Global COVID-19 Twitter dataset and language models for sentiment analysis and topic modelling

Rohitash Chandra

School of Mathematics and Statistics
UNSW Sydney

Seminar: University of Tasmania, and University of Melbourne

Overview

- Language modelling with deep learning
- Sentiment analysis during the rise of novel COVID-19 cases in India.
- Topic modelling to compare the three major waves in India.
- Anti-vaxxer sentiments worldwide with the goal of comparing major events/countries.
- Data release: Global dataset of major Twitter active countries such as the UK, Brazil, USA, India, Japan, Indonesia, Australia, and Indonesia.

UNSW THE UNIVERSITY OF NEW SOUTH WALES SYDNEY • AUSTRALIA

Sentiment Analysis

- Sentiment analysis is part of natural language processing, text analysis, computational linguistics, and biometrics.
- Sentiment analysis typically use language models to identify, extract, quantify, and study affective states (emotions).
- Sentiment analysis is used for marketing and advertising, analysing customer reviews and surveys, social media analysis, and applied to various domains including medicine and public health.

Topic Modelling

- Topic modelling is used for discovering abstract "topics" that occur in a collection of documents or text corpus.
- Topic modelling used text-mining tool for discovery of hidden semantic structures in a text body.
- A document typically concerns multiple topics in different proportions; hence, it is a challenge for naturally language processing algorithms and models given ambiguity in expressions (Twitter and social media) .
- Topic modelling is useful in social media analysis, understanding public behaviour during events such as elections. It has been useful for marketing and advertising.

Recurrent Neural Networks

At times, we only need to look at recent information to perform the present task, such as a language model trying to predict the next word based on the previous ones. In such cases, where the gap between the relevant information and the place that it's needed is small, recurrent neural networks (RNNs) can learn to use past information.

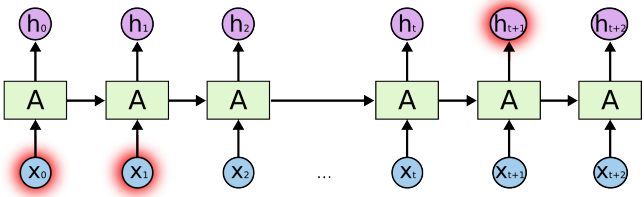


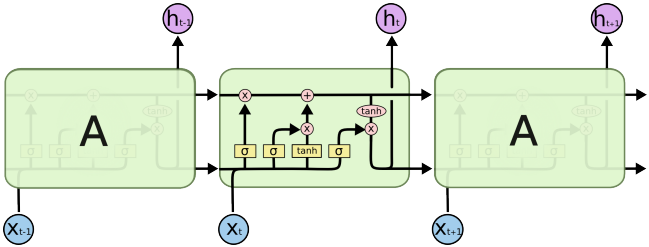
Figure: Vanishing gradient problem - long term dependencies [1]

[1] Source: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

LSTM

There are also cases where we need more context where the gap between the relevant information and the point where it is needed to become very large. As that gap grows, RNNs have difficulty to learn to connect the information. The problem is known as the vanishing gradient problem where RNNs have difficulty to learn long term dependencies in sequences.

Long short-term memory (LSTM) networks address the vanishing gradient problem with memory cells.



LSTM

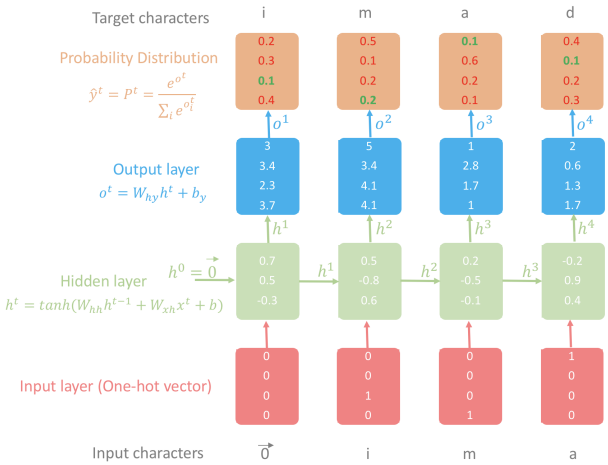


Figure: LSTM for Language modelling [3]

Transformers

- A transformer is an extended LSTM model that adopts the mechanism of attention which mimics cognitive attention to enhance important parts of the data while fading the rest.
- Transformers also use an encoder-decoder architecture and have mostly been used for NLP tasks such as translation and text summarising.
- In comparison to conventional RNNs, transformers do not require data to be processed in a sequential order since the attention operation provides context for any position in the input sequence.

Transformers

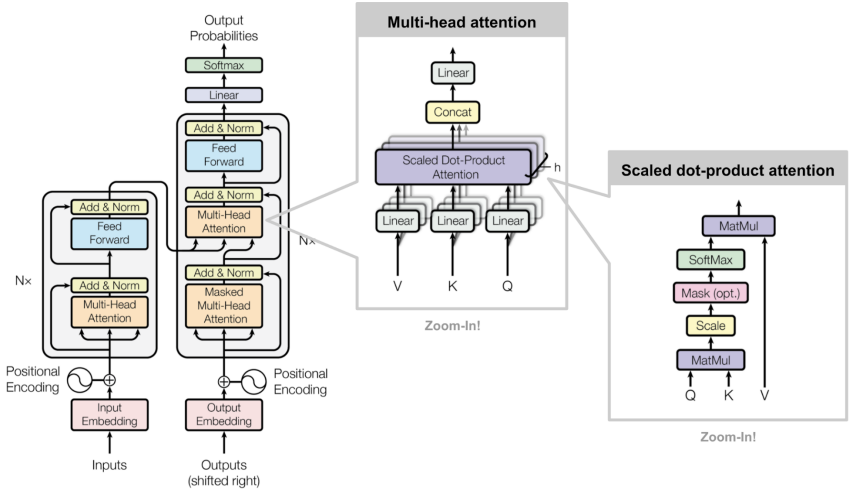


Figure: Transformers - specialised LSTM networks with attention mechanism.

BERT

Developed by Google, bidirectional encoder representations from Transformers (BERT) is pre-trained on a large corpus for language tasks.

- The original BERT has two models: 1.) BERT-base features 12 encoders with 12 bidirectional self-attention heads, and 2.) BERT-large features 24 encoders with 16 bidirectional self-attention heads.
- These are pre-trained from unlabeled data extracted from a corpus with 800 million words and English Wikipedia with 2,500 million words, respectively.
- Word2vec and GloVe are content-free models that generate a single word embedding representation for each word, whereas BERT takes into account the context for each occurrence of a given word which makes BERT one of the best language models.

BERT

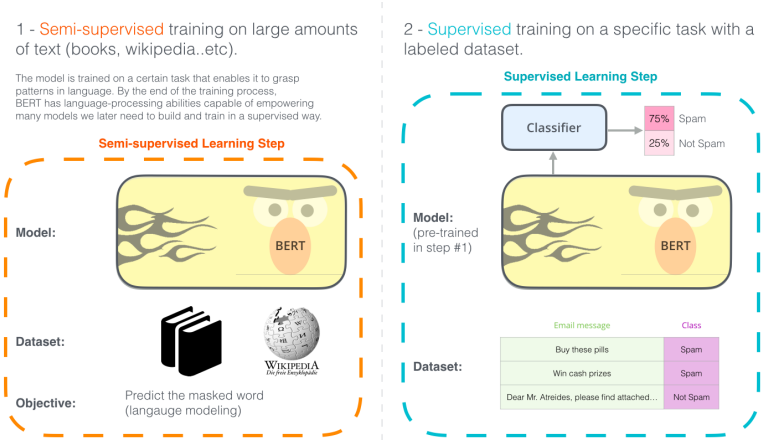


Figure: BERT for Language modelling [4]

[4] Source: https://jalammar.github.io/illustrated-bert/

Social Media and Language Modelling

- Advancements of deep learning-based language models have been promising for sentiment analysis with data from social networks such as Twitter.
- Social media plays a crucial role in shaping the worldview during election campaigns. Social media has been used as a medium for political campaigns and a tool for organizing protests; some of which have been peaceful, while others have led to riots.
- Previous research indicates that understanding user behaviour, particularly in terms of sentiments expressed during elections can give an indication of the election outcome.

COVID-19 Sentiment Analysis

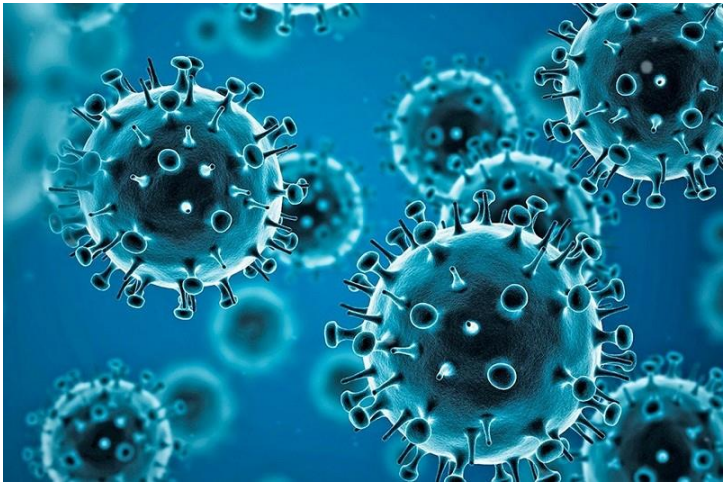


Figure: Visualisation of coronavirus. Source: WHO

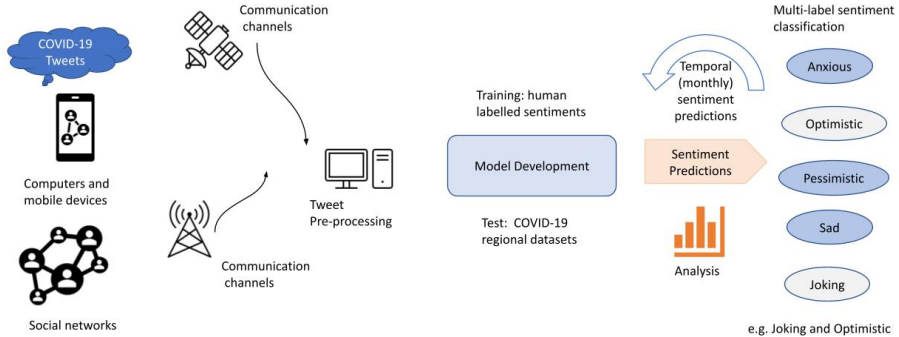
COVID-19 Sentiment Analysis

- The COVID-19 pandemic is a catastrophic event that has raised a number of psychological issues such as depression given abrupt social changes and lack of employment.
- During the rise of COVID-19 cases with stricter lock-downs, people have been expressing their sentiments in social media. This can provide a deep understanding of human psychology during catastrophic events.
- We present a framework that employs deep learning-based language models via long short-term memory (LSTM) recurrent neural networks for sentiment analysis during the rise of novel COVID-19 cases in India.
- The framework features LSTM language model with a global vector embedding and state-of-art BERT language model.

COVID-19 Sentiment Analysis

- We review the sentiments expressed for selective months in 2020 which covers the first major peak of novel cases in India.
- Our framework utilises multi-label sentiment classification where more than one sentiment can be expressed at once.

COVID-19 Sentiment Analysis



COVID-19 Sentiment Analysis

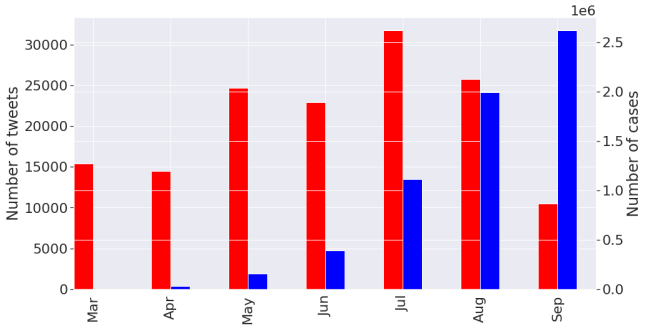


Figure: Data analysis - Tweets in India. The red bars indicate the number of tweets while the black bars indicate the number of novel cases.

COVID-19 Sentiment Analysis

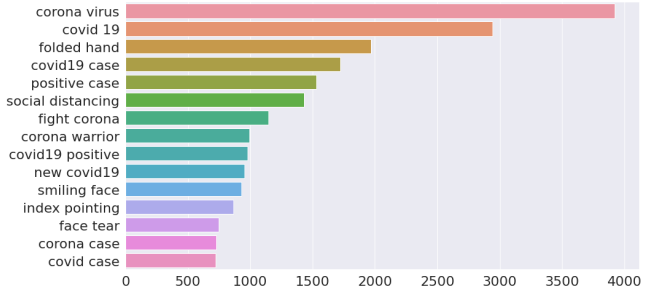


Figure: Top bi-grams

COVID-19 Sentiment Analysis

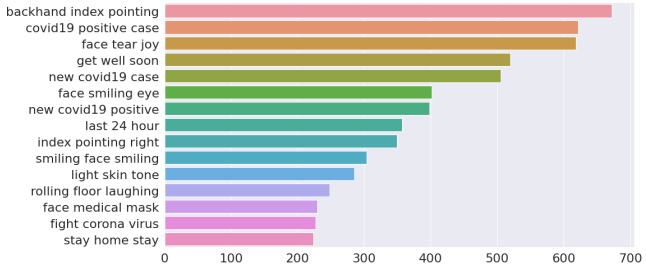


Figure: Top tri-grams

COVID-19 Sentiment Analysis

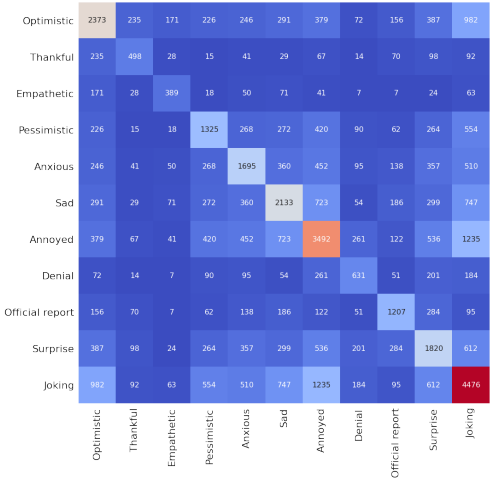


Figure: Senwave hand-labelled data for training

COVID-19 Sentiment Analysis

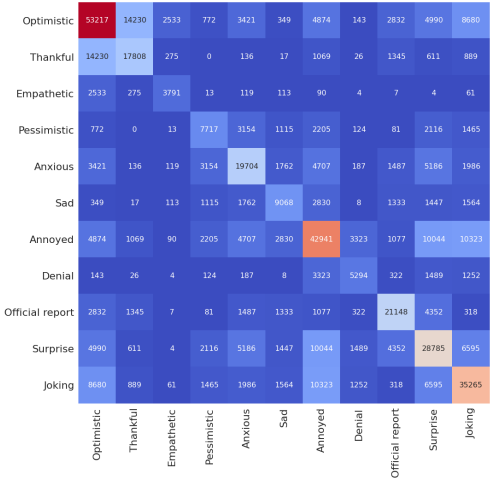


Figure: Predictions - India

COVID-19 Sentiment Analysis

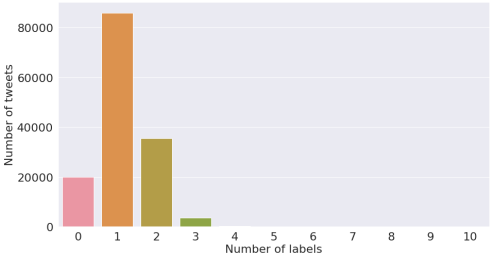


Figure: Multiple sentiments at once.

COVID-19 Sentiment Analysis

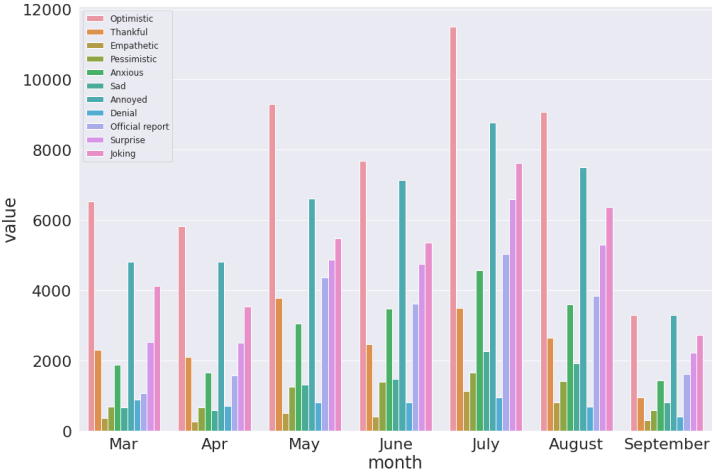


Figure: Monthly predictions - India

COVID-19 Sentiment Analysis

- Our investigation revealed that majority of the tweets have been "optimistic", "annoyed" and "joking" that expresses optimism, fear and uncertainty during the rise of the COVID-19 cases in India.
- The number of tweets significantly lowered towards the peak of new cases. Furthermore, the optimistic, annoyed and joking tweets mostly dominated the monthly tweets with much lower number of negative sentiments expressed.
- We found that most tweets that have been associated with "joking" were either "optimistic" or "annoyed", and minority of them were also "thankful". In terms of the "annoyed" sentiments in tweets, mostly were either "surprised" or "joking".
- These predictions generally indicate that although the majority have been optimistic, a significant group of population has been annoyed towards the way the pandemic was handled by the authorities.

Anti-vaxxer sentiments during COVID-19

Figure: Source: https://www.latimes.com/

Anti-vaxxer sentiments during COVID-19

Figure: Source: https://www.boredpanda.com

Anti-vaxxer sentiments during COVID-19

- We analyse the sentiments from beginning of COVID-19 pandemic and study the behaviour during the planning, development and deployment of vaccines expressed in tweets worldwide using sentiment analysis framework with LSTM network and global vector for word representation (GloVe) embedding.
- We train the model using Senwave sentiment analysis dataset which features 10,000 tweets during COVID-19 with 10 sentiments labelled by 50 experts.
- Furthermore, we use the pre-trained BERT language model developed by Google for comparison.
- We use the trained model to predict sentiments associated with the term vaccine from tweets worldwide spread weekly for about two years since beginning of COVID-19. In this way, we provide analysis of monthly anti-vaccine sentiments over the course of the COVID-19 pandemic.

Anti-vaxxer sentiments during COVID-19

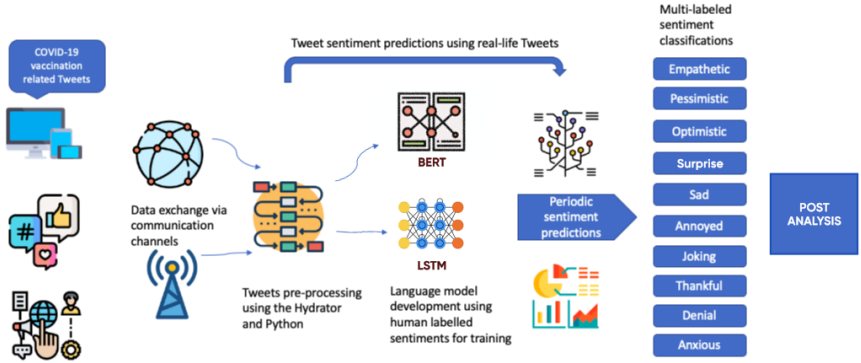


Figure: Framework

Anti-vaxxer sentiments during COVID-19

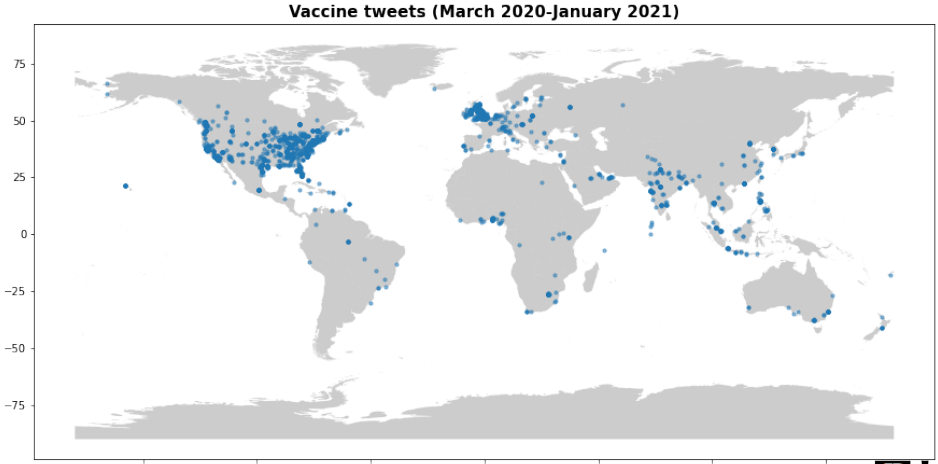


Figure: Framework

Anti-vaxxer sentiments during COVID-19

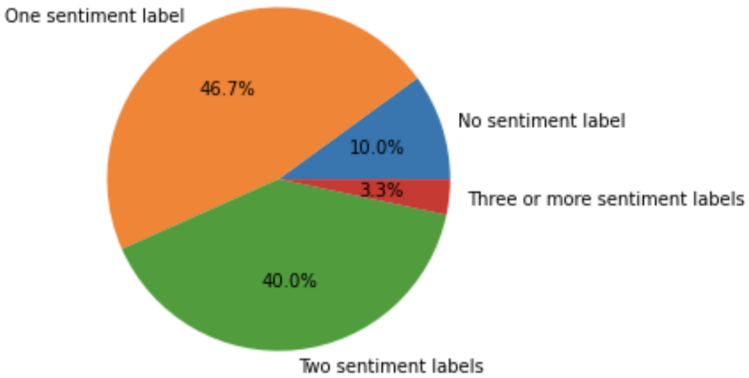


Figure: Framework

Anti-vaxxer sentiments during COVID-19

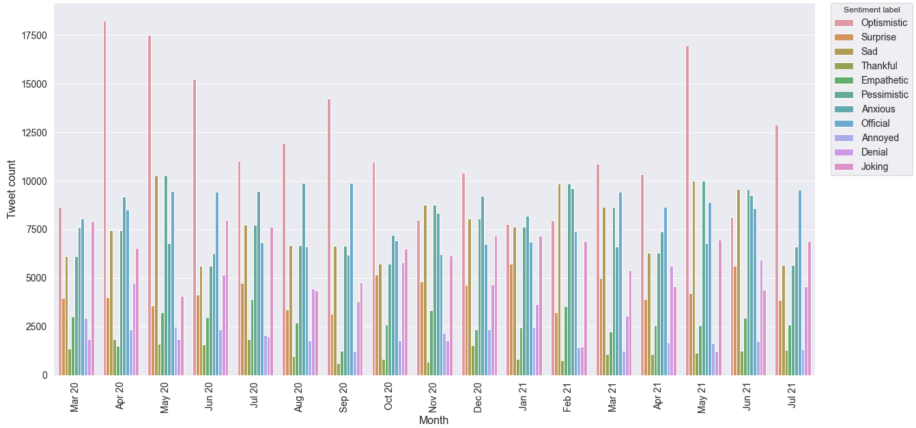


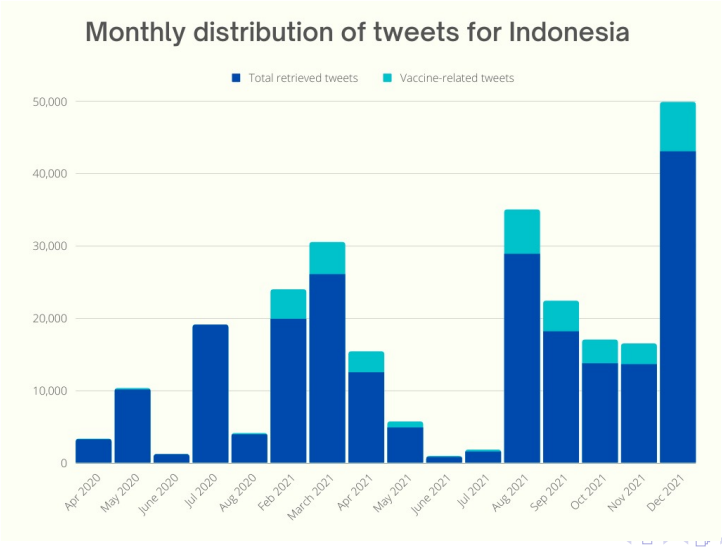
Figure: Framework

Anti-vaxxer sentiments during COVID-19

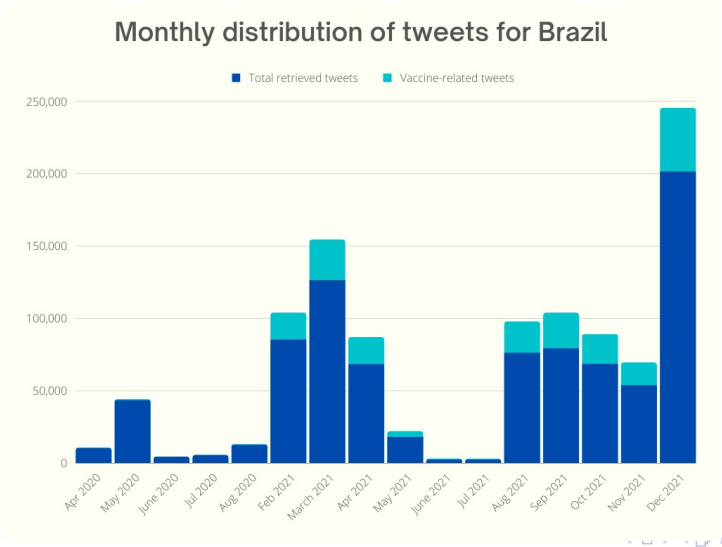
Sentiments	Sample tweet	Score
Sad	*"please move we need a vaccine for corona we need to complete our normal lives"*	0.13
Empathetic, Pessimistic, Sad	*"no hell no i am vulnerable and the only way i would consider risking the aura of life with the corona virus vaccine is if the total process were proven definitely to me and i could trust it"*	0.05
Sad, Anxious	*"please take care and seriously think about others this stuff is real we do not have a vaccine it is still an awful disease to catch and even if you do not get it you might pass it on"*	-0.38
Annoyed, Denial	*"do not get distracted from the main agenda behind this pandemic mass genocide through vaccines and a one world government run by fascist dictatorship peirs is distracting you"*	-0.11
Official	*"not taking a covid vaccine i got the flu vaccine twice in my life those were the only two years i got the flu"*	-0.06
Annoyed	*"i will murder anyone who tries to force me or my child to inject a coronavirus vaccine"*	-0.38
Annoyed, Pessimistic	*"maybe trump should investigate bill gates and the cdc thoroughly regarding the vaccines change the laws so people can sue vaccine companies people do not trust vaccines these days no one wants it"*	-0.21
Denial	*"never what they can do via vaccine is too scary and they have no idea of the impact "*	-0.05

Figure: Framework

Anti-vaxxer sentiments during COVID-19



Anti-vaxxer sentiments during COVID-19



Anti-vaxxer sentiments during COVID-19

- We find more tweets being identified with the ‘optimistic’ and ‘empathetic’ labels when vaccine development progress has been announced and when transmission has been reasonably contained. Vice versa, when negative news concerning the quality and effectiveness of COVID-19 vaccinations was revealed, we observe an increasing tendency for negative emotion labels including ‘sad’, ‘annoyed’ and ‘angry’ to be associated with a greater amount of tweets.
- This sends out an important message to public health professionals and disease control bodies that they should increase the transparency and timeliness of information release to actively keep the public informed to minimise the spread of misinformation about vaccines, as an effort to restore faith in scientific evidence and reduce ungrounded anti-vaccine sentiments on social media.

COVID-19 Topic Modelling

- India witnessed a number of events during these waves, which included regional elections, farmers protests, and roll-out of vaccines. Our goal is to see how these affected the topics covered in COVID-19 tweets during the respective timelines.
- We use deep learning based language framework for COVID-19 topic modelling taking into account data from COVID-19 emergence, which includes three distinct peaks (waves) in India as a case study.

COVID-19 Topic Modelling

Corpus	# Tweets	# Words	#Cases	#Major Variants	#Mutants
First Wave	159312	2468018	10302012	Alpha	E484Q and E484K
Second Wave	187531	12274701	24207004	Delta, Delta Plus	K417N
Third Wave	172583	4589488	7652375	Omicron	

Table: Dataset Statistics (India)

COVID-19 Topic Modelling

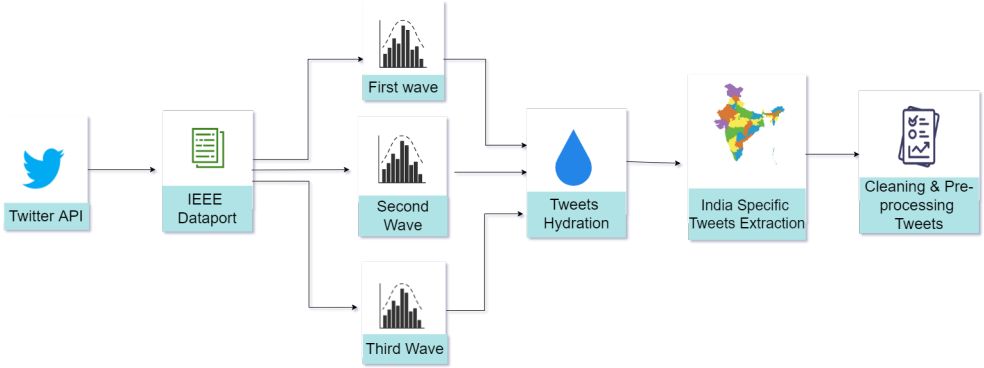


Figure: India specific dataset extraction from global COVID-19 tweets dataset

COVID-19 Topic Modelling

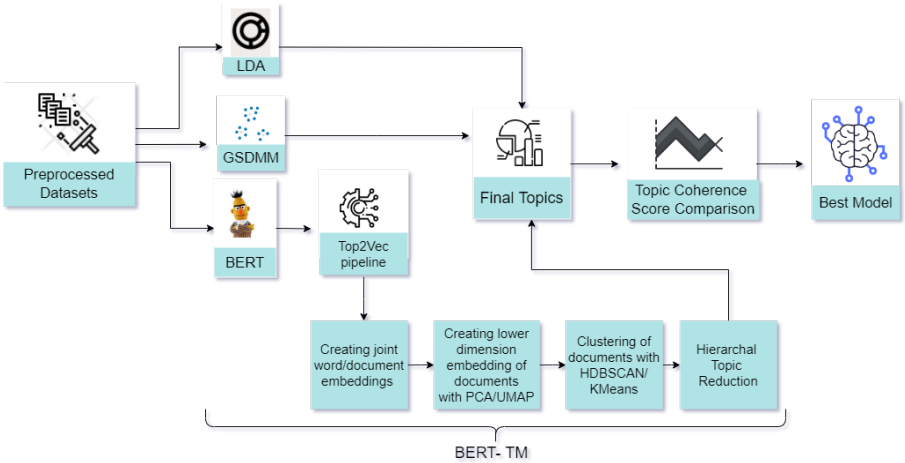


Figure: Framework for COVID-19 topic modeling.

COVID-19 Topic Modelling

We use Twitter data from India and also compare with our earlier works that looked at sentiment analysis of the first wave in India. We note that we refer to the three distant peaks as waves. Our goal is to extract and study the various topics emerging in the three different waves and discuss the relationship to emerging events and issues in the media during the respective time-frames.

COVID-19 Topic Modelling

We need a way to measure how coherent and understandable the topics are for humans, and hence a score is needed.
Topic coherence (TC) is measured by normalised point-wise measure information (NPMI) which considers a pair of words (w_i, w_j) from the top N (set to 50) words of a given topic:

$$\text{NPMI}\left(w_i, w_j\right)^{\gamma} = \left(\frac{\log \frac{P\left(w_i, w_j\right)+\epsilon}{P\left(w_i\right) \cdot P\left(w_j\right)}}{-\log\left(P\left(w_i, w_j\right)+\epsilon\right)} \right)^{\gamma}$$

where, the joint probability $P\left(w_i, w_j\right)$, i.e the probability of the single word $P\left(w_i\right)$ is calculated by the Boolean sliding window approach. γ is the time range.

COVID-19 Topic Modelling

Model	Datasets					
	First wave		Second wave		Third wave	
	#Topics	TC- NPMI	#Topics	TC- NPMI	#Topics	TC- NPMI
BERT TM	58	0.69442				
GSDMM	60	0.38958				
LDA	58	0.34419				

COVID-19 Topic Modelling

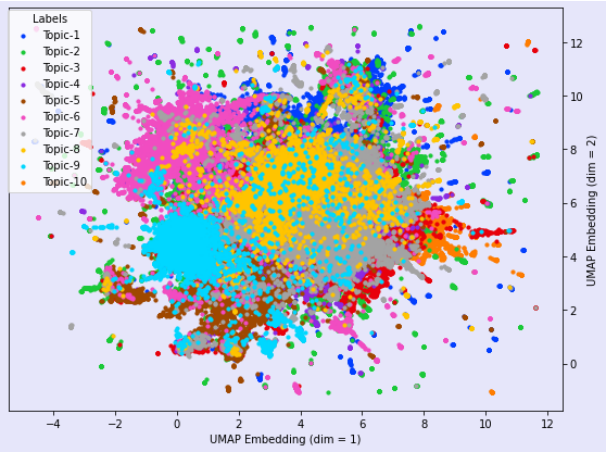


Figure: UMAP embedding showing topic overlap in first wave

COVID-19 Topic Modelling

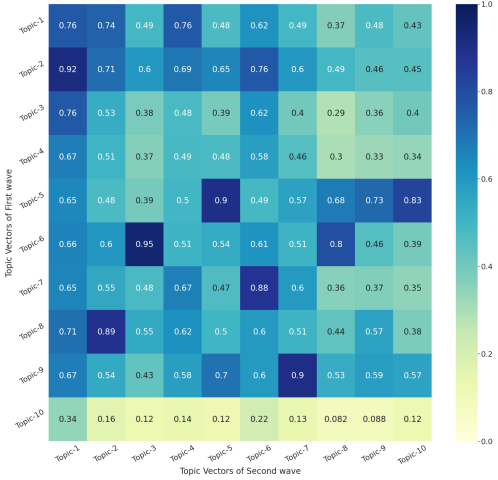


Figure: Heatmap showing correlation of topics in first and second wave

COVID-19 Topic Modelling

#Topic	First wave #Topic Words	Second wave #Topic Words
Topic 1	lockdown', 'locked', 'lockdowns', 'lockdownindia', 'lockdow', 'blocked', 'lock', 'unlock', 'lack', 'prevent	breakingnews', 'covidindia', 'covidvaccine', 'covidsecondwave', 'coronaupdate', 'covid', 'murshidabad', 'covaxin', 'catastrophic', 'coronavaccine'
Topic 2	kejriwal', 'amitabhbachchan', 'suspected', 'lakh', 'tested','suffered', 'examination', 'bharat', 'haryana', 'hence'	governments', 'govt', 'corruption', 'kejriwal', 'governance', 'parliament', 'protests', 'politician', 'shameful', 'arrestramdev'
Topic 3	rather', 'hence', 'pathetic', 'thane', 'facepalming', 'toh','amitabhbachchan', 'worry', 'than', 'suspected'	indiahelp', 'hindustan','bharat' 'kejriwal', 'diwali','healthyindia', 'crore', 'pakistan', 'gandhi', 'bangladesh','caste'
Topic 4	coronaupdate', 'coronaupdates', 'corona', 'coronawarriors', 'coronaindia', 'coronalockdown', 'coronavaccine', 'coronapandemic','coronil', 'coron'	suffered', 'volunteer', 'disaster', 'cancelboardexam', 'unemployment', 'helping', 'sorry', 'disasters', 'donation', 'poverty'
Topic 5	coronavirus', 'chinesevirus', 'chinavirus', 'uhanvirus', 'vaccine', 'vaccines', 'vaccination', 'virus', 'viruses', 'viruse'	vaccineforall', 'vaccinemaitri', 'vaccinated', 'vaccinezehad', 'vaccine', 'vaccinequity', 'getvaccinated', 'immunization', 'epidemic', 'viral'
Topic 6	indian', 'india', 'hindu', 'indians', 'hindustan', 'bharat','kejriwal', 'hindus', 'hindi', 'indi'	kejriwal', 'thanking', 'blessed', 'jharkhand' 'jammuandkashmir', 'diwali', 'gratitude', 'ahmedabad', 'murshidabad', 'haryana', 'celebrated'
Topic 7	appreciate', 'gratitude', 'blessed', 'grateful', 'appreciated', 'blessing', 'thankful', 'bless', 'blessings', 'helping'	medical', 'hospitals', 'hospitalized', 'healthcare', 'ambulance', 'doctors', 'patients', 'ambulances', 'clinical', 'cure'
Topic 8	governments', 'government', 'govt', 'govts', 'modi', 'modigovernment', 'parliament', 'politicians', 'authorities', 'ministers'	'vaccinate', 'unvaccinated', 'indiahelp', 'india', 'indian', 'getvaccinated', 'hindu', 'indians', 'immunization', 'healthyindia'
Topic 9	hospitals', 'hospital', 'medical', 'patients', 'healthcare','clinical', 'nurse', 'doctors', 'nurses', 'doctorsday'	'vaccinated', 'vaccinequity', 'unvaccinated', 'immunization', 'epidemic', 'chinesevirus', 'flu', 'coronavirus', 'immunity', 'contagious'
Topic 10	havoc', 'wait', 'coz', 'hours', 'that', 'apne', 'kumar', 'zany','amitshah', 'blessing'	'masks', vaccinations', 'wearmask', 'immunization', 'stayhome', 'lockdown', 'coronavirus', 'flu', 'immunity', 'covaxin'

Figure: Table of topics in first and second wave.

COVID-19 Topic Modelling

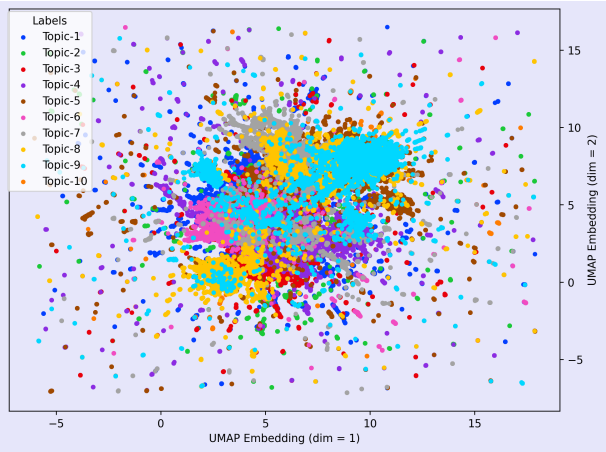


Figure: UMAP embedding showing topic overlap in second wave

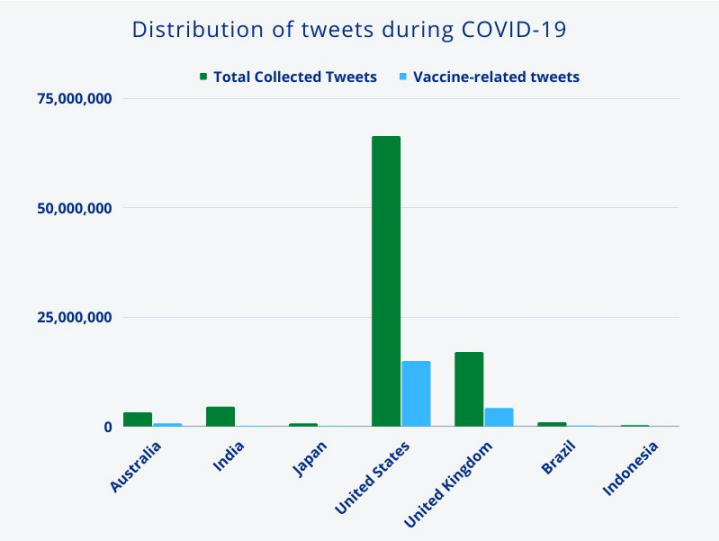
COVID-19 Topic Modelling

- UMAP embedding in the Figures show an overlap in topics, though major clusters are shown.
- Heatmap shows high correlation in certain combination of topics, in first and second waves. eg. Topic 1 second wave Topic 2 first wave.
- We will have further analysis about nature of these topics for during the respective waves by comparison with COVID-19 novel case trend, with media sources (Newspapers).

Global COVID-19 Twitter dataset

- The major events during COVID-19 resulted in active social media usage and harnessing the knowledge from the data from platforms such as Twitter can be beneficial for researchers.
- Hence we provide easy access to tweets from the emergence of COVID-19 capturing major Twitter active countries such as the UK, Brazil, USA, India, Japan, Indonesia, Australia, and Indonesia.

Global COVID-19 Twitter dataset



Global COVID-19 Twitter dataset

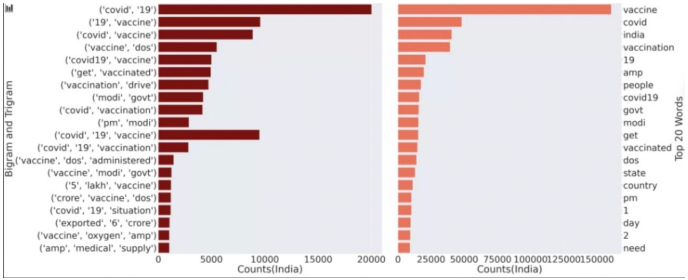


Figure: Framework

Global COVID-19 Twitter dataset

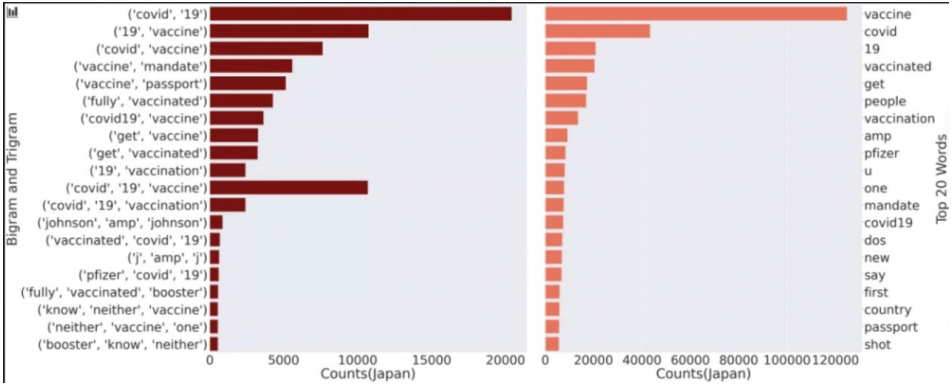
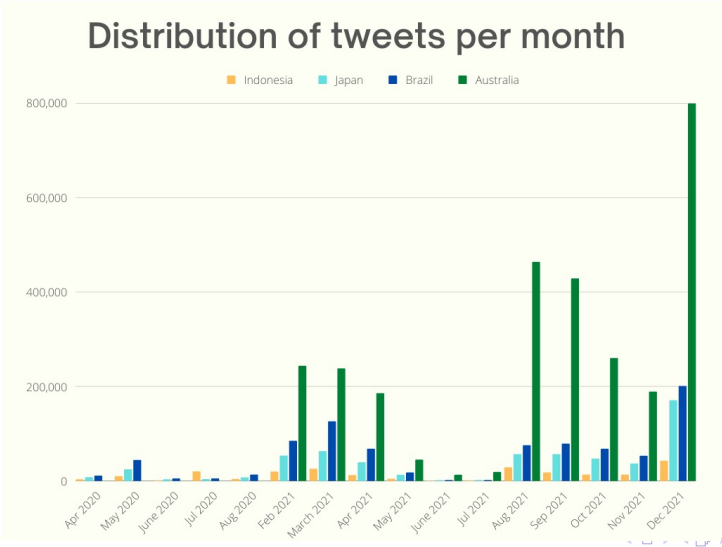


Figure: Framework

Global COVID-19 Twitter dataset



Global COVID-19 Twitter dataset

- We extracted extracted from Tweet IDs: Rabindra Lamsal, "Coronavirus (COVID-19) Tweets Dataset", IEEE Dataport, March 2020, doi: https://dx.doi.org/10.21227/781w-ef42.
- We obtained minimum 800,000 global tweets for a day (20-30 MBs). Due to the large time required in hydration of tweets from IDs, we selected only three days per week (Monday, Wednesday and Friday) from March 2020- February 2022.
- Hydrator CSVs obtained limit to around 2GB per day (JSON 20 GB).
- COVID-19 by country USA(15GB), UK(5GB), India(2GB) and Australia, Indonesia, Brazil, Japan (¡1GB).
- Further, we obtained country specific antivaxxer tweets for antivaccine sentiments.

Global COVID-19 Twitter dataset

- Janhavi Lande, Yashwant Kaurav, Cathy Yu, amp; Rohitash Chandra. (2022). Global COVID-19 Twitter dataset. Kaggle. https://doi.org/10.34740/KAGGLE/DS/2397387 (2 GB zipped)
- Note USA dataset is being uploaded (15GB but zipped as 5GB)

Acknowledgements

(a) Janhavi Lande (IIT Guwahati)

(b) Yashwant Singh Kaurav (IIT Delhi)

(c) Cathy Yu (UNSW)

(d) Arti Pillay (Fiji National University)

(e) Aswin Krishna (IIT Guwahati)

Acknowledgements

- Janhavi Lande (IIT Guwahati), Yashwant Singh Kaurav (IIT Delhi), Arti Pillay (Fiji National University)
- Aswin Krishna (IIT Guwahati), Ritij Saini (IIT Bombay)
- Mukul Ranjan (IIT Guwahati), Venkatesh Kulkarni (IIT - Guwahati) (Language models for Hindu texts)
- Cathy Jiaxin Yu (UNSW) and Prof. Seshadri Vasan (Western Australia - Department of Health)

  * Indian Institute of Technology (IIT)

References

- Chandra R; Krishna A, 2021, 'COVID-19 sentiment analysis via deep learning during the rise of novel cases', PLoS ONE, vol. 16, pp. e0255615, http://dx.doi.org/10.1371/journal.pone.0255615
- Chandra R; Saini R, 2021, 'Biden vs Trump: Modeling US General Elections Using BERT Language Model', IEEE Access, vol. 9, pp. 128494 - 128505, http://dx.doi.org/10.1109/ACCESS.2021.3111035
- Chandra R; Kulkarni V; Rathi, S. 2022, 'Semantic and sentiment analysis of the Bhagavad Gita translations using BERT-based language models', arXiv (to appear January 2022)
- Chandra R; Rajan M; . 2022, 'Topic modelling in Hindu philosophy via BERT: Mapping the Upanishads with the Bhagavad Gita', arXiv (to appear January 2022)

Software and Data Availability

- Please note all projects have open software and data published via Github.
- https://github.com/sydney-machine-learning