Mining Public Opinion on COVID-19 Vaccines using Unstructured Social Media Data

Chad A. Melton

Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville, TN, United States

University of Tennessee Health Science Center-Oak-Ridge National Laboratory (UTHSC-ORNL) Center for Biomedical Informatics, Department of Pediatrics, College of Medicine, Memphis, TN, United States

Outline

1) Introduction

a) Purpose of Research

b) Research questions

c) Objectives and broad background.

2) **Description of data, harvesting, and preparation**.

3) **Description, Background, Methods, Results, Remarks, and Limitations for each**:

a) Topic Modeling

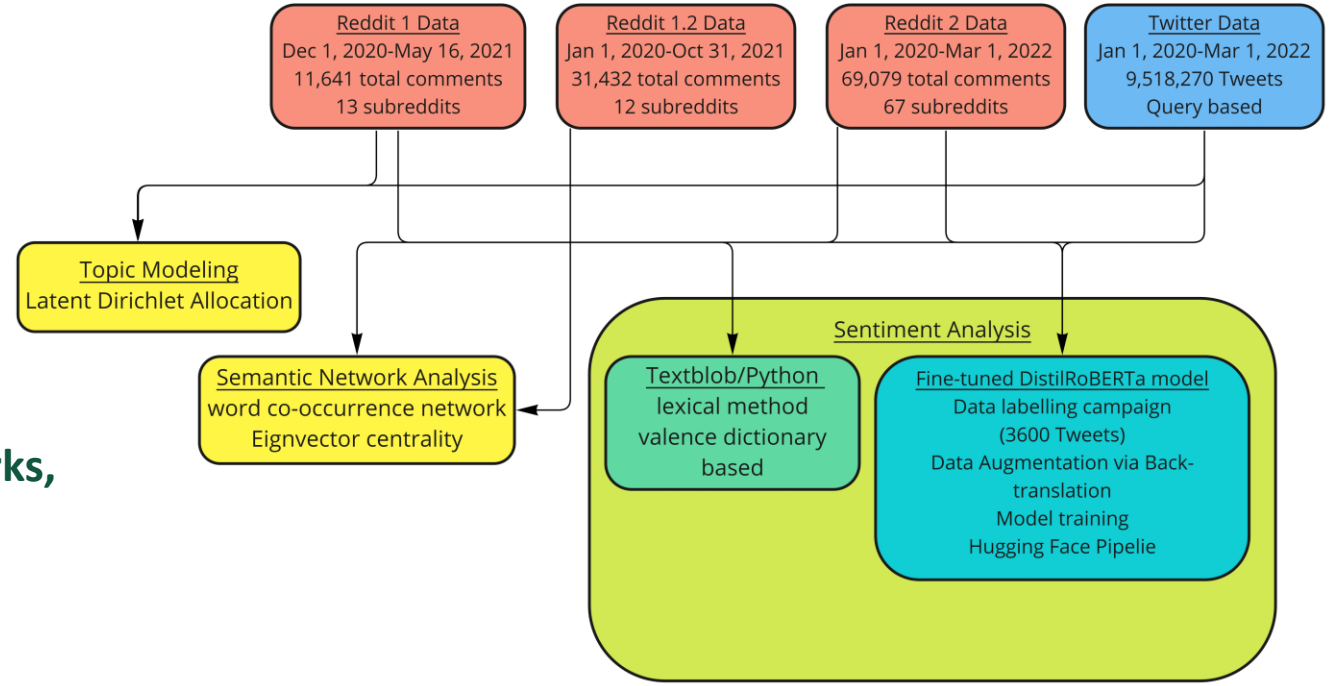
b) Network Analysis

c) Sentiment Analysis

4) **Conclusions**

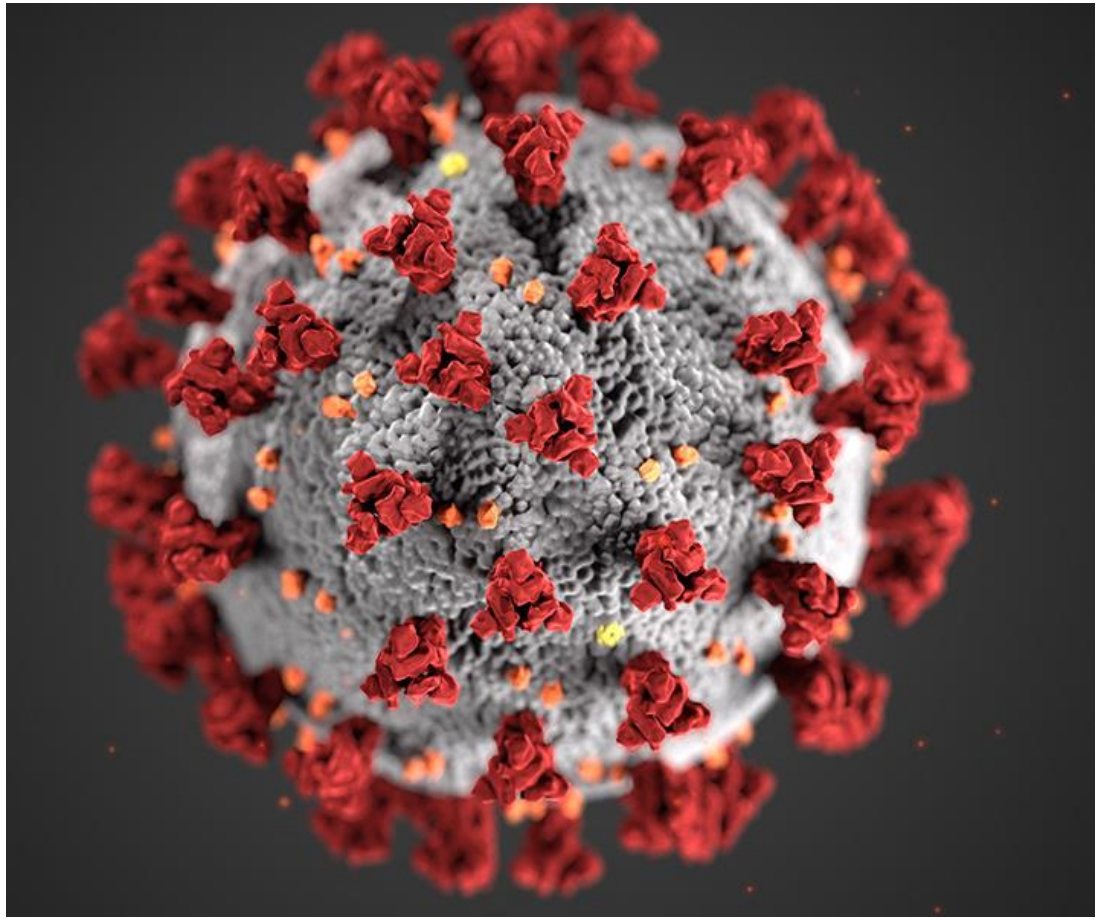
5) **Contributions of research**

6) **Future work**



Purpose of Research

- The COVID-19 pandemic fueled the most rapid vaccine development and distribution in history
- The spread of misinformation → low vaccination rates →morbidities and untimely deaths
- ~4 billion social media users → information gathering
- Sentiment Analysis, Topic Modeling, and Network analysis
- Public health messaging, education, digital interventions, and public health policies



Research questions?

Sentiment change?

Can changes be correlated with events?

Reddit vs Twitter?

Similar topics across platforms?

What are people talking about?

How much mis/disinformation is out there?

Is my approach generalizable?

Specific Objectives

Objective I. Detect misinformation/disinformation

1) topic modeling and semantic network analysis

Objective II. Document how public discourse has diffused throughout the pandemic

1) Semantic analysis through topic modeling and semantic network analysis

Objective III. Document how sentiment evolved during the course of a pandemic

1) Sentiment analysis

Public health surveillance from social media during disease outbreaks

EID	N	Studies
H1N1 (swine flu)	15	Atlani-Duault et al,[13] Biswas,[3] Chew and Eysenbach,[14] Collier et al,[15] Ding and Zhang,[16] Freberg et al,[9] Gao et al,[17] Kim and Liu,[18] Liu and Kim,[19] Luoma-aho et al,[20] Nerlich and Koteyko,[21] Pandey et al,[22] Signorini et al,[23] Tausczik et al,[24] Tirkkonen and Luoma-aho[25]
Ebola virus	10	Basch et al,[26] Househ,[27] Lazard et al,[28] Nagpal et al,[29] Odlum and Yoon,[30] Pathak et al,[31] Seltzer et al,[32] Strekalova,[33] Towers et al,[34] Wong et al[35]
H7N9 (bird flu/avian flu)	2	Fung et al,[36] Vos and Buckner[37]
West Nile virus	1	Dubey et al[38]
EHEC	1	Gaspar et al[39]
MERS-CoV	1	Fung et al[36]
Measles	1	Mollema et al[40]

Courtesy of Tang et al., (2018)

Data

- ~450 million users (50% in the US)
- Topic-based communities called “subreddits”
- Community regulations
- "upvote" or "downvote“, and share.
- Large character limit

Data (continued)

- Individualized rather than community based
- ~ 400 million accounts
- Very few restrictions on posting and shared content
- Users can “like” posts, share posts, retweet posts, or comment

Reddit data sets

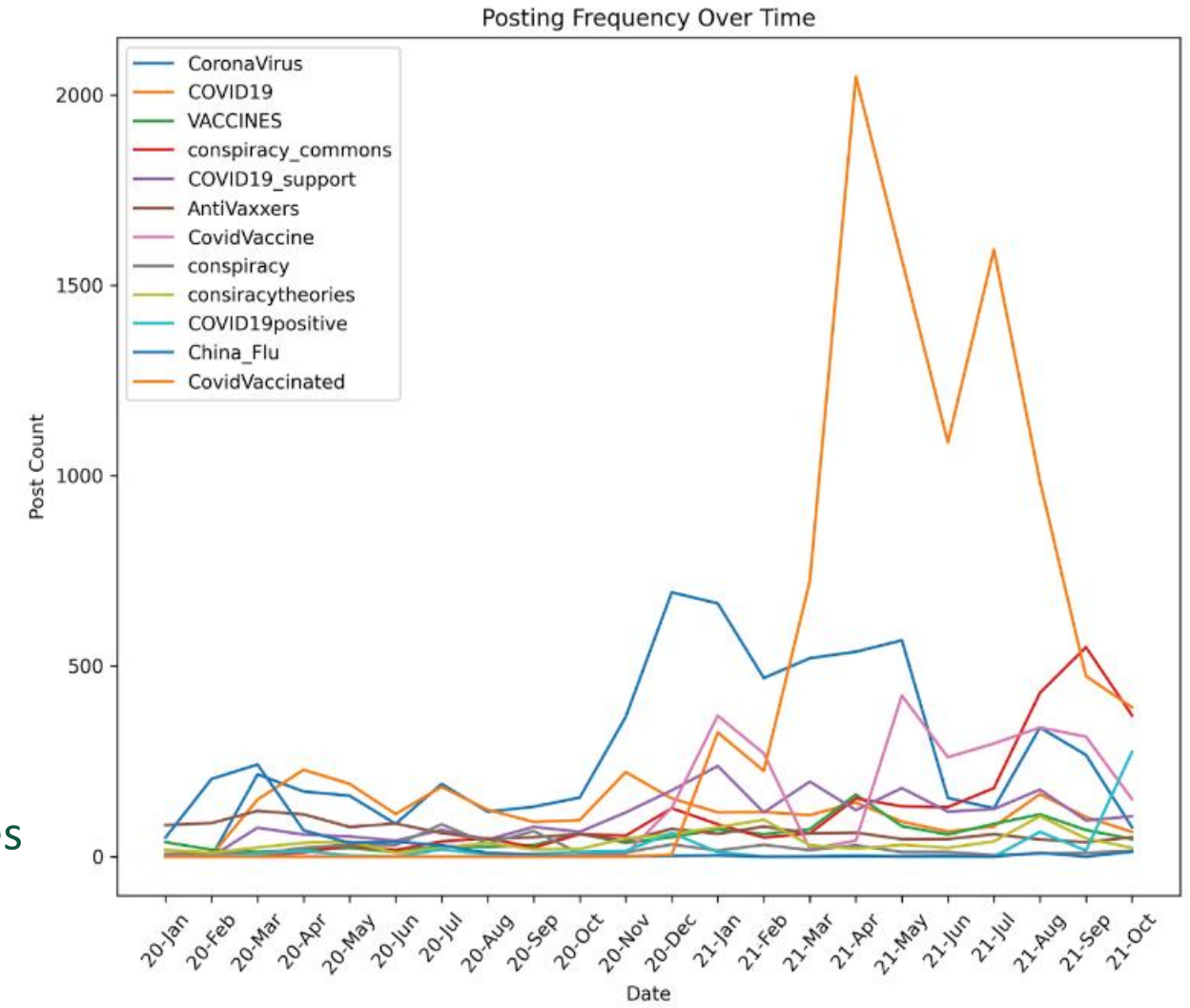
Reddit 1.0

- 18000 posts/comments from 13 subreddits
- Cleaned and queried for topic relevance
- Sorted by time (December 1, 2020 –May 16, 2021)
- Finalized data set contained 11,641 total entries
- 8021 unique authors

Reddit data sets

Reddit 1.2

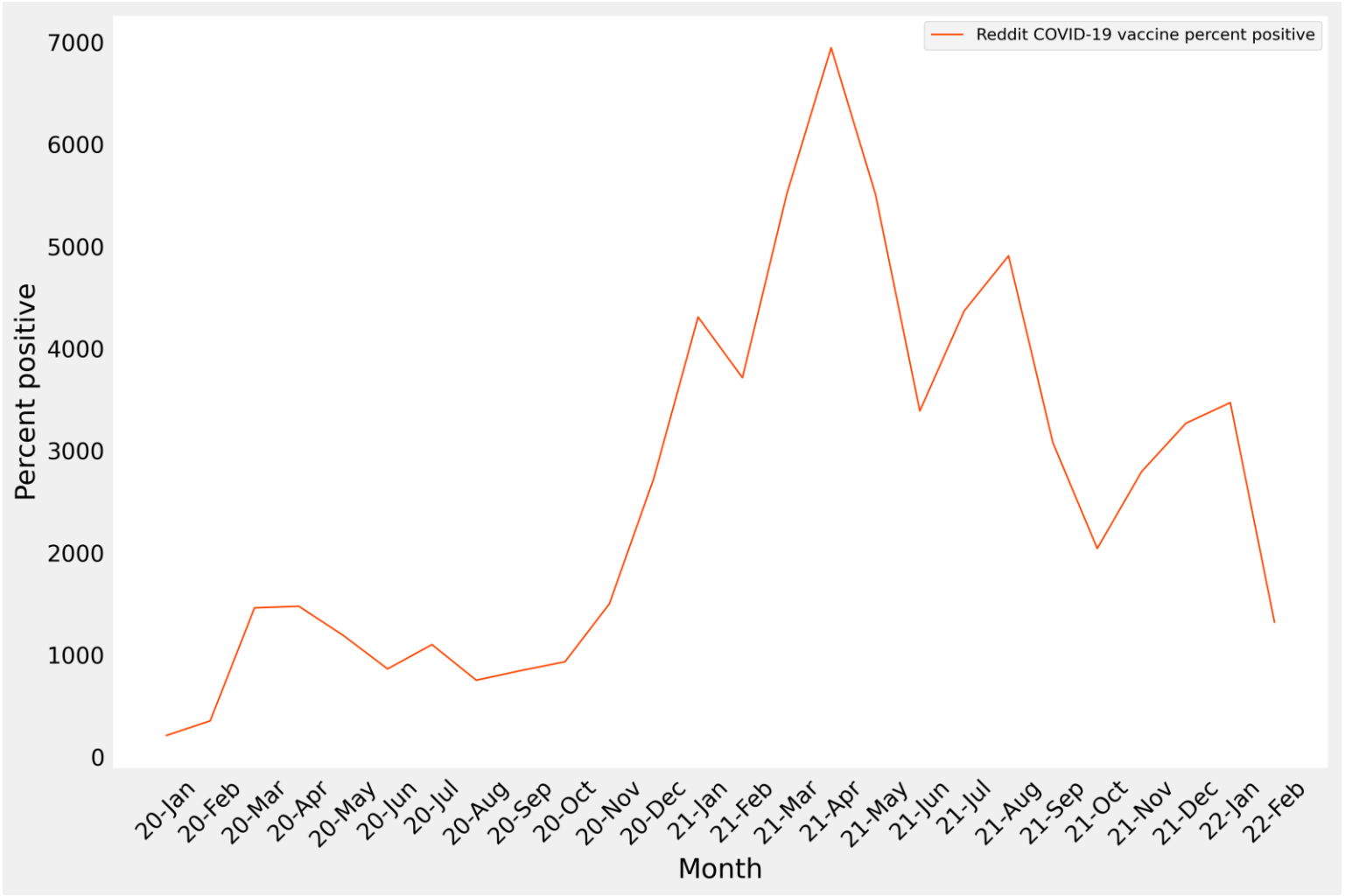
- Harvested 300,000 posts from 12 subreddits
- Cleaned and queried for topic relevance
- Sorted by time (January 1, 2020 – October 31, 2021)
- Finalized data set contained 31,432 total entries
- 20429 unique authors



Reddit data sets

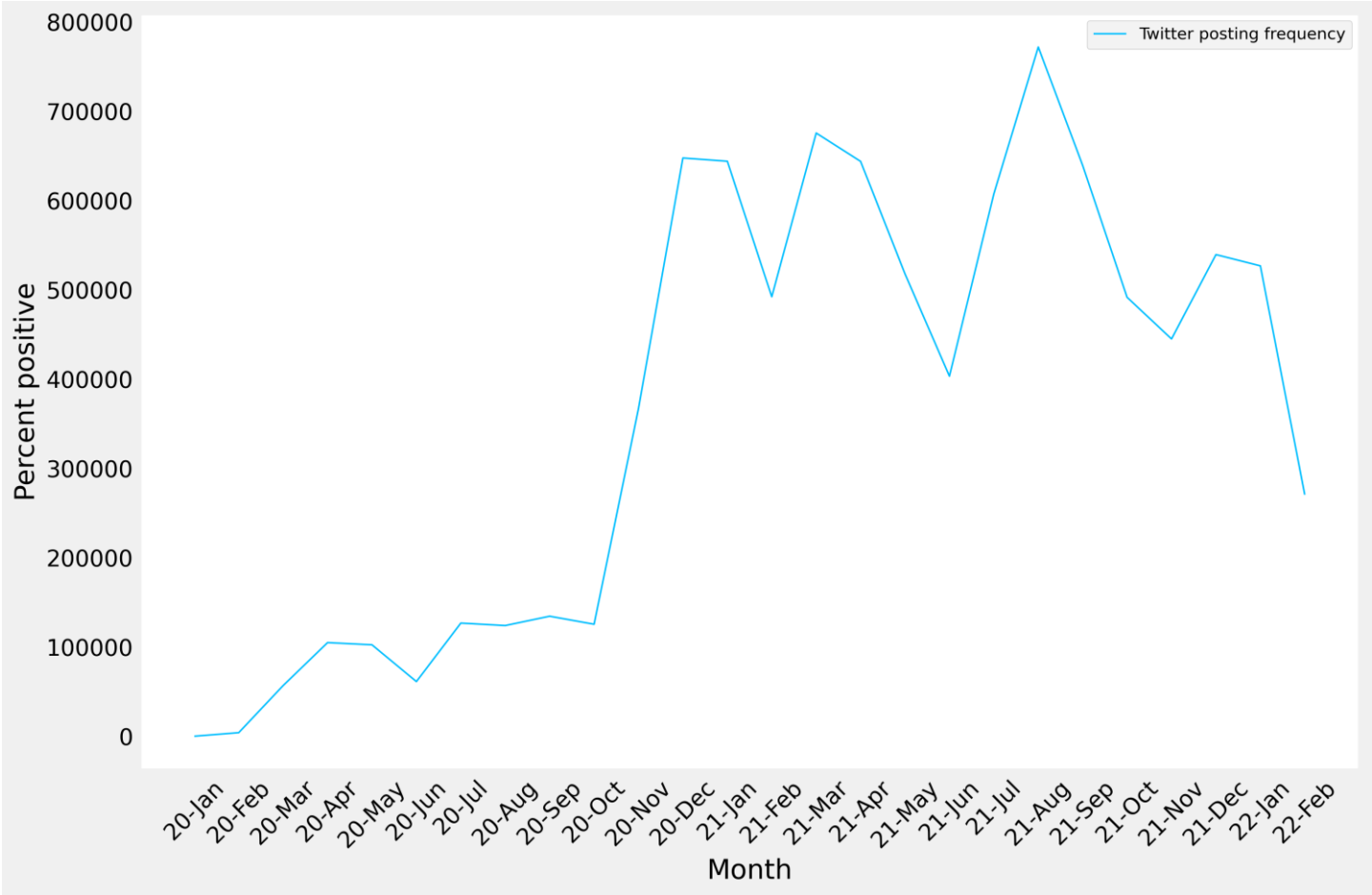
Reddit 2.0

- Harvested 579,241 posts/comments from 67 subreddits
- Cleaned and queried for topic
- Sorted by time (January 1, 2020 – March 1, 2022)
- Finalized data set contained 69,079 total entries
- ~10,000 unique authors



Twitter data

- Harvested ~13 million Tweets
- Cleaned and queried for topic relevance
- Sorted by time (January 1, 2020 – March 1, 2022)
- Finalized data set contained 9,518,270 total entries
- 3,006,075 unique authors



Comparison of Reddit and Twitter

- Twitter: Large number of Tweets
- Excellent for breaking news or facilitate early outbreak detection
- Lots of cleaning
- Reddit: plentiful in longer texts → topic modeling and semantic analysis
- Possible echo chamber effect

Topic Modeling

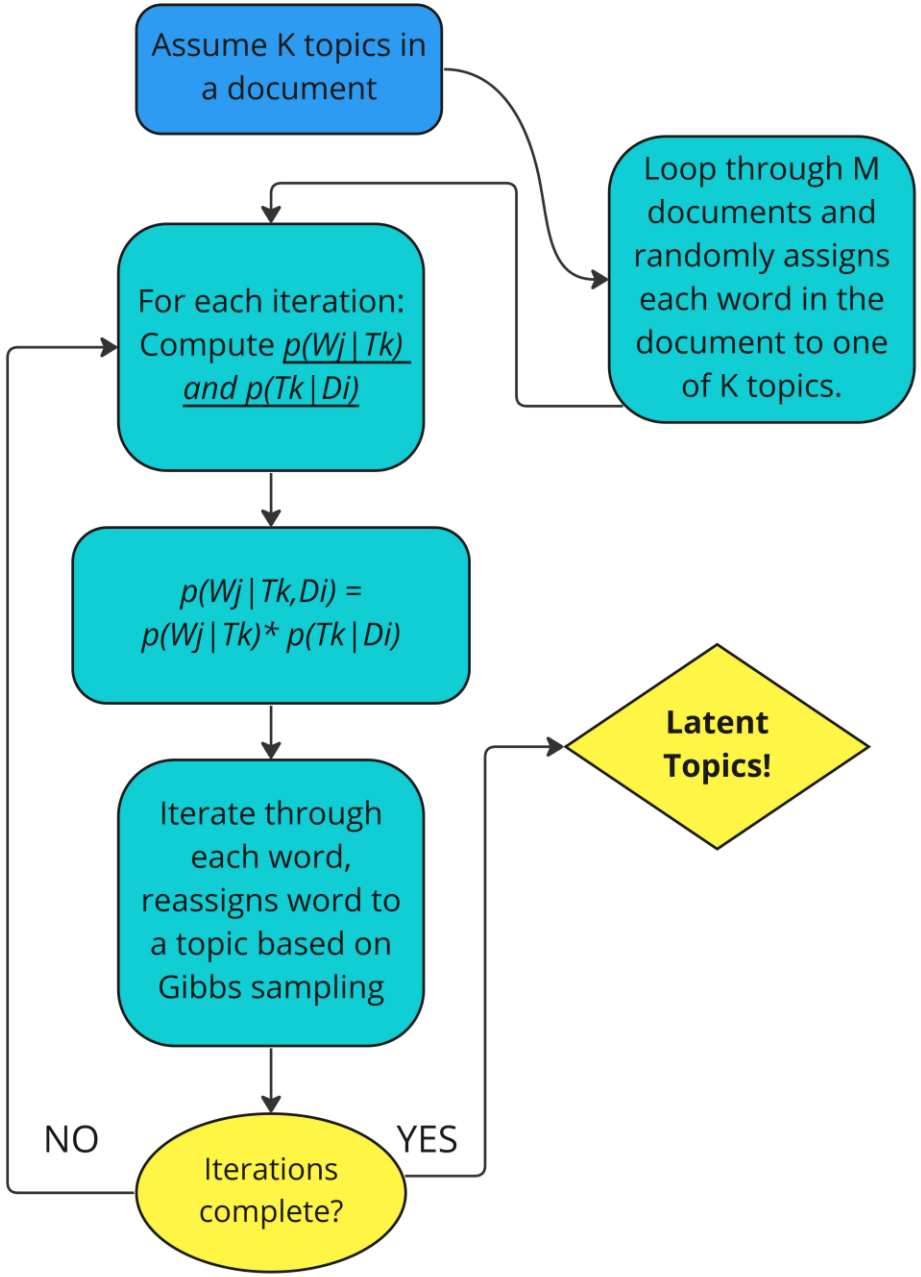
Topic modeling is a (typically) unsupervised machine learning method that is built to detect a topic or sets of topics within a corpus or corpora.

- Latent Dirichlet Allocation (LDA): Blei et at., 2002

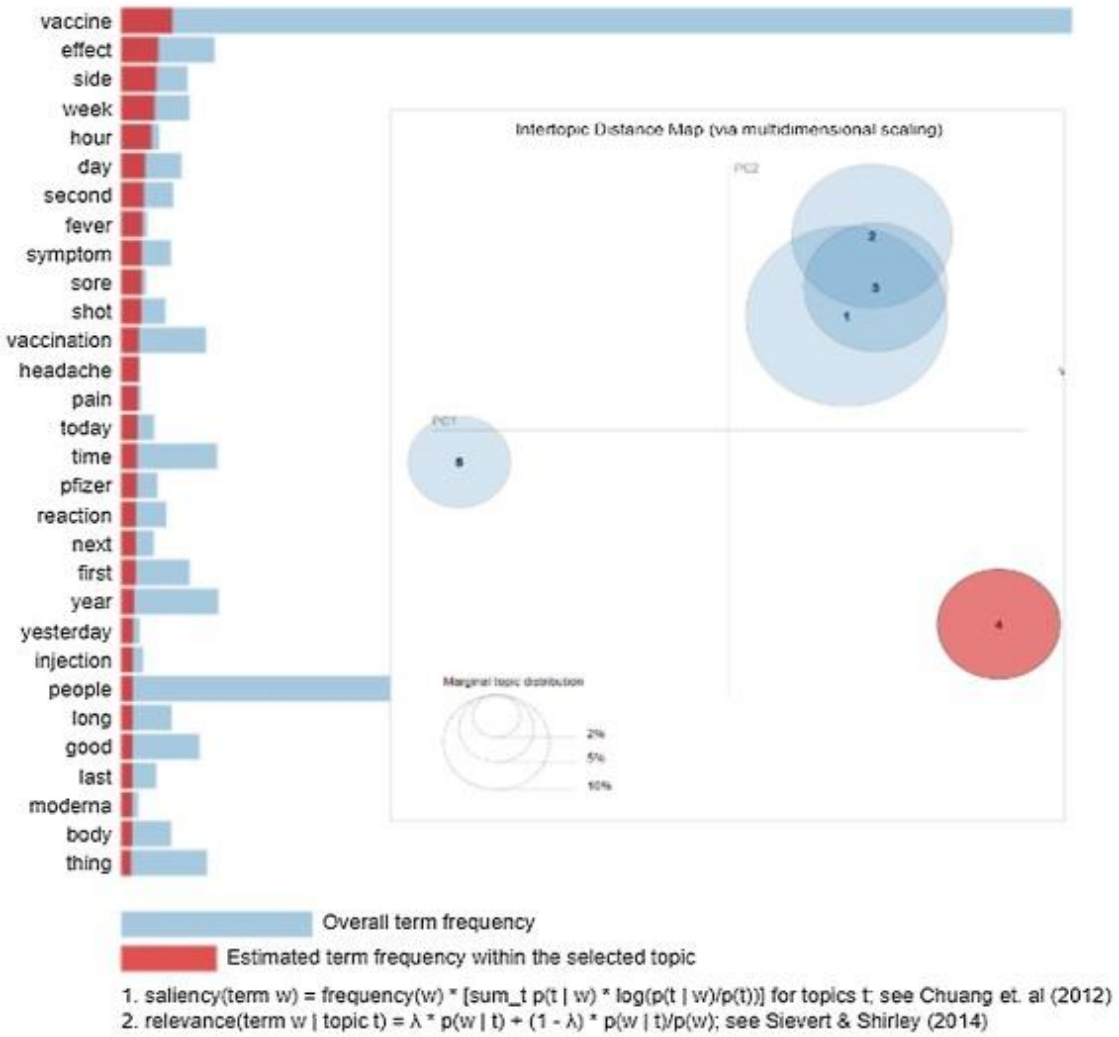
- Many diverse use cases.

LDA process

Gibbs Sampling



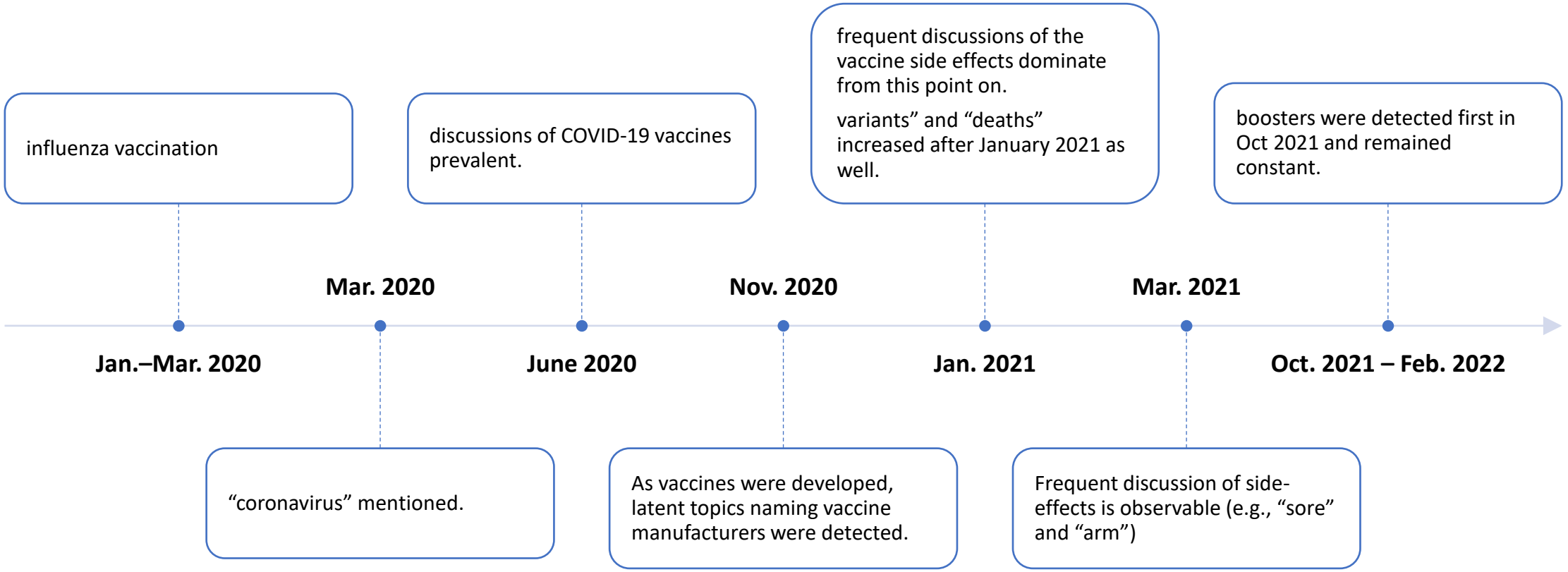
Results: Reddit 1.0 data



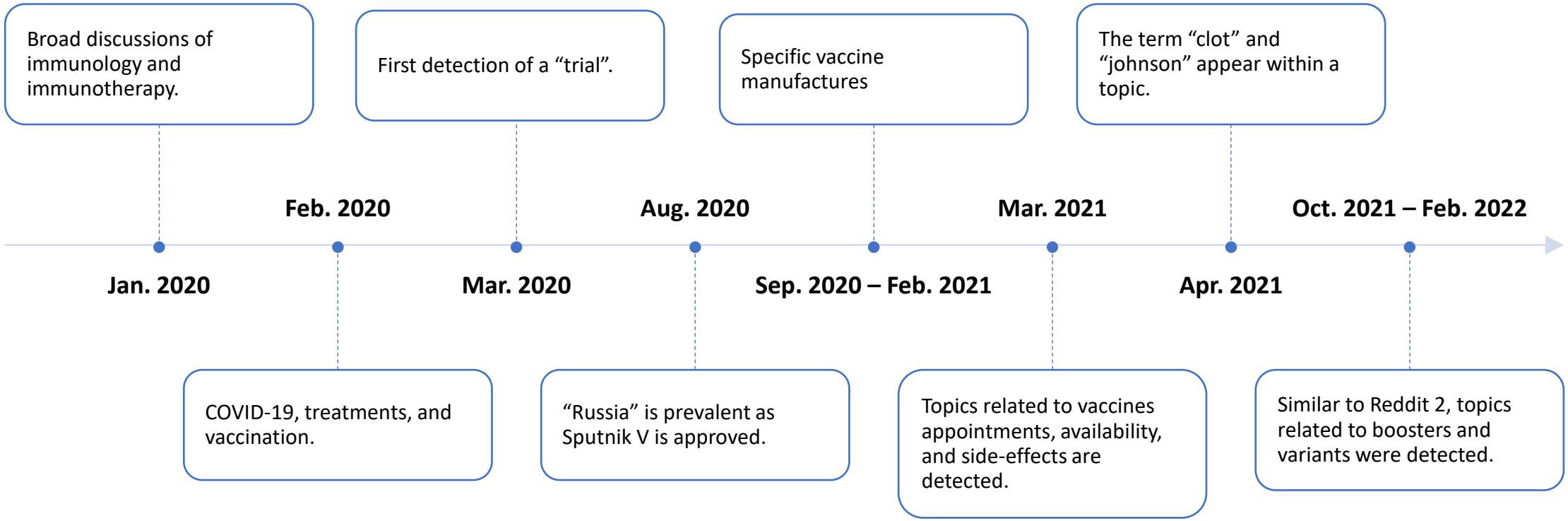
- *Topics 1-4:* vaccines, safety concerns, efficacy, and potential side effects
- *Topic 5*: broader terms, information (i.e., *news, source, question*), vaccine safety concerns
- Topic 5: Mentioning of autism

Example of “Topic 4”. The blue rectangles are representative of overall term frequency and the red rectangles represent frequency within Topic 1. See github.com/Cheltone/NLP_Reddit for an interactive display of LDA topics. Spheres represent relative topic distribution.

Reddit 2.0: Jan 2020-Feb 2022



Jan 1, 2020-Mar 1, 2022



Remarks and limitations

- Topics reflect changes in the pandemic and vaccine development
- Twitter: more broadly focused on major events
- Reddit: more terms related to side-effects
- Vaccine hesitancy and deaths were detected in all
- Minimal detection of conspiracy theories

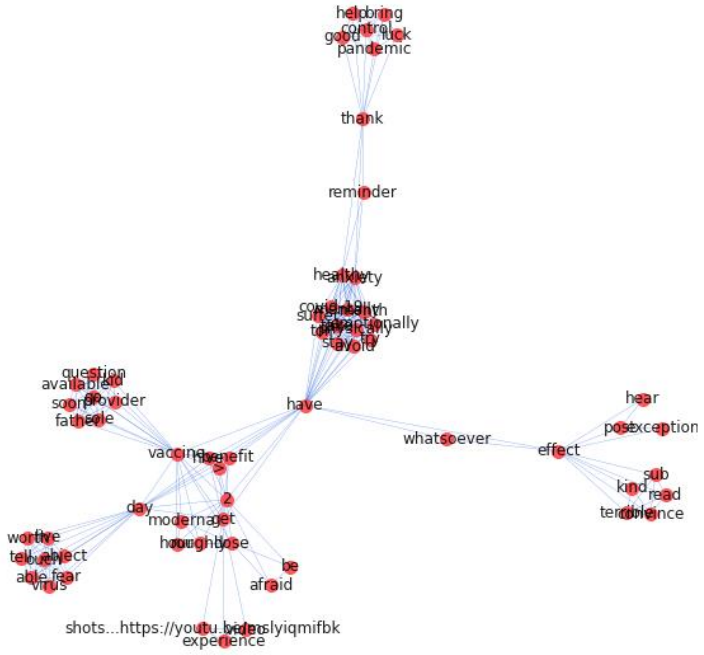
Challenging to evaluate

You get out what you put in.

Methods: Semantic Network Analysis

Network analysis: provide insight into data that may not be observable upon the surface

- *Nodes = words*
- *Edges = lines connecting words*

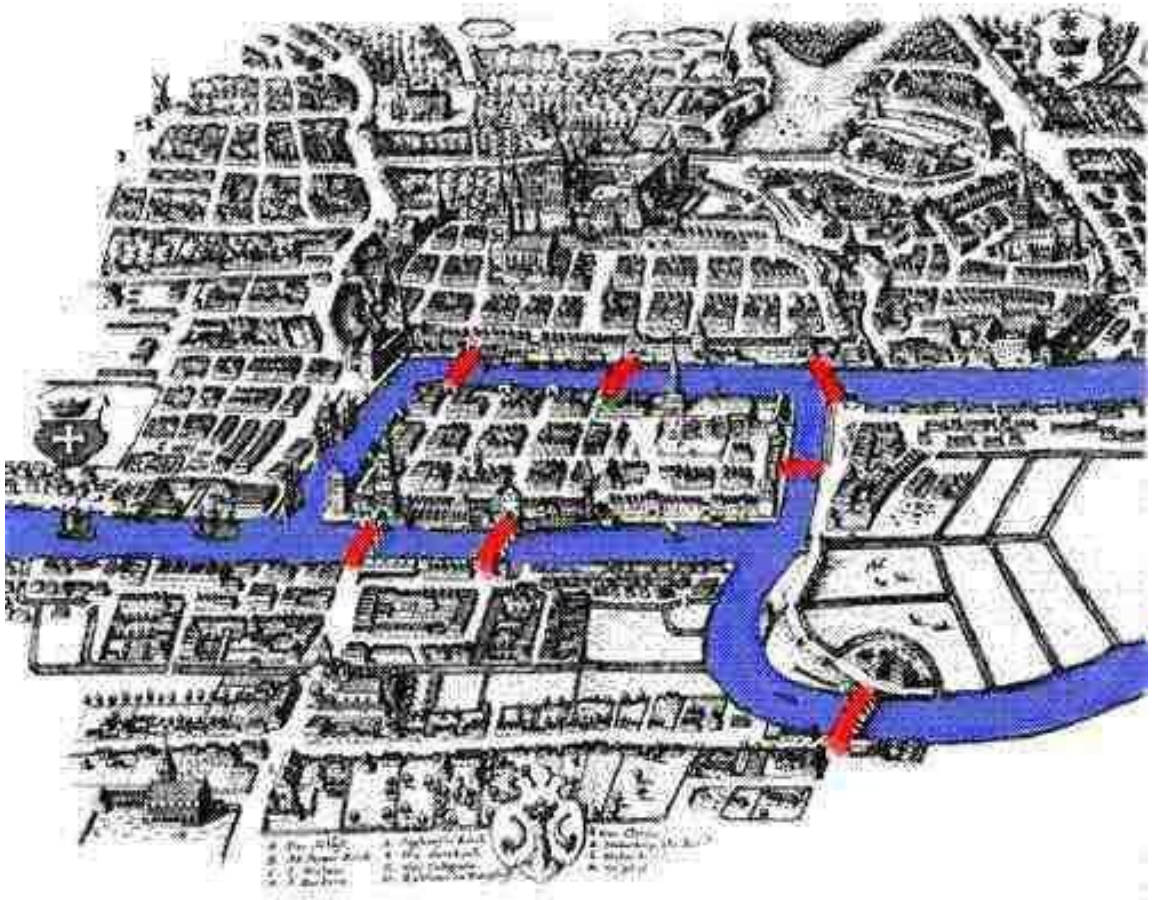


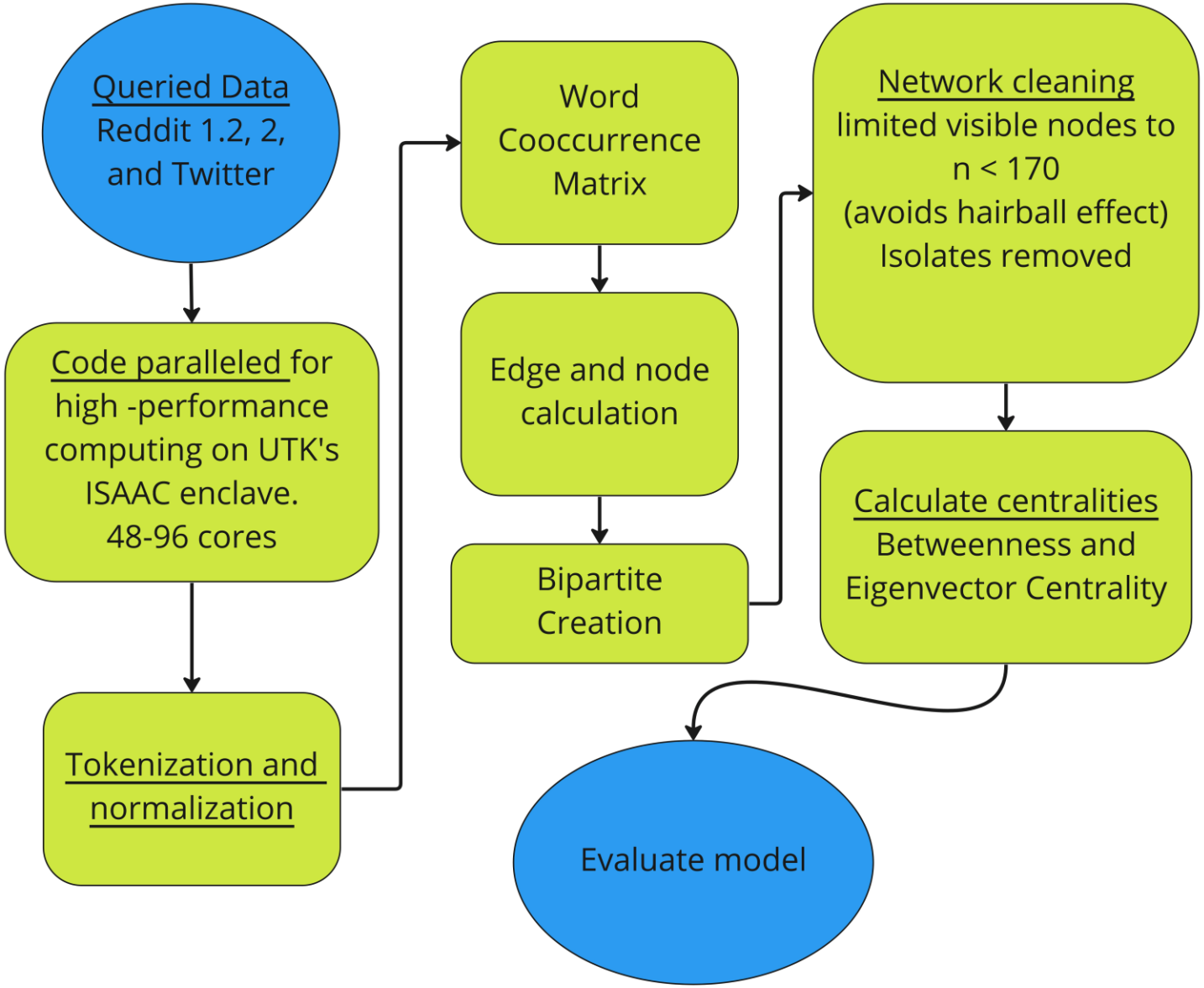
Network Analysis

- **Network analysis** has been used for a wide variety of problems across many scientific domains

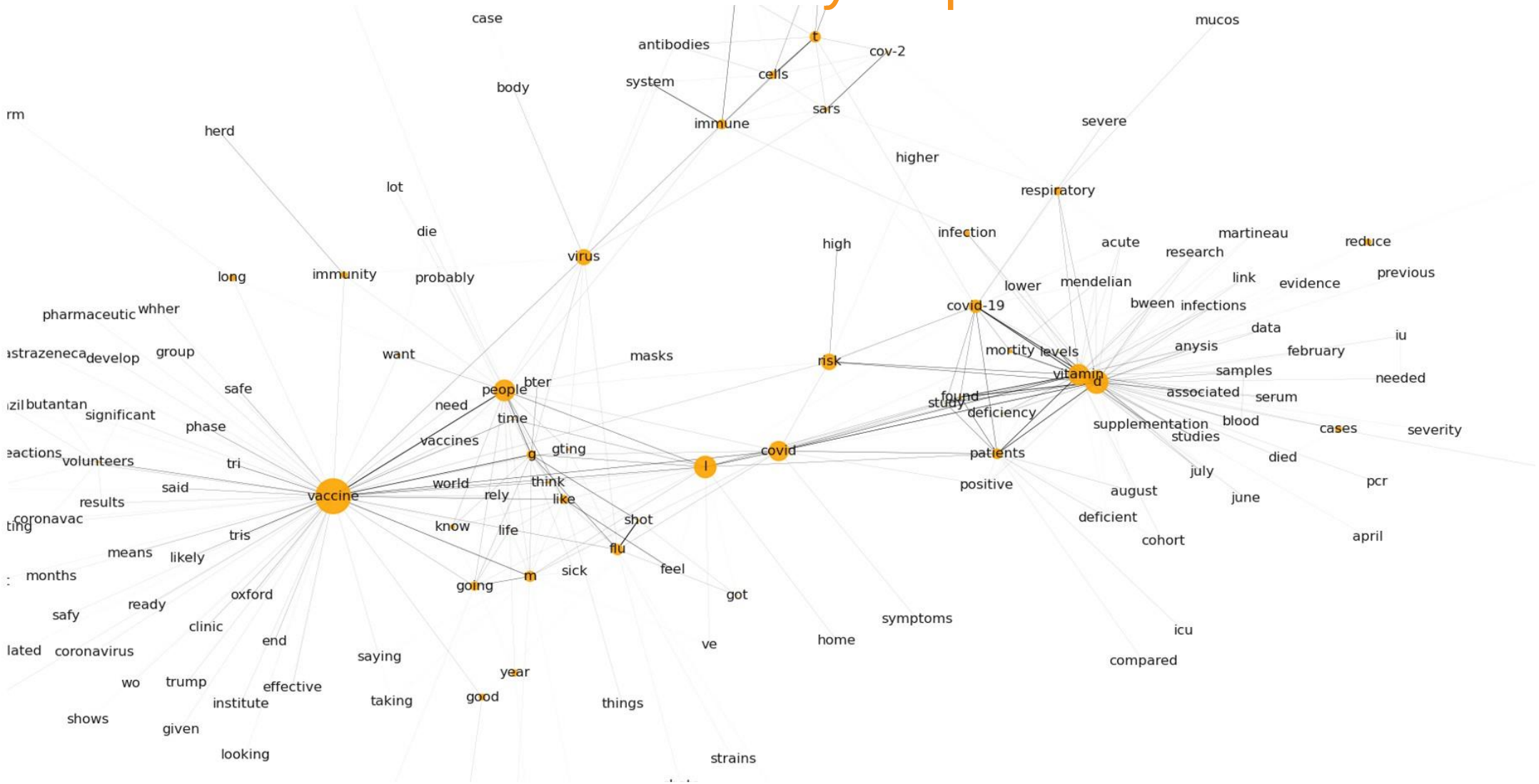
- Health: COVID-19 comorbidity networks of patients in Korean hospitals. (Kyong-Shin et al., 2021)

- **Computational semantic analysis**
- Luo et al (2021): Chinese vs American perceptions of COVID vaccination
- Network analysis of COVID-19 vaccines appears to be relatively unexplored

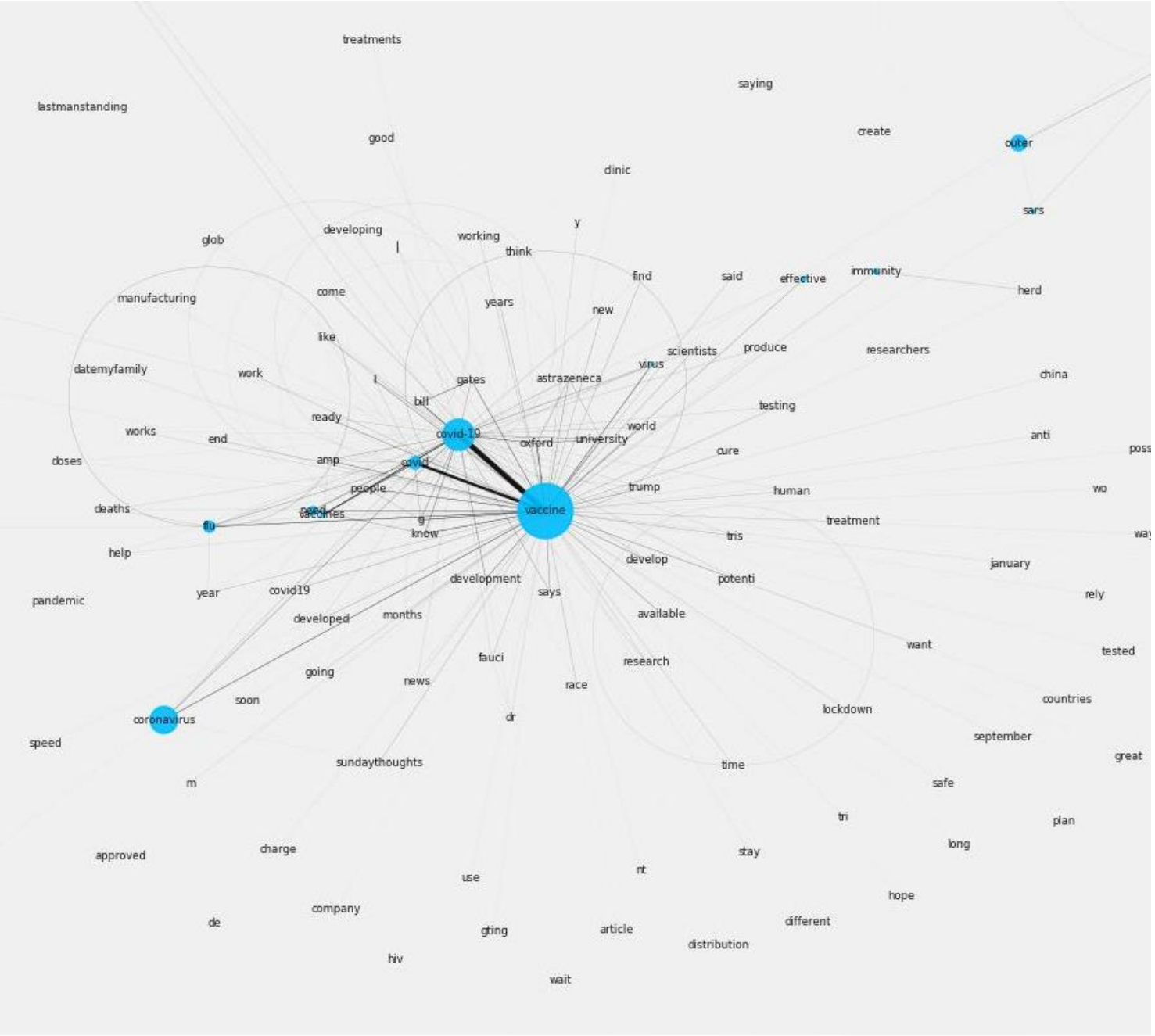



Queried Data
Reddit 1.2, 2, and Twitter
Code paralleled for high -performance computing on UTK's ISAAC enclave.
48-96 cores
Tokenization and normalization
Word Cooccurrence Matrix
Edge and node calculation
Bipartite Creation
Network cleaning
limited visible nodes to n < 170
(avoids hairball effect)
Isolates removed
Calculate centralities
Betweenness and Eigenvector Centrality
Evaluate model

Reddit 1.2 Betweenness Centrality September 2020

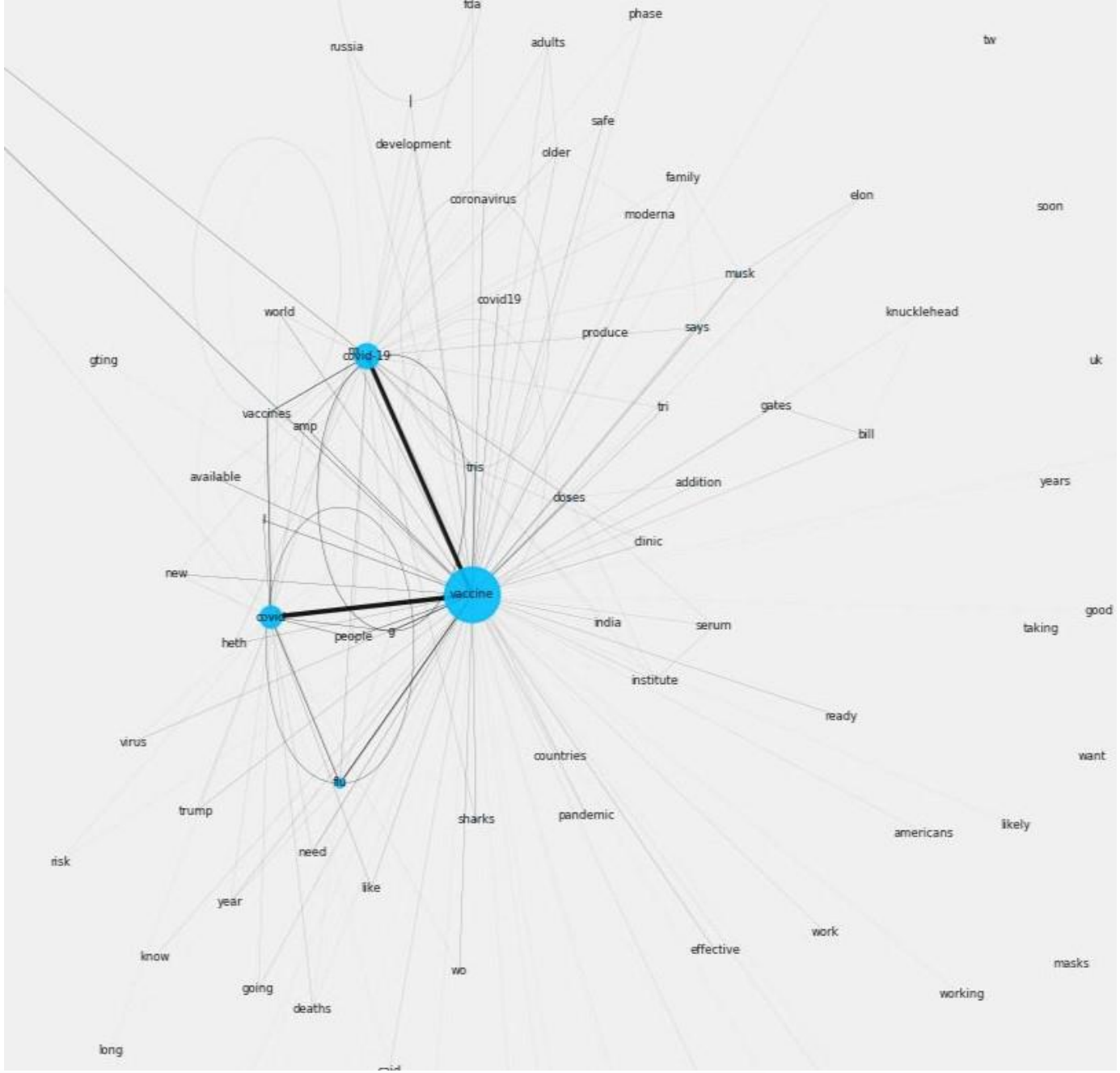


Twitter
April 2020

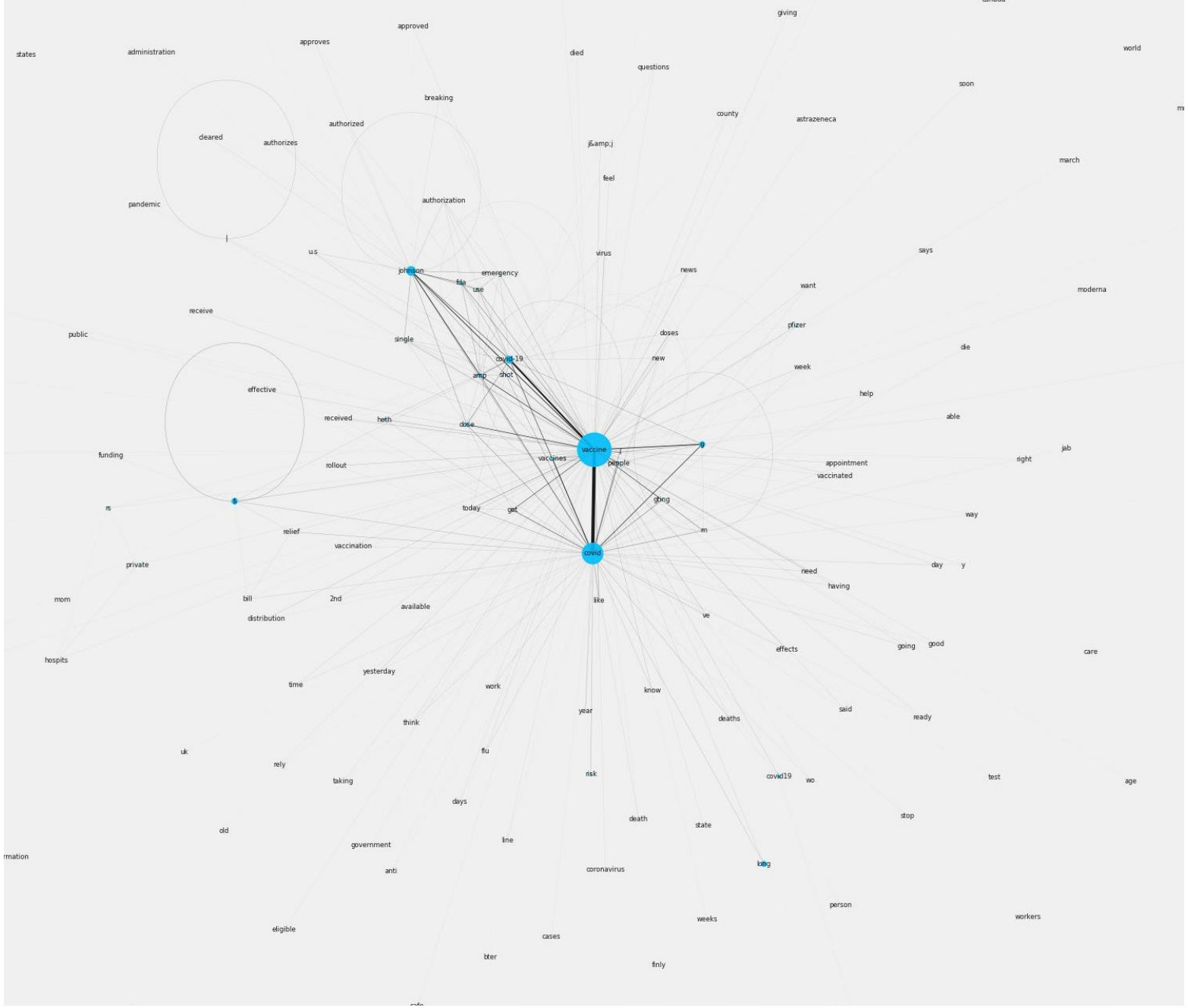


Ask me for my dropbox link!

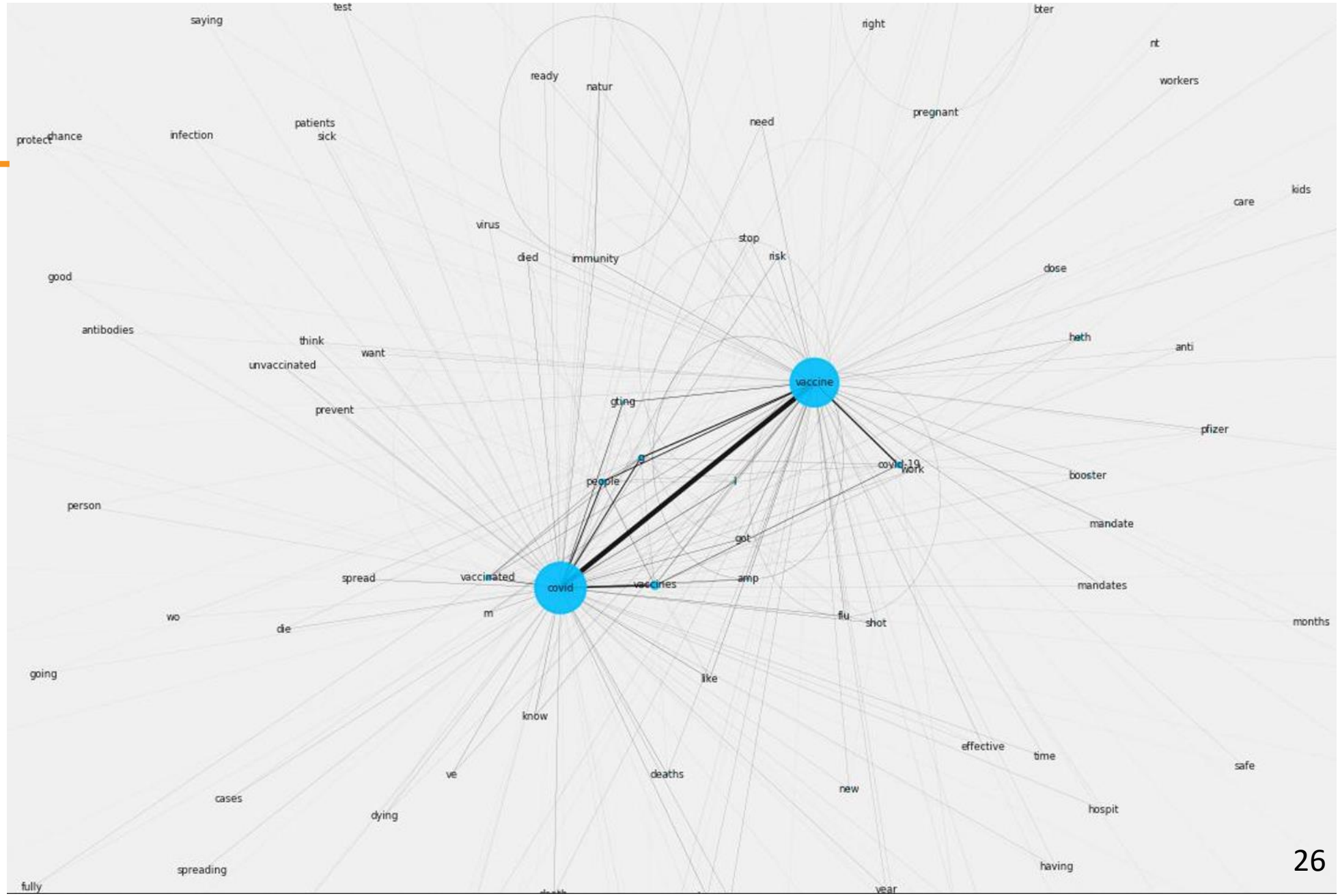
Twitter
Sept 2020



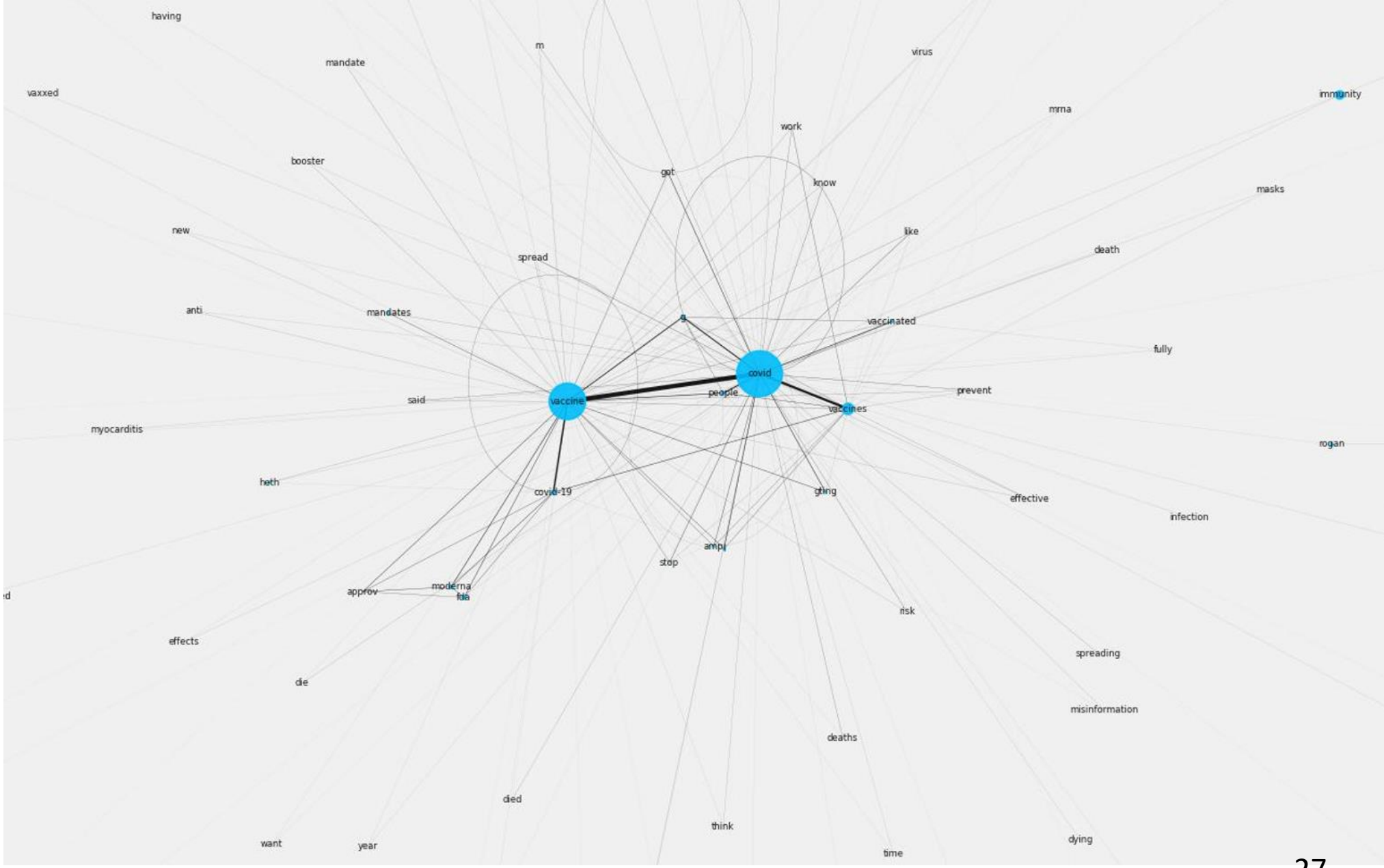
Twitter
Feb 2021



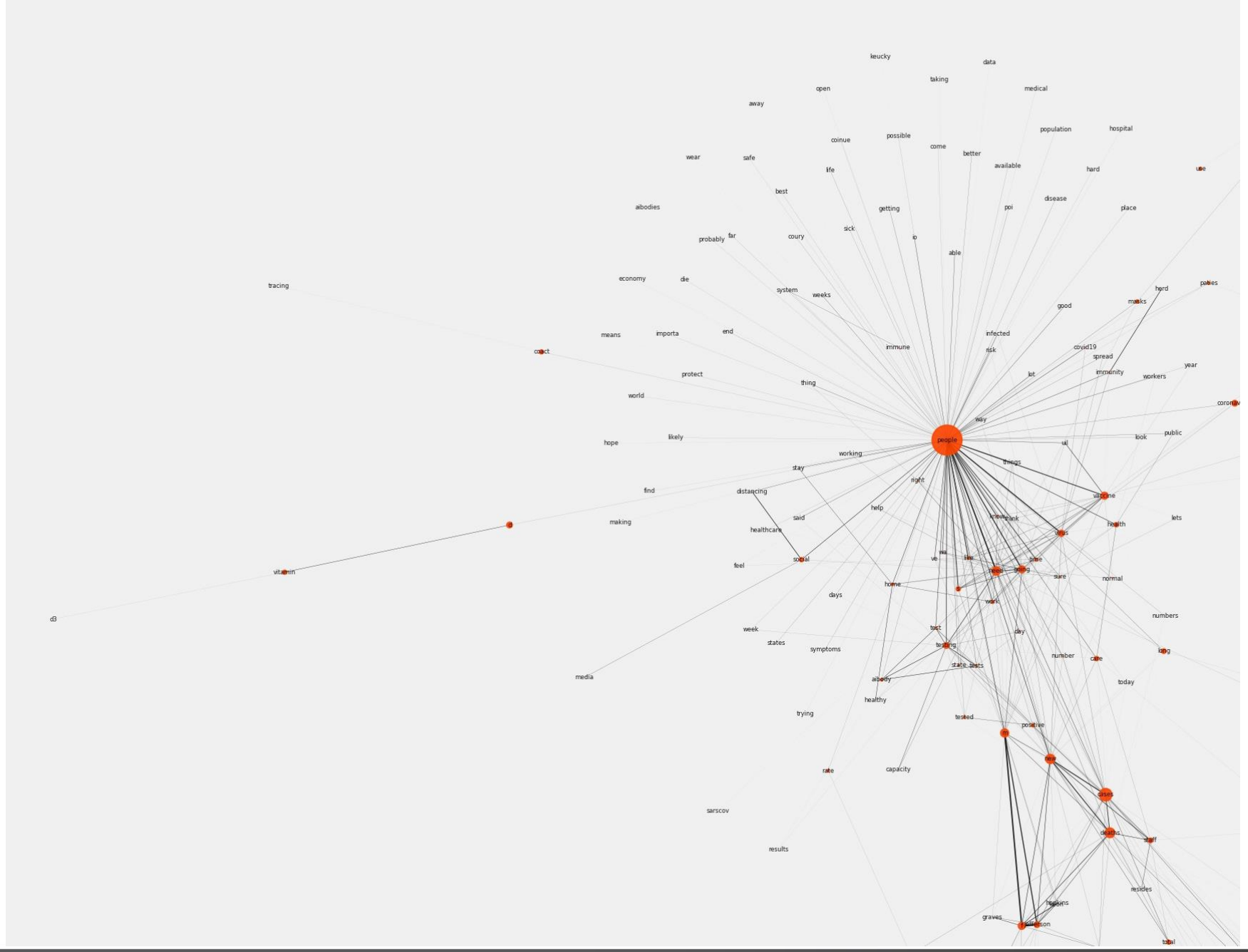
Twitter
Sept 2021



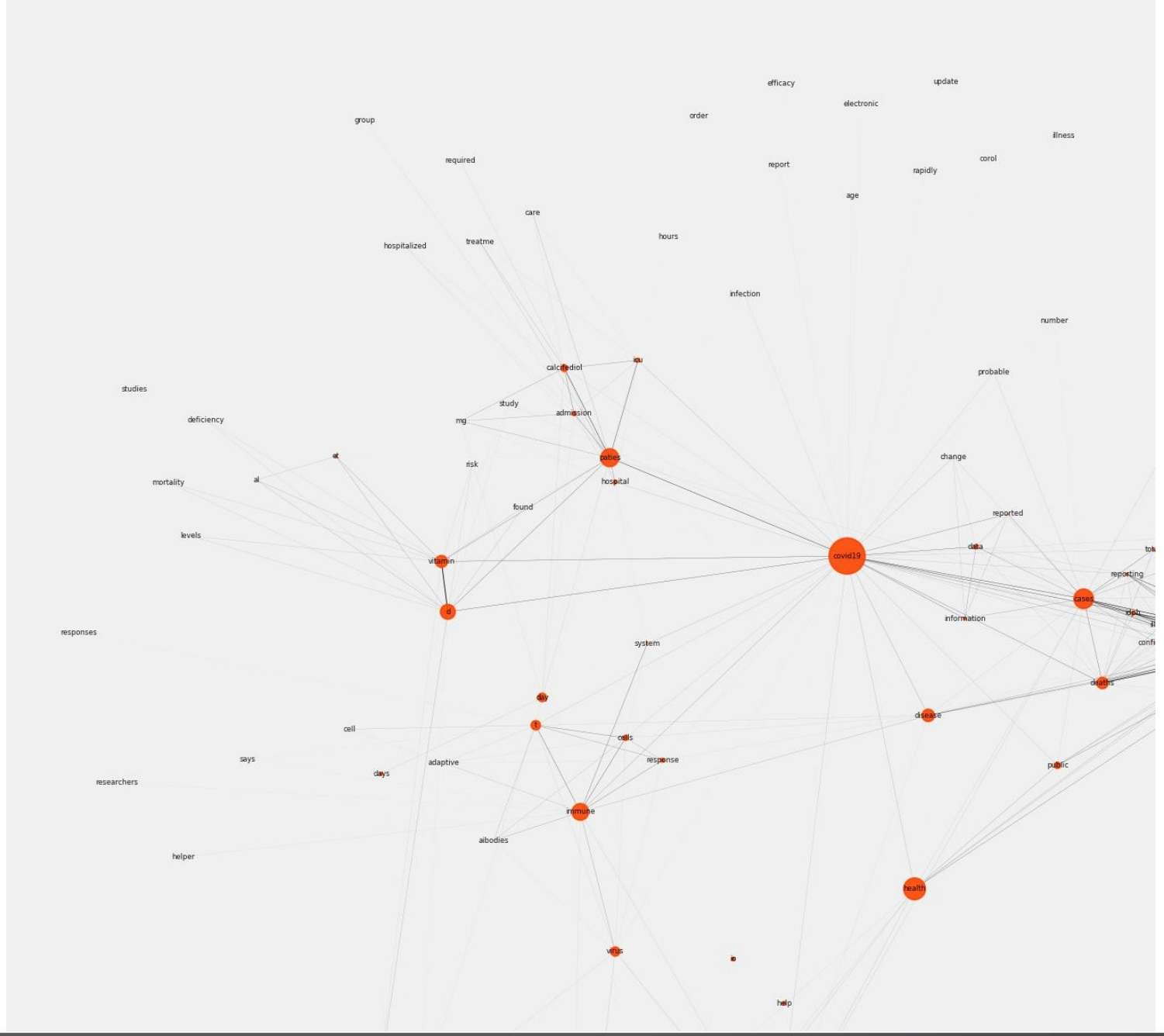
Twitter
Jan 2022



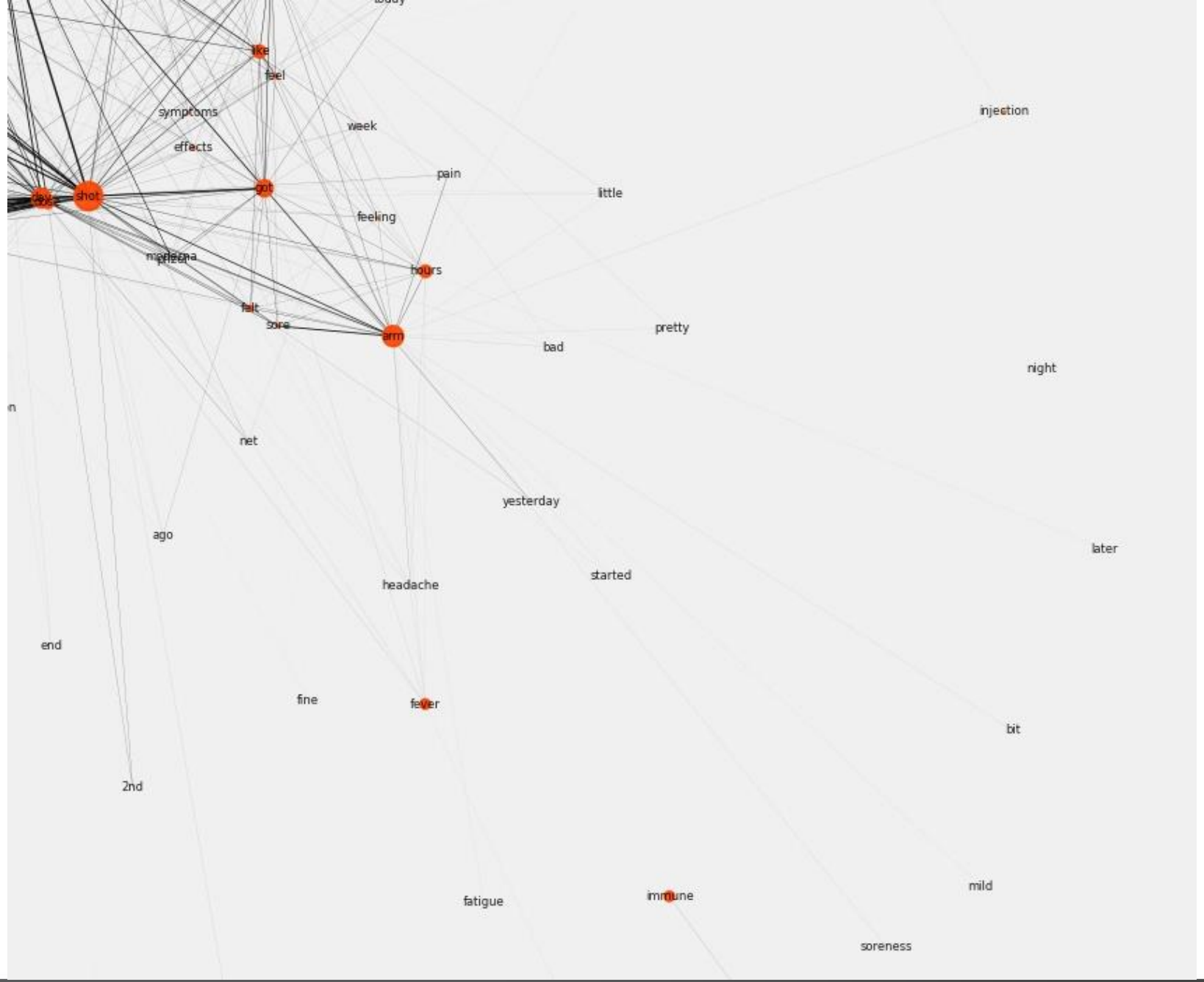
Reddit 2
April 2020



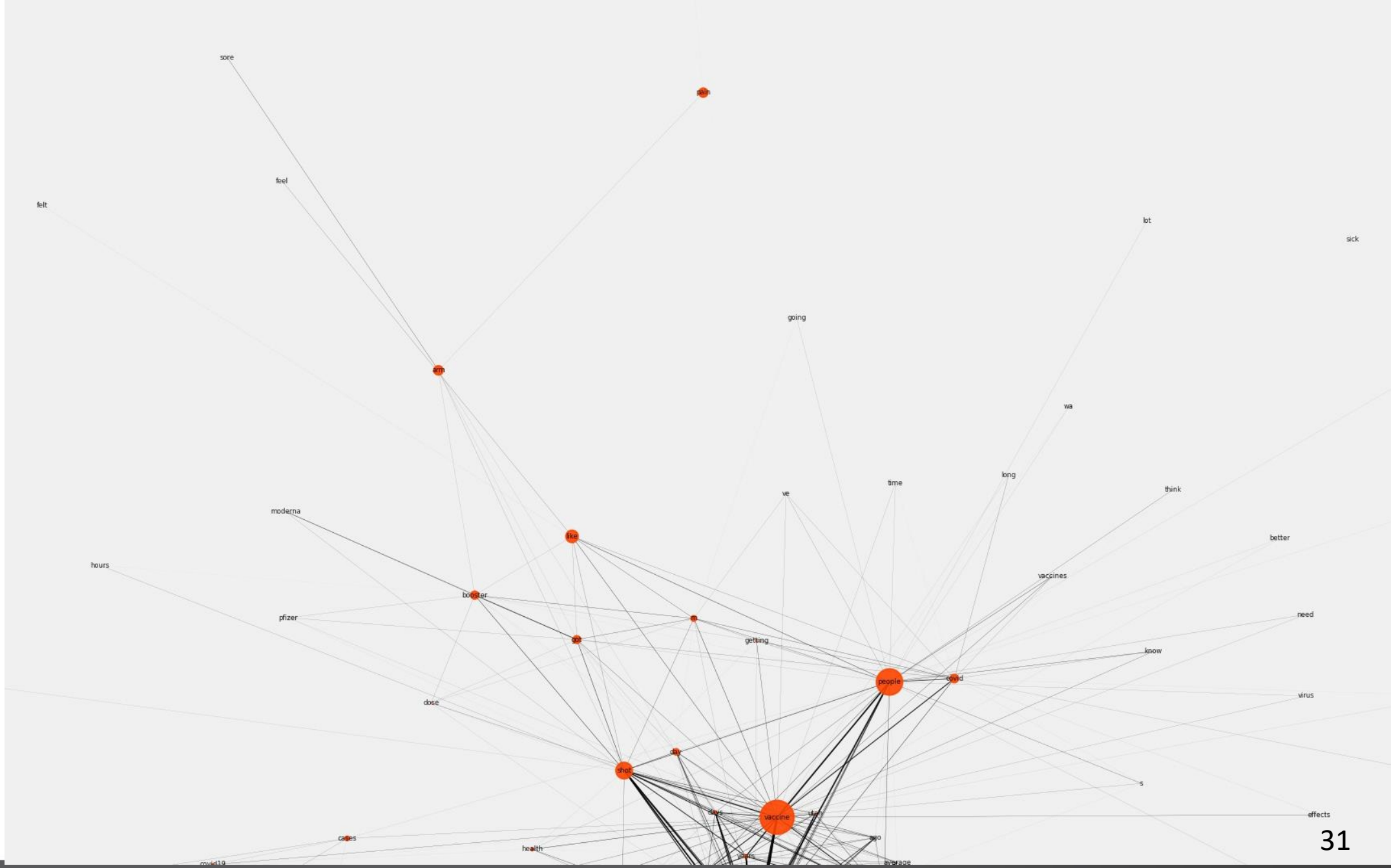
Reddit 2 September 2020



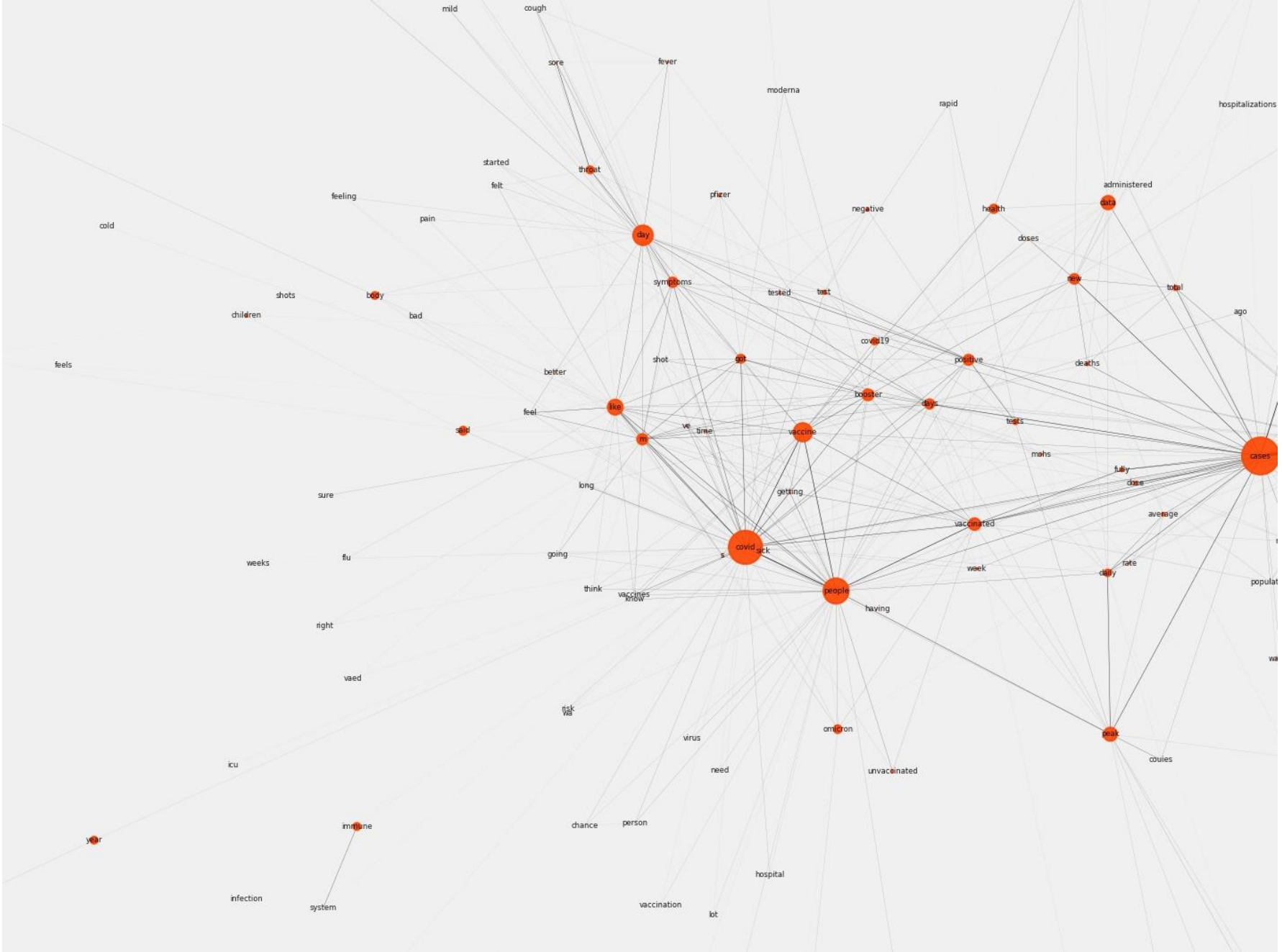
Reddit 2 March 2021



Reddit 2 November 2021



Reddit 2 January 2022



Network analysis

Network analysis

- Networks reflect the events throughout the pandemic
- Terms related to misinformation were observed in both platforms
- Network structure difference between platforms
- More terms related to side-effects on Reddit than Twitter

Should be used in concert with centrality measures

Challenging to evaluate (similar to LDA)

Careful when cleaning data

Methods: Sentiment Analysis

Sentiment analysis is the practice of extrapolating the sentiment of a subject, idea, event, or phenomena by classifying written texts as positive or negative

Two methods

- TextBlob (traditional)

- BERT (Bidirectional Encoders Representations for Transformers)

Methods: Sentiment Analysis

Sentiment analysis is the practice of extrapolating the sentiment of a subject, idea, event, or phenomena by classifying written texts as positive or negative

Two methods

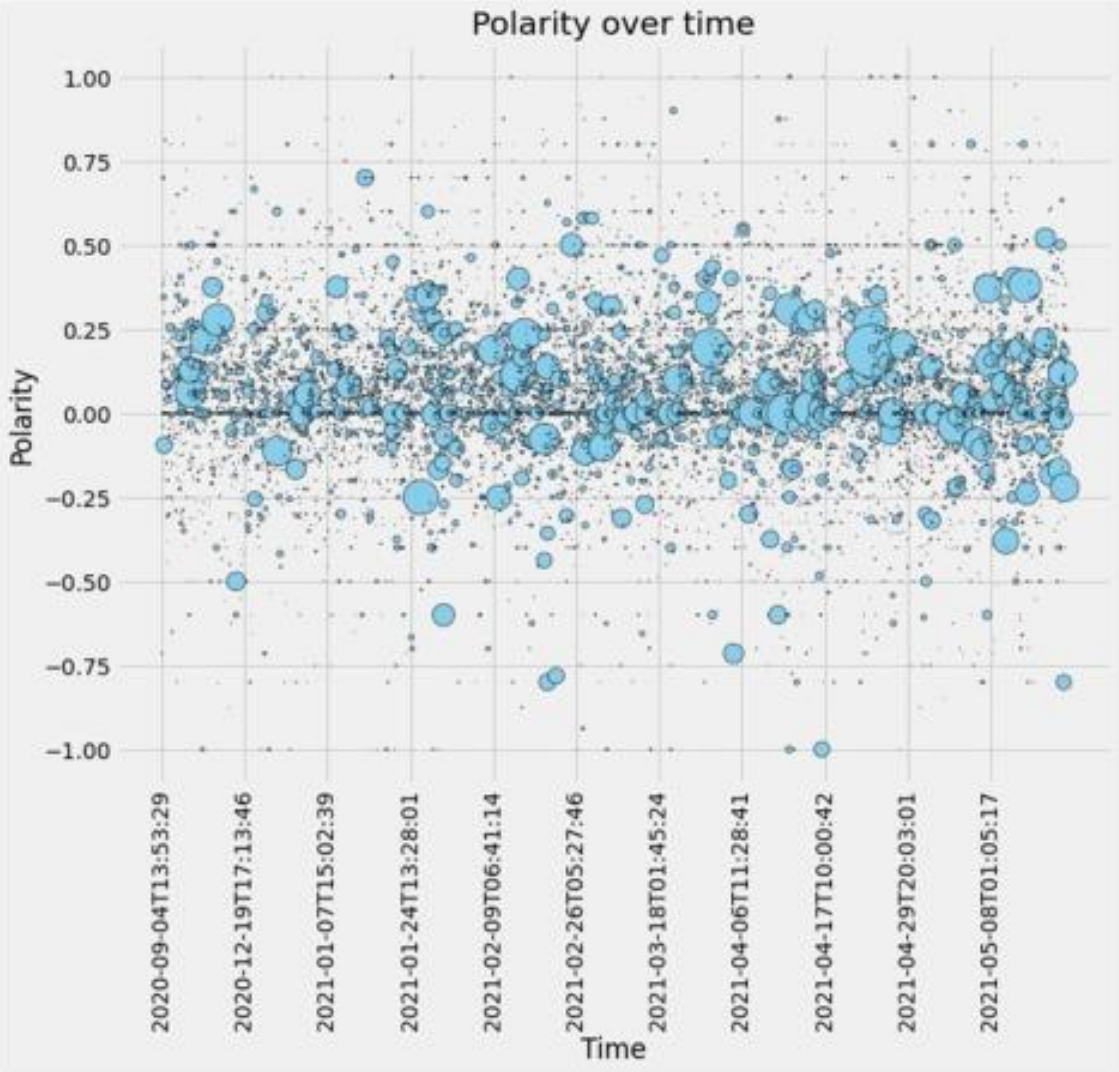
- TextBlob (traditional)

- BERT (Bidirectional Encoders Representations for Transformers)

Background: *Sentiment Analysis*

- Historically: Restaurant reviews, product reviews, and political campaigns
- Influenza outbreak detection (Alessa and Miad., 2019)
- Public opinion on mask wearing (Sanders et al., 2021)
- Correlated increased measles infections and negative social media posts (Raghupathi et al., 2020)
- HPV vaccine misinformation in social media (Du et al., 2021)

TextBlob Method: Sentiment Analysis



- Melton, C. A., Olusanya, O. A., Ammar, N., & Shaban-Nejad, A. (2021). Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence. *Journal of Infection and Public Health*, *14*(10), 1505-1512.

- 56.68% posts measured positive
- 27.69% were negative
- 15.63% neutral.

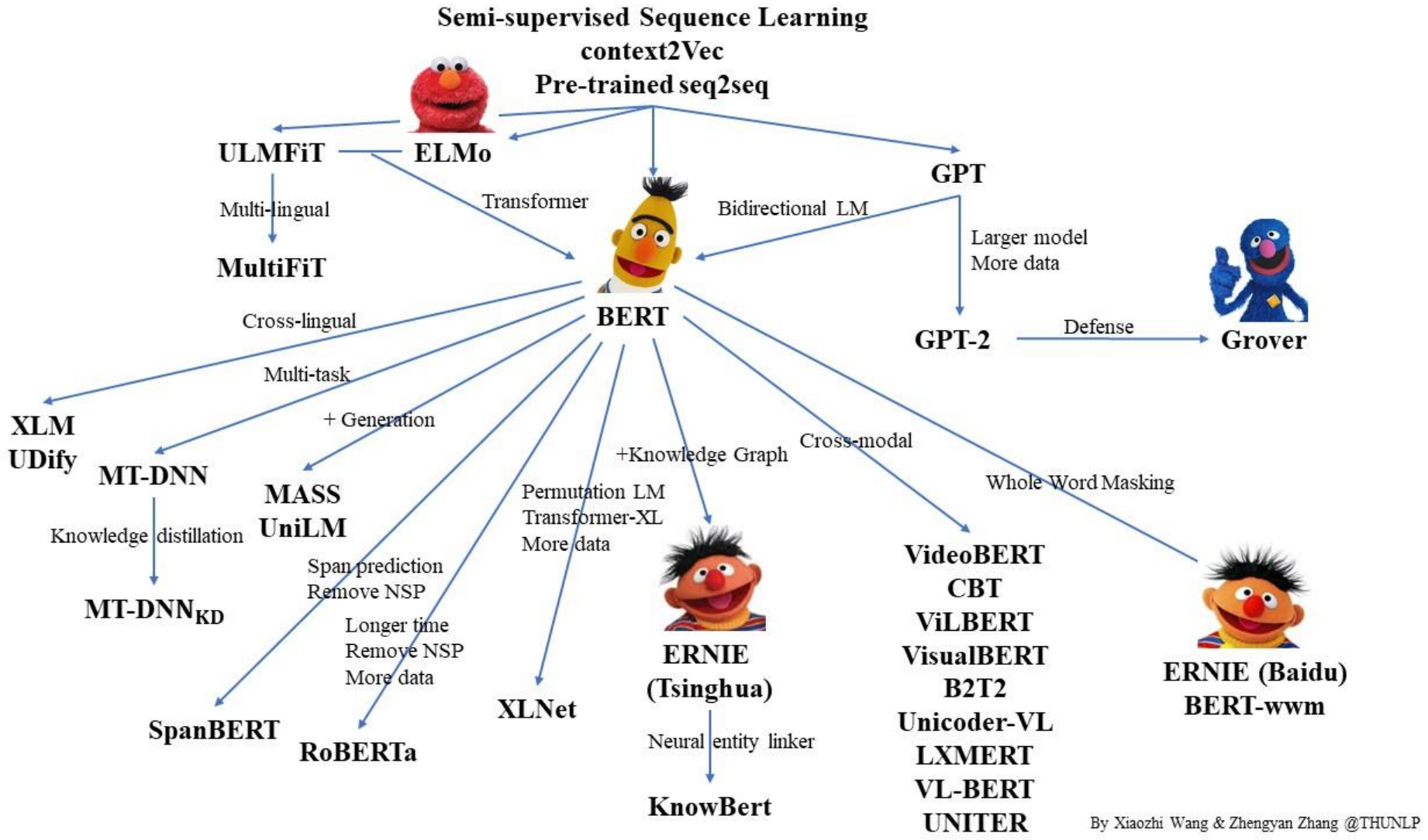
- Ranged from 53.05% (May 2021) to 59.49% (Jan 2021)

- Conclusion: Overall more positive than negative and didn't change substantially

BERT Intro

Bidirectional Encoder Representation from Transformers

- If you used Google ever, you have used BERT
- Transformer-based algorithm for NLP at Google (Devlin et al., 2018)
- Propelled NLP into a new era
- Trained on all of *Wikipedia* and the *Brown Corpus* over four days of 16 cloud-based TPUs (Tensor Processing Units)


Semi-supervised Sequence Learning
context2Vec
Pre-trained seq2seq
ULMFiT
ELMo
GPT
Multi-lingual
MultiFiT
Transformer
Bidirectional LM
Larger model
More data
BERT
GPT-2
Defense
Grover
Cross-lingual
Multi-task
+ Generation
XLM
UDify
MT-DNN
MASS
UniLM
+Knowledge Graph
Cross-modal
Whole Word Masking
Permutation LM
Transformer-XL
More data
Knowledge distillation
MT-DNN$_{KD}$
Span prediction
Remove NSP
Longer time
Remove NSP
More data
SpanBERT
RoBERTa
XLNet
ERNIE
(Tsinghua)
Neural entity linker
KnowBert
VideoBERT
CBT
ViLBERT
VisualBERT
B2T2
Unicoder-VL
LXMERT
VL-BERT
UNITER
ERNIE (Baidu)
BERT-wwm
By Xiaozhi Wang & Zhengyan Zhang @THUNLP

BERT Architecture

Word embeddings → stacked multiple encoders

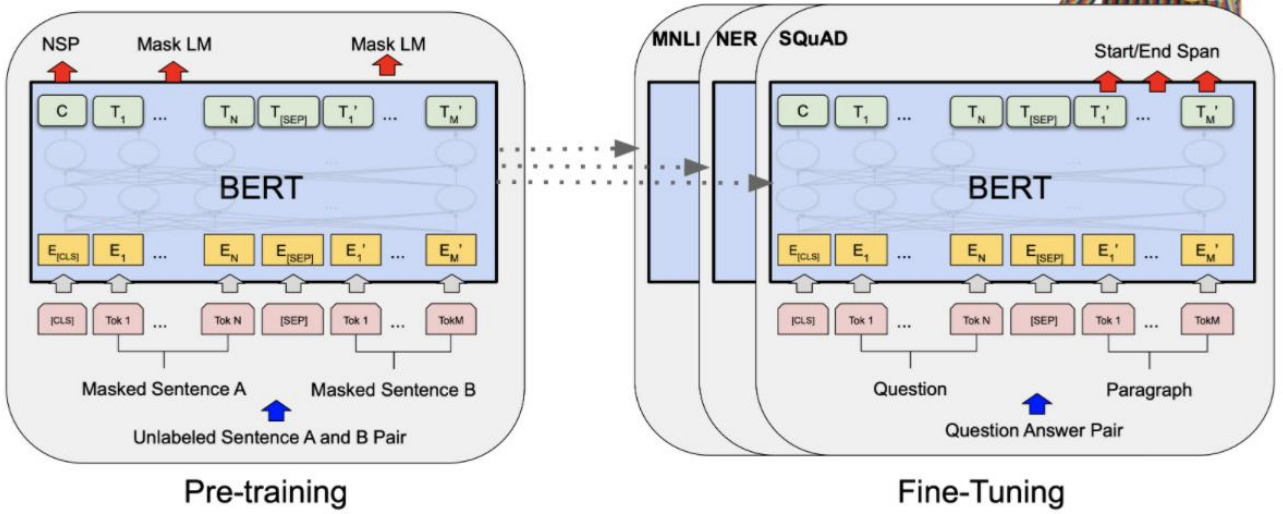
Mask language modeling and next sentence prediction

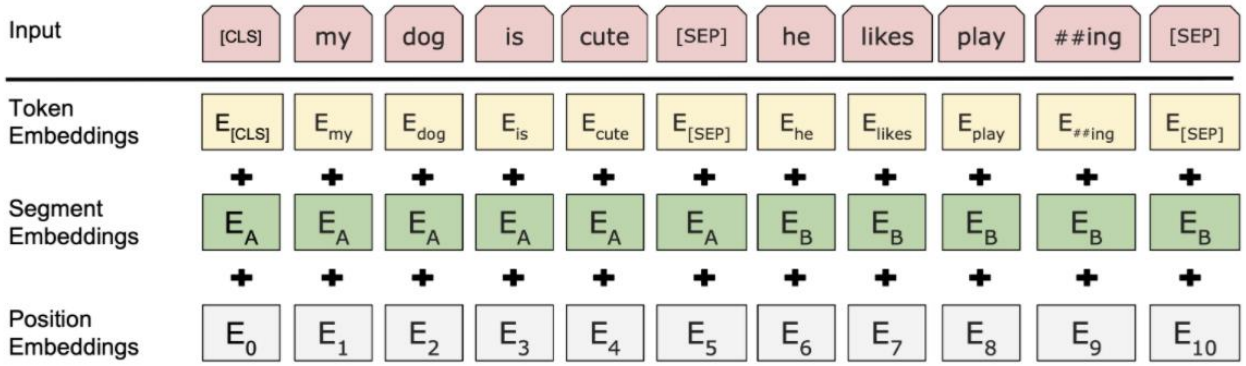
"My [MASK1] is [MASK2] and happy."

MASK1= "dog" or "world"
Mask2 = "fat" or "pizza"

Give some probability for a prediction

Fine-tuned!





RoBERTa and DistilRoBERTa

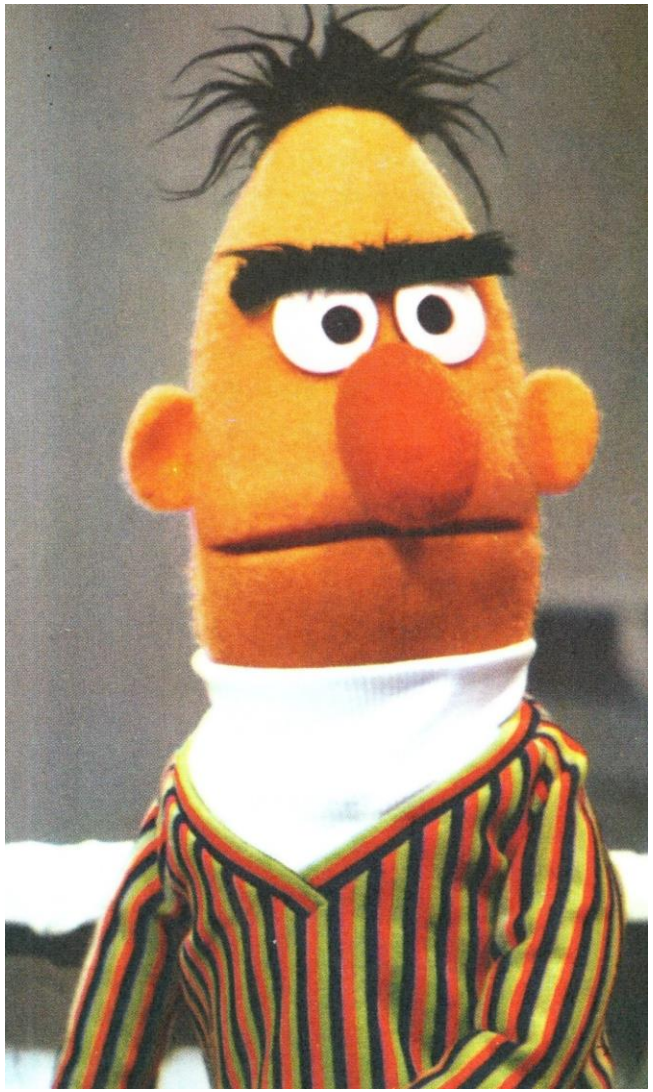
Robustly **O**ptimized **B**idirectional **E**ncoder **R**epresentation from **T**ransformers 2019

RoBERTa More robust than BERT

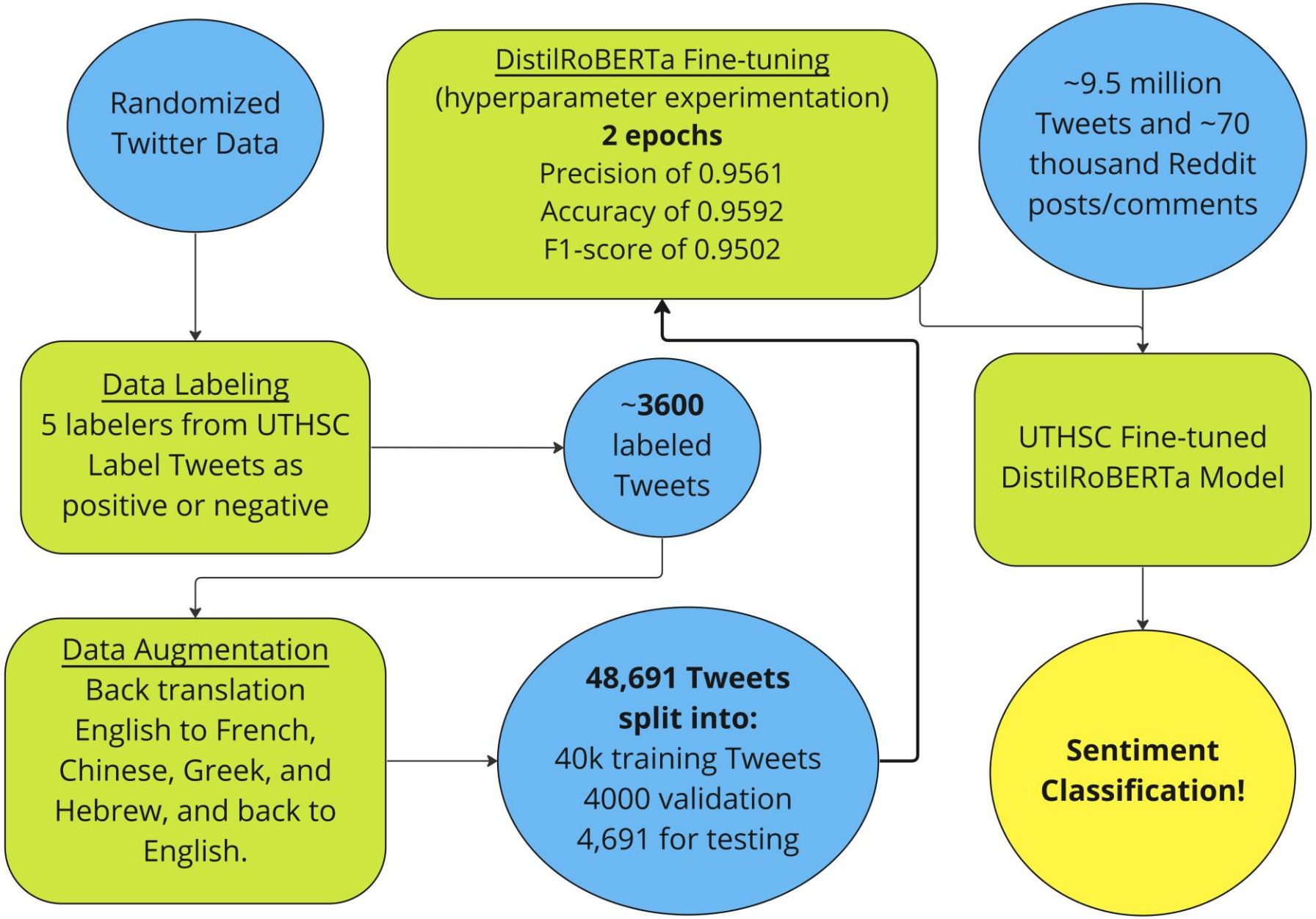
- Trained on 160 GB of text compared to 16 GB of BERT
- Dropped the next sentence prediction but added dynamic token masking
- Increased performance between 2-20 percent

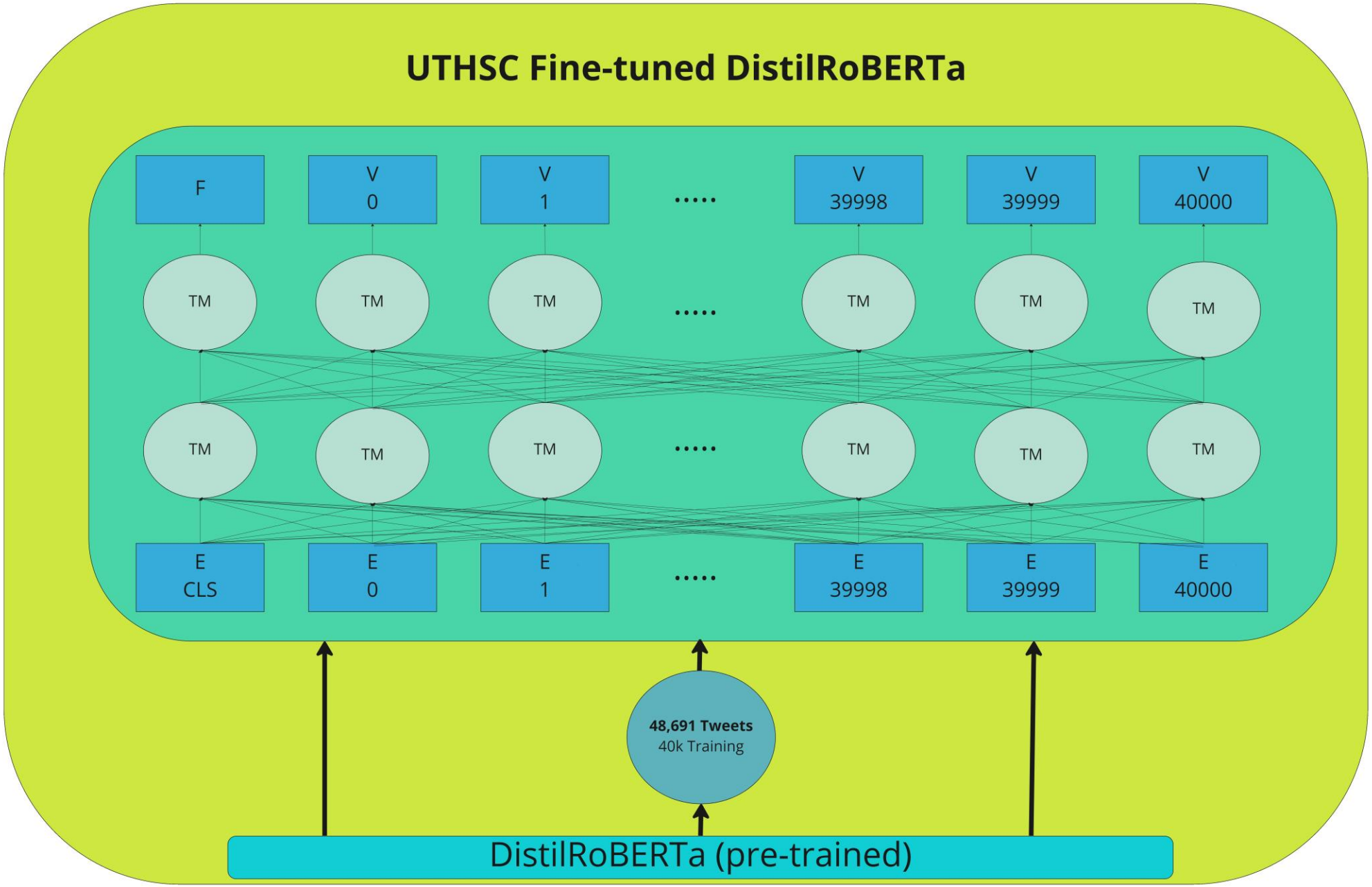
DistilRoBERTa

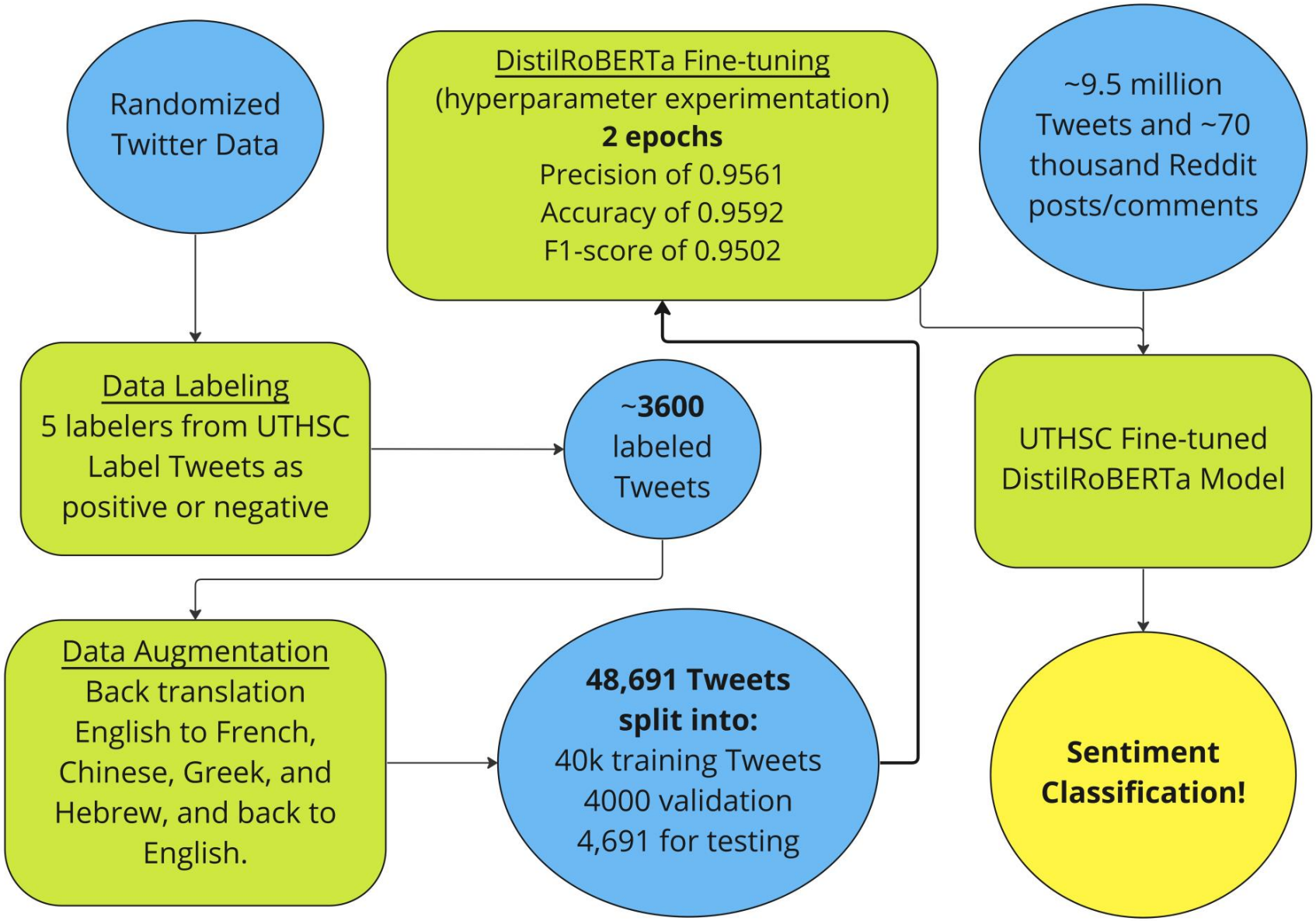
- Optimized version of RoBERTa (2x speed)
- Trained on 40 GB from OpenWebTextCorpus.
- 97% of BERT's performance



Sanh et al., 2019


Randomized Twitter Data
DistilRoBERTa Fine-tuning
(hyperparameter experimentation)
2 epochs
Precision of 0.9561
Accuracy of 0.9592
F1-score of 0.9502
~9.5 million Tweets and ~70 thousand Reddit posts/comments
Data Labeling
5 labelers from UTHSC
Label Tweets as positive or negative
~3600 labeled Tweets
UTHSC Fine-tuned DistilRoBERTa Model
Data Augmentation
Back translation English to French, Chinese, Greek, and Hebrew, and back to English.
48,691 Tweets split into:
40k training Tweets
4000 validation
4,691 for testing
Sentiment Classification!


UTHSC Fine-tuned DistilRoBERTa
F
V 0
V 1
V 39998
V 39999
V 40000
TM
E CLS
E 0
E 1
E 39998
E 39999
E 40000
48,691 Tweets
40k Training
DistilRoBERTa (pre-trained)


Randomized Twitter Data
DistilRoBERTa Fine-tuning
(hyperparameter experimentation)
2 epochs
Precision of 0.9561
Accuracy of 0.9592
F1-score of 0.9502
~9.5 million Tweets and ~70 thousand Reddit posts/comments
Data Labeling
5 labelers from UTHSC
Label Tweets as positive or negative
~3600 labeled Tweets
UTHSC Fine-tuned DistilRoBERTa Model
Data Augmentation
Back translation English to French, Chinese, Greek, and Hebrew, and back to English.
48,691 Tweets split into:
40k training Tweets
4000 validation
4,691 for testing
Sentiment Classification!

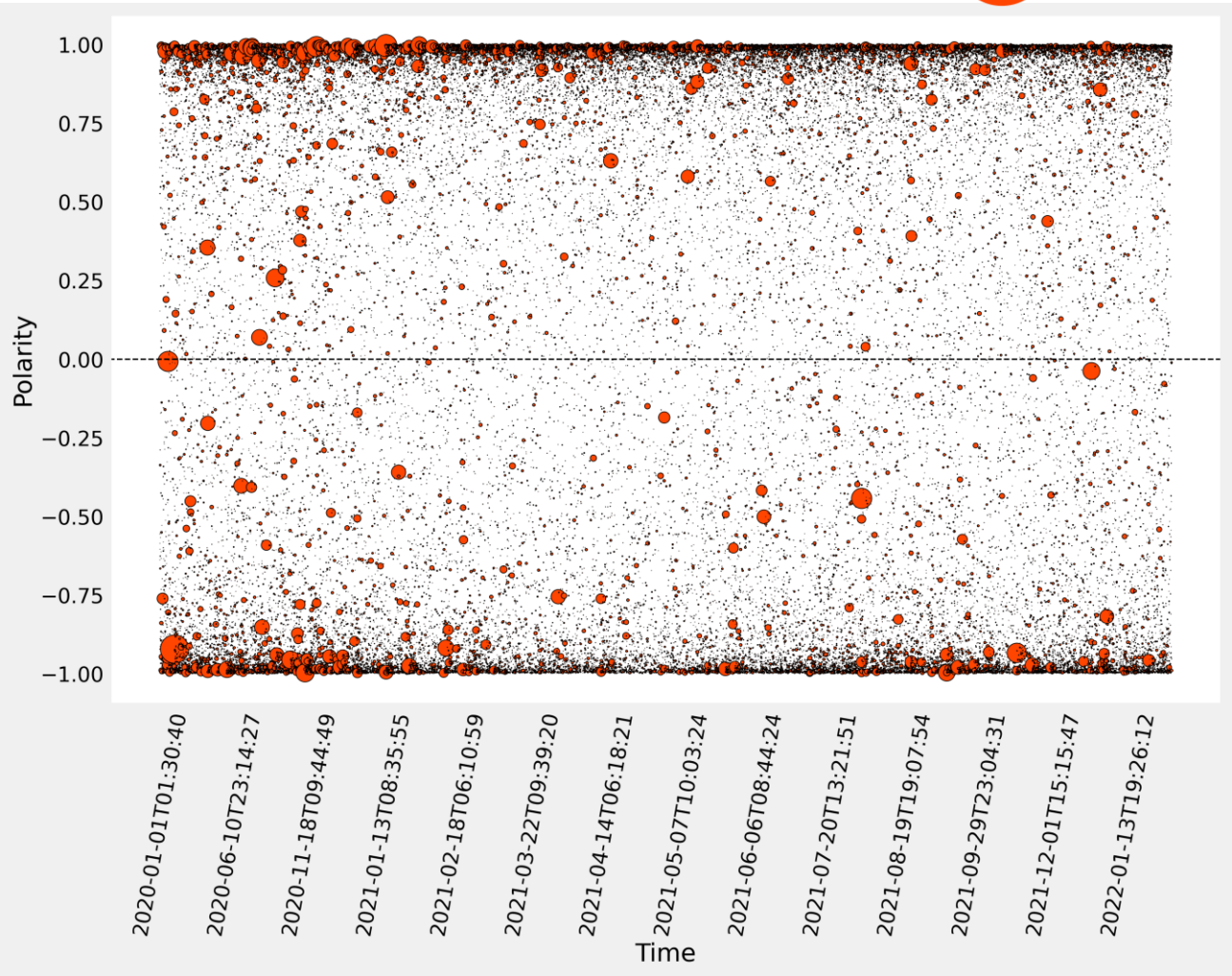
Reddit results

37.7% (n=25,646/ 67,962) were classified as negative

62.3% (n=42,316/ 67,962) were classified as positive

The maximum positive rating occurred in April 2021 (6611/9044, 73.1 %)

The minimum positive rating occurred in February 2020 (n=170/351, 48.4%)



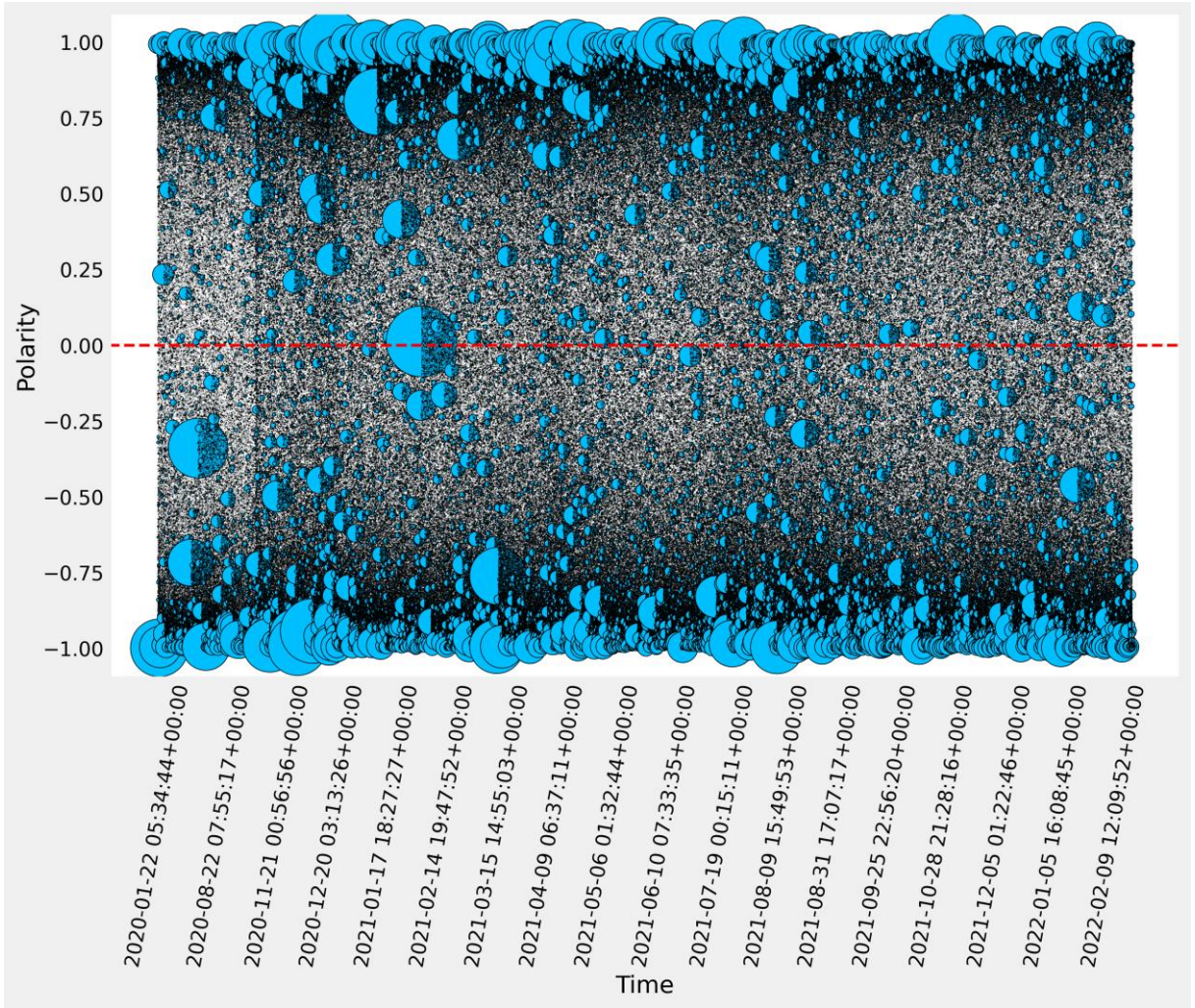
Twitter results

(n=5,215,830/ 9,518,270, 54.8%) negative

(n=4,302,440/ 9,518,270, 45.2%) positive

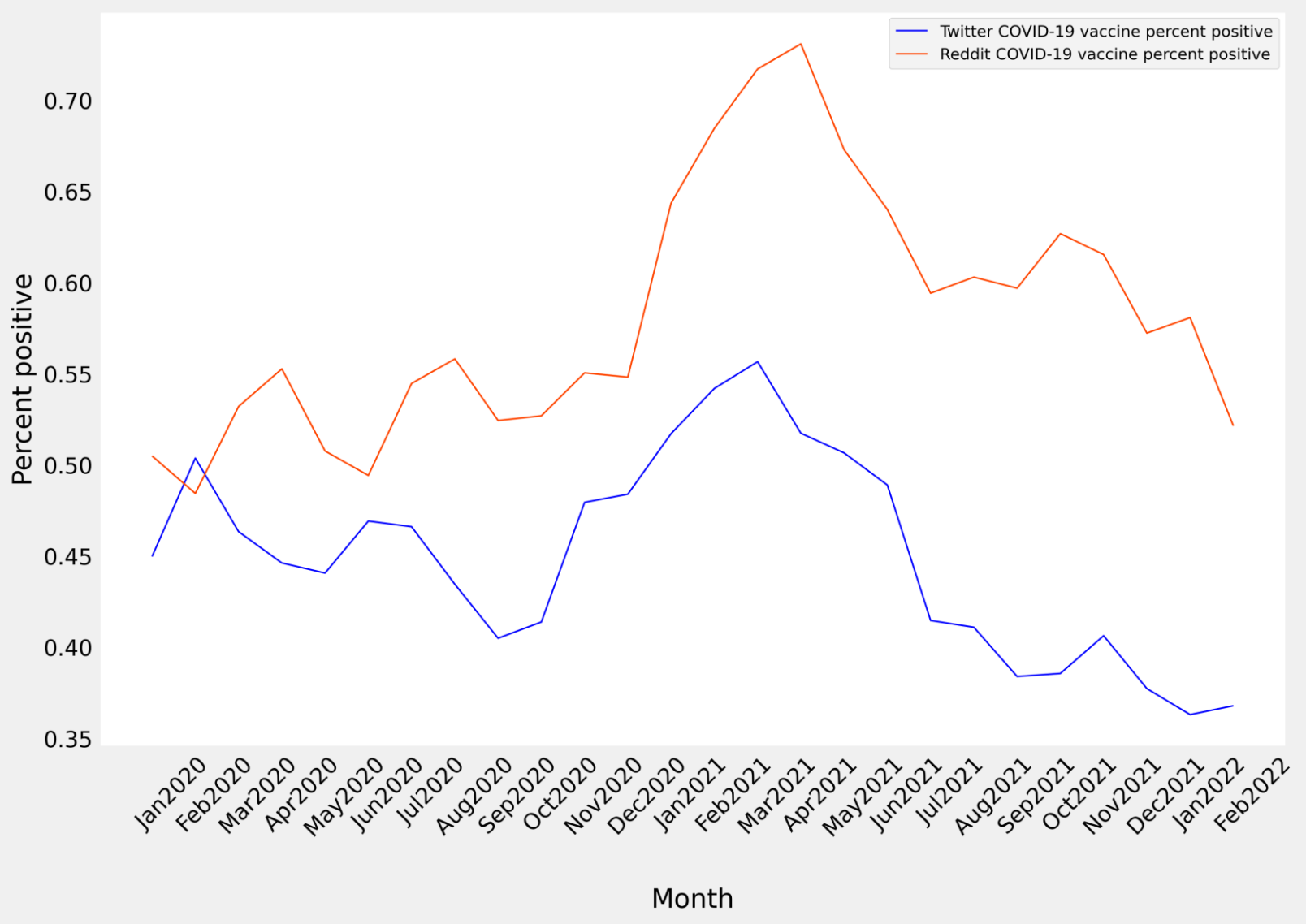
The maximum positive rating occurred in March 2021 (n=375,789/675,274 55.6%)

The minimum positive rating occurred in January 2022 (191,159/526,582, 36.3%)



Tweet polarity from the DistilRoBERTa model fine-tuned to COVID-19 vaccine. Polarity and the corresponding confidence probability are represented on the y-axis, and time is represented on the x-axis. Tweets are represented as light blue circles. Circle size indicates the number of likes per tweet—larger circles indicate more likes and smaller circles indicate fewer likes.

Fine-tuned sentiment analysis results-overall



Comparison of Reddit and Twitter continued

- Reddit tended to follow slightly behind Twitter in some instances
- Reddit posts versus Tweet character limit (i.e., 280 vs 10,000 characters, respectively)
- Shorter posts can be reactionary in nature → driving negative sentiment quickly
- Reddit users take advantage of the longer character limit
- Valuable source when considering the development of public health messaging and education campaigns

Limitations

- Long standing problems in natural language processing.

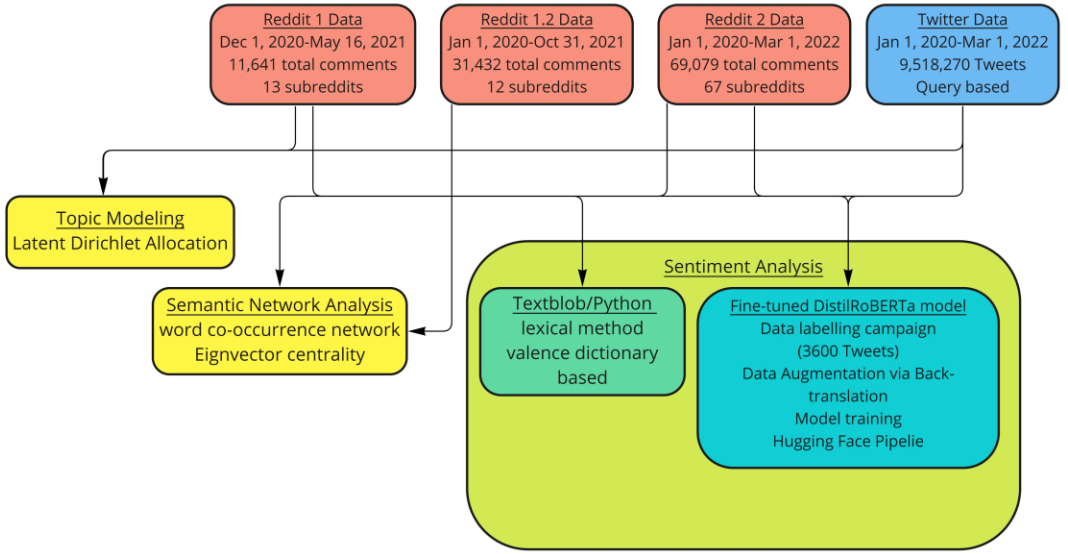
- Augmented data can cause overfitting. Parameters must be monitored closely.

- Unintended bias inherited from labeling data:

““Looking forward to being treated like the plague for refusing the – gene-therapy – vaccine. As a proud introvert, I can't wait for people to avoid me!”, “ (Positive/0.59)

Overall recap

- Explore public discourse related to COVID-19 vaccination expressed on social media
- Multi-modal approach including sentiment analysis, topic modeling, and semantic network analysis
- Leveraged in the fight against COVID-19 and other disease outbreaks
- Health policy, EDUCATION, decision-making, program implementation, and precision health promotion



Implications

- Data science, and especially NLP can be employed for near real-time public health surveillance and historical pandemic research
- Misinformation detection across social media platforms
- Specific use cases for various social media platforms
- Proof of concept: guide public health officials at city, state, and federal levels to develop interventions and education campaigns

Contributions

- Four data sets collected throughout the COVID-19 pandemic totaling over 13.5 million text entries
- Contributed sentiment labeled data and augmented data sets
- Publicly available code repositories featuring state of the art NLP tools
- POC for Multimodal analysis
- Framework to be used for near real-time monitoring of public health discourse
- 11 publications (lead author on 6) and over 60 citations

Publications

a. **Melton, C. A.**, White, B. M., Davis, R. L., Bednarczyk, R. A., & Shaban-Nejad, A. (2022). Fine-Tuned Sentiment Analysis of COVID-19 Vaccine Related Social Media Data: A Comparative Study. *Journal of Medical Internet Research (preprint)*.

b. **Melton C.A**, Olusanya O.A., Ammar N, and Shaban-Nejad A. Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence. J Infect Public Health. 2021 Oct;14(10):1505-1512. doi: 10.1016/j.jiph.2021.08.010. Epub 2021 Aug 14. PMID: 34426095; PMCID: PMC8364208.

c. **Melton C.A.**, Olusanya O.A., and Shaban-Nejad A. Network Analysis of COVID-19 Vaccine Misinformation on Social Media. Stud Health Technol Inform. 2021 Nov 18;287:165-166. doi: 10.3233/SHTI210839. PMID: 34795104.

d. **Chad A Melton**, Olufunto A. Olusanya, Nariman Ammar, and Arash Shaban-Nejad. "Sentiment Analysis of the Covid-19 Vaccines on Social Media", Submitted to the 18th World Congress on Medical and Health Informatics. To appear in Proc. of the 2021 World Congress on Medical and Health Informatics (MedInfo 2021), Virtual, Oct 2-4, 221. Studies in health technology and informatics, IOS Press.

e. **Chad A. Melton**, Jintae Bae, Olufunto A. Olusanya, Jon Hael Brena, Eun Kyong Shin, and Arash Shaban-Nejad. "Semantic Network and Content Analysis of COVID-19 Vaccine Related Social Media Text". To appear in Proceedings of the AAAI International Workshop on Health Intelligence (W3PHIAI 2022). Vancouver, Canada. Feb 22- March 1, 2022.

f. **Melton, C. A.**, Hughes, D. C., Page, D. L., & Phillips, M. S. (2020). Temporal multispectral and 3D analysis of Cerro de Pasco, Peru. *Science of the Total Environment*, *706*, 135640.

g. White, B.M. , **Melton, C. A.**, Davis, R. L., Bednarczyk, R. A., & Shaban-Nejad, A. (2022). "Exploring Celebrity Influence on Public Attitude Toward the COVID-19 Pandemic: Social Media Shared Sentiment Analysis". BMJ Health & Care Informatics. (In review)

h. Nariman Ammar, Olufunto Olusanya, **Chad Melton**, Lokesh Chinthala, Xiaolei Huang, and Arash Shaban-Nejad. "From Personal Health Coaches to Digital Personal Health Librarians: HPV Vaccine Education and Promotion". in Proc. of International Workshop on AI in Health: Transferring and Integrating Knowledge for Better Health (AIHEALTH-WWW-2021) workshop at Web Conference 2021, ACM Press, April 2021.

i. Nariman Ammar, Briana White, **Chad Melton**, Lokesh Chinthala, Xiaolei Huang, and Arash Shaban-Nejad. (2022)."A Digital Personal Health Coaching Platform for HPV Vaccine Promotion and Education". PEC Innovation (In review).

j. Olusanya O.A.,, White, B. M., Amuchi, B., **Melton, C.A.**, Shaban-Nejad, A. (2022). Perceptions and Misinformation on COVID-19 Mask Mandate During Tennessee School Board Meetings: A Qualitative Analysis Perceptions and Misinformation on COVID-19 Mask Mandate During Tennessee School Board Meetings: A Qualitative Analysis. Patient Education and Counseling. (In review).

k. Olusanya, O. A., White, B., **Melton, C. A.**, & Shaban-Nejad, A. (2022). Examining the Implementation of Digital Health to Strengthen the COVID-19 Pandemic Response and Recovery and Scale up Equitable Vaccine Access in African Countries. *JMIR Formative Research*, *6*(5), e34363.

Conference Presentations

- MediInfo 2021
- European Federation for Medical Informatics STC 2021
- W3 PHIAI 2022

Future Work

- Open-source public health software
- Incorporate smaller time frames
- Fine-tuned emotion classifier with DistilRoBERTa
- Misinformation detection from a labeled data set
- Incorporate geographic and demographic information into analysis
- Deploy for use with other disease outbreaks

Acknowledgements

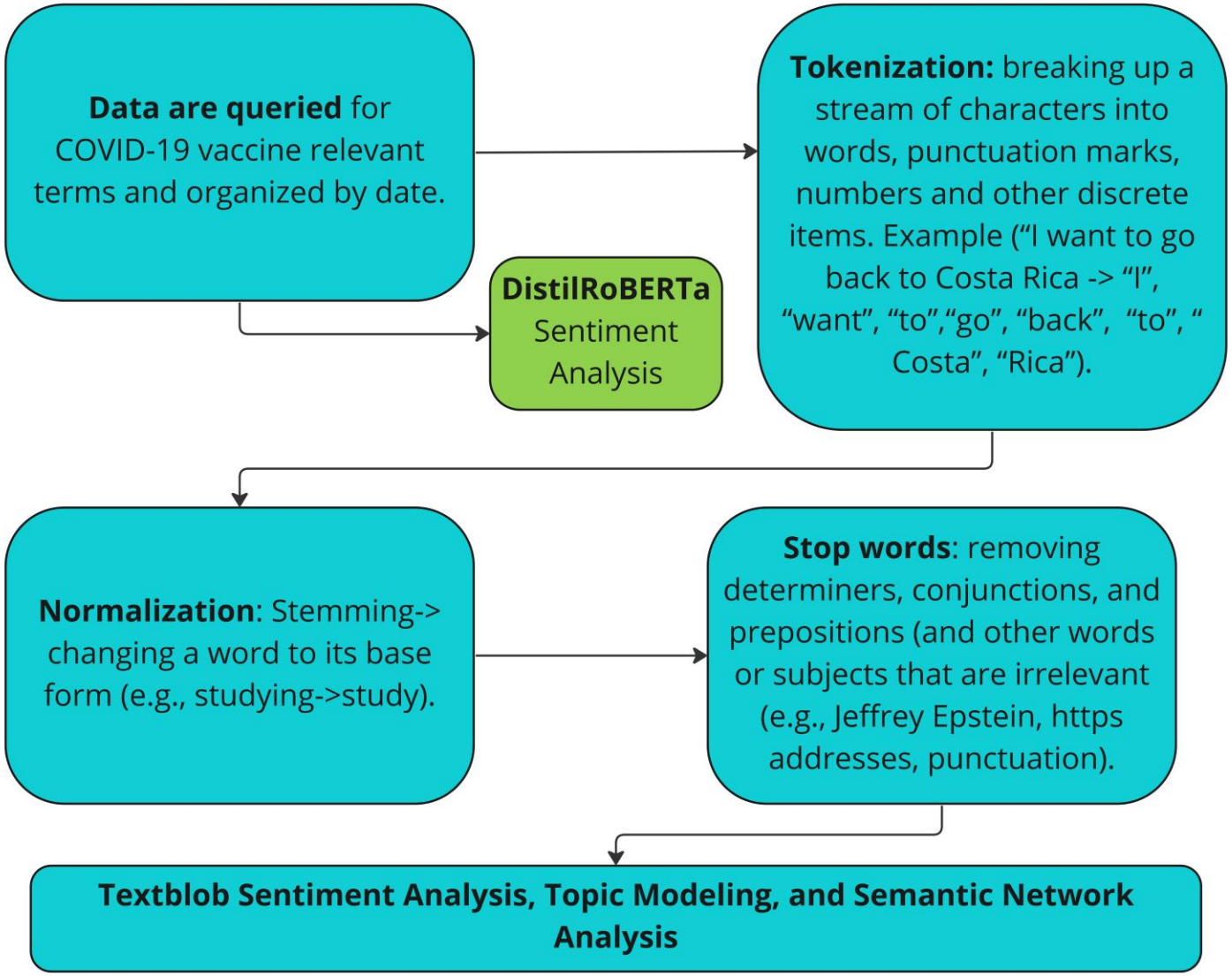
First and foremost, I would like to thank Arash Shaban-Nejad for opportunity to work under his guidance and support. I could not have hoped for a better advisor and mentor. I would like to thank my committee – Drs. Kathleen Brown, Chuanren Liu, and Eun Kyong Shin for their superb guidance and for helping make this research much better. This work would not be possible without your input. I would like to sincerely thank the National Cancer Institute for funding my research. I would also like to thank past and current members of the CMBI research group at UT Health Science Center, and the Bredesen Center staff from the University of Tennessee for their knowledge, support, and the opportunity to pursue this path.

I would like to offer sincere appreciation to Brianna White, Olufunto Olusanya, Whitney Brakefield, Lokesh Chinthala, Nariman Ammar, Akram Mohammed, Nicole Siuti, Peyton Vanhook, Charinjit Chaturlal, Budhu Bhaduri, David Page, April Morton, Melissa Dumas, Mark Littman, Russell Zaretzki, Audrey Martin, Nathanial Raymond and the Y'all Real Crew, Henry Simpson, BC 2019 cohort, Subhadeep Chakraborty, Salman Khan, James Sousa, Mike Seal, Carl Sagan, and Richard Berendzen. This work would not have been possible without your inspiration, friendship, lessons, and guidance. I would like to thank my family and grandparents for their encouragement and providing interesting experiences that made me curious about the universe that we find ourselves a part of. Finally, I would like to thank my wonderful wife Alison for hanging in there with me over the last 8 years with high spirits. I am forever grateful for your love and support.

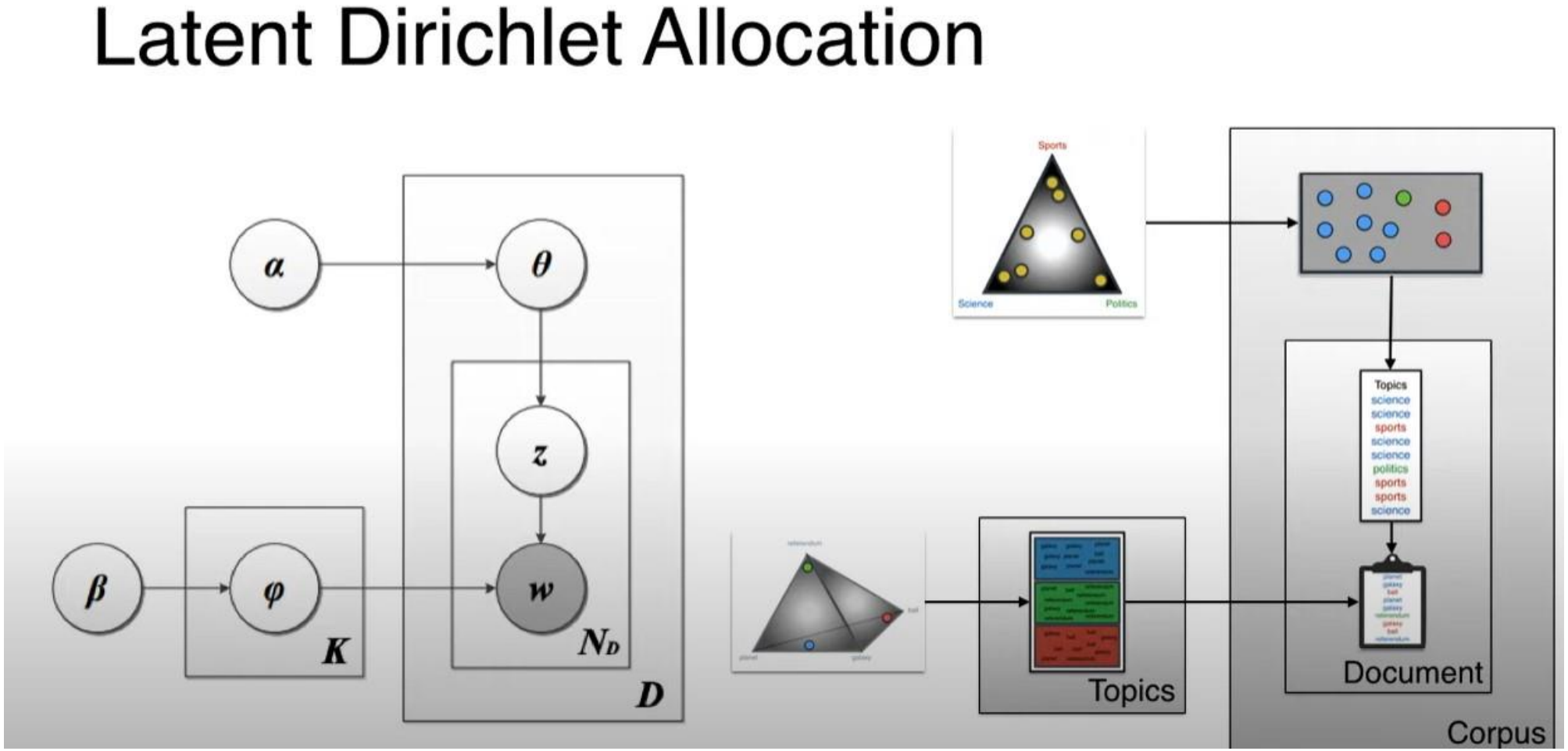
Thank you!

Questions?

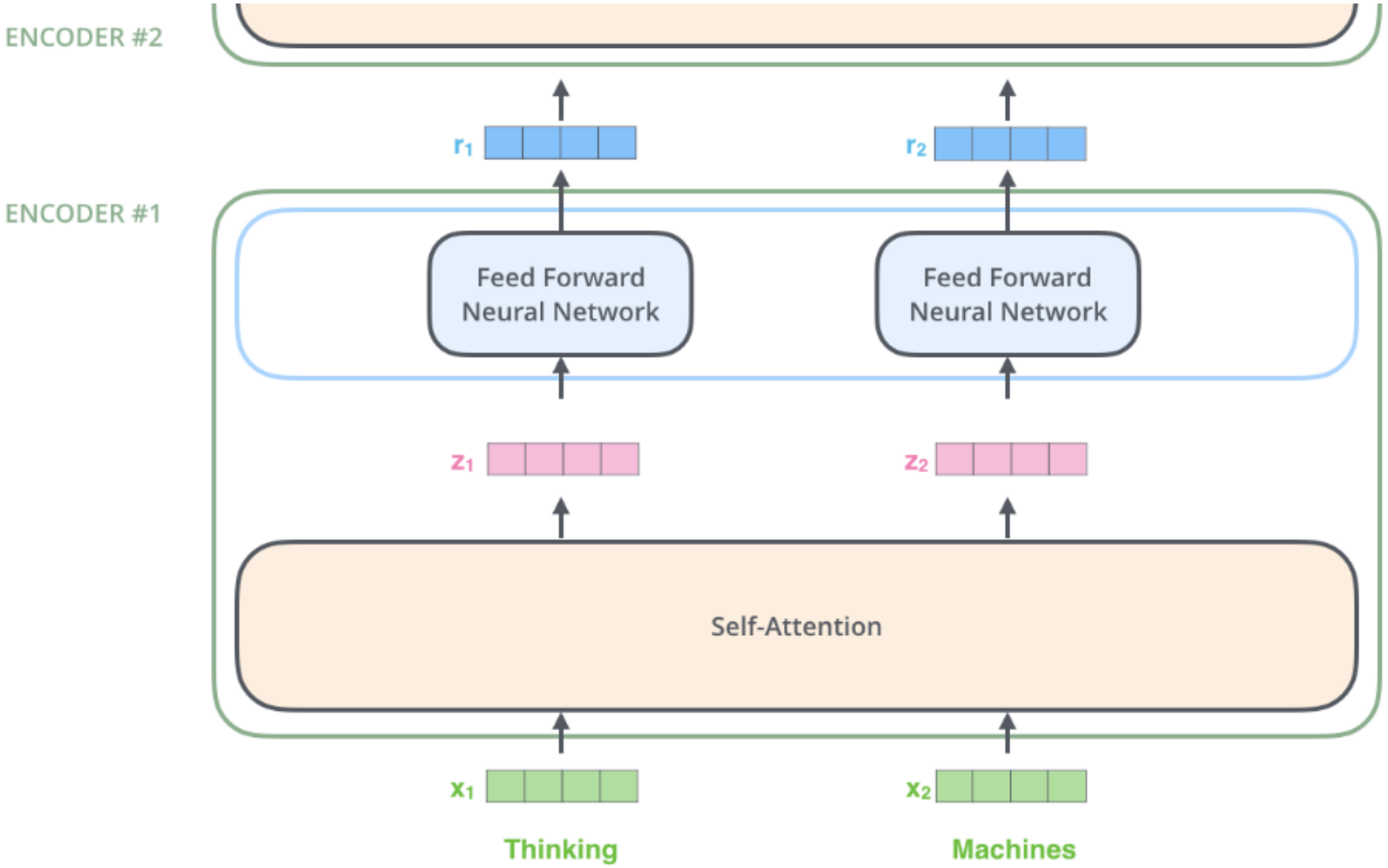
Data preparation



LDA process


Latent Dirichlet Allocation
α
θ
z
w
β
φ
K
N_D
D
Sports
Science
Politics
Topics
science
science
sports
science
science
politics
sports
sports
science
Topics
Document
Corpus

Transformer



Courtesy of
Jay Allamar

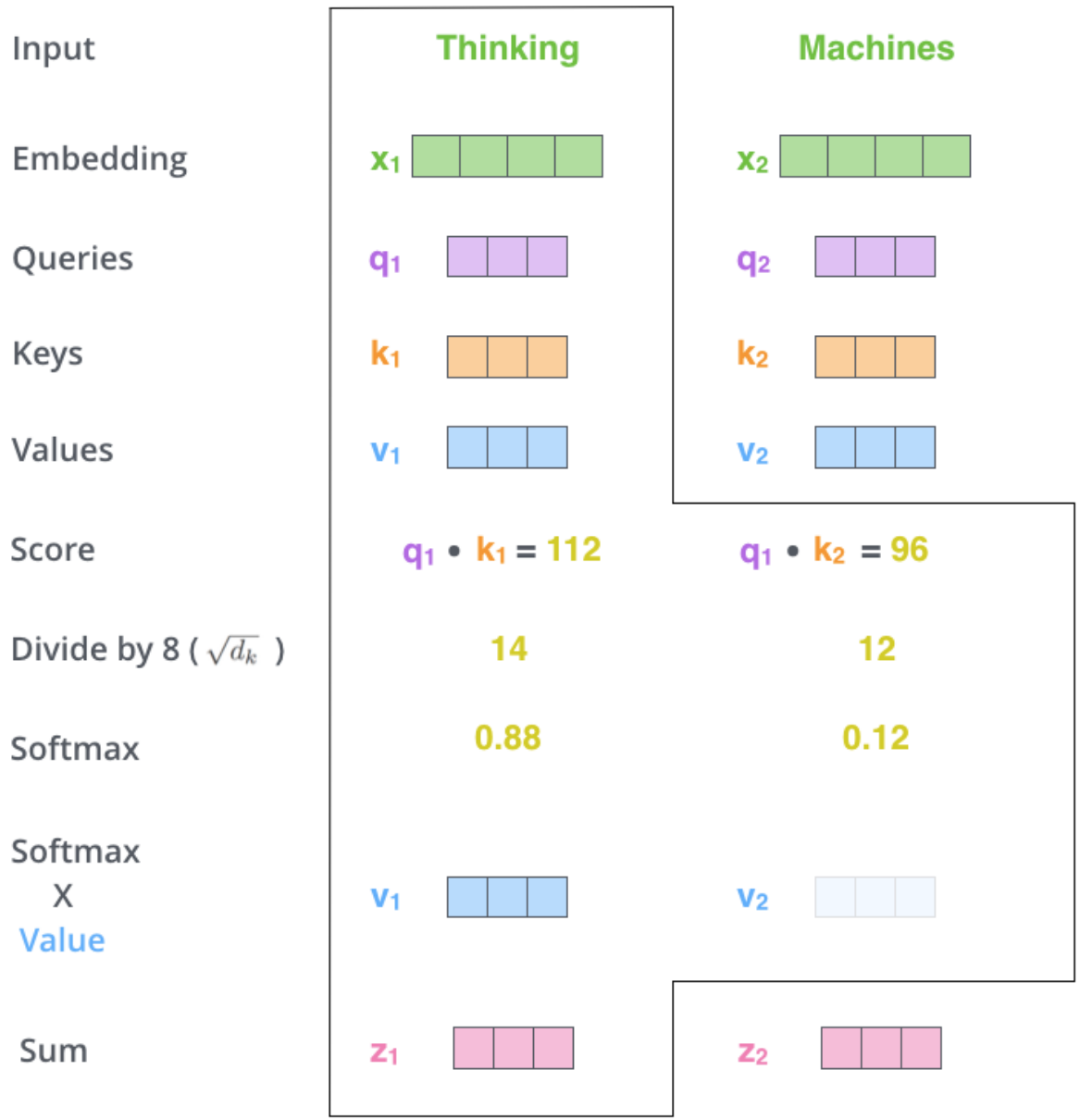
Dot-product attention

$$A(Q, K, V) = softmax(QK^T)V$$

Attention is all you need!

This is done multiplie times at once (e.g., Multihead attention and creates multiple heads.

Vaswani et al., 2018



Reddit 1.2

CovidVaccinated, Coronavirus, CovidVaccine, conspiracy_commons, COVID19_support, COVID19, AntiVaxxers, VACCINES, conspiracytheories, China_Flu, COVID19positive, conspiracy.

Reddit 2 subreddits

China_Flu, CoronavirusCanada, CoronavirusNewYork, Coronavirus_PH, Coronavirus_KY, Coronavirus_NC, Coronavirus_BC, Coronavirus_Ireland, COVID19_Ohio, COVID19_Maine, COVID19_Testimonials, COVID19positive, modernavaccine, COVID19, Coronavirus, CoronavirusUS, HealthAnxiety, CoronavirusUK, COVID19_support, PfizerVaccine, Covid19VaccineRats, GotTheVaccine, CovidAnxiety, COVIDVaccineTalk, covidlonghaulers, AskDocs, COVID19_support, CoronavirusCA, CoronavirusWA, CoronavirusNewYork, CoronavirusCanada, CoronavirusOC, CoronavirusAZ, CoronaVirus, CoronaVirusTX, CoronaVirusUT, Coronavirus_NZ, CoronavirusTN, CoronavirusAlabama, CoronavirusAL, CoronavirusMichigan, CoronavirusMissouri, CoronavirusMA, CoronavirusVA, CoronaVirusWV, CoronavirusOR, CoronavirusOregon, CoronavirusAZ, CoronavirusGA, CoronavirusSC, CoronavirusNJ, coronavirusnewmexico, CoronavirusNE, CoronavirusIllinois, CoronavirusIdaho, Coronavirushawaii, CoronavirusArmy, CoronavirusLouisiana, CoronavirusNV, Coronavirusnevada, CoronavirusMO, CoronavirusMontreal, CoronavirusMontana, CoronavirusMN, CoronavirusWI, CoronavirusPA, CoronavirusNH, CoronavirusKansas, CoronavirusOH, CoronavirusOhio)nt very happy.

Reddit 1 subreddits

Vaccines, CovidVaccine, CovidVaccinated, AntiVaxxers, vaxxhappened, antivaccine, conspiracy, conspiracytheories, NoNewNormal, conspiracy_commons, COVID19, COVID, and coronavirus

Query term

COVID vaccine, vaccine, vaccination, immune, immunity, COVID vaccination, corona vaccine, COVID19 vaccination, COVID-19 vaccination, coronavirus vaccination, coronavirus vaccine, COVID-19 vaccine, coronavirus vaccine, coronavirus vaccination, Moderna, Pfizer, J&J, Johnson & Johnson, COVID vax, corona vax, covid-19 vax, covid19 vax, coronavirus vax

Centrality

Network Supplementary

- ***Betweenness centrality*** measures the number of times a node lies on the shortest path between other nodes. (i.e., nodes bridge a network together. Ideal for communication networks).

- ***Eigenvector centrality*** measures node influence based on quantity of connections to other nodes (ideal for social networks) .

- ***Degree centrality*** is determined by number of links held by each node (how many connections… frequency of connections for nodes).

Betweenness Centrality of *node X* $=$ $\sum_{\text{for all pairs of nodes}}$ $\frac{\text{fraction of shortest paths that go through } \textit{node X}}{\text{all shortest paths between every node}}$

Degree centrality: number of nearest neighbours

$$C_D(i) = k(i) = \sum_j A_{ij} = \sum_j A_{ji}$$

Normalized degree centrality

$$C_D^*(i) = \frac{1}{n-1} C_D(i)$$

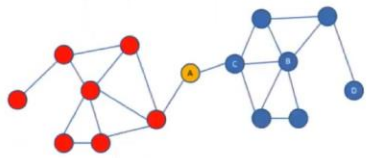
High centrality degree -direct contact with many other actors
Low degree - not active, peripheral actor

Betweenness centrality: number of shortest paths going through $\sigma_{st}(i)$

$$C_B(i) = \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

Normalized betweenness centrality

$$C_B^*(i) = \frac{2}{(n-1)(n-2)} C_B(i)$$



Remarks and limitations

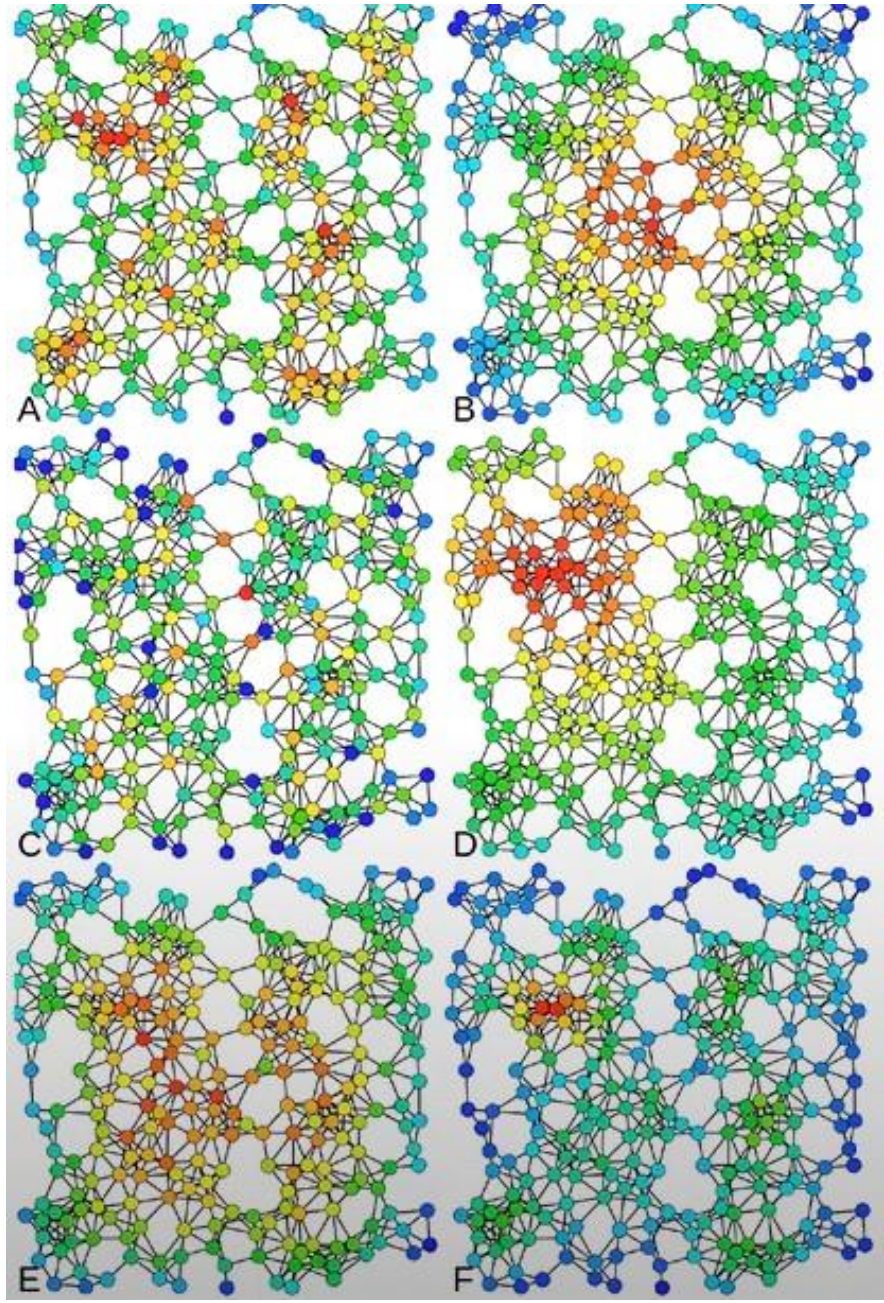
Limitations

LDA topic modeling quality is often challenging to evaluate.

Metrics exist but often don’t make sense to humans.

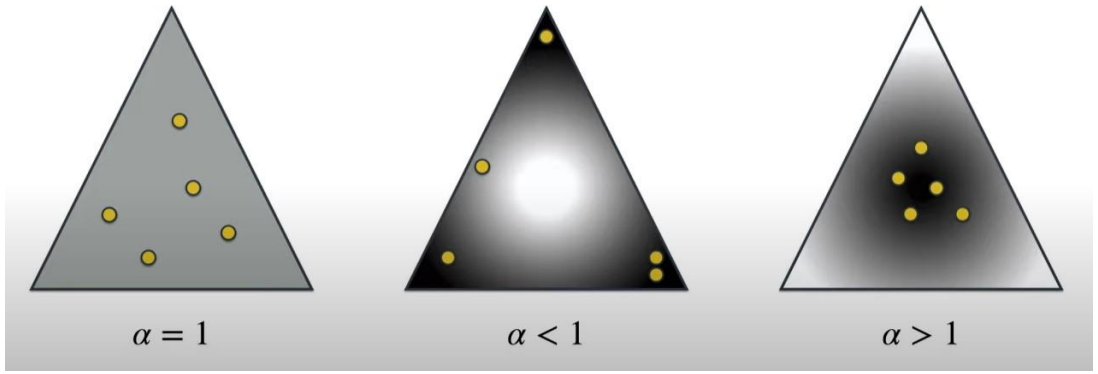
Fundamentally necessary to inspect returned topics as well as data content.

You get out what you put in.



- A) Degree centrality
- B) Closeness centrality
- C) Betweenness central
- D) Eigenvector centrali
- E) Katz centrality
- F) Alpha centrality

Dirichlet Distributions



$$P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta) = \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) P(W_{j,t} \mid \varphi_{Z_{j,t}})$$

- Total probability of the LDA model
- Dirichlet distribution of topics over terms
- Dirichlet distribution of documents over topics
- prob of a topic appearing given document
- prob of word appearing given a topic

Combined Data Set (December 1, 2020 - May 16, 2020)

Topic Number	Latent Topics
1	vaccine , people , effect , time , many , thing , year , death , month , good
2	vaccine , effect , side , week , hour , day , second , fever , symptom , sore
3	vaccine , people , dose , mask , thing , group , datum , year , immunity , efficacy
4	vaccine , virus , people , immune , system , year , antibody , vaccination , immunity , body
5	vaccine , question , contact , concern , action , people , source , news , moderator , answer
	December 2020
1	vaccine, virus, immune, system, question, cell, protein, infection, symptom, body
2	vaccine, dose, trial, group, first, efficacy, datum, case, day, participant
3	vaccine, people, year, thing, effect, time, virus, long, side, good
	January 2021
1	vaccine, dose, effect, people, side, day, second, week, first, shot
2	vaccine, people, virus, year, time, good, immunity, immune, risk, case
	February 2021
1	vaccine, dose, second, effect, day, side, week, people, first, hour
2	vaccine, people, virus, immune, vaccination, immunity, time, antibody, cell, mask
	March 2021
1	vaccine, people, virus, mask, year, thing, immunity, good, time, immune
2	vaccine, vaccination, dose, death, question, effect, week, day, people, concern
	April 2021
1	vaccine, people, mask, thing, year, effect, time, vaccination, virus, death
2	vaccine, people, virus, vaccination, immune, t, immunity, death, effect, time
	May 2021
1	vaccine, people, effect, side, time, second, shot, week, death, day
2	vaccine, people, mask, virus, vaccination, immunity, risk, year, thing, t

Bidirectional Encoder Representation from Transformers

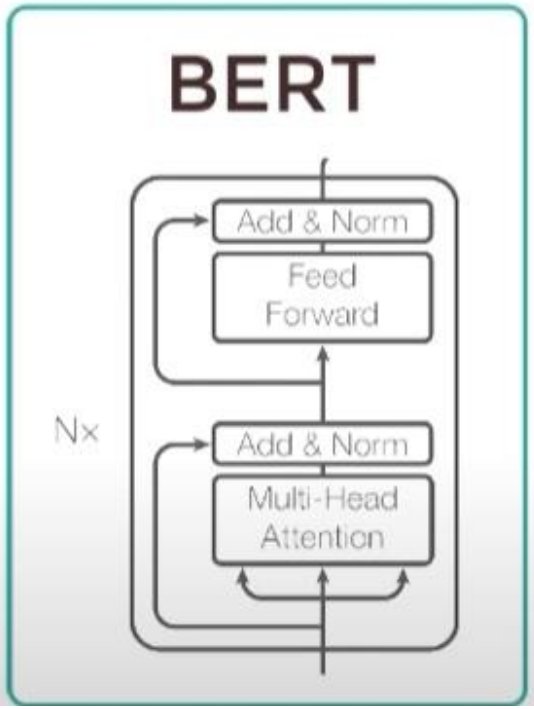
Pretraining (Pass 1) : "What is language? What is context?"

Masked Language Model (MLM)

The [MASK1] brown fox [MASK2] over the lazy dog.

Next Sentence Prediction (NSP)

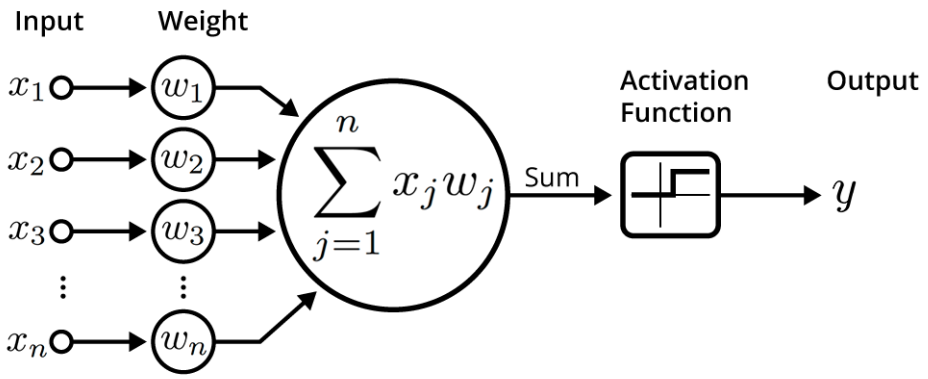
A: Ajay is a cool dude.
B: He lives in Ohio



[MASK1] = quick
[MASK2]= jumped

Yes. Sentence B follows sentence A

ADD CENTRALITY #’s



An illustration of an artificial neuron. Source: Becoming Human.

Fine-tuning

Step	Epoch	Training loss	Validation loss	Precision	Accuracy	F_1-score
500	0.4	0.5903	0.4695	0.7342	0.7728	0.7890
1000	0.8	0.3986	0.3469	0.8144	0.8596	0.8684
1500	1.2	0.2366	0.1939	0.9313	0.9260	0.9253
2000	1.6	0.1476	0.1560	0.9207	0.9452	0.9465
2500	2.0	0.1284	0.1167	0.9561	0.9592	0.9592

Natural language processing (**NLP**) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

Betweenness Centrality Twitter

20-Feb		20-Mar		20-Apr		20-May		20-Jun	
Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality
covid19	0.66666667	vaccine	0.51674648	vaccine	0.68124884	vaccine	0.78338583	vaccine	0.93333333
persists	0.5	covid19	0.34145238	covid19	0.23455916	covid19s	0.28086253	covid	0.33333333
vaccine	0.5	vaccines	0.17882803	coronavirus	0.1821511	covid	0.1273824	people	0
flu	0	cough	0.16510827	migvax	0.17629966	world	0.11066625	amp	0
influenza	0	amp	0.12924966	virus	0.0889196	like	0.09074573	covid19	0
20-Jul		20-Aug		20-Sep		20-Oct		20-Nov	
Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality
vaccine	0.71818182	vaccine	0.723809524	vaccine	0.78204829	vaccine	0.511111111	vaccine	0.808322891
covid	0.16363636	covid	0.114285714	covid	0.12629855	covid	0.322222222	covid	0.159200084
covid19	0.05454545	covid19	0.080952381	long	0.09569402	covid19	0.022222222	covid19	0.016760652
vaccines	0.00909091	vaccines	0.004761905	study	0.08758865	vaccines	0.011111111	vaccines	0.00161863
moderna	0	willing	0	covid19	0.08379602	amp	0	canada	0.00093985
20-Dec		21-Jan		21-Feb		21-Mar		21-Apr	
Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality
vaccine	0.72109637	vaccine	0.691797797	vaccine	0.727337093	vaccine	0.65963595	vaccine	0.705557543
covid	0.236943062	covid	0.279742547	covid	0.24619152	covid	0.3021266	covid	0.270435301
tweetstorm	0.029519071	vaccines	0.026444619	johnson	0.029493525	covid19	0.04027713	covid19	0.048580799
covid19	0.022063755	new	0.016129032	covid19	0.011163847	austiexas	0.02402337	long	0.015384615
vaccines	0.000451446	eu	0.016129032	fda	0.002307853	reallyopen	0.02402337	vaccines	0.004951302
21-May		21-Jun		21-Jul		21-Aug		21-Sep	
Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality
vaccine	0.564197531	vaccine	0.514335332	covid	0.562348789	covid	0.560897436	covid	0.558730499
covid	0.40893592	covid	0.45623938	vaccine	0.419961056	vaccine	0.434418146	vaccine	0.425035781
vaccines	0.027219283	covid19	0.064772614	vaccines	0.011241329	vaccines	0.010167183	vaccines	0.022833476
covid19	0.012637068	vaccines	0.035849575	people	0.003921139	covid19	0.00355147	pregna	0.01459854
long	0.012345679	booster	0.017241379	vaccinated	0.001115979	people	0.001825632	covid19	0.005343853
21-Oct		21-Nov		21-Dec		22-Jan		22-Feb	
Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality
covid	0.58918948	covid	0.55753479	covid	0.74877594	covid	0.64300936	covid	0.722121395
vaccine	0.3879491	vaccine	0.416248497	vaccine	0.213942	vaccine	0.34260163	vaccine	0.276567856
vaccines	0.0202975	vaccines	0.02143341	vaccines	0.02511483	vaccines	0.0279256	deaths	0.034482759
big	0.01190476	lamb	0.016122477	people	0.00479403	available	0.0240946	immunity	0.034482759
people	0.0022859	marcus	0.016122477	vaccinated	0.0005988	covid19	0.01718157	vaccines	0.007360355

Betweenness Centrality Reddit 2

20-Feb		20-Mar		20-Apr		20-May		20-Jun	
Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality
people	0.886243	people	0.977273	people	0.75764843	people	0.690122	county	0.658339
cases	0.373016	masks	0.060606	need	0.11864629	cases	0.163288	covid19	0.362731
flu	0.206349	health	0.060606	new	0.05711421	deaths	0.107036	paties	0.201991
virus	0.206349	home	0.060606	virus	0.04135024	going	0.08438	cases	0.127786
days	0.198413	m	0.060606	information	0.03570646	f	0.053489	information	0.052632

20-Jul		20-Aug		20-Sep		20-Oct		20-Nov	
Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality
mr	0.468795	cases	0.459695	covid19	0.334713	people	0.517012	vaccine	0.463758
county	0.374193	county	0.252723	people	0.265462	paties	0.202589	people	0.333681
people	0.33707	covid19	0.215686	couy	0.166878	vaccine	0.183526	m	0.119952
cases	0.277505	male	0.17756	health	0.12001	covid19	0.130604	cases	0.11553
covid19	0.195287	female	0.149237	vaccine	0.10804	like	0.094442	virus	0.081981

20-Dec		21-Jan		21-Feb		21-Mar		21-Apr	
Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality
vaccine	0.790704225	vaccine	0.392361	vaccine	0.496097	vaccine	0.482418	vaccine	0.362386
m	0.166827632	doses	0.200452	county	0.159794	shot	0.10924	shot	0.175872
people	0.155305164	county	0.180369	total	0.08593	people	0.101514	people	0.08463
cases	0.113829645	total	0.153824	m	0.072666	new	0.095858	m	0.06979
days	0.106036217	new	0.110287	immune	0.063585	m	0.076774	days	0.061929

21-May		21-Jun		21-Jul		21-Aug		21-Sep	
Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality
vaccine	0.371508	vaccine	0.334411	vaccine	0.512088	vaccine	0.514286	vaccine	0.485507
people	0.164159	people	0.118134	people	0.107698	people	0.096657	people	0.117955
shot	0.110992	shot	0.10318	vaccinated	0.061056	vaccinated	0.088388	including	0.078276
day	0.08498	couy	0.089772	day	0.056043	shot	0.056204	shot	0.075622
vaccinated	0.055578	m	0.06159	shot	0.04509	day	0.042034	covid	0.062854

21-Oct		21-Nov		21-Dec		22-Jan		22-Feb	
Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality	Node	Centrality
vaccine	0.241747	vaccine	0.364944	vaccine	0.296053	cases	0.329593	people	0.413338
vaccinated	0.1557	people	0.217826	people	0.280222	covid	0.265788	peak	0.28101
fully	0.09848	vaccinated	0.166677	covid	0.176362	people	0.153051	cases	0.226465
shot	0.07618	shot	0.085868	booster	0.125368	day	0.093247	vaccine	0.14582
got	0.056463	like	0.047621	cases	0.121072	vaccine	0.081507	covid	0.130853

Gibbs Sampling

Gibbs Equation

$$p(z_{d,n} = k | \vec{z}_{-d,n}, \vec{w}, \alpha, \lambda) = \frac{n_{d,k} + \alpha_k}{\sum_i^K n_{d,i} + \alpha_i} \frac{v_{k,w_{d,n}} + \lambda_{w_{d,n}}}{\sum_i v_{k,i} + \lambda_i}$$

- Number of times document *d* uses topic *k*
- Number of times topic *k* uses word type $w_{d,n}$
- Dirichlet parameter for document to topic distribution
- Dirichlet parameter for topic to word distribution
- How much this document likes topic *k*
- How much this topic likes word $w_{d,n}$

Death stats

Lifetime odds of death for selected causes, United States, 2020	
Cause of Death	**Odds of Dying**
Heart disease	1 in 6
Cancer	1 in 7
COVID-19	1 in 12
All preventable causes of death	1 in 21
Chronic lower respiratory disease	1 in 28
Opioid overdose	1 in 67
Suicide	1 in 93
Motor-vehicle crash	1 in 101
Fall	1 in 102
Gun assault	1 in 221
Pedestrian incident	1 in 541
Motorcyclist	1 in 799
Drowning	1 in 1,024
Fire or smoke	1 in 1,450
Choking on food	1 in 2,745
Bicyclist	1 in 3,396
Sunstroke	1 in 6,368
Accidental gun discharge	1 in 7,998
Electrocution, radiation, extreme temperatures, and pressure	1 in 14,705
Sharp objects	1 in 26,744
Cataclysmic storm	1 in 35,074
Hot surfaces and substances	1 in 50,341
Hornet, wasp, and bee stings	1 in 57,825
Dog attack	1 in 69,016
Lightning	Too few deaths in 2020 to calculate odds
Railway passenger	Too few deaths in 2020 to calculate odds
Passenger on an airplane	Too few deaths in 2020 to calculate odds