of topic j of all content published on Day i

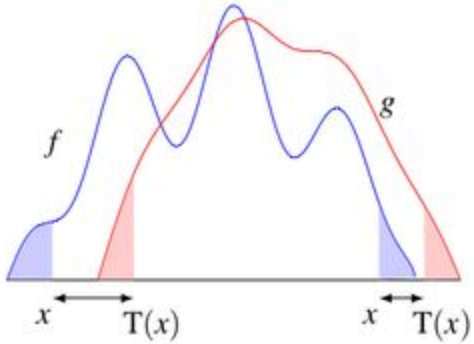
Distance Functions



Frobenius Norm

$$||A||_F = [\sum_{i,j} abs(a_{i,j})^2]^{1/2}$$

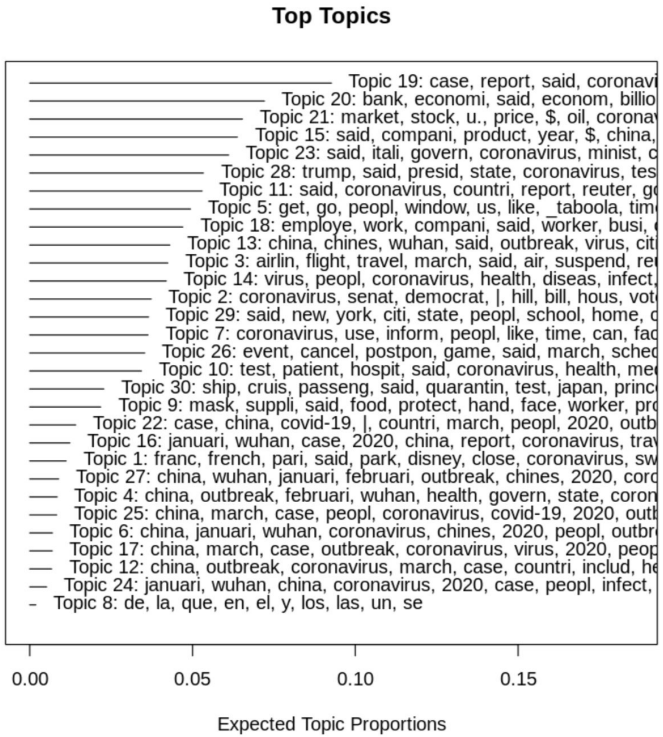
where $A = \mu$ - ν



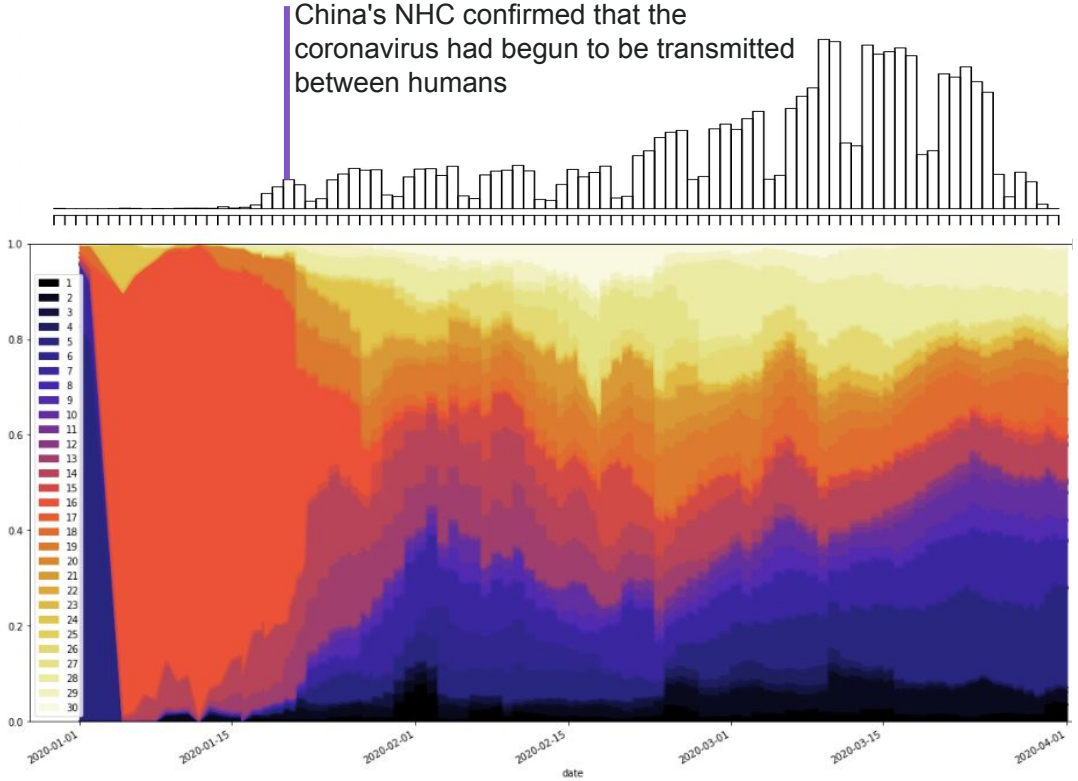
Wasserstein Distance

$$W_p(\mu,\nu) := \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M\times M} d(x,y)^p \, \mathrm{d}\gamma(x,y) \right)^{1/p},$$

Result - STM topics

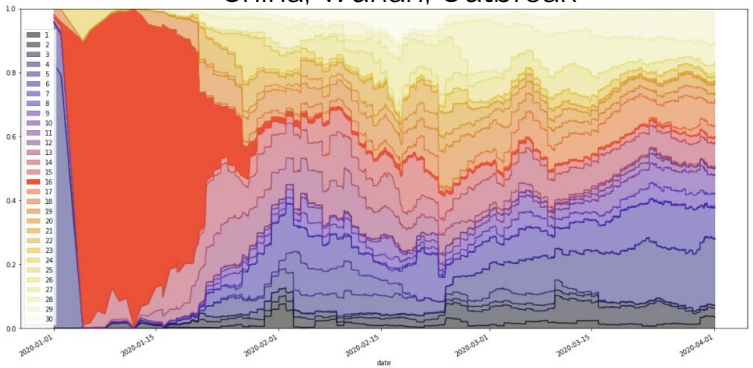


Results - Topic Distribution

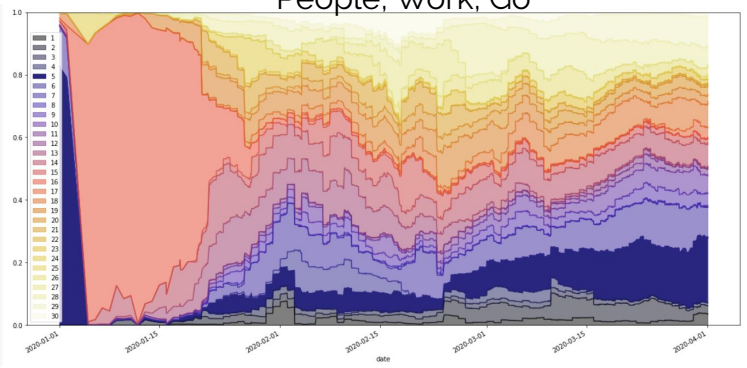


Results - Plot by Topic

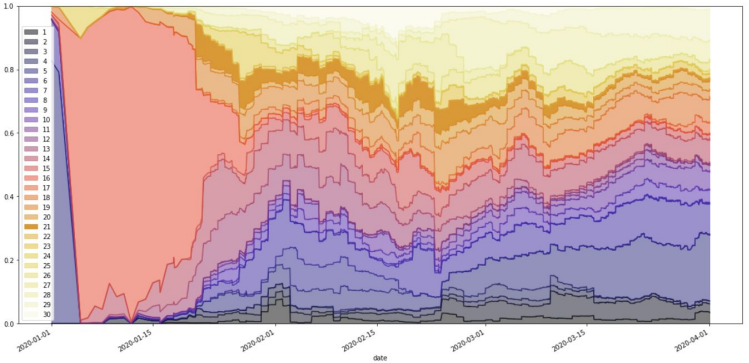
China, Wuhan, Outbreak



People, Work, Go



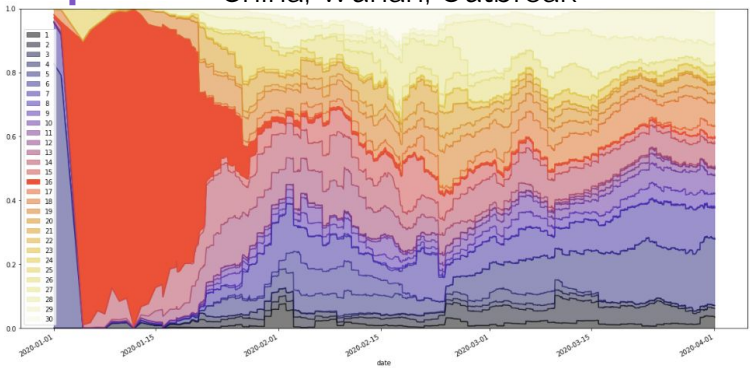
Market, Stock, Economy



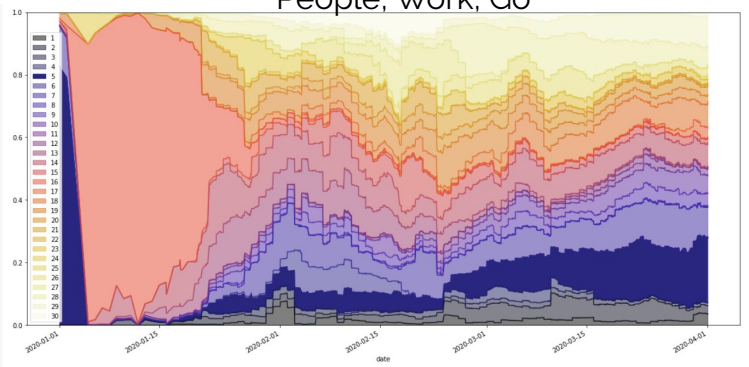
Jan 1, 2020

Xinhua News, the Huanan Seafood Market in Wuhan was closed on 1 January 2020 for "renovation"

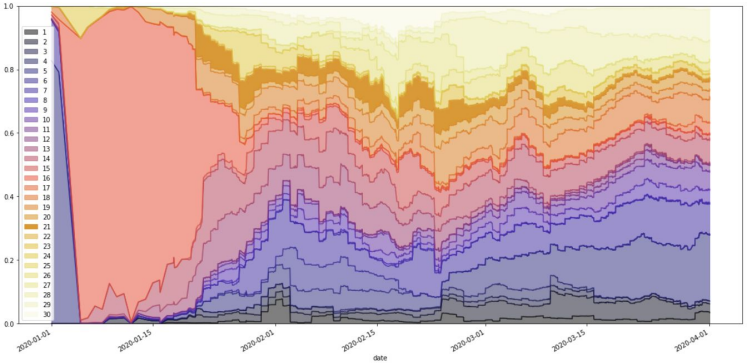
China, Wuhan, Outbreak



People, Work, Go



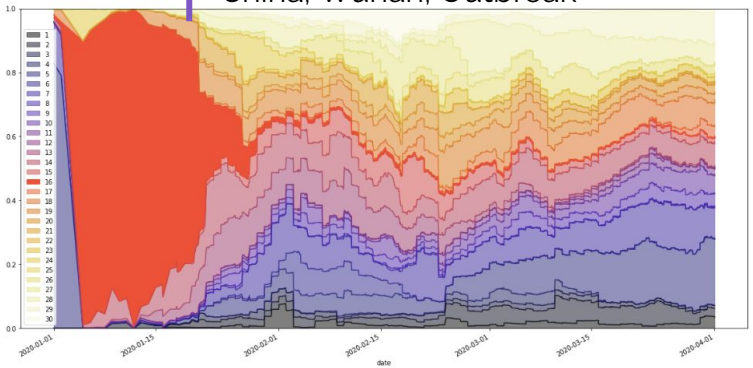
Market, Stock, Economy



Jan 20, 2020

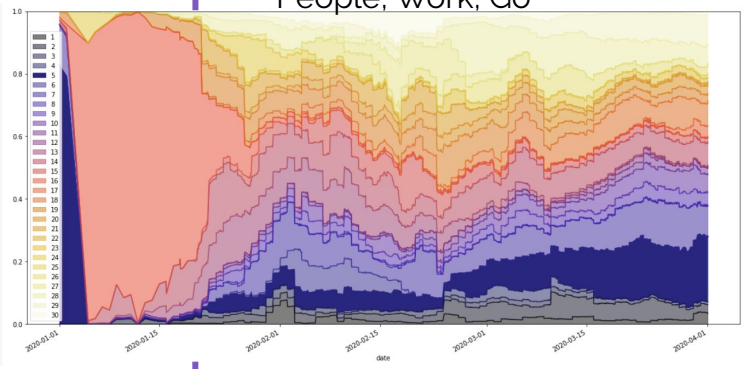
China's NHC confirmed that the coronavirus had begun to be transmitted between humans

China, Wuhan, Outbreak

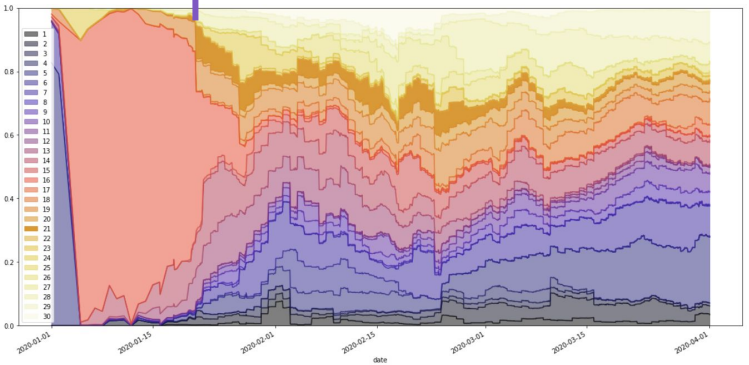


China's NHC confirmed that the coronavirus had begun to be transmitted between humans

People, Work, Go



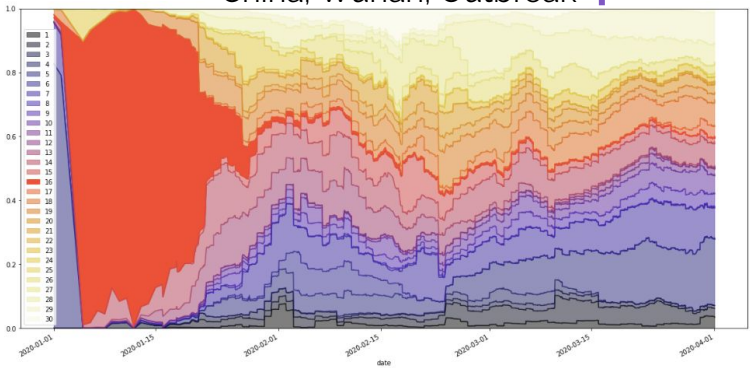
Market, Stock, Economy



March 22, 2020

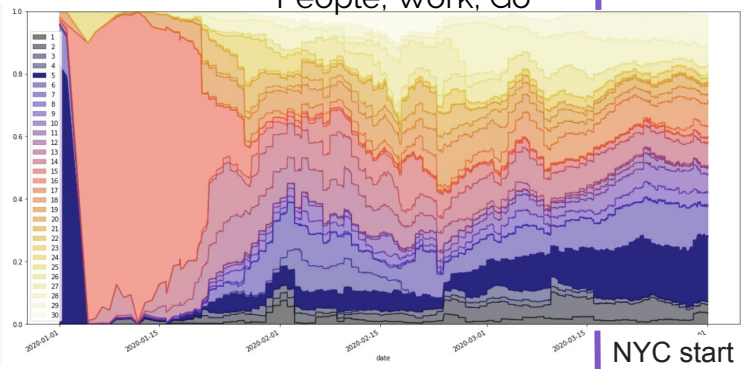
China, Wuhan, Outbreak

NYC start stay-at-home order



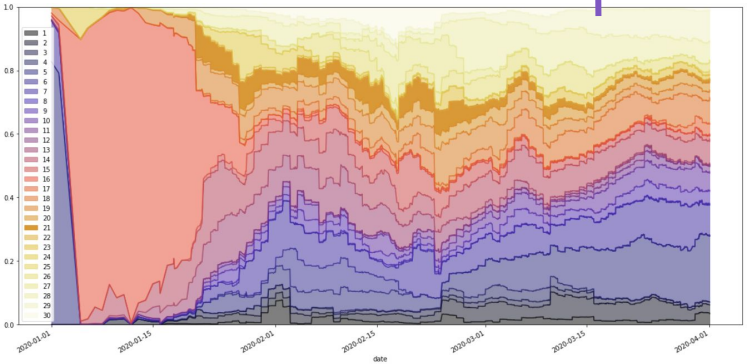
People, Work, Go

NYC start stay-at-home order



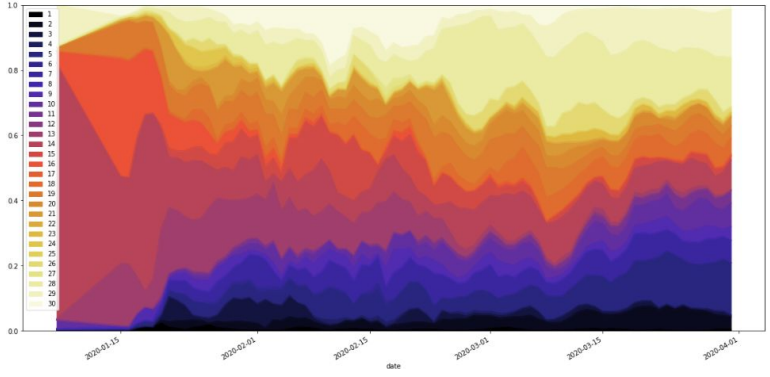
Market, Stock, Economy

NYC start stay-at-home order

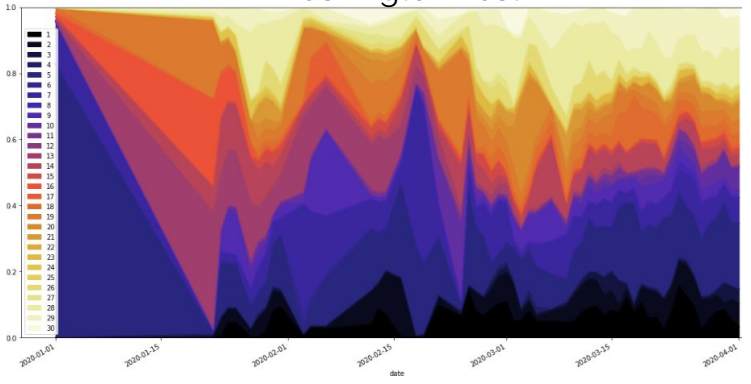


Results-Plot by Source

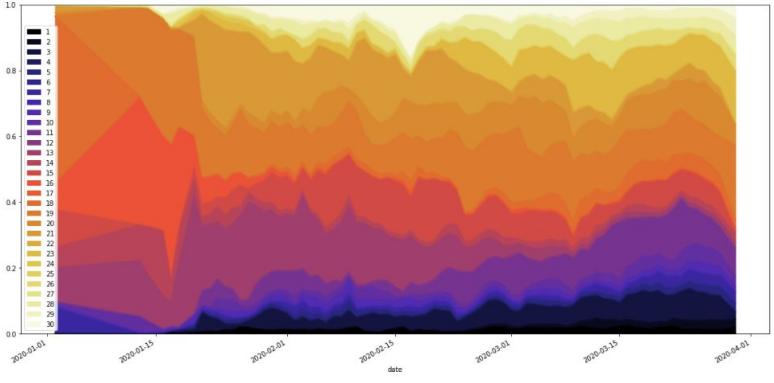
CNN



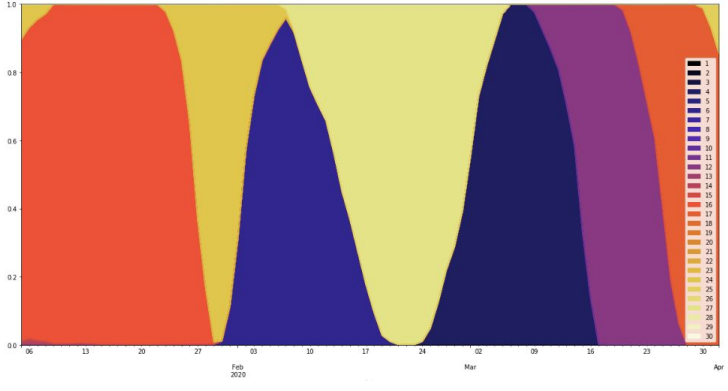
Washington Post



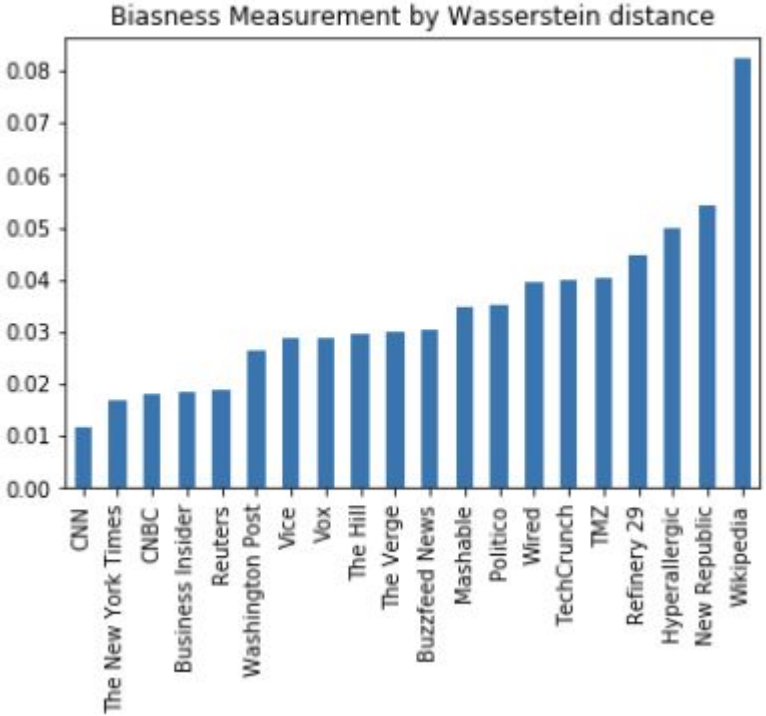
Reuters

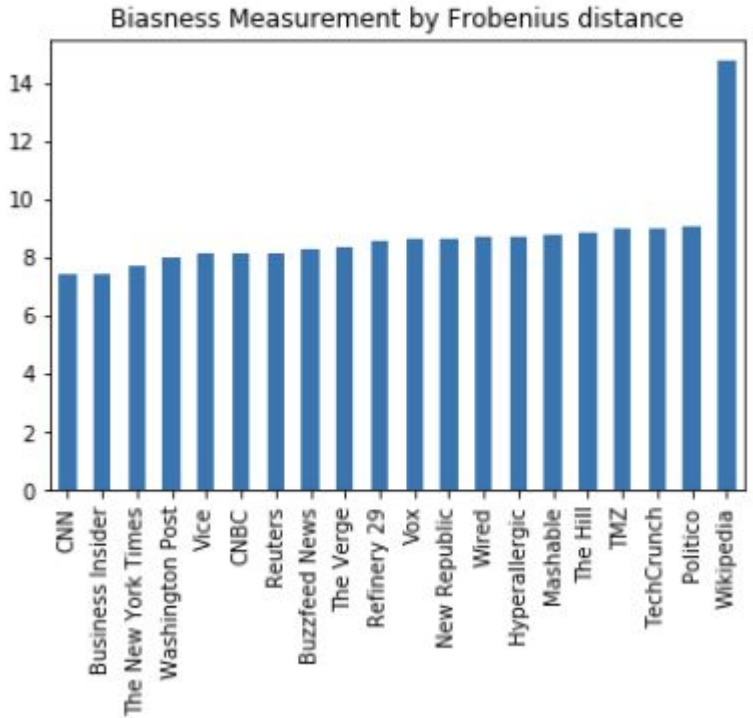


Wikipedia



Results-Biasness Measurements





Conclusions

- CNN is the most trustworthy source by both metrics
- Wikipedia is the least trustworthy source
- In general: Traditional News Agency > Online media > Wikipedia
- Result may varies if we consider more Wikipedia articles

Future Directions

- Extend the temporal scope as well as number of Wikipedia articles
- Explore Topic Modeling by deep word embeddings from BERT or XLNet
- More fine-grained semantic analysis of each topic
- Specifically look into the Russian articles since they have the most pageviews

Thank you!