2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Sentiment on Dissemination about COVID-19 of Mainstream Media Around the World: What Information are They Delivering?

Ruoyu Shen, *Department of Media and Information, Michigan State University*

Abstract— **The purpose of this article is to research the sentiment and topic classification about COVID-19 of mainstream social media in the United States to interpret what information the American public receives toward the COVID-19, and what are the perspectives of News and articles on epidemics in different topic fields. The study will extract unigrams to trigrams of different articles to judge the sentiments of articles, and use region-related keywords, dates, and topics extracted by classification as independent variables to measure the differences between disparate features. The result shows that news related to the business and health fields are more frequent (48.2% and 20.8% respectively). It also reveals that news regarding entertainment and technologies has a lower rate to be negative during the pandemic (5.6% and 11.1% respectively). With time flows during the research period, the sports news has a trend to be more negative, and a trend to be more positive for entertainment news and technology news.**

I. INTRODUCTION

2020 is an unusual year due to the epidemic of COVID-19. Up until September 14, 2020, there are 6,503,030 confirmed cases, 193,705 death cases in the United States, which ranked 1st of the whole world. Although the growth rate of confirmed cases has been smoother than in July, it's still significant and will not vanish in a short period [1]. In such a situation, the economy, army, agriculture production, health of the population, and other industries, as well as organizations, are all under the negative effect. The unemployment rate increased sharply; some huge corporations have bankrupted due to the loss of purchasing power. Thus, some civilians in the United States show their concerns about life under the epidemic. But on the other hand, some people still refuse to adopt efficient methods to keep safe, for example: staying at home, wearing masks, or not participating parties. This leads to a question: what attitude is the mainstream under the COVID-19? Which field in life is more promising to be improved soon in the public's perspective? This study is designed to address these questions by researching public media content.

There are already some other studies that have conducted similar research recently. Abd-Alrazaq et al. analyzed what Tweeters are concerned about during COVID-19 and identified the main topics of relevant Twitter by classifying approximately 2.8 million Twitter [2]. A study about the information dissemination status on Twitter during COVID-19 in Europe has been done previously by Pobiruchin, Zowalla, and Wiesner [3]. Li et al. conducted multiple methods to assess social media buzzes when the COVID-19 started to spread, and the media platforms are mostly in China [4]. As we can see, there are very rare studies that are targeted at the news. This study will put eyes on the news. The study will extract unigrams and trigrams of different articles and use word bags to generate the sentiment tags of articles. The word-bag is also used to obtain the topics distribution, combined with date, as an independent variable to measure the differences between disparate features.

II. METHODS

As the final goal of this study is to classify the topics of news published by mainstream media, the core method will be revolved around and be simplified to the classification problems to grasp the best maneuverability of the result yielding. The database used in this study is Jenna's public database [5], downloaded from a public database webpage, Kaggle. The data in Jenna's database was collected by information fetch with the help of a crawler program from 65 selected influential news websites and focused on no-medical aspects of COVID-19. English is the only language that has been included in the database, and the 8-month period of the database has a range from January 13, 2020 to September 14, 2020. Around 224,000 articles have been kept in the database. To maintain the only most important features, Title, Content, and Time have been used for the research. Titles and contents in this database are combined and formed to the sparse matrix in terms of word-bags by reducing the dimension before the training.

Although the topic tags have already been automatically generated in Jenna's database, it has poor precision, and the concept of each classified category is too wide. A more detailed and precise classification is preferred for this project. To train a model for classifying the topic of research databases, another public database from Kaggle is used to achieve this. Artem's dataset has a size of 108,774 entries, with 8 different classifications of the topics [6], which is way more ideal for our research target. 3 columns in the database will be used for training: Link, Title, and Topic. The news content will be fetched from the link that the database provided by an application developed with the help of the Beautiful Soup [7] from their original webpages. The logic of handling the content in the page is fetching all plain texts after the title appears between the <p></p> tags within <article></article> tags, and if <article> tag is not existed, then fetch <p></p> tags within <body></body> tags. The title of articles will be combined to fetch content before the training.

The model for predicting the sentiment will be trained using Chaithanya's dataset [8], downloaded from Kaggle as well. It is a dataset with the plain text from Twitter, and Reddit. There are 190,000 samples collected in this database and are fetched from Twitter and Reddit's official websites. The sentiment tags in these databases have 3 states: positive (1), negative (-1), and neutral (0). As the negative news is the emphasis in this study, data with a neutral tag has been reassigned with a positive (1) tag and regarded as a piece of positive news.

After preprocessing the data, the training process will be conducted. Because the essence for getting the sentiments is a binary classification problem, the Logistic Regression model will be used for this model to train the vectorized words from the plain text [9], as Model 1 shown in Figure 1. Meantime, a Naïve Bayes classifier will also be trained as a baseline to compare the performance with Logistic Regression. Model 2 is another model used for classifying the topics. The content sequence is not critical in the topic classification process, so the word-bags combining with Random Forest will be used as the classifier [10]. 80% - 20% dataset splitting is applied for both models, which means use 80% data entries as training data, and the rest of 20% as testing data. The training and fitting process is used in Python 3.0, and uses PyTorch and SkLearn packages to build the model [11][12].

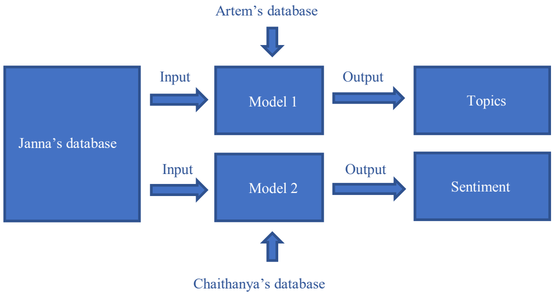

Figure 1. Training and analysis process for the study

As Model 2 shown in Figure 1 is trained by content from social media platforms and used to apply to the news contents, here we have an assumption that a social media-based dataset trained model can mostly be used as the predictor for news, as they have similar content patterns. A possible error with the assumption mentioned above is the news has more literature words, while posts on Twitter will be mixed with literature and oral words, but this error can be weakened for a large training dataset.

To keep the consistency for two models, Model 1 and Model 2 are all trained with word-bags. The process of training and evaluating each model is shown in figure 2. The output of model 1 contains multiple labels, thus the use of losses computed by loss functions for model evaluation is feasible and relatively precise. The equation for computing the loss of the model used in this study is:

$$loss = \sum_{i=1}^{n} \frac{(1-prob_{i,a})}{n} \qquad (1)$$

Among the equation (1), *prob* is the prediction matrix that contains each sample's probability on all labels, *i* represents the *i*-th sample, and *a* represents the real label of the sample. Model 2 is a typical Logistic Regression model with the penalty function of L1 and has only 2 probable labels in output. It is designed to use *AUROC* and accuracy to evaluate the performance of the model. *AUROC* is a tool for measuring non-equilibrium in classification and is popularly used to evaluate a binary classifier. It's highly robust regardless of the ratio of positive labels to negative labels. To sum up, it's an ideal method in this case as an evaluation tool. The equation for computing the *AUROC* is shown as [13]:

$$AUROC = \frac{\sum_{i \in positive} rank_i - \frac{M(1+M)}{2}}{M \times N} \qquad (2)$$

In equation (2), rank represents the *i-th* sample's score rank from highest to lowest, *M* means the volume of positive samples, and *N* means the volume of negative samples. It's a simplified method to calculate the rough value of *AUROC* but still has high fidelity, however, it significantly reduces computation complexity.

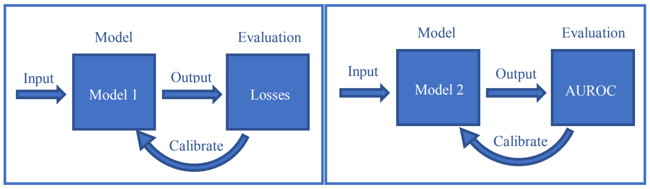

Figure 2a. Random Forest Model Figure 2b. Naïve Bayas Model

Figure 2. Model training and evaluating process

III. RESULTS

After training two models with their training datasets respectively, the models have been evaluated and tuned for the best performance by using the test dataset as the input and comparing the real classification and outputted classification from the models. In the evaluation process, as shown in Table 1, the Random Forest got a 77.6% accuracy and a loss of 0.52. As the baseline, a random prediction model for this dataset has a 12.5% accuracy and a loss of 0.86. For the Logistic model, the AUROC is the indicator to judge the performance of the model. After exploring several hyperparameters of the built-in functions, the accuracy of the best performance Logistic Regression model is 85.8%. For comparison, the Naïve Bayes model as the baseline has a 69.5% accuracy, and a random prediction model will only have a 50% accuracy. The Logistic Regression model has a better AUROC than Naïve Bayes as well, which are 0.90 and 0.79 respectively (Table I), the visualized AUROC is shown in Figure 3.

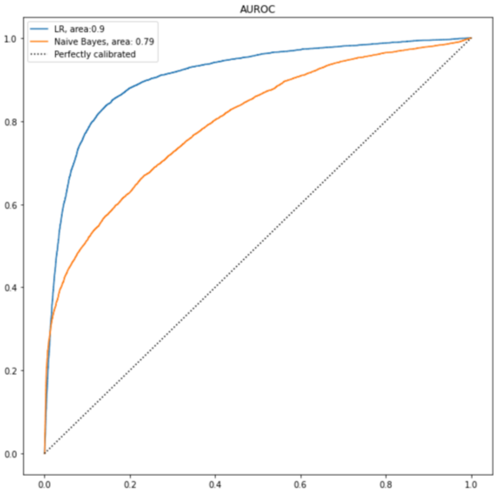

Figure 3. AUROC curves and areas

TABLE I. MODEL PERFORMANCES

Feature	Models	Performances		
		Accuracy	*Loss*	*AUROC*
Topic	Random Forest	77.6%	0.52	-
	Random Prediction	12.5%	0.86	-
Sentiment	Logistic Regression	85.8%	-	0.90
	Naïve Bayes	69.5%	-	0.79
	Random Prediction	50.0%	-	-

From the percentage of different topics shown in Figure 4, we found that news with tags "BUSINESS" and "HEALTH" have much higher volumes than other tags (48.2% and 20.8% respectively). From this fact, we can conclude that among news from mainstream media during the database recorded period, business issues and health issues are mentioned more frequently than other themes. It is not hard to detect that the "SCIENCE" tag of the topic in the main dataset has only 0.2% percentage, which means only 448 samples. This will possibly lead to a shortage of samples in analyses and lead to biased results, thus, discussion about the "SCIENCE" related topic will be excluded in this research.

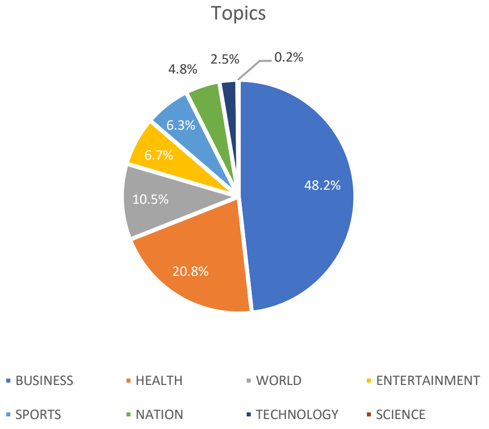

Figure 4. Ratios of topics in main dataset

Figure 5 shows the sentiment shares. The difference the between amounts of positive news and negative news is huge (238,960 positive news and 32,666 negative news), and this situation also occurs in every single topic in the sentiment research by different topics. The most important reason that leads to this phenomenon is that neutral sentiment news is also included in the positive sentiment category in this study. As the negative sentiment news is more likely to get insightful information about the public's attitude from the research and it also attracts more attention to the public [14], the result makes sense and can be dug deeper by combining with the feature of the topic.

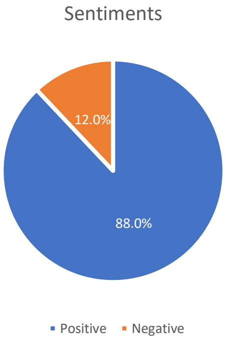

Figure 5. Ratios of sentiments in main dataset

Research on the ratio of sentiments on the different topics is essential to know the public is more pessimistic about which fields during the COVID-19 epidemic. The cross-comparison study shows more information to us (TABLE II). We find that "WORLD" and "NATION" tags have a significantly higher percentage on negative sentiment than average negative ratio (19.3% and 18.2% respectively) on sentiment. Oppositely, "ENTERTAINMENT" and "TECHNOLOGY" have a significantly lower negative ratio (5.6% and 7.6% respectively) than the average on sentiment. It indicates that general worldwide related topics and general national related news are more likely to bring negative messages to the public. Inversely, entertainment news and new technology releases convey more about positive attitudes.

The "BUSINESS" and "SPORTS" tags also wear a relatively lower percentage of negative sentiment than the average. It reveals that the public also believe in a sanguine future on business and sports matches, although COVID-19 has a severe negative impact on small business [15].

TABLE II. SENTIMENT DISTRIBUTION BY TOPICS

Topic	Sentiment	Features	
		Counts	*Percent*
BUSINESS	Positive	116,543	89.0%
	Negative	14,390	**11.0%**
HEALTH	Positive	49,608	87.8%
	Negative	6,903	12.2%
WORLD	Positive	23,101	80.7%
	Negative	5,540	19.3%
ENTERTAINMENT	Positive	17,296	94.4%
	Negative	1,035	**5.6%**
SPORTS	Positive	15,138	88.9%
	Negative	1,891	**11.1%**
NATION	Positive	10,598	81.7%
	Negative	2,367	18.3%
TECHNOLOGY	Positive	6,254	92.4%
	Negative	513	**7.6%**

* Bold: lower than average negative ratio.

The trend of sentiment changes over time flows can also reveal lots of fun facts on COVID-19. The curve of negative sentiment percentages by month is shown in Feature 6. The total trend of the negative sentiment ratio is smooth, but it varies among different topic tags. The topic tags of "TECHNOLOGY" and "ENTERTAINMENT" have an obvious downward trend as time goes by, which means the public gradually becomes more positive about these topics. In other words, we can imply that the public gradually feels more confident about new technologies continuously emerging out, and they gradually relax their vigilance on or adapt to the epidemic and then attend more entertainment events and gain happiness from them. By contrast, the "SPORTS" topic presents a sharp upward trend in the percentage of negative sentiment. It reveals the fact that the public's concern is more about sports-related problems with the time flows.

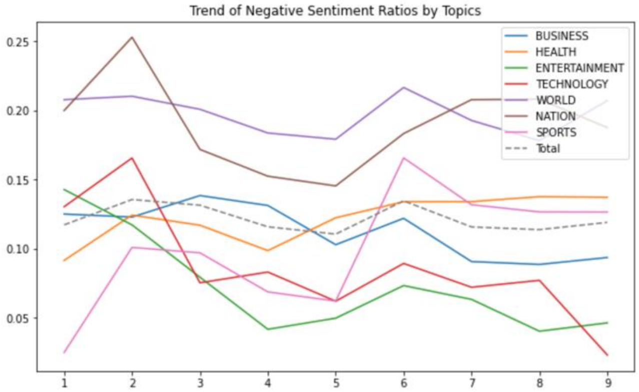

Figure 6. Trend of Negative Sentiment Ratios by Topics

IV. Discussion

Relatively more business and health-related news are released during the research period, which indicates these two fields have more incidents happen during the COVID-19 period. Or in other words, the public cares more about business and health affair's development. Entertainment has been affected by the pandemic from the early stage, and the sports industry is suffering a heavy setback the reason that athletics and managers have been infected by the COVID-19 virus lockdown [16]. While the truth is that the public is still wild about entertainment news and has a trend to become more positive about it. It's a piece of rational advice for the government that develops and encourages other kinds of entertainment activities for the public to improve their satisfaction with life during COVID-19 and enhance the faith-facing pandemic. But it's necessary to make sure it's under control and avoid any risky actions, like face-to-face activities, etc. Sports industry entrepreneurs should also find a way to adapt to the current situation.

Fortunately, the decline of negative sentiment rate for health-related and technology-related topics shows media platforms are gradually delivering more positive information to help to build optimistic attitudes toward the development of a vaccine for COVID-19 and faith that the epidemic will be overcome by human beings among the public, or the public is no more showing low-confidence toward future under the effect of COVID-19. Business activities have been affected by the COVID-19 badly, while the public still takes a positive view about it. Building strong faith in the public has been proven to be an efficient way to recover from a pandemic [17][18], and the more individuals who have a positive attitude on the epidemic, the higher possibility that human beings can sooner overcome the COVID-19.

References

[1] CDC, "Coronavirus Disease 2019 (COVID-19) in the U.S.," *Centers for Disease Control and Prevention*, Mar. 28, 2020. https://www.cdc.gov/covid-data-tracker (accessed Sep. 14, 2020).

[2] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah, "Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study," *J. Med. Internet Res.*, vol. 22, no. 4, p. e19016, 21 2020, doi: 10.2196/19016.

[3] M. Pobiruchin, R. Zowalla, and M. Wiesner, "Temporal and Location Variations, and Link Categories for the Dissemination of COVID-19-Related Information on Twitter During the SARS-CoV-2 Outbreak in Europe: Infoveillance Study," *J. Med. Internet Res.*, vol. 22, no. 8, p. e19629, 28 2020, doi: 10.2196/19629.

[4] J. Li, Q. Xu, R. Cuomo, V. Purushothaman, and T. Mackey, "Data Mining and Content Analysis of the Chinese Social Media Platform Weibo During the Early COVID-19 Outbreak: Retrospective Observational Infoveillance Study," *JMIR Public Health Surveill*, vol. 6, no. 2, p. e18700, 21 2020, doi: 10.2196/18700.

[5] L. Janna, "Covid-19 Public Media Dataset by Anacode," 2020. https://kaggle.com/jannalipenkova/covid19-public-media-dataset (accessed Oct. 25, 2020).

[6] B. Artem, "Topic Labeled News Dataset," 2020. https://kaggle.com/kotartemiy/topic-labeled-news-dataset (accessed Oct. 25, 2020).

[7] L. Richardson, "Beautiful soup documentation." 2007.

[8] K. A. Chaithanya, "Twitter and Reddit Sentimental analysis Dataset," 2020. https://kaggle.com/cosmos98/twitter-and-reddit-sentimental-analysis-dataset (accessed Oct. 25, 2020).

[9] J. A. Nelder and R. W. M. Wedderburn, "Generalized Linear Models," *Journal of the Royal Statistical Society. Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972, doi: 10.2307/2344614.

[10] A. Liaw and M. Wiener, "Classification and Regression by RandomForest," *Forest*, vol. 23, Nov. 2001.

[11] A. Paszke *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *arXiv:1912.01703 [cs, stat]*, Dec. 2019, Accessed: Jul. 23, 2021. [Online]. Available: http://arxiv.org/abs/1912.01703

[12] L. Buitinck *et al.*, "API design for machine learning software: experiences from the scikit-learn project," *arXiv:1309.0238 [cs]*, Sep. 2013, Accessed: Jul. 23, 2021. [Online]. Available: http://arxiv.org/abs/1309.0238

[13] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982, doi: 10.1148/radiology.143.1.7063747.

[14] R. F. Baumeister, E. Bratslavsky, C. Finkenauer, and K. D. Vohs, "Bad is Stronger than Good," *Review of General Psychology*, vol. 5, no. 4, pp. 323–370, Dec. 2001, doi: 10.1037/1089-2680.5.4.323.

[15] A. W. Bartik, M. Bertrand, Z. Cullen, E. L. Glaeser, M. Luca, and C. Stanton, "The impact of COVID-19 on small business outcomes and expectations," *Proc Natl Acad Sci U S A*, vol. 117, no. 30, pp. 17656–17666, Jul. 2020, doi: 10.1073/pnas.2006991117.

[16] M. Drewes, F. Daumann, and F. Follert, "Exploring the sports economic impact of COVID-19 on professional soccer," *Soccer & Society*, vol. 0, no. 0, pp. 1–13, Aug. 2020, doi: 10.1080/14660970.2020.1802256.

[17] I. M. Rosenstock, "The Health Belief Model and Preventive Health Behavior," *Health Education Monographs*, vol. 2, no. 4, pp. 354–386, Dec. 1974, doi: 10.1177/109019817400200405.

[18] P. Halligan and M. Aylward, *The power of belief: psychosocial influences on illness, disability and medicine*. Oxford: Oxford University Press, 2006. Accessed: Dec. 15, 2020. [Online]. Available: http://orca.cf.ac.uk/3256