Building an NLU model to Analyze Evolution of Topics in Sets of Text Documents

Authors: Zain Ali, Jun Ma

Define the problem:

1. What is the question your project is attempting to answer?

Our main goal for this project was to train the most efficient topic evolution model on large datasets consisting of news articles. Efficiency was determined based on the following criteria: model perplexity, topic coherence, topic diversity, and topic interpretability.

2. Where does your project fit within the broader conversation/controversy surrounding your topic?

This project collected data that was associated with the current Covid-19 pandemic which brings a new and relevant addition to existing research on topic evolution models. In contrast to most existing research, we took a more globalized approach and tracked the development of topics like ‘anxiety’, ‘mental health’, and ‘lockdown’ in the context of countries like India and Brazil.

What success would look like

1. What are you trying to accomplish?

We are aiming to find and map the evolution of sub topics regarding mental health over a span of several months prior to and after the start of the Covid-19 pandemic in India and Brazil. In doing so, we fine tuned models like BERTopic and DETM (Dynamic Embedded Topic Model) in order to optimize topic coherence, diversity, and interpretability.

2. What is the outcome you hope to achieve?

The ultimate goal was to create a topic evolution model that would effectively extract interpretable topics from any given dataset. By looking at local topic representations at each time stamp, we can determine the change/stability in subtopics that are being associated with mental health.

The Data

1. Where does it come from? What bias might be present in the data?

We collected the news through Opoint, the news data vendor. It was based on certain keywords and a time interval of 11-12 months. There are no large concerns of bias as this data has good coverage and is representative for the reporting. Getting news articles from a data aggregator instead of limited and certain channels prevents political bias. Additionally, bias is not a large issue for this project as the task has not been to map the evolution of the anxiety reporting but to do topic evolution in a given dataset.

What were some of the other issues with the dataset (missing values, limitations, etc.)? How did you deal with those issues?

It might be harder to train topic models on full text of the news articles since multiple paragraphs may report different themes. That's why full-text articles are segmented and relevant content around search keywords are extracted. Since the data coming from the news aggregator is in the form of text and date pairs, there weren't any missing values.

Your Solution/Model

1. What statistical model did you use? How does it work?

One of the models that we used was the BERTopic model. BERTopic utilizes c-TF-IDF to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions. It also allows word embedding models like Word2Vec, Fasttext, and Bert to be utilized when training the model. It creates the topic evolution visualization using Plotly.

Another model we used was the Dynamic Embedded Topic Model or D-ETM, which extends the D-LDA and vanilla ETM. Each topic is represented as a vector varying over discretized time slots, which allows the topic to vary smoothly. As the name suggested, it trains the model in a word embedding space. It visualizes the topic evolution by figuring out top words whose embedding agrees with the topic's embedding. We've tried embedding models like Word2Vec and Fasttext.

2. Did you try any other models before settling on your final one?

Instead of trying various models, we focused on fine tuning both the topic models and word embeddings in order to create the most efficient topic evolution model.

Impact/Next Steps

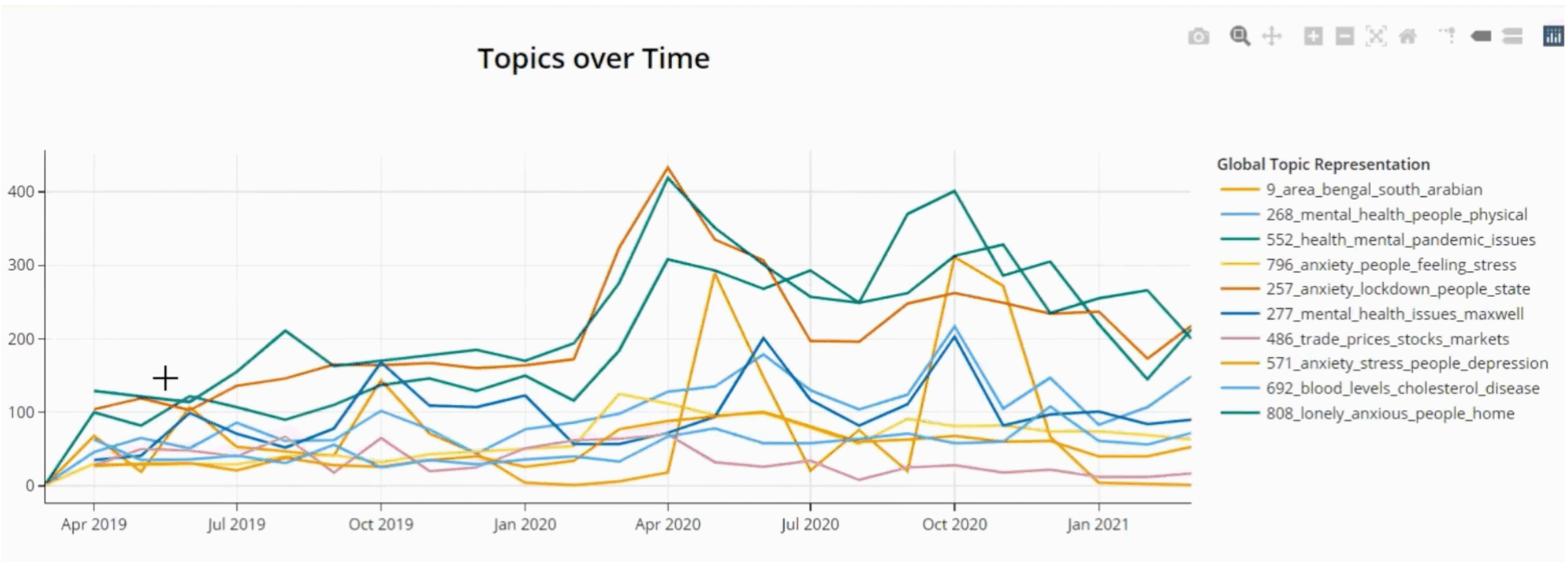
1. What were your results / what results do you expect?

After training embedding models like Word2Vec, Fasttext, and Bert, we decided to use two pre-trained Fasttext models (English & Portuguese) from which we fine-tuned on our datasets to create our final embedding model. This model for BERTopic was trained with 100 epochs for both datasets as this was where there was a convergence of loss per epoch. For BERTopic, I set the number of topics = 10, and the timestamps were sorted by month for the 'Topics Over Time' visualization.

2. What decisions will be made as a result of your work?
3. What work is left to be done?
4. Will this work be relevant in short/medium/long term?

Results for BERTopic:

Model trained on Indian News Dataset



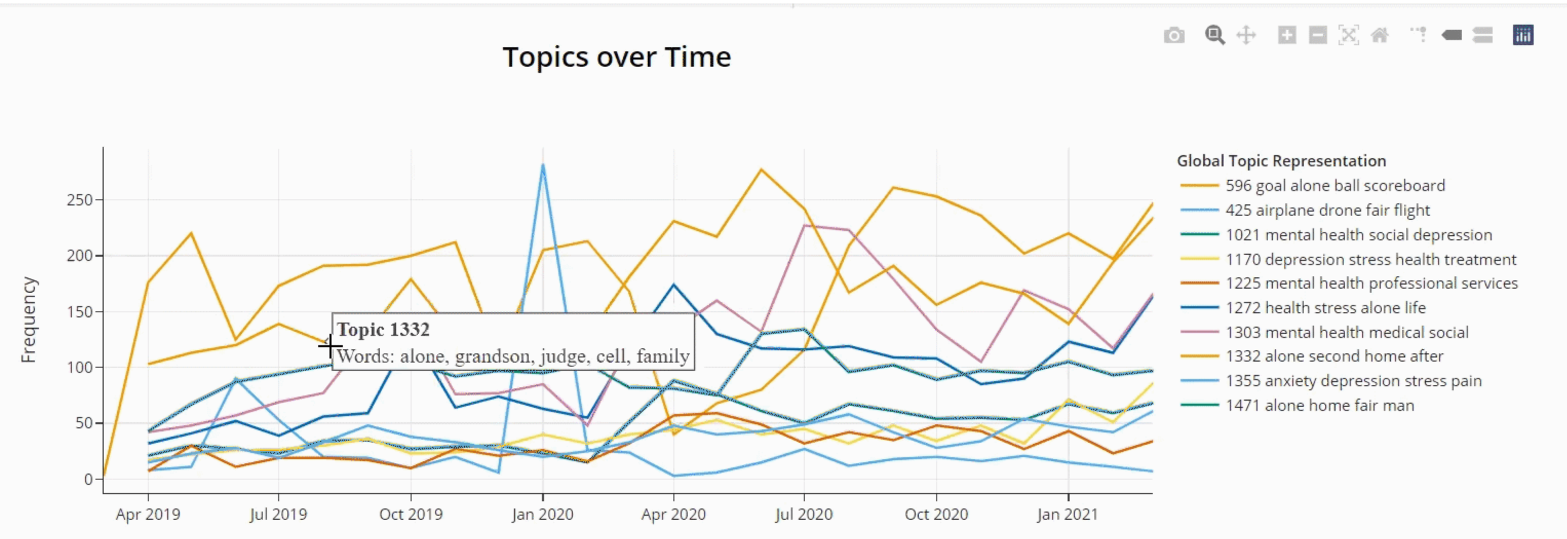
Major Representative Topics about Mental Health with Local Representations:

	2019-03-01	induced, printable, archana, handicapped, gyna...
	2019-04-01	modesty, mental, health, strongly, governed
	2019-05-01	mental, health, address, publishing, issues
	2019-06-01	health, mental, fortis, sciences, behavioural
	2019-07-01	mental, health, people, care, fortis
	2019-08-01	tighter, mass, wider, proposed, shootings
	2019-09-01	mental, health, suicide, people, awareness
	2019-10-01	mental, health, awareness, issues, important
	2019-11-01	mental, health, issues, players, cricket
	2019-12-01	health, mental, humbled, dedicate, pageants
	2020-01-01	health, mental, surrounding, fortis, physical
	2020-02-01	mental, health, care, people, physical
268	**2020-03-01**	health, mental, fortis, important, parikh
	2020-04-01	health, mental, care, support, counselling
	2020-05-01	mental, health, support, care, people
	2020-06-01	mental, health, important, issues, people
	2020-07-01	mental, health, resources, physical, issues
	2020-08-01	mental, health, people, important, workplaces
	2020-09-01	mental, health, journey, awareness, issues
	2020-10-01	health, mental, care, issues, awareness
	2020-11-01	mental, health, issues, people, important
	2020-12-01	health, mental, current, issues, people
	2021-01-01	health, mental, employees, important, needs
	2021-02-01	mental, health, support, wellbeing, people
	2021-03-01	mental, health, betterup, harry, royal

	2019-03-01	carries, involves, resilience, expert, wellness
	2019-04-01	health, mental, care, services, problems
	2019-05-01	health, mental, issues, turner, team
	2019-06-01	health, mental, institute, university, issues
	2019-07-01	health, mental, services, care, institute
	2019-08-01	health, mental, issues, services, physical
	2019-09-01	health, mental, services, institute, issues
	2019-10-01	mental, health, illness, awareness, youth
	2019-11-01	health, mental, research, associated, evidence
	2019-12-01	health, mental, issues, sources, confirmed
	2020-01-01	mental, health, university, researchers, general
	2020-02-01	health, mental, institute, issues, neurosciences
552	**2020-03-01**	coronavirus, health, mental, quarantine, people
	2020-04-01	health, mental, pandemic, covid, general
	2020-05-01	health, mental, covid, pandemic, crisis
	2020-06-01	health, mental, covid, issues, pandemic
	2020-07-01	health, mental, issues, pandemic, covid
	2020-08-01	health, mental, sushant, covid, pandemic
	2020-09-01	sushant, health, singh, lawyer, chakraborty
	2020-10-01	health, mental, countries, pandemic, covid
	2020-11-01	mental, health, covid, pandemic, guidelines
	2020-12-01	outpatient, centers, institute, centre, virology
	2021-01-01	institute, genomics, biology, centre, cellular
	2021-02-01	health, mental, pandemic, issues, covid
	2021-03-01	health, mental, pandemic, issues, covid

	2019-04-01	anxiety, associated, likely, having, disorder
	2019-05-01	anxiety, stress, symptoms, bacteria, known
	2019-06-01	anxiety, stress, depression, self, manage
	2019-07-01	anti, anxiety, effects, benzodiazepines, menop...
	2019-08-01	anxiety, sleep, stress, benefits, exercise
	2019-09-01	anxiety, stress, regular, antidepressant, pati...
	2019-10-01	people, anxiety, stress, loneliness, imaging
	2019-11-01	anxiety, stress, frequent, evidence, generalised
	2019-12-01	anxiety, stress, cause, disorder, thoughts
	2020-01-01	anxiety, thoughts, stress, depression, feeling
	2020-02-01	anxiety, sleep, stress, avoid, deprivation
	2020-03-01	anxiety, stress, fear, lead, scares
571	**2020-04-01**	anxiety, stress, fear, people, lockdown
	2020-05-01	anxiety, stress, people, individuals, uncertainty
	2020-06-01	anxiety, covid, stress, disrupted, increased
	2020-07-01	anxiety, stress, people, adolescents, issues
	2020-08-01	anxiety, stress, disorders, increased, people
	2020-09-01	anxiety, issues, psychiatric, substance, compa...
	2020-10-01	anxiety, stress, symptoms, withdrawal, related
	2020-11-01	anxiety, stress, children, genesight, stressful
	2020-12-01	anxiety, stress, depression, increased, disorder
	2021-01-01	anxiety, stress, symptoms, overworking, reduce
	2021-02-01	anxiety, stress, sleep, self, insomnia
	2021-03-01	anxiety, stress, burnout, sleep, reduce

Model trained on Brazil News Dataset



Major Representative Topics about Mental Health with Local Representations:

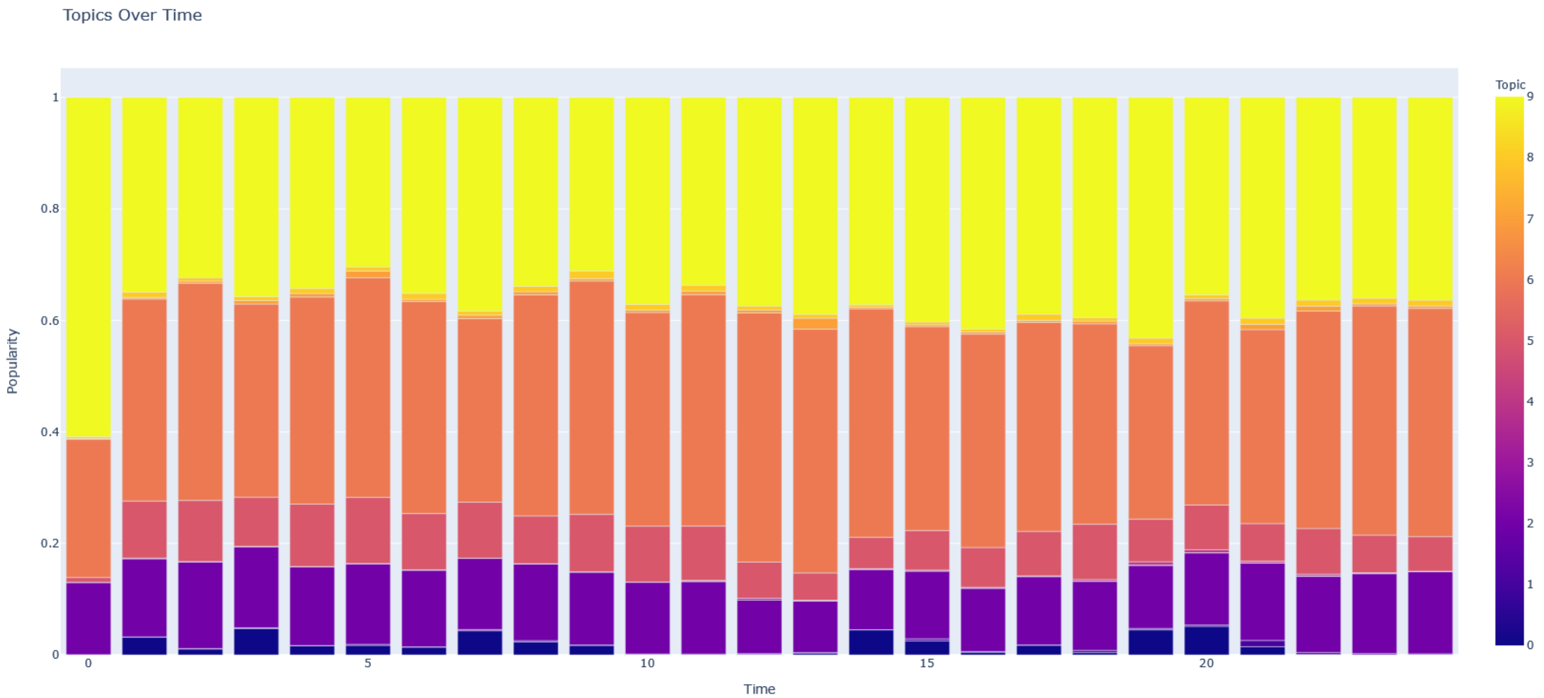
1303	2019-04-01	health, mental, center, depression, fair
	2019-05-01	health, medical, mental, clinic, education
	2019-06-01	health, mental, medical, technical, queen
	2019-07-01	health, mental, university, social, depression
	2019-08-01	health, mental, medical, technical, life
	2019-09-01	health, mental, medical, yellow, depression
	2019-10-01	health, mental, priority, motorcycle, network
	2019-11-01	mental, medical, technical, administrative, nurse
	2019-12-01	medical, health, genetics, mental, technical
	2020-01-01	health, mental, medical, white, technical
	2020-02-01	medical, health, mental, technical, woman
	2020-03-01	health, mental, social, economy, anxiety
	2020-04-01	health, mental, social, service, art
	2020-05-01	health, mental, social, stress, service
	2020-06-01	health, mental, social, confinement, promotion
	2020-07-01	store, mental, telegram, youtube, facebook
	2020-08-01	health, mental, store, telegram, channel
	2020-09-01	health, mental, pandemic, social, yellow
	2020-10-01	health, mental, life, network, depression
	2020-11-01	health, mental, social, physical, work
	2020-12-01	health, mental, ministry, politics, pandemic
	2021-01-01	health, mental, alcohol, white, years
	2021-02-01	health, mental, swan, professionals, pandemic
	2021-03-01	health, mental, professional, public, magic

1170	2019-04-01	prevention, work, vascular, depression, treatment
	2019-05-01	dementia, risk, depression, diabetes, life
	2019-06-01	stress, depression, treatment, anxiety, disorder
	2019-07-01	anxiety, disorder, mood, health, insomnia
	2019-08-01	depression, anxiety, coffee, health, illness
	2019-09-01	depression, prevention, low, behaviors, health
	2019-10-01	patience, anxiety, insomnia, depression, infla...
	2019-11-01	depression, risk, tea, mushroom, brain
	2019-12-01	anxiety, problems, stress, depression, treatment
	2020-01-01	depression, treatment, disorders, hyperactivit...
	2020-02-01	depression, suffering, treatment, anxiety, alc...
	2020-03-01	consumption, health, depression, isolation, so...
	2020-04-01	anxiety, bed, stress, sleep, health
	2020-05-01	stress, sleeping, food, percentages, sedentary
	2020-06-01	depression, stress, potential, study, anxiety
	2020-07-01	anxiety, depression, stress, breakdown, health
	2020-08-01	mask, stress, male, anxiety, soap
	2020-09-01	depression, anxiety, alcohol, illness, health
	2020-10-01	alcohol, anxiety, stress, simple, depression
	2020-11-01	depression, anxiety, illness, cases, stress
	2020-12-01	coffee, depression, anxiety, consumption, truth
	2021-01-01	depression, parents, child, infant, physical
	2021-02-01	stress, anxiety, depression, health, treatment
	2021-03-01	anxiety, depression, stress, candle, health

1225	2019-04-01	psychology, live, bank, dam, algar
	2019-05-01	health, mental, attention, assistance, rights
	2019-06-01	health, mental, psychosocial, hospital, patients
	2019-07-01	health, assistance, assistance, reception, psy...
	2019-08-01	health, mental, disorders, depression, social
	2019-09-01	professionals, safety, prevention, health, pat...
	2019-10-01	health, mental, filters, men, miscellaneous
	2019-11-01	rehabilitation, health, islamic, mental, chris...
	2019-12-01	depression, rodents, patients, health, aids
	2020-01-01	health, mental, self, facebook, imperative
	2020-02-01	memory, health, cognitive, mental, cognitive
	2020-03-01	health, doctors, mental, nurses, precarious
	2020-04-01	health, professionals, groups, clinical, social
	2020-05-01	health, mental, social, professional, countries
	2020-06-01	health, mental, professional, anxiety, serious
	2020-07-01	health, mental, beds, services, pandemic
	2020-08-01	health, mental, public, educational, social
	2020-09-01	health, mental, family, social, beds
	2020-10-01	health, mental, services, countries, problems
	2020-11-01	health, mental, schools, children, education
	2020-12-01	health, mental, services, interactive, local
	2021-01-01	mental, health, teenagers, mental, children
	2021-02-01	health, medical, social, mental, promotion
	2021-03-01	health, mental, anxiety, family, parents

Results for D-ETM:

Model trained on Indian News Dataset



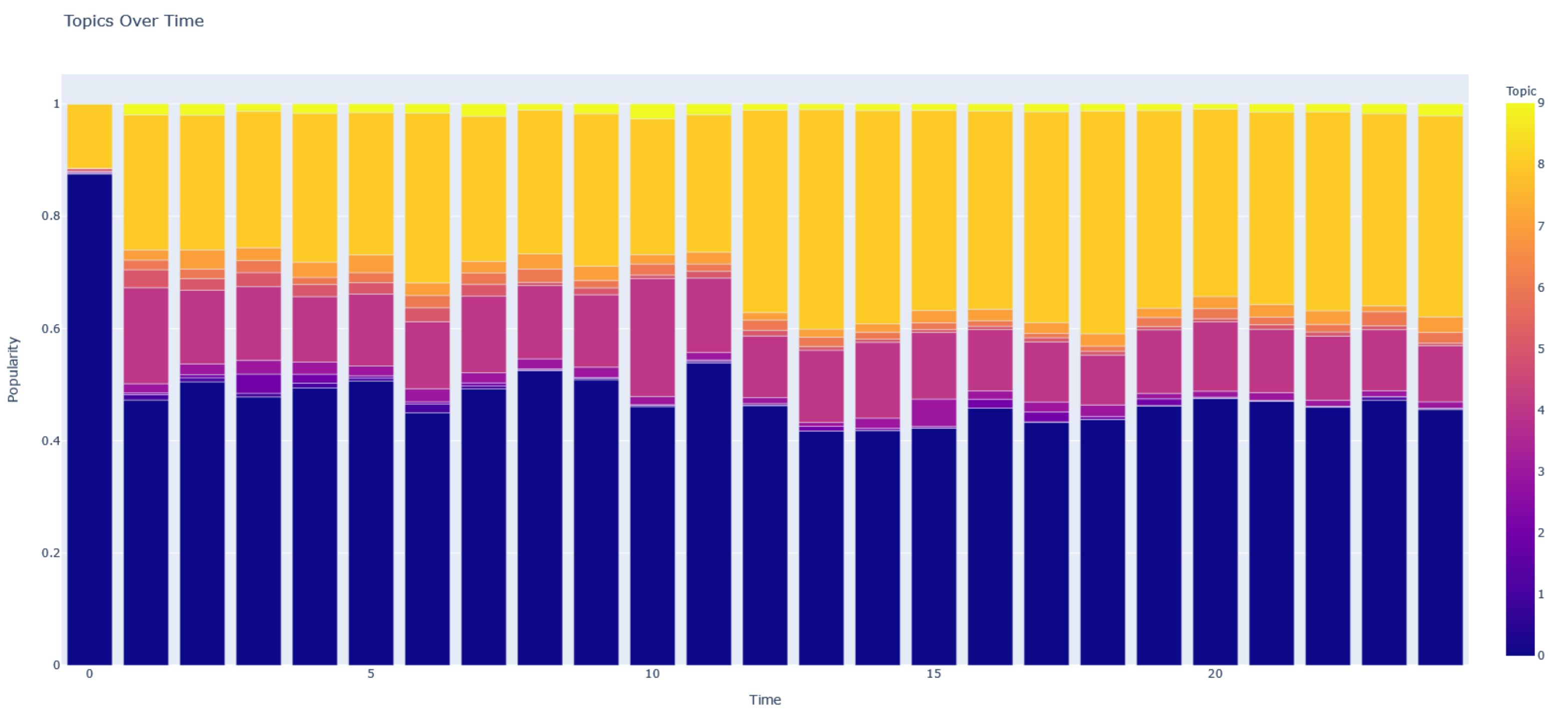
Major Representative Topics about Mental Health with Local Representations:

	2019-03	'ora', 'mendes', 'bieber', 'benny', 'demi', 'bea
	2019-04	'ora', 'mendes', 'bieber', 'benny', 'demi', 'bea
	2019-05	'ora', 'mendes', 'bieber', 'demi', 'benny', 'bea
	2019-06	'mendes', 'ora', 'bieber', 'demi', 'benny', 'bea
	2019-07	'mendes', 'ora', 'bieber', 'demi', 'benny', 'bea
	2019-08	mendes', 'bieber', 'ora', 'demi', 'benny', 'beat
	2019-09	'mendes', 'bieber', 'ora', 'demi', 'benny', 'bea
	2019-10	'mendes', 'bieber', 'ora', 'demi', 'benny', 'bea
	2019-11	'mendes', 'bieber', 'ora', 'demi', 'benny', 'bea
	2019-12	'mendes', 'bieber', 'demi', 'ora', 'benny', 'bea
	2020-01	'mendes', 'bieber', 'demi', 'ora', 'benny', 'bea
	2020-02	'mendes', 'bieber', 'demi', 'ora', 'beatles', 'be
Topic 3	2020-03	'ora', 'mendes', 'prince', 'harry', 'rita', 'benny
	2020-04	'padukone', 'tedros', 'deepika', 'ghebreyesus'
	2020-05	'prince', 'dravid', 'william', 'harry', 'oprah', 'r
	2020-06	'padukone', 'deepika', 'prince', 'william', 'fou
	2020-07	'foundation', 'bachchan', 'padukone', 'amitak
	2020-08	'director', 'institute', 'national', 'psychiatrist',
	2020-09	'national', 'director', 'institute', 'department'
	2020-10	'cent', 'national', 'founder', 'foundation', 'dir
	2020-11	cent', 'percent', 'founder', 'anytime', 'conven
	2020-12	cent', 'percent', 'mm', 'progression', 'cpi', 'inc
	2021-01	cent', 'percent', '™', 'complimentary', '2025',
	2021-02	cent', 'percent', 'items', 'respondent', 'labs', '
	2021-03	cent', 'kohli', 'percent', 'pinnacle', 'virat', 'tes

	2019-03	olympic, tokyo, games, phelps, hong kong, olympics
	2019-04	olympic, tokyo, games, phelps, hong kong, olympics
	2019-05	olympic, tokyo, games, phelps, hong kong, olympics
	2019-06	olympic, tokyo, games, phelps, hong kong, olympics
	2019-07	olympic, tokyo, games, phelps, hong kong, olympics
	2019-08	olympic, tokyo, games, phelps, hong kong, olympics
	2019-09	olympic, tokyo, games, phelps, hong kong, cancel
	2019-10	olympic, tokyo, games, phelps, hong kong, cancel
	2019-11	olympic, tokyo, games, phelps, cancel, 2020
	2019-12	olympic, tokyo, games, phelps, cancel, 2020
	2020-01	olympic, tokyo, games, phelps, cancel, 2020
	2020-02	olympic, tokyo, games, phelps, cancel, 2020
Topic 7	2020-03	tokyo, cancel, olympic, 2020, hong kong, facebook
	2020-04	hug, israeli, shetty, snack, lesson, lockdown
	2020-05	energetics, business, economy, economic, global, notwithstanding
	2020-06	energetics, business, country, security, global, nation
	2020-07	energetics, business, global, new, limited, country
	2020-08	energetics, study, datum, research, report, report
	2020-09	energetics, study, research, datum, report, survey
	2020-10	study, research, energetics, survey, accord, state
	2020-11	study, research, accord, researcher, university, energetics
	2020-12	institute, research, study, national, accord, laboratory
	2021-01	institute, research, study, national, accord, researcher
	2021-02	research, study, accord, institute, researcher, national
	2021-03	accord, research, study, institute, base, department

	2019-03	anxiety, health, mental, people, child
	2019-04	anxiety, health, mental, people, child
	2019-05	anxiety, health, mental, people, child
	2019-06	anxiety, health, mental, people, child
	2019-07	anxiety, health, mental, people, child
	2019-08	anxiety, health, mental, people, child
	2019-09	anxiety, health, mental, people, child
	2019-10	anxiety, health, mental, people, child
	2019-11	anxiety, health, mental, people, child
	2019-12	anxiety, health, mental, people, child
	2020-01	anxiety, health, mental, people, child
	2020-02	health, anxiety, mental, people, child
Topic 9	2020-03	anxiety, health, mental, people, coronavirus
	2020-04	health, mental, anxiety, lockdown, people, covid-19
	2020-05	health, mental, people, lockdown, covid-19, pandemic
	2020-06	mental, health, people, issue, covid-19, lockdown
	2020-07	mental, health, people, issue, covid-19, child
	2020-08	mental, health, people, issue, pandemic, covid-19
	2020-09	mental, health, issue, people, pandemic, covid-19
	2020-10	mental, health, issue, people, pandemic, covid-19
	2020-11	mental, health, issue, people, covid-19, pandemic
	2020-12	mental, health, issue, people, covid-19, energetics
	2021-01	mental, health, issue, people, covid-19, energetics
	2021-02	mental, health, people, issue, child, work
	2021-03	mental, health, issue, people, child, family

Model trained on Brazil News Dataset



Major Representative Topics about Mental Health with Local Representations:

	2019-03	country', 'year', 'govern', 'day', 'week', 'audience'
	2019-04	country', 'year', 'govern', 'day', 'week', 'audience'
	2019-05	country', 'govern', 'day', 'week', 'audience', 'brazil'
	2019-06	country', 'govern', 'day', 'audience', 'week', 'brazil'
	2019-07	country', 'govern', 'audience', 'week', 'brazil', 'ministry'
	2019-08	country', 'govern', 'week', 'audience', 'brazil', 'usa'
	2019-09	country', 'govern', 'audience', 'usa', ' federal'
	2019-10	govern', 'police', 'country', 'coup', 'corporation', 'detain'
	2019-11	police', 'hurt', 'fall', 'country', 'man', 'fire'
	2019-12	police', 'hurt', 'dead', 'report', 'fall', 'victim'
	2020-01	country', 'authority', 'inform', 'japan', 'police', 'rule'
	2020-02	police', 'dead', 'inform', 'victim', 'man', 'accident'
Topic 4	2020-03	'police', 'inform', 'victim', 'accident', 'military', 'dead'
	2020-04	dead', 'wake up', 'inform', 'city', 'register', 'victim'
	2020-05	wake up', 'inform', 'city', 'dead', 'morning', 'register'
	2020-06	'inform', 'dead', 'wake up', 'Wednesday', 'morning', 'disclose'
	2020-07	country', 'usa', 'economy', 'global', 'week', 'world'
	2020-08	usa', 'country', 'american', 'economy', 'global', 'governing'
	2020-09	country', 'usa', 'brazil', 'global', 'world', 'week'
	2020-10	country', 'usa', 'global', 'week', 'election', 'brazil'
	2020-11	usa', 'country', 'govern', 'american', 'election', 'economy'
	2020-12	country', 'ministry', 'governing', 'population', 'health'
	2021-01	country', 'usa', 'national', 'population', 'January', 'govern'
	2021-02	country', 'january', 'govern', 'usa', 'economy', 'brazil'
	2021-03	country', 'hospital', 'brazil', 'pandemic', 'population', 'school'

	2019-03	'depression', 'anxiety', 'mental', 'greeting', 'problem'
	2019-04	'depression', 'anxiety', 'mental', 'greeting', 'problem'
	2019-05	'depression', 'anxiety', 'mental', 'greeting', 'helping'
	2019-06	depression', 'anxiety', 'mental', 'problem', 'greeting'
	2019-07	'depression', 'anxiety', 'mental', 'greeting', 'helping'
	2019-08	'depression', 'anxiety', 'problem', 'mental', 'greeting'
	2019-09	depression', 'anxiety', 'problem', 'mental', 'helping'
	2019-10	'depression', 'anxiety', 'problem', 'helping', 'looking for'
	2019-11	anxiety', 'depression', 'problem', 'helping', 'looking for'
	2019-12	anxiety', 'depression', 'problem', 'helping', 'having'
	2020-01	anxiety', 'depression', 'problem', 'greeting', 'mental'
	2020-02	anxiety', 'depression', 'problem', 'greeting', 'helping'
Topic 8	2020-03	anxiety', 'depression', 'greet', 'mental', 'quarantine'
	2020-04	greet', 'mental', 'depression', 'isolation', 'pandemic'
	2020-05	greet', 'mental', 'depression', 'pandemic', 'isolation'
	2020-06	greet', 'mental', 'depression', 'pandemic', 'isolation'
	2020-07	greet', 'mental', 'pandemic', 'depression', 'isolation'
	2020-08	salute', 'mental', 'pandemic', 'social', 'psychological'
	2020-09	mental', 'greet', 'pandemic', 'social', 'psychological'
	2020-10	greeting', 'mental', 'pandemic', 'social', 'working'
	2020-11	greet', 'mental', 'pandemic', 'work', 'social', 'care'
	2020-12	greet', 'mental', 'pandemic', 'work', 'social', 'care'
	2021-01	greet', 'mental', 'pandemic', 'social', 'work', 'care'
	2021-02	mental', 'anxiety', 'greet', 'depression', 'problem'
	2021-03	mental', 'greet', 'depression', 'anxiety', 'problem'

5. What decisions will be made as a result of your work?

From the results above, we can visualize the main underlying themes of mental health issues like anxiety, stress, and depression that the Indian public has been going through before and after the Covid-19 pandemic. The biggest indicator being topic **552**, which peaked in the month of April only a month after the Covid pandemic hit the US and once again rose in September when Covid cases increased in India. Other topics **(9, 808)** discuss triggers for stress and anxiety like monsoons, exams, and lockdown loneliness. In Brazil, topics **(1303, 1272)** peak in April and July, 2020 indicating the rise in discussion surrounding illnesses like depression and receiving medical help for mental health struggles.

6. What work is left to be done?

It would be beneficial to continue tracking the rise in mental health issues and awareness as well as other topics that might be prevalent in countries that are suffering the most from the Covid-19 pandemic (up until 2023-2024). The pandemic has shifted society in several ways, and mapping mental health issues before and after the end of the pandemic would provide a greater understanding of what services to provide to people who are still struggling.

7. Will this work be relevant in short/medium/long term?

This work will be relevant in the short term as it provides a correlation between the Covid-19 pandemic and a rise in discussion surrounding mental health issues like depression and anxiety as well as ways to solve them. In the long term, these topic evolution models can help visualize whether mental health remains a much more relevant topic years after the pandemic in countries like India and Brazil.

Links:

BERTopic on Indian News Data:
https://github.com/zain711/dcipher/blob/main/IndianNews.ipynb

BERTopic on Brazil News Data:
https://github.com/zain711/dcipher/blob/main/BrazilNews.ipynb

Pre-trained Fasttext embedding models:
https://storage.googleapis.com/dcipher-staging-trained-models/public/FastText/english.bin
https://storage.googleapis.com/dcipher-staging-trained-models/public/FastText/portuguese.bin