How Personal Perceptions Of COVID-19 Have Changed Over Time

Tianyi Sun

Supervised by Prof. Maria Gini

UNIVERSITY OF MINNESOTA
Driven to Discover®

Abstract

Used emotional responses and comments to COVID-19 pandemic to analyze people's perceptions towards COVID-19. Based on the trend of perceptions, we predicted the trend for the next month.

Contribution

- estimated the trend in sentiment changes towards COVID-19; extracted five main topics from the dataset; predicted the trend of the sentiments and topics for the next 31 days;
- estimated the health condition of the active authors in Reddit; and
- gave suggestions for helping people in the pandemic.

Outline

- Abstract
- Contribution
- Why Perception
- How to get Perception
- Details
  - Dataset
  - Text Preprocessing
  - Sentiment Analysis
  - Topics Extraction
  - Sequential Prediction
- Results & Conclusion
- Open Questions

Why Perceptions?

- Variation of existing Factors and Emergence of new Factors, include natural factors and humanity factors, would influence on the accuracy of prediction of spread trend.

- Perception→ Behavior→ Humanity factors→ Rate of spread

How to get Perceptions?

- Sentiment Analysis
- Topic Extraction
- Sequential Prediction

Dataset

- The first ground truth dataset of emotional responses toward COVID-19.
  - **Only** used for training sentiment classification model
  - a survey in England, in April
  - 5000 texts (2500 short; 2500 long)
  - Labeled with 8 sentiments (all -)
  - + 5 sentiments (+, neutral) → avoid misclassification
- A time series dataset of Reddit comments toward COVID-19.
  - Used for classifying sentiments, topic extraction, and prediction
  - 409,476 texts
  - from January 1st to April 17th, 2020

Text Preprocessing

Note: the order is important

- Remove URLs
- Remove mentions, i.e. ‘@name’
- Convert to lower case
- Demojized the emogies
- Correct misspellings
- Expand contractions
- Remove punctuations

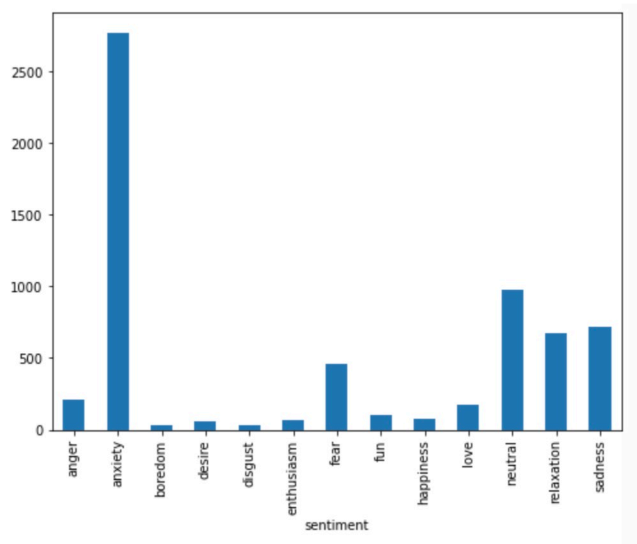
Sentiment analysis

- Task type: Multi-class classification.
- Methods:
  - Machine learning algorithms
    - Naive Bayes
    - Linear SVM
    - Logistic Regression
    - Linear SVC
    - LSTM
  - Transformers
    - BERT (Google)
    - RoBERTa (Facebook)
    - XLNet (Google)
    - DistilBERT
- Evaluation Method: Accuracy Score

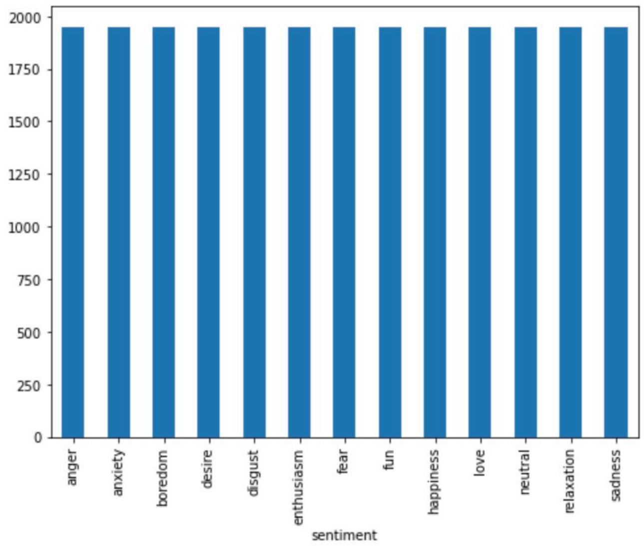
Before modeling:

– 1. The collection of preprocessed texts
– → a matrix of token counts with fixed size
– → a matrix of normalized TF-IDF terms.

- 2. Balancing the imbalanced data by oversampling.
- **Before balancing**



After balancing



Sentiment classification Modeling Evaluation Result

Train test split: 3:1

Model Name:	Accuracy Score:
Naive Bayes	0.8440
Linear SVM	0.7989
Logistic Reg	0.9528
Linear SVC	0.9218
Random Forest	0.7337
LSTM	0.9534
BERT	0.8538
RoBERTa	0.7410
XLNet	0.6547
DistilBET	0.8987

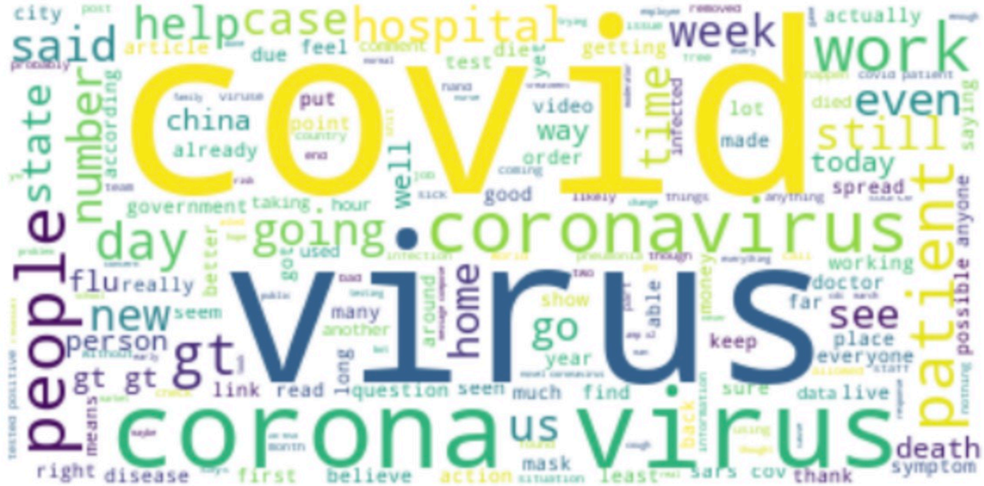
Table 1: Model evaluation for sentiment classification

Topics Extraction

- Task type: Unsupervised learning.
- Methods: Latent Semantic Indexing (LSI), Random Projections (RP), **Latent Dirichlet Allocation (LDA) (optimal)**
  - LDA is a probabilistic extension of LSI. The advantages of LDA is that it can allocate topics of any texts.

Before Modeling

- tokenized each preprocessed text to a list of words.
- lemmatized and stemmed each word into their original form.
- removed stop words,
- further removed words other than nouns, verbs, adjectives, and adverbs.
- visualized the top 50 most frequent words to remove words quiet frequent, but not useful for extracting topics, e.g. ‘coronavirus’, ‘corona’, and ‘covid’.



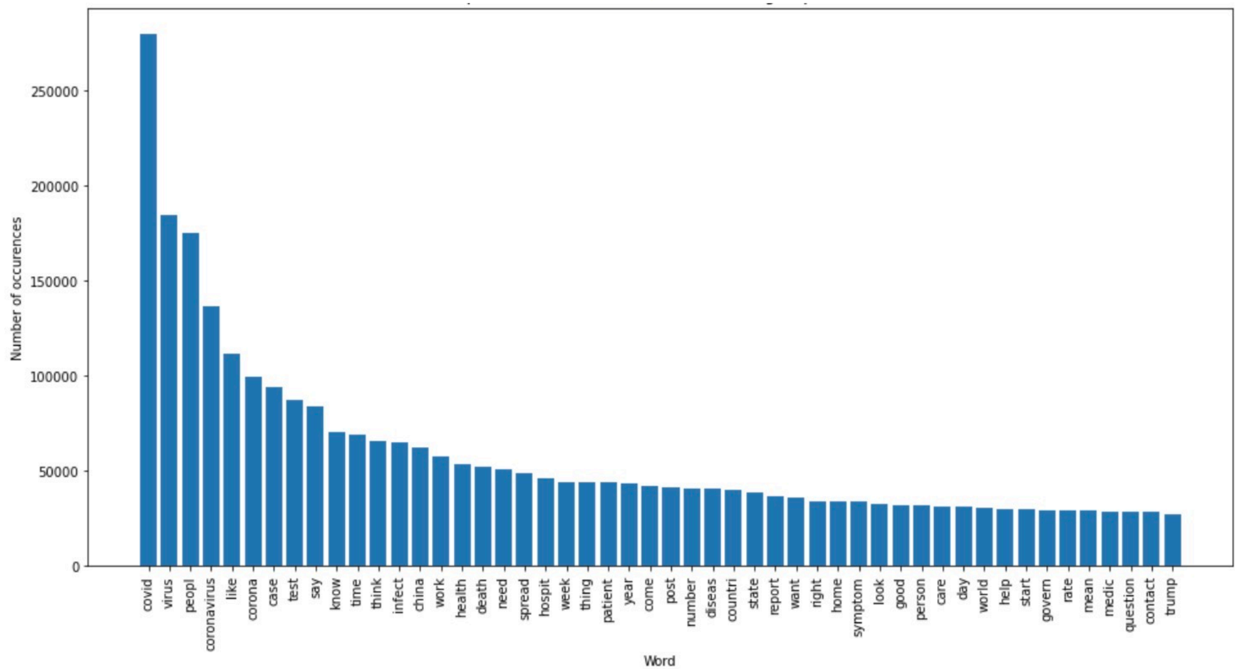


Figure 1: The top 50 most frequently appeared words in COVID-19 related Reddit Comments dataset

Topic Extraction Result

```
1: 0.008*"press_confer" + 0.007*"walk" + 0.007*"regardless" + 0.007*"recov" + 0.006*"reddit_vote_thread" + 0.006*"figur" + 0.006*"drink" + 0.006*"medicin" + 0.005*"booster" + 0.005*"port_entri" + 0.005*"account" + 0.005*"petri_dish" + 0.005*"swap" + 0.004*"week" + 0.004*"join" + 0.004*"strain" + 0.004*"overeact" + 0.004*"late" + 0.004*"pathogen" + 0.004*"advoc"

2: 0.009*"parma" + 0.008*"prensa" + 0.008*"cali" + 0.005*"pero_ahora" + 0.005*"exchang" + 0.005*"zijn" + 0.004*"farm_wild_anim" + 0.004*"blanket_term" + 0.004*"valu" + 0.004*"morgu" + 0.004*"korea_center_diseas" + 0.003*"camera" + 0.003*"constitut" + 0.003*"schedul" + 0.003*"autonom" + 0.003*"shouldn" + 0.003*"scar" + 0.003*"pharma" + 0.003*"plagu_plagu" + 0.002*"chines_govern"

3: 0.008*"sequenc_genom" + 0.006*"inflam" + 0.005*"discoveri" + 0.005*"malaysia_director_general" + 0.004*"34" + 0.004*"cover" + 0.004*"undermin" + 0.004*"world" + 0.004*"compar" + 0.004*"epidemiolog" + 0.004*"fever_cough_troubl_breath" + 0.004*"februari" + 0.004*"él" + 0.004*"18" + 0.003*"individu" + 0.003*"south_korean" + 0.003*"share" + 0.003*"reason" + 0.003*"sever_acut_respiratori_syndrom" + 0.003*"countri"

4: 0.009*"necessarili" + 0.008*"slight_shadi" + 0.008*"spread" + 0.007*"posit" + 0.007*"nasti" + 0.007*"american" + 0.007*"board" + 0.006*"kitti" + 0.005*"provid" + 0.005*"develop" + 0.005*"countri" + 0.005*"adopt" + 0.005*"genet" + 0.005*"problem" + 0.004*"definit" + 0.004*"honest" + 0.004*"mayb" + 0.004*"762" + 0.004*"figur" + 0.004*"panic"

5: 0.036*"live" + 0.028*"blood_panel" + 0.022*"high_temperatur" + 0.021*"guidanc_forthcom" + 0.020*"govern" + 0.020*"confirm" + 0.020*"overlap" + 0.019*"world_health_organ" + 0.018*"degre" + 0.018*"crucial" + 0.018*"recoveri" + 0.017*"leav" + 0.016*"merit" + 0.016*"quick_googl_search" + 0.015*"concern" + 0.015*"thread" + 0.014*"depart_homeland_secur" + 0.012*"larg" + 0.012*"googl" + 0.011*"scientif"
```

Topic Number	Topic Name	Terms
1	**Recovering Strategies**	walk, recov, figur, drink, medicin, booster, petri dish, overeact,pathogen
2	**Source of Disease**	farm wild anim, prensa, morgu, camera, autonom, scar, pharma, plagu
3	**Infected Symptoms**	sequenc genom,inflam, discoveri, undermin, epidemiolog, fever cough troubl breath, reason, sever acut repiratori
4	**Route Of Spread**	spread, posit, nasti, board, countri, genet, figur, panic
5	**Future Precaution**	live, blood panel, high temperatur, guidanc forthcom, govern, confirm, world health organ, recoveri, leav, depart homeland secur

Sequential Prediction

- Task: Time Series Prediction
- Methods:
  - Autoregressive Integrated Moving Average (ARIMA) + Grid Search optimal hyper parameter
  - Seq2seq — Encoder-Decoder LSTM
- Evaluation Method: RMSE

Before Prediction:

Convert dataset into the formate used for prediction.

Text dataset with a column of sentiment labels and a column of topic labels

$\longrightarrow$

A **numerical** dataset for Sentiments prediction:
Columns are the 13 sentiments,
rows are dates, and
values are total num of texts per day.

A **numerical** dataset for Topics prediction:
Columns are the 5 topics,
….

Sequential Prediction Modeling Evaluation Results

- LSTM with look back value (LSTMLB), LSTM with Window Method (LSTMWM), LSTM with Time Steps (LSTMTS), and LSTM with Memory Between Batches(LSTMM).
- Optimizer: adam.

Model Name	RMSE Sentiment Prediction	RMSE Topic Prediction
LSTMLB	64.56	125.73
LSTMWM	56.71	170.96
LSTMTS	67.62	159.57
LSTMM	277.72	242.68
ARIMA	27.07	85.71

Table 2: Model's evaluation results for topics trend prediction and sentiments trend prediction.

Optimal Hyper parameter of ARIMA model by Grid Search minimum RMSE for prediction of each sentiment trend and each topic trend

Topic	Hyperparameters	RMSE
anxiety	ARIMA(1, 0, 0)	89.839
relaxation	ARIMA(0, 2, 2)	52.292
sadness	ARIMA(0, 2, 2)	54.768
neutral	ARIMA(0, 1, 1)	51.042
fear	ARIMA(0, 1, 1)	56.136
anger	ARIMA(6, 0, 0)	21.300
love	ARIMA(0, 1, 2)	5.983
fun	ARIMA(4, 1, 0)	6.253
desire	ARIMA(2, 0, 0)	3.456
enthusiasm	ARIMA(0, 1, 1)	5.435
happiness	ARIMA(2, 0, 0)	3.232
disgust	ARIMA(2, 0, 0)	1.4816
boredom	ARIMA(0, 1, 1)	0.699

Table 4: Optimal Hyperparameter of ARIMA model for prediction of each sentiment evaluated by RMSE

Topic	Hyperparameters	RMSE
Topic1	ARIMA(6, 1, 0)	97.049
Topic2	ARIMA(1, 0, 1)	63.308
Topic3	ARIMA(8, 1, 1)	83.525
Topic4	ARIMA(8, 1, 0)	93.691
Topic5	ARIMA(10, 0, 0)	90.984

Table 6: Optimal Hyperparameter of ARIMA model for prediction of each topic evaluated by RMSE

Prediction of sentiments

Results:

- COVID-19 caused **attention on Jan 19th**.
- Num of sentiments **increased until Mar 1st**, after that each of the 13 sentiments became stable.
- The **descending order** of the averaged number of sentiments throughout the entire timeline: **anxiety, relaxation, sadness, neutral, fear, anger, love, desire, fun, enthusiasm, happiness, disgust, and boredom.**

Conclusion:

The increasing num of anxiety comments →
People were paying attention to COVID-19, which is good.

But the continuously increasing trend of anxiety comments is not good.→
Governments and WHO should build up confidence by providing more information useful for people to avoid from infected.

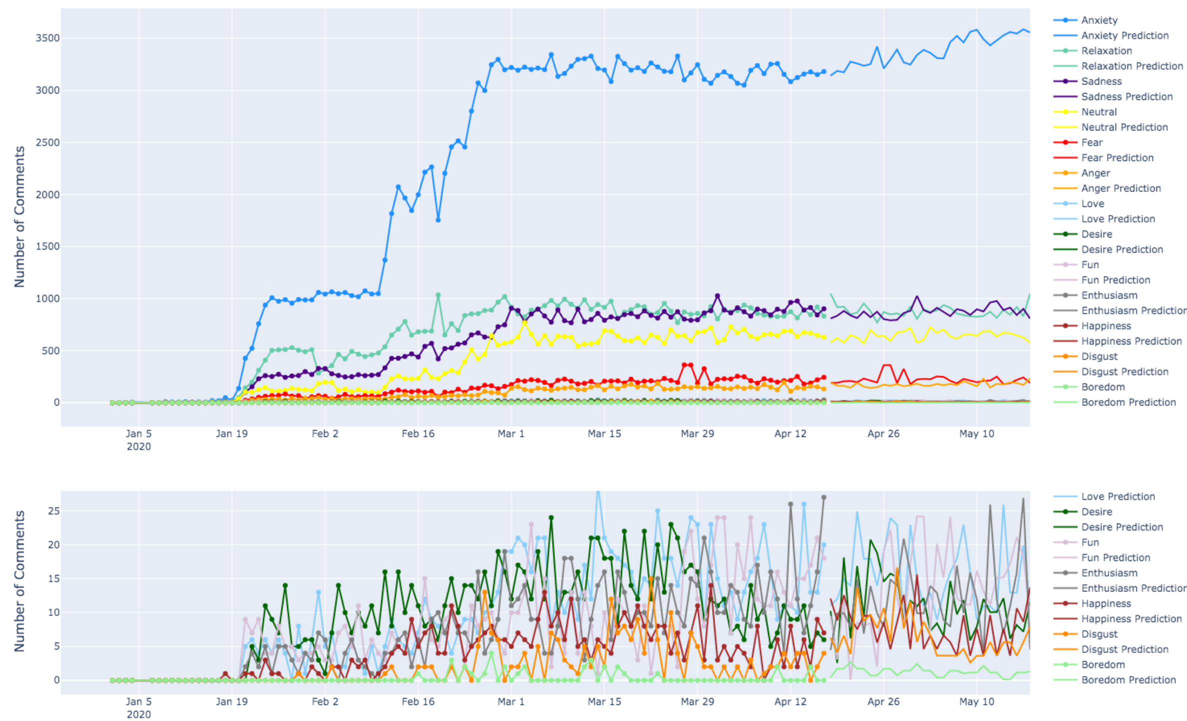


Figure 1: Sentiments multi-class classification using LSTM and predictions using ARIMA. Training and predicted data are distinguished by lines with/without daily markers

Prediction of topics

Results:

- The 5 topics in **descending order**: **Infected Symptoms, Future Precaution, Source of Disease, Route of Spreads, and Recovering Strategies.**
- The **two remarkable growth** of all the five topics were started on January 19th and February 11th, respectively.
- After March 1st, each of the five topics became stable. The prediction shows that Future Precaution has a growing trend, and other topics are fluctuating around their previous values.

Conclusion:

The great attention to Infected Symptoms and Future Precaution, and the prediction of increasing trend on Future Precaution → detecting and preventing COVID-19 are hot topics.

The low number of comments about Recovering Strategies →
most authors were not infected with COVID-19.

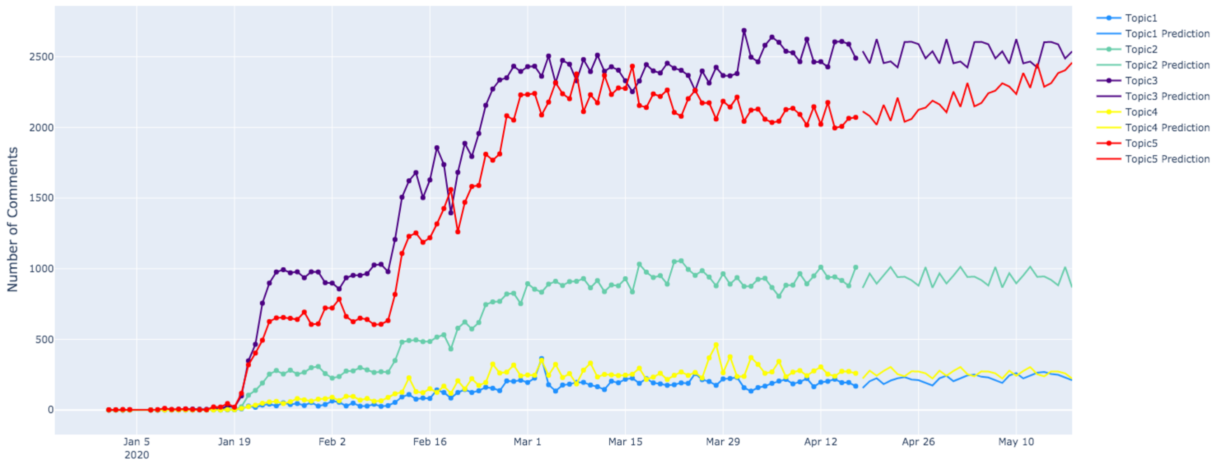


Figure 2: Topics extracted using LDA and prediction using ARIMA. Training and prediction data are distinguished by lines with/without daily markers

Future Work

- Keep the ‘author’ feature in the loop to study how an individual's sentiment and topic changed over time in order to find the correlation of topics and sentiments based on time.

Open Questions

1. Texts related to COVID-19 on social media are not convincing, since exaggerated sentiments are hard to classify. **Reliable COVID-19 text data are still limited.**
GPT-2 (trained a large-scale unsupervised language model which generates coherent paragraphs of text)

2. Improving Language Understanding. Eg, convert LDA to semi-supervised learning model.

3. We cannot **explore people's thoughts.** Facial expressions, physical and mental activities might be good indicators of thoughts.

University of Minnesota

Driven to Discover®

Crookston Duluth Morris Rochester Twin Cities

The University of Minnesota is an equal opportunity educator and employer.