How Personal Perceptions of COVID-19 Have Changed Over Time

Tianyi Sun
University of Minnesota
sun00234@umn.edu

Abstract

In this work, we used datasets on emotional responses and comments to COVID-19 to analyze people's perceptions towards COVID-19. Based on the trend of perceptions, we predicted the trend for the next month.

Introduction

Analysing people's perceptions toward COVID-19 is crucial for controlling the pandemic. In this work, we used the first ground truth dataset of emotional responses to COVID-19 (ER) (Kleinberg, van der Vegt, and Mozes 2020) and a time series dataset of COVID-19 related Reddit comments (TSRC) for perception analysis and perception trend prediction of the next 31 days. Two approaches we used for perception analysis are sentiment analysis and topic extraction.

Our contributions are: (1) we estimated the trend in sentiment changes towards COVID-19, extracted five main topics from the dataset, and predicted the trend of the sentiments and topics for the next 31 days; (2) we estimated the health condition of the active authors in Reddit; and (3) we give suggestions for helping people in the pandemic.

Background

Sentiment Analysis. Sentiment Analysis is a multi-class classification problem. Models used to solve it are Naive Bayes, Linear Support Vector Machine(Linear SVM), Logistic Regression(Log Reg), Linear Support Vector Classification(Linear SVC), Random Forest, Long Short-Term Memory(LSTM) neural networks, Bidirectional Encoder Representations from Transformers(BERT) (Devlin et al. 2018), a Robustly Optimized BERT Pretraining Approach(RoBERTa),an autoregressive pretraining method (XLNet) and a distilled version of BERT(DistilBERT).

Topic Extraction. Topic extracting is an unsupervised learning task. Popular topic extraction methods include Latent Semantic Indexing (LSI), Random Projections (RP), and Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003). LDA is a probabilistic extension of LSI. The advantages of LDA is that it can allocate topics of any texts. We used LDA for topic extraction.

Sequential Prediction. Auto-regressive integrated moving average (ARIMA) is a popular statistical method to analyze time series data and predict future trends. In addition, to address the "gradient vanishing and exploding" problem in standard RNNs, we used LSTM networks, instead of standard RNNs, to construct the Encoder-Decoder LSTM for sequence to sequence prediction.

Experiments and Intermediate Results

Datasets and Preprocessing. We used ER as the training set for sentiment analysis. It contains 5,000 texts, labeled with eight types of emotions. Because of the lack of neutral and positive sentiments, we added texts labeled with another five sentiments into the training set to avoid misclassification. The dataset we used to classify sentiments and extract topics was TSRC. It contains 409,476 texts from January 1st to April 17th, 2020. To preprocess the data set, we removed URLs and mentions, such as "@name", converted the text to lower case, demojized the emojies, corrected misspellings, expanded common contraptions, and removed punctuation.

Sentiments Analysis. We converted the collection of preprocessed texts into a matrix of token counts with fixed size, and transformed it into a matrix of normalized TF-IDF terms. We set up Naive Bayes, Linear SVM, Log Reg, Linear SVC, Random Forest, LSTM, BERT, RoBERTa, XL-Net and DistilBERT models for sentiments classification and compared the performance using the accuracy score. The results(see1) shows LSTM model is the most accurate. Thus we trained the LSTM model on the entire training set and then used it to classify sentiments on the TSRC.

Topics Extraction. First, we tokenized each preprocessed comment to a list of words in TSRC. Second, we lemmatized and stemmed each word into their original form. Third, we removed stop words, and words other than nouns, verbs, adjectives, and adverbs. Fourth, we counted and visualized the top 50 most frequent words to remove words, e.g. "coronavirus", "corona", and "covid". We set up and fine-tuned LDA model to extract topics from TSRC. Our interpretation of the five topics are: Topic1: Recovering Strategies. Topic2: Source of Disease. Topic3: Infected Symptoms. Topic4: Route of Spreads. Topic5: Future Precaution.

Sequential Prediction. For sentiments prediction, we reconstructed the TSRC into a 106 by 13 dataset, where each

row represents a day from January 1st to April 17th in chronological order, each column represents a sentiment, and each value represents a daily sum of texts. Similarly, we prepared the topics prediction dataset, getting a 106 by 5 dataset, where each column represents a topic. Next, we predicted each sentiment independently. We Grid Search to find the optimal hyperparameter of ARIMA model who has the least RMSE. In addition, we compared different LSTMs based on RMSE, including LSTM with look back value (LSTMLB), LSTM with Window Method (LSTMWM), LSTM with Time Steps (LSTMTS), and LSTM with Memory Between Batches(LSTMM). The optimizer for each model was *adam*. We used the same modeling methods and approach for topics trend prediction. The result(see Table2) shows the ARIMA model is optimal for both tasks. Thus, we used the ARIMA model for the next 31 days' sentiments and topics trend prediction(visualized in Figure1 and 2).

Model Name	Accuracy	Model Name	Accuracy
Naive Bayes	0.8440	Liear SVM	0.7989
Logistic Reg	0.9528	LinearSVC	0.9218
Random Forest	0.7337	LSTM	0.9534
BERT	0.8538	RoBERTa	0.7410
XLNet	0.6547	DistilBERT	0.8987

Table 1: Model evaluation for sentiment classification

Model name	RMSE (Sentiment)	RMSE (Topic)
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.

Results

COVID-19 suddenly caused people's attention on January 19th (see Figure1). All of the 13 sentiments increased gradually until March 1st, 2020. After that each of the five topic became stable. The descending order of the averaged number of sentiments throughout the entire timeline were anxiety, sadness, relaxation, neutral, anger, fear, desire, enthusiasm, disgust, happiness, love, fun, and boredom. The prediction shows the number of anxiety is gradually increasing, while the other sentiments fluctuate around previous status.

The topics that people mainly talked about in Reddit are Infected Symptoms, Future Precaution, Source of Disease, Route of Spreads, and Recovering Strategies in descending order. 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.

Conclusions and Future Work

The large number of anxiety comments indicates that people were paying attention to COVID-19, which is good. but

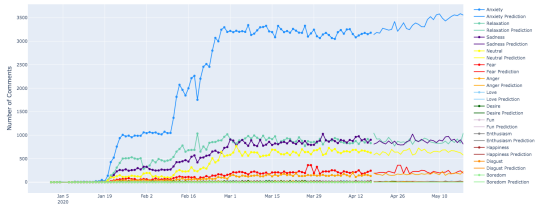

Figure 1: Sentiments multi-class classification using LSTM and predictions using ARIMA. Training and predicted data are distinguished by lines with/without daily markers

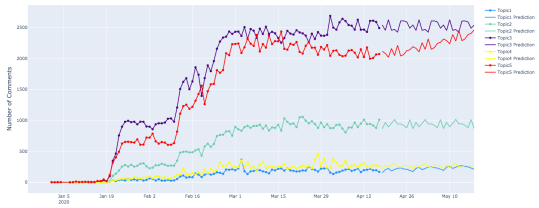

Figure 2: Topics extracted using LDA and prediction using ARIMA. Training and prediction data are distinguished by lines with/without daily markers

the increasing trend of anxiety comments is not good. The governments and WHO should provide more information to increase people confidence. The great attention to Infected Symptoms and Future Precaution, which is predicted to increase, indicate that detecting and preventing COVID-19 are hot topics. The low number of comments about Recovering Strategies indicates that most authors were not infected with COVID-19. We would include the "author" feature to study how an individual's sentiment and topic changed over time.

Surveys for COVID-19 are limited. 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. Even though using methods from natural language processing we can extract the meaning behind a sequence of words, we cannot explore people's thoughts. Facial expressions, physical and mental activities might be good indicators of thoughts. Thus, cooperating with people in physiology and psychology is needed.

Acknowledgements: I would like to thank Prof. Maria Gini who supervised this project.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3: 993–1022.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Kleinberg, B.; van der Vegt, I.; and Mozes, M. 2020. Measuring Emotions in the COVID-19 Real World Worry Dataset. arXiv:2004.04225.