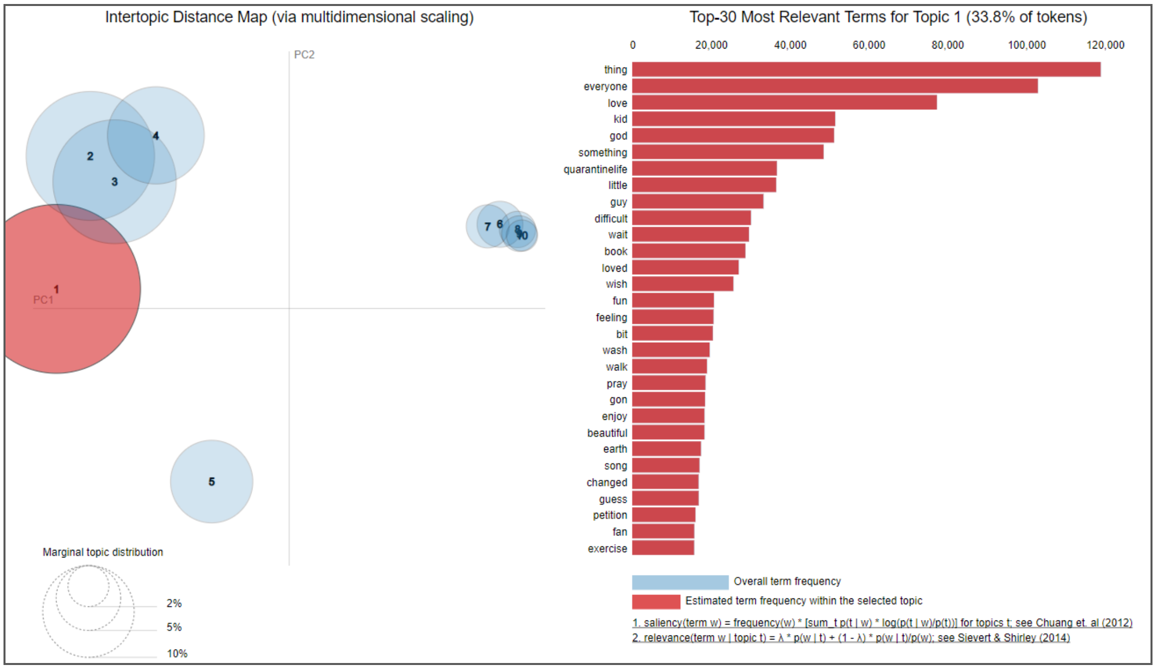
Discussing the Covid19 Pandemic

Using Unsupervised Machine Learning (Latent Dirichlet Allocation) with Natural Language Processing to Conduct Topic Modeling



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

By:

SHREYAS CHITRANSH

BSc (Honours), MSc

A final project submitted in part fulfilment for the Diploma in DataScience Bootcamp at BrainStation

27th June 2021

Introduction

The Covid19 pandemic swept across the world starting in March 2020 and was continuing throughout the time this report was generated in June 2021.This is the first time that the world not only saw a lot of fear, grief, and sadness but also some of the best solidarity amongst the global citizens. This project was conducted to find out the Covid19 pandemic related themes in depth, that were being discussed globally. The unprecedented impacts of the pandemic in terms of loss of human lives and health were evident, however they also caused a chain reaction that disrupted most of the industries, and in turn the world economy. Therefore, the focus was to illicit the subtle yet, profound effects on people's lives and livelihood. The main goal was to generate a truthful epilogue of the people's perceptions which can form the basis of 'scientific analysis and debate' with a view to learn from this unfortunate event of human history.

The project was conducted by analyzing Twitter data from the first major surge of global Covid19 cases, in March/April 2020. It conducted Topic Modeling using the Latent Dirichlet Allocation (LDA) model and generated a summary list of topics discussed by the mass English speaking Twitter users.

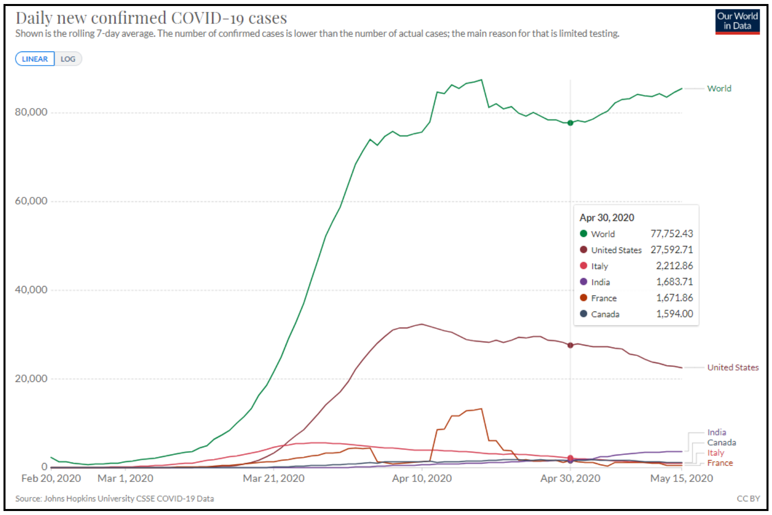


Figure 1: First Major Global Surge of Covid19 - March/April 2020

Data Collection

Hydrating Tweets from Twitter directly using the Tweet ID was attempted as a data collection method however, the twitter application of the author kept getting suspended every 2-3 days (followed by appeal and reactivation). This hassle in data collection meant that the Kaggle dataset was a better alternative. In retrospect, it is obvious that next time a simpler explanation for the application use will be given so it doesn't get caught in the Twitter Machine Learning algorithm of automatic suspension.

The Kaggle dataset provided the Twitter data as .csv files which contained Tweets from 29th March 2020 - 30th April 2020. The data consisted of all Tweets from users who used the following hashtags: `#coronavirus`, `#coronavirusoutbreak`, `#coronavirusPandemic`, `#covid19`, `#covid_19`, `#epitwitter`, `#ihavecorona`. It should be noted that from 11 April onwards, 2 additional hashtags, namely `#StayHomeStaySafe`, `#TestTraceIsolate`, were also added into the Tweet collection set. This ensured that the major 'hashtags' incorporated in discussions of Covid19 were utilized. It would have been preferable to have more data from March 2020 as only the last 3 days of the month were covered. This would have given a better representation of the ramp up during the first major surge of the pandemic.

Data Cleaning

The data was cleaned by removing duplicates and null values followed by filtering for English Tweets. In general, the data was quite clean, however it would've been preferred to have Tweets with more Covid19 related 'hashtags'. A final dataset of 81,333 English Tweets was selected which represented the timeframe chosen.

Exploration of the Tweet texts gave some great insights. Majority of the Tweets comprised of 15-35 words (100-280 characters) with a long tail where they extended to upwards of 90 words (850 characters). These outliers were found to be Tweets where the number of 'mentions' was high and/or the tweets contained shared website link(s).

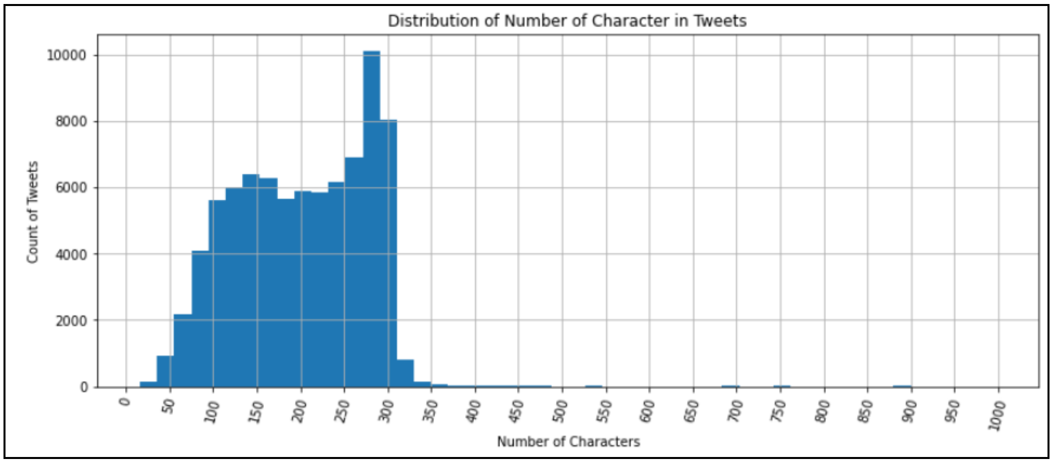


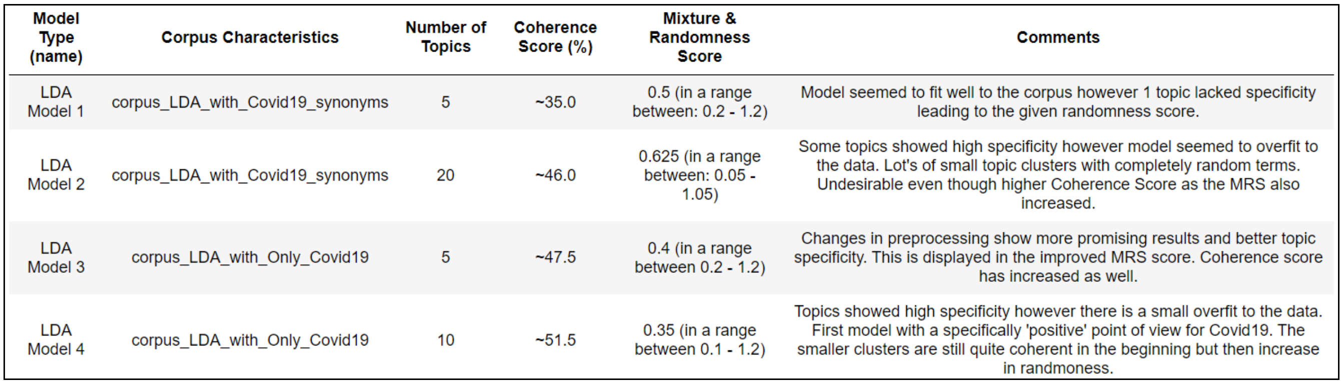
Figure 2: The Long Tail of Character Lengths for Tweets

The text data exploration concluded that removal of 'stop words', 'links', '&amp', 'dashes', 'mentions' and 'dots' is necessity for effective Topic Modeling.

Topic Modeling

In the modeling workflow the optional hyperparameters were not usable since they required a-priori knowledge about the number of topics or word probability expected. Therefore, the input data preprocessing methodology and number of topics were changed as methods to yield improved results. The model efficacy was evaluated using 2 methods. The Model Coherence Score, which evaluates the degree of semantic similarity between high scoring words in the topic, and a self devised method called Mixture and Randomness Score (MRS) which evaluates the amount of randomness in a model using the randomness of words in the top 30 terms associated with the topic.

For models 1 & 2 a corpus was generated using Tweet preprocessing where the 'stop words', 'links', '&amp', 'dashes' and 'dots' were removed, and the text tokenized as well as lemmatized. The use of PyLDAvis interactive visualization proved to be an excellent tool for analyzing and determining topics. While Model 1 was generated using 5 topics, Model 2 was generated after finding the optimum number of topics (20) using coherency as a measure. Lessons learnt from models 1 & 2 were used to change the preprocessing methodology as a means to improve models 3 & 4, wherein the corpus was generated with additional steps of converting all words to lowercase while unifying all the Covid19 synonyms. Model 3 was generated using 5 topics and Model 4 was generated after finding the optimum number of topics (being 10) that gave the highest coherence score. This was chosen while keeping in mind that a balance between coherence score and number of topics is required to maintain a low MRS. The resulting model 4 had the highest Coherence score with the lowest MRS compared to all previous models. The models and their respective evaluation parameters, as well as observations and issues can be seen in Figure 3 below. It shows the successive improvement of model performance over preceding ones.



Model Type (name)	Corpus Characteristics	Number of Topics	Coherence Score (%)	Mixture & Randomness Score	Comments
LDA Model 1	corpus_LDA_with_Covid19_synonyms	5	~35.0	0.5 (in a range between: 0.2 - 1.2)	Model seemed to fit well to the corpus however 1 topic lacked specificity leading to the given randomness score.
LDA Model 2	corpus_LDA_with_Covid19_synonyms	20	~46.0	0.625 (in a range between: 0.05 - 1.05)	Some topics showed high specificity however model seemed to overfit to the data. Lot's of small topic clusters with completely random terms. Undesirable even though higher Coherence Score as the MRS also increased.
LDA Model 3	corpus_LDA_with_Only_Covid19	5	~47.5	0.4 (in a range between 0.2 - 1.2)	Changes in preprocessing show more promising results and better topic specificity. This is displayed in the improved MRS score. Coherence score has increased as well.
LDA Model 4	corpus_LDA_with_Only_Covid19	10	~51.5	0.35 (in a range between 0.1 - 1.2)	Topics showed high specificity however there is a small overfit to the data. First model with a specifically 'positive' point of view for Covid19. The smaller clusters are still quite coherent in the beginning but then increase in randmoness.

Figure 3: Summary of the Models and their Performance

Results

The final model (Model 4) gave a better representation to most topics, and some of the smaller topics were able to identify underlying subtle discussions taking place. There was 1 completely random topic still present, which showed that even the best models cannot classify everything. In the final model, the main themes of the different topics (in order of prevalence) are given below with a reasoning:

1. **Quarantine Silver Lining** - showing how people were showing solidarity in the face of adversity with words such as ‘love’, ‘family’, ‘exercise’, ‘fun’ and ‘enjoy’;
2. **Business and Financial Impacts** – because it out to be more severe than expected;
3. **Covid19 Tracking** – this was an imperative out of the necessity to monitor the spread of Covid19;
4. **Covid19 Medical Advances** – medical response for prevention, treatment and cure;
5. **USA Republican Politics** – as a consequence of the US election with an outspoken president;
6. **Indian Politics w/ Randomness** - India being the 3rd largest English Tweeting community with a proactive Prime Minister who used fight against Covid19 as a political tool;
7. **USA Democratic and Canadian Politics** – as in 5;
8. **Random** – contained words that couldn’t fit into other topics;
9. **Negative Lockdown Aspects** – response of the suffering masses due to emotional/financial strains;
10. **American Conspiracy Theories** – as a political and infodemic fallout from vested interests.

Conclusions

This workflow has shown that one can successfully utilize Topic Modeling with Twitter data to identify the main topics of discussion. This has a lot of utility as it can summarize the 1000's of Tweets into separate, comprehensive, identifiable topics within a small timeframe. Similar approaches have been used to identify the degree of severity of Covid19 using the speed and velocity Tweet generation. They have also attempted to get a sentiment about Covid 19 or conduct network analysis for the Tweets. However, no product was found which could illuminate the subtle topics being discussed in-depth which can form the basis of ‘scientific analysis and debate’.

Examples of its use can be for government entities, influencers and promoters who get thousands of Tweets related to their policies or products. Using this workflow, they can automatically get to know a summary of the mass opinions and feedbacks being shared, to provide better services to their stakeholders. This MVP can be custom tailored for, and scaled to the specific requirements of individuals, entities or organizations depending on their requirements. Once tailored, it can result in a quicker turn around time to track the trending opinions in more detail. In the future the project aims to use a bigger dataset with data from later months leveraging cloud computing and testing out other modeling types such as LDA Mallet and preprocessing with TF-IDF instead of Bag of Words. These may show the evolution of pandemic discussion and hopefully other topics such as the environmental improvements due to decreased travelling in the pandemic.