Pandemic Data Visualization, Spread Simulation and Sentiment analysis

Bettina (19BCE1482), Deepak Balaji (19BCE1816), Makesh Srinivasan (19BCE1717) and Narendra G (19BCE1082)

ABSTRACT

The COVID pandemic has forced virtually to consider the way we all operate and re-evaluate every conception and misconceptions about their own everyday lives. Simple things like walking on the street and meeting people seem very daunting now. This, with no doubt, changed people, their opinions about healthcare and public services. NLP can help understand, evaluate and put some numbers on this seemingly vague topic of “Opinion Gauging”. Data that is not utilised is no good to anyone, and not everyone has the tools or the necessary knowledge to analyse and comprehend data from rows and columns; that is where data science and machine learning tools helps us understand the world through the tools of data visualisation and graphs. The idea is to create a platform where people from all backgrounds can come and utilise the tools we provide on our website, to understand and prepare for the situation we are in today.

INTRODUCTION

The introduction of social media helped, no doubt connect people. The distance factor never mattered since at the tap of a button, a message goes around the world. However, social media has also helped people be more vocal about their thoughts, opinions and feelings. This makes it ideal for people to be profiled easily: track what they do and tell how they feel. This project will focus more on the latter part.

Surveys and actual physical research and a lot of footwork was needed earlier to gather people’s thoughts on some topics. But to identify topics of most interest took a lot more footwork itself. Study groups had to be formed, questions needed to be framed in a way that makes it easier for all participants to follow and answer. This cost a lot too. Also, these were necessary, because it is based on this feedback that companies or governments take necessary steps to amend problems or fulfil needs.

But, with the advent of the internet, the workload has reduced at least in a way. People seem to be very vocal about their

passions, disinterests, views and opinions. This data is made available to the public. Thus, analyzing people's feelings or views on topics has gotten easier. Twitter has a huge database of what is essentially peoples' thoughts.

LITERATURE SURVEY

There have been numerous implementations of sentiment analysis, especially using Machine Learning. A paper by Adil Rajput[1] looked into Natural Language Processing in sentiment analysis and clinical analytics. This was done with clinical intentions. Researchers and practitioners need access to data that reflects the true mentality of a subject towards certain things. One of the ways to go about this is to get data from social groups and families of the subjects. This however is second hand information and also requires a lot of physical effort. Another alternative is "Opinion Mining", a sentiment analysis domain. This helps researchers collect data needed as mentioned before. Then, NLP methods applied on this would help summarize the mentality of a subject.

As mentioned before, data is available widely online now. It is only a matter of analyzing them effectively. Twitter provides public developers access to the insides of peoples' minds, or their "tweets" to be precise. Though they only make an insignificant amount of their data available, they still result in millions of tweets. They can be virtually about anything.

Coming to analysis, Machine Learning is a popular choice. KNN classification, RNN (Recurrent Neural Networks) and SVMs (Support Vector Machines) are widely used in sentiment analysis. This requires text to be processed and Natural Language Processing before being fed into the model. Natural Language processing in and of itself does do a good job of sentiment analysis.

With the use of Machine Learning, overfitting is of course a possibility. But this can be minimized by tweaking hyperparameters. There is however a caveat when it comes to training a model. A model trained on regular formal languages and documents fails when it comes to understanding text data like tweets [2]. This has to do with the structure of these languages. The use of emoticons, abbreviations and slang in tweets make it harder for analysis. Also, context is needed a lot. The positioning of a word has a lot to do with the meaning of the word itself as well as the entire sentence. Ali Shariq[4] and

colleagues have tried using RNNs models in detecting emotions using COVID related tweets. Certain pitfalls of using RNNs have been overcome by their models by using LSTMs (Long Short Term Memory).

Another method, which is implemented in this project, is using natural language processing techniques. NLP techniques work by extracting lexicons specific to the subject and consider the sentiment they contain. Each of these sentiments are presented in numerical forms or polarities. Most Natural Language Processing techniques achieve better results than Machine Learning models. In fact, the Obama administration used Natural Language Processing in order to gauge public opinion on certain topics. This is in part due to feature extraction. As said earlier, they work on subject specific lexicons or words, and depending on what they are, the overall sentiment of the sentence is deduced. Feature extraction aids this by looking for words that have sentiment attached to the subject. A paper by T Nasukawa [3] reported around ninety percent accuracy in analyzing and extracting sentiment from news articles and web pages.

Many news channels report on the pandemic at different locations and each of them are dependent on each other for information and news. However, there is more chance for fake news to be spread, intentionally or unintentionally, and a domino effect with an endless ripple of misinformation and disinformation is circulated. Considering the magnitude of the severity of the pandemic, especially in India, misinformation of data and news must be avoided thoroughly. This was the motivation behind creating our project with fully functional and extremely informative modules with data visualisation and simulation. There are plenty of websites and dashboard applications which serve the same purpose, but not all of them cater to the needs of the people and provide information specific to their location. Moreover, we are also providing sentiment analysis and a simulator that can provide a unique perspective and an insight into the way we analyse the pandemic in India. But, if the user wishes, he or she can analyze virtually any topic discussed in any part of the world.

A paper [5] written by Joao L. D. Comba, from the Universidade Federal do Rio Grande do Sul explores the prospect of the dashboard applications. Joao L. D. Comba has developed a dashboard with several graphics and visualisations tools and also incorporated information from social media

and news. A graphics simulation of the action of spreading of virus directly or indirectly over a blockade is also provided on the software. Pharmacological treatment, trends, Choropleth maps, geographical and regional maps are also integrated into the website. They also developed a search engine to also address the queries of the users but we developed a chatbot that can do the same but with more interactivity. Although the bot is not as powerful as the one in this software, it is still in beta testing and for now serves as a virtual help-desk, with minimal and simple functionalities. The interface of the website is very lucid and visually appealing, and as a result it is easier to use the software.

Mathematical models and differential equations are also used in collaboration with machine learning and deep learning to identify if the virus is present in a human being based on the x-ray and other data. [6]Another paper by Quoc-Viet Pham, Nguyen, Hunyh, Hwang and Pubudu N. Pathirana at National Research Foundation of Korea (NRF) Grant funded by the Korea Government (MSIT), explores the possibilities of the same. The researchers claim that instead of using RT-PCR and other traditional methods to identify respiratory based viruses, DL or ML models can be implemented due to its fast and effective way of classifying and learning from existing models to predict the presence of the virus. Medical image processing can also be used to clearly identify the COVID-19 virus from a CT and X-Ray scans of the patient's chest and lungs. Convolutional neural network (CNN) was used to train the model and was effective in helping the doctors determine which patients to test next. The model was able to acquire an overall accuracy of 93.3%. They also claim that this can be increased to 99+ by using more powerful models and CT scans because the x-rays are not as accurate and have low sensitivity and by using CT scans the overall accuracy can be increased. However, the latter is highly time consuming and expensive when compared to a X-Ray scan.

The source of data is also reputable and extremely reliable; in addition to the dataset, web-scraping is also performed on Twitter and social media. The data hub hosted by Tableau and the top 100 R-resources organised by Soetewey also are incorporated into the software. The information is also cross verified by ranking them and comparing parallelly. For the sixth paper, an open source dataset of 13,975 images of 13,870 patients was used to train a CNN

model. The accuracy was around 93% when faster and less reliable methods were used (x-ray scans); however, they showed that it was possible to increase this to 99+ by using CT scans and other more sophisticated models. For a diagnostic and a predictive model, when trained with a dataset of 568 CT images and tested on 100 samples, it shows promising results. The overall opacity of the scans were used to determine how dense and clogged the person's lungs were from the scans; the more clogged it is the more the chance of having the virus. The POHO, LHOS, POO and LSS were 0.98, 0.96, 0.97 and 0.96 respectively. Another method using DL involving two parameters - percentage of infection and volume of infection and a pre-trained ResNet50V2 model was used to estimate the uncertainty of the diagnosis of a Bayesian Deep Learning classifier and also a RF (random-forest) model with 63 features.

In another paper published by Buyannemekh Munkhbat at MIT [7], they explored the possibilities of modelling the spread of the virus using Evolving Contact Network Algorithm (ECNA). It is one of the most effective modelling techniques because it is as powerful as modelling the entire population with high computational requirements but only at a fraction of the cost of such sophisticated tools. This is because the ECNA technique generates contact networks of the infected persons and their immediate contacts and the network evolves as the number of infected grows further; hence the name Evolving Contact Network Algorithm. This paper also explores the possibilities of the Agent-based network models and it takes into account the individual level complexities and relations and contacts of the individuals unlike the traditional differential-equation based computational model. Their research work shows the interaction of nodes (humans) with each other under various settings and results of the spread and extent of the spread are computed and represented graphically. They tested for various sizes and densities of the population and inferred those results. They concluded that ECNA performs better than ABNM in simulating the disease that has low-prevalence in a large population as the need to generate the full contact network is eliminated before the simulation. ECNA performs faster when a number of various iterations are simulations are required in short time-periods.

In the ECNA algorithm, the methods used were predominantly based on graphs and trees data structures. They were used to represent the population and each node

represented the human. The edges determined the network and connections between the humans. The infected and the non-infected were allowed to interact via general interactions and the evolution of the network was the result of such simulation. They used MATLAB and a module called NetworkX from Python3 to run the algorithm. They also used Java based applications and software to simulate agent-based simulation-environment and models.

As for the simulation performed by the researchers at MIT, the dataset was produced by themselves and the tests were performed on the same. This is because this is a pure mathematical concept and their objective was to develop a model, and not to implement or test it in the real world. Different iterations and simulations were run with different population densities and sizes to understand the spread of the virus in the network, and they were all randomly generated or designed by the researchers.

METHODOLOGY

The language of choice would be Python as Machine Learning and NLP techniques needed are easily accessible. Its syntactic simplicity only makes it a better candidate. TextBlob and VADER are NLP modules built on NLTK that help in tokenizing, embedding and sentiment analysis. Both these modules use NLP techniques to extract sentiment. The result in TextBlob is just a single floating point polarity ranging from -1 to 1. The polarity is directly correlated with the kind of sentiment, i.e., the more positive the polarity, the more positive is the sentiment of the statement towards the subject.

Before the text is parsed and analyzed, they are tokenized, or broken into simple words or symbols. Lexemes will also need to be generated. These can be considered to be the most basic form of the words. For instance, the word “run” can take the forms “ran” or “running”. The word “run” is the basis of these, known as a lemma. This lexical analysis (or also known as “stemming”) is done after the text is preprocessed as mentioned earlier. Third, comes syntactical analysis. As the phrase says, it enforces the meaning of a sentence. Just like programming languages, natural languages have rules to follow. They are understood better under the light of grammars like Context Free and Context Sensitive. Production rules here help form the meaning. To understand it easier: a word in one part of the sentence may mean

something different if it were in some other part of the sentence. The production rules help in interpreting what the context of the word is and thus makes sure the sentence is meaningful (that is in its grammatical structure, not the meaning conveyed by the sentence). Then comes the semantics. A sentence can convey many meanings depending on the context. For over a century, computer scientists and linguists have worked on many theories to formalize the process. In NLP methods, choice of corpus is key. It's a collection of text that belongs to a language and contains data related to it.

TextBlob and VADER have certain differences although their means to the ends are similar. VADER works better in analyzing text with abbreviations and emoticons. This is very good for tweets. Whereas TextBlob works much better for formal languages. However, in addition to polarity, Textblob also has a subjectivity scale, which is a number that illustrates how highly opinionated the statement analyzed seems to be. VADER just provides polarity scores in three values: positive, neutral and negative. It is based on these three that a final sentiment is deduced. Thus, depending on the data to be visualized, the appropriate module needs to be selected.

The Data visualization part of the Pandemic is done with the help of Folium Maps, specifically folium heatmap and folium bubble map. The CoViD-19 Data set acquired from the API particularly for India was processed with python pandas . To map the dataset in folium with the coordinates of every state's shape, a state-wise shape file was used , Once these two datasets were processed , the scale for the heat map had to be set, for this purpose the 'Confirmed Cases' in the dataset was converted to logarithmic scale. Folium then can plot a heatmap on the given shape coordinates and export it as a html file.

In order to, show the user the magnitude of a worst-case scenario of a virus based on randomized simulation with graph-based connectivity, for which the user can choose the R0 (Basic Reproduction Number) of a virus to simulate and then is allowed to make any type of graph to model either a community within a city/an interconnected city graph , the user can then choose to assign weights to each edge of the graph/ each node. The weight on each edge from a given source node determines the probability of virus transmission from that source node, whereas the weight on the node itself determines the probability of death on that node (Which may depend upon

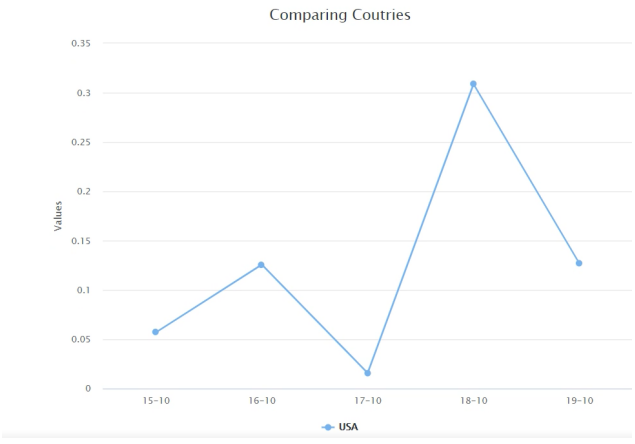
pre-existing health conditions). In each cycle of transmission , from a list[Infected nodes] a proportion of the neighbourhood nodes are chosen based on weighted probabilities of the edges connecting them and added to the newly infected nodes, this cycle continues till the user's choice or till every node is affected. A final graph of the affected nodes, dead nodes, recovered nodes is displayed to the user. The visualization for this simulation is done through the NetworkX library in python which can be used to depict graph networks combined with matplotlib.

DATASET

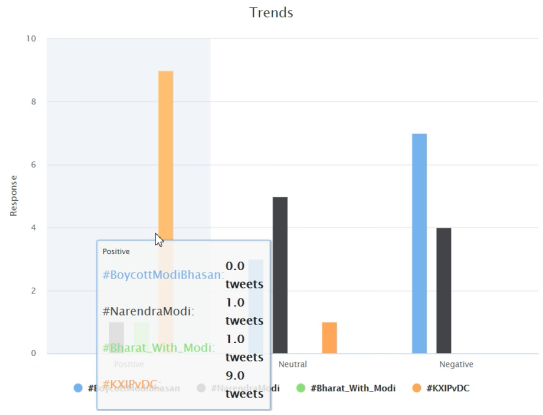
Twitter is essentially the go to candidate as it contains one of the largest databases of opinions and thoughts. The character limit forces users to be concise and to the point most of the time. Though twitter offers only a very small percentage of the data it collects, they still amount to millions. The only thing that holds one back is the rate at which requests can be made to the API and number of tweets extracted at a time. The preprocessing of the text to be analyzed is taken care of by the module chosen and sentiment on a specific topic is tallied. This helps one understand how people are responding to subjects like COVID or any other topics currently in discussion in the locality.

For data visualization modules on the dashboard and other ones on the website the source of data is from an API that is updated on a daily basis. This is procured and using Python SQLite library, a database was created using MySQL. This is the primary dataset for our software and almost all modules. Data analysis and predictive modelling are done on the server side and the output is sent to the website as an object (as HTML file) to reduce burden to the client side of the software. This enables the website to load faster and reduces lagging, thereby making it easier and convenient to use the website.

DISCUSSION AND RESULTS



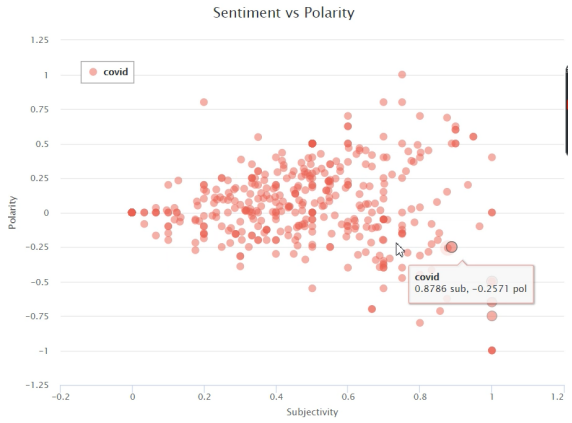
The user can choose any topic and country to analyze. The plots show various features. The first is a plot of the sentiment and feelings towards a topic by people over the last week. The Y axis shows the polarity and the X axis, the date. The more positive or higher the Y coordinate, the more positive the attitude towards the topic and the lower,



the more negative. The user can also see various other topics popular in the country or region selected and can compare them and how they are perceived by people.

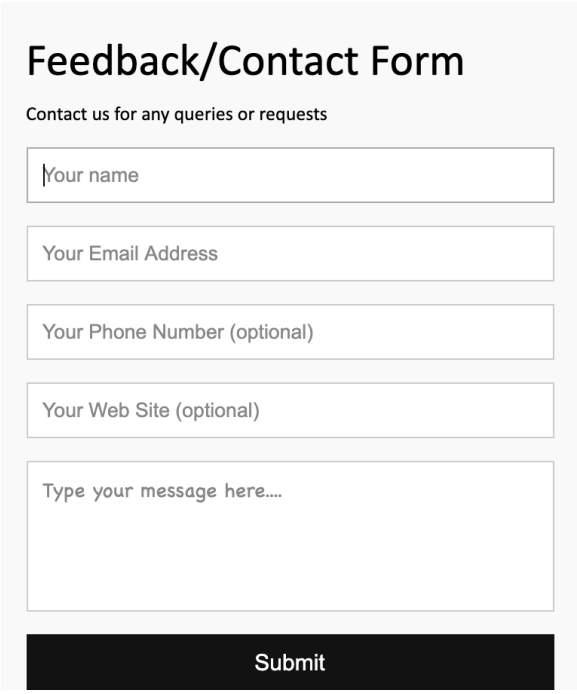
Subjectivity can also be compared with polarity with a scatterplot, on a specific topic. This shows how subjective a topic is and gives a sense of how people respond to it.

The dashboard also presents demographic maps that show the spread of the virus among the population in India. One map displays the spread as a heat map with respect to time- displays the magnitude or the severity of the spread using colours on a two dimensional map - and another map that displays the spread using circles around the most affected regions, with respect to time. They are both interactive and can demonstrate the spread automatically when the user presses the play button. They are connected to the database mentioned in the



dataset section, and therefore they are always up to date with the current situation and provide the users with the most recent information.

The trends curve can show the ways in which the virus has affected the population over time, and the user can choose which state in India to analyse. After a state is chosen, the number of current or affected patients, cured, or deceased patients are made available in this module.



The experts forum presents current news related to the pandemic from reputed and reliable sources, and also provides a place for experts to voice their opinions. There is also an option for the users to provide feedback or contact the administration via the website for any queries or requests that they might have with respect to the website or the dataset.

FUTURE ENHANCEMENTS

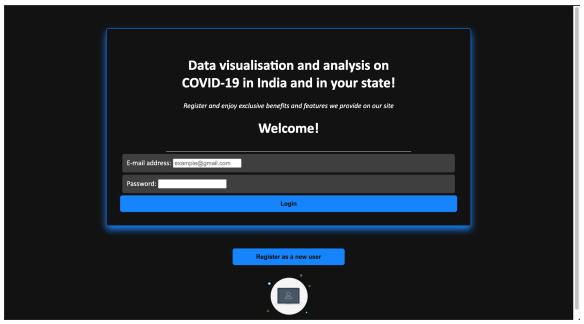
The current work relies on Natural Language Processing to gauge opinions and sentiments to topics of interest. As illustrated, various research papers have found NLP to dominate traditional ML methodologies in some aspects. However, developing an RNN or SVM and comparing final results firsthand would be better than taking the word of papers for gospel. The language used in tweets is not formal. They tend to involve slang and emoticons. Textblob does not perform too well with such features. There is also pre-processing that takes quite some time. We were unable to create a Chat-bot application on the platform due to time constraints and lack of experience. We have ideas for it but as far as implementation is concerned we were unable to fulfil that requirement (deliverable). In a software development process this is a common scenario, there will be cases when the developers are unable to create some modules, they discuss with the clients and decide on what to do. We have all the UML diagrams and everything as far as object oriented software development is concerned. *However, the assumed scenario here could be that the clients and the developers (us) had a meeting and decided that the implementation deadline can be extended.* We have plans set in place but the implementation of the application in code is taking some time. This is something that we would be interested in working on in the future.

CONCLUSION

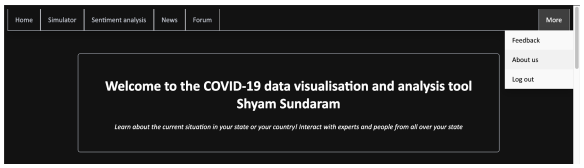
Thus, using NLP techniques, we have analyzed the sentiment of the people towards certain topics in certain places. Looking at the impact of the current pandemic is also aided by maps. We also visualised and graphed various plots to understand the situation better through data science and machine learning. Provision of a simulator also helps people test out their own epidemic with initial conditions of their choice to analyse how one could prevent such pandemic in the future.

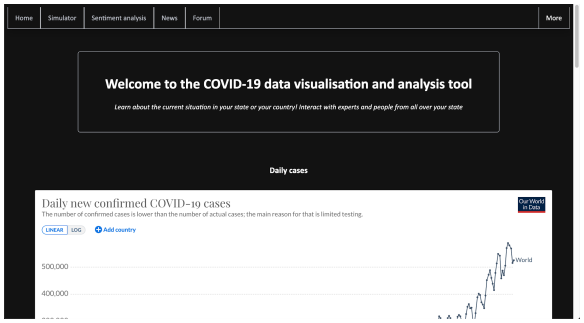
THE UI

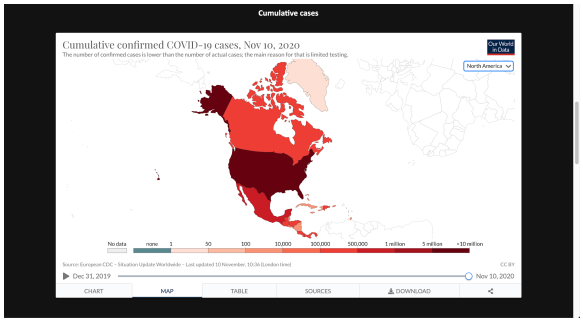
The login and registration module:



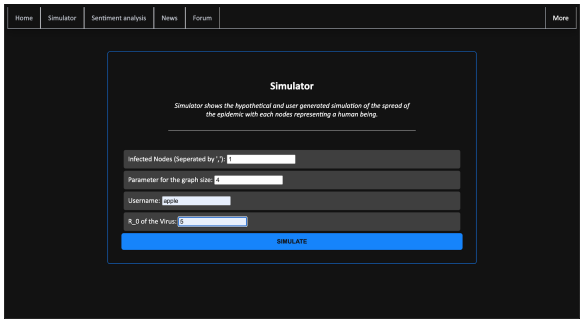
Dashboard (data visualisation):

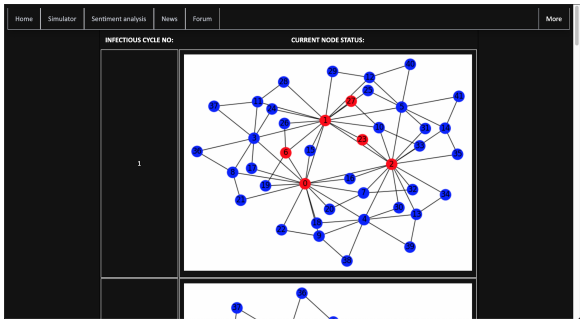




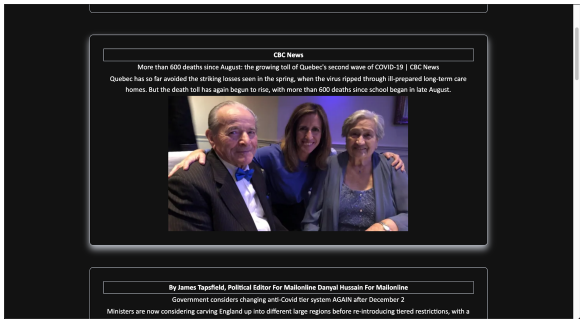


Simulator (data visualisation):

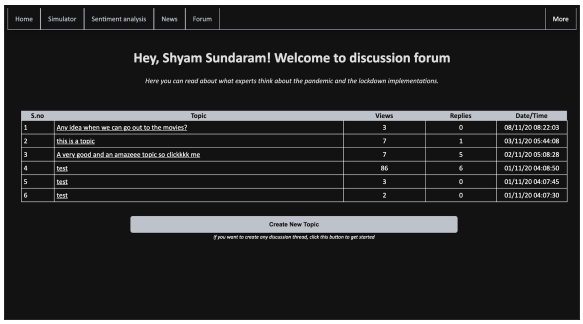




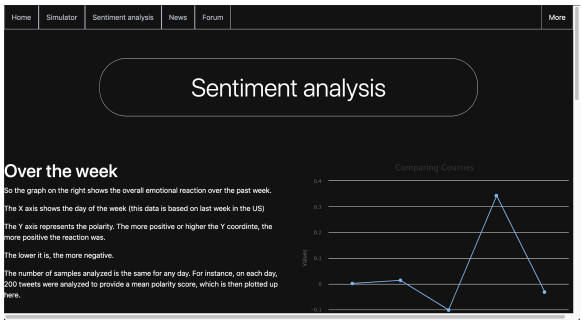
News module (Web Scraping):

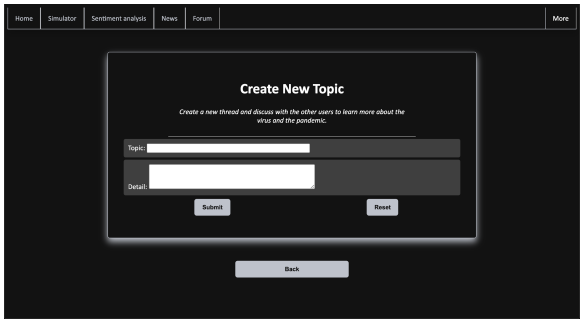


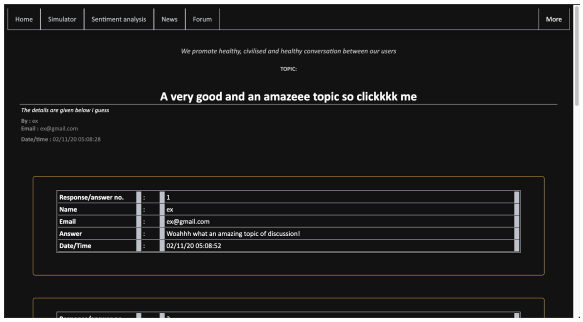
Discussion forum:



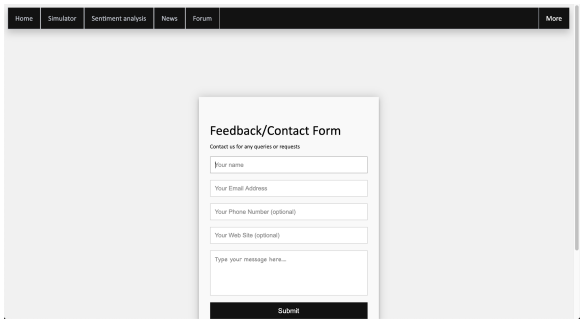
Sentiment analysis (ML and NLP):







Feedback:



REFERENCES

[1] Natural Language Processing, Sentiment Analysis and Clinical Analytics, Adil Rajput, Assistant Professor, Information System Department, Effat University, Jeddah, Saudi Arabia

[2] B. Pang, L. Lee and S. Vaithyanathan, “Thumbs up?: sentiment classification using Machine Learning techniques”, in Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol 10, pp. 79-86, Association for Computational Linguistics, July 2002.

[3] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, “Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques”, in Data Mining 2003, ICDM 2003. Third IEEE International Conference, pp. 427-434, IEEE, 2003.

[4] Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets, Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati and Rakhi Batra

[5] J. L. D. Comba, "Data Visualization for the Understanding of COVID-19," in Computing in Science & Engineering, vol.

22, no. 6, pp. 81-86, 1 Nov.-Dec. 2020, doi: 10.1109/MCSE.2020.3019834.

[6] Q. Pham, D. C. Nguyen, T. Huynh-The, W. Hwang and P. N. Pathirana, "Artificial Intelligence (AI) and Big Data for Coronavirus (COVID-19) Pandemic: A Survey on the State-of-the-Arts," in IEEE Access, vol. 8, pp. 130820-130839, 2020, doi: 10.1109/ACCESS.2020.3009328.

[7] Munkhbat, Buyannemekh, "A Computational Simulation Model for Predicting Infectious Disease Spread using the Evolving Contact Network Algorithm" (2019). Masters Theses. 790.