Covid-19 News Sentiment Classification

Alan Luo
tl2905@nyu.edu

Oliver Z. Wang
o.wang@nyu.edu

Qiuhao Zhang
qz2016@nyu.edu

Abstract

The Covid-19 pandemic has profoundly impacted our social and economic activities. In order to better understand the public emotion on the topic, we conducted a sentiment analysis on 122,826 Covid-19 related news between January 2020 and September 2020. A fine-tuned VADER model and a modified BERT model are used to approach this task. We first identified the words that show different semantic scores in Covid-19 news compared to general topics and the words with high semantic polarities specific to Covid-19. We then tested different aggregation methods to arrive at the article-level sentiment, and it turned out that averaging over sentences is the best granularity. The results show that the VADER model achieved a macro-F1 score of 0.61 on this ternary classification task, while the BERT model achieved a macro-F1 score of 0.66. Finally, we calculated the daily average news sentiment scores and identified that it leads some of the S&P 500 Index patterns by 2-3 weeks.

1 Introduction

Since the first identified case of Covid-19 in Wuhan, China, in December 2019, the disease has spread to over 190 countries and regions and caused a worldwide pandemic. In addition to the public health crisis it brought along, the pandemic has also fundamentally reshaped our social behaviors and economic activities. The public is undergoing a complicated feeling about this pandemic. Suppose we can distinguish people's emotions and perspectives on the topic. In that case, the policymakers will be better informed to make the right decision, and the economists and investors may better forecast the economic trend in the near future. Aguilar et al. (2020) managed to predict Spanish GDP with news-based sentiment indicators.

Sentiment analysis has been an essential topic in the field of natural language processing. Depending on the text sources, sentiment analysis can be performed on the news, social media postings, online reviews, etc. There are multiple approaches to sentiment analysis, including lexicon-based approaches and machine learning-based approaches, which will both be covered in this article. Shapiro et al. (2020) evaluated both lexicon models and machine learning models on text sentiment analysis of economic news and demonstrated its superiority to survey-based consumer sentiment index.

VADER (Valence Aware Dictionary for sEntiment Reasoning) is a lexicon-based sentiment analysis model that aims to overcome the challenges of performing the task on micro-blog-like social media content. It generalizes better across contexts than benchmarks such as LIWC, ANEW, and many machine learning techniques. Hutto and Gilbert (2014) developed VADER by first constructing and validating a sentiment lexicon customized for their target context of social media. Then they qualitatively derived five generalizable grammatical and syntactical heuristics that capture the sentiment intensity of the texts. For instance, these rules take factors like degree modifiers, capitalization, and punctuation into consideration. The VADER model combines these two components and outperforms human raters on sentiment analysis of tweets. More importantly, VADER also performs better on New York Times opinion news articles than all of the seven established lexicon models used in the experiment. Therefore, we believe it has great potential for Covid-19 related news articles.

Devlin et al. (2018) introduced BERT (Bidirectional Encoder Representations from Transformers), a language representation model with pre-trained data that can be used for different types of tasks. Munikar et al. (2019) added a dropout layer and a softmax classifier layer to the existing BERT model and proved that the model works well on the sentiment classification task of movie reviews SST-2 and SST-5 (Socher et al., 2013).

Despite VADER and BERT's advantages, they may not perfectly fit the Covid-19 related news. Many terms with positive sentiment scores in general-purpose lexicons may have negative semantic meanings within Covid-19 news. For example, the word "positive" is associated with positive sentiment in most contexts but carries a mostly negative meaning in phrases like "positive cases" and "tested positive," which are common in reports on the pandemic. Therefore, the vanilla models need to be fine-tuned to achieve the best performance on this specific topic.

In this article, we retrieved Covid-19 related news data from GDELT (Global Database of Events, Language, and Tone) (Leetaru, n.d.) from January 2020 to September 2020. By manually labeling 714 news articles, we identified the words that show different semantic scores than general topics, and the specific words to the Covid-19 topic. We then created a modified VADER out of these words and calculated article-level sentiment scores using different aggregating methods. In addition, we also trained a modified BERT model on sentence-level with two sets of training data: 1553 human-labeled sentences and 19,000 augmented sentences labeled by the modified VADER. Finally, we computed the daily average sentiment scores during the time span and identified some turning points of the public sentiment related to certain social events.

2 Method

2.1 Problem Statement

Given that VADER is initially built to conduct sentiment analysis on short snippets of text on general topics, we would like to find an appropriate method to compute sentiment scores of entire articles, which tend to be longer texts than what VADER is optimized for. We also need to develop proper modifications of the VADER lexicon so that it is more attuned to articles on the Covid-19 pandemic.

For the modified BERT model, we would like to know if the existing model with further training on Covid-19 related news is good enough to produce results. We also need to augment the size of training data with limited hand-labeled data.

2.2 Lexicon-based Approach - VADER

Our lexicon-based approach uses the VADER model as the starting point. VADER has its own lexicon with each word assigned a sentiment score (Hutto and Gilbert, 2014). To customize VADER for Covid-19 specific news, we updated the sentiment scores for words that show a different sentiment than usual in these articles. Besides, we appended new words with high frequency and strong sentiments in our data to VADER's lexicon.

With the updated VADER model, we scored over 714 articles that we manually labeled using different aggregation methods, such as by full article score, by average sentence scores, and by average paragraph scores, etc. Then we computed accuracy metrics for these scored articles using our manual labels. Since the VADER sentiment score is a real number between -1 and 1, we optimized the boundaries for negative, neutral, and positive predictions that yields the highest macro-F1 scores.

2.3 Machine Learning Approach - BERT

BERT (Devlin et al., 2018) is a multi-layer bidirectional Transformer encoder model. It accepts a sentence as a token sequence for the input and produces a sequence embedding for the outputs. This model has been pre-trained on two datasets: BooksCorpus (Zhu et al., 2015) and English Wikipedia.

Munikar et al. (2019) added a dropout layer and a softmax classifier layer to the model. The dropout layer is used to avoid the model from overfitting on our new dataset (Srivastava et al., 2014). The softmax classifier layer is used to generate the classification results. Figure 1 shows the architecture of the new model.

Our machine learning approach uses Covid-19 related news data to train a similar model as Munikar et al. (2019). Since the data do not come with labels, and we have limited hand-labeling resources, we used our customized VADER model from 3.3 to augment the size of training data.

3 Experiments

We conducted two sets of experiments on the modified VADER model and the modified BERT model. Although we used the same dataset for these experiments, we processed the data differently for each approach. However, the machine learning approach does incorporate some results from the experiment

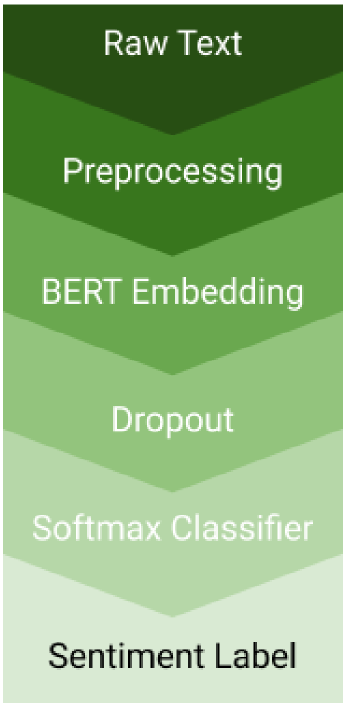

Figure 1: The structure of BERT model.

with the VADER model to facilitate data processing and hyperparameter tuning.

3.1 GDELT Dataset

The GDELT Project (Leetaru, n.d.) tracks global news every day from nearly every country in more than 100 languages. The project also extracts themes, events, locations, and so on from the news.

Between January 1st, 2020, and September 30th, 2020, GDELT Project collected about 23 million news worldwide.

Each news is associated with one or more themes, such as GENERAL_HEALTH (any general discussion of human, animal, plant, etc. health, from hygiene to hospitals, doctors to disease), and TAX_ETHNICITY (any mention of a major ethnicity from around the world). We collected news with at least one theme in TAX_DISEASE_CORONAVIRUS, WB_2167_PANDEMICS, or HEALTH_PANDEMIC. That gives us about 7 million news articles.

Since the GDELT database does not contain the news body, we built a scraper (Myers and McGuffee, 2015) to collect news body. We only collected news from the three most frequent websites: Msn.com, Yahoo.com, and Reuters.com, which leave us a total of 394,257 news. We randomly sampled the news entries and scraped 122,826 news. Table 1 shows the number of articles from each source.

Source	Count
Msn.com	37,993
Yahoo.com	39,228
Reuters.com	45,605
Total	122,826

Table 1: News sources and counts.

The GDELT Project dataset includes a tone indicator for each article, but it is computed using a naive lexicon-based approach (Leetaru, n.d.).

For this project, we manually labeled two subsets from the data we collected.

We first randomly sampled 714 articles and labeled them into three categories: negative, neutral, and positive. Based on subsequent analysis, we noticed that both models work better on sentence-level instead of article-level. Thus, we randomly sampled 1553 sentences from this 714 articles and labeled them into the three sentiment categories.

Table 2 shows how many articles and sentences in each category. Please notice that both subsets have more negative samples than neutral and positive ones due to the news theme. On average, the general tone of Covid-19 news is negative.

Category	Article Count	Sentence Count
Negative	374	634
Neutral	159	593
Positive	181	326
Total	714	1553

Table 2: Article-level and sentence-level categories and counts.

3.2 Evaluation Metrics

Because our predictions and labels are continuous and ordinal, respectively, Spearman's Rho Correlation Coefficient is appropriate to measure the correlation between them. Since we are interested in the model's performance in all prediction classes, we computed the F1 scores for negative, neutral, and positive predictions, as well as a macro-F1 score. Because the three classes' distribution of data is highly skewed, we took the weighted average of the three F1 scores as the macro-F1. Also, we calculated the Coefficient, R^2, to provide another measurement of our model's overall ability to predict news sentiment.

3.3 Lexicon-based Approach - VADER

With the labeled news articles, we computed pointwise mutual information (PMI) scores between each word and sentiment category (Shapiro et al., 2020). We then calculated a custom score (S) for each word, which is the difference between the positive PMI and the negative PMI:

$$\mathrm{PMI} = \log\left(\frac{p(w,c)}{p(w)p(c)}\right) \tag{1}$$

$$S(w) = \mathrm{PMI}(w, positive) - \mathrm{PMI}(w, negative) \tag{2}$$

After computing the custom scores, we plotted them against VADER scores for frequent words in our corpus, as shown in Figure 2. We observed that most words in the figure cluster near the diagonal, indicating that the most frequent words in our dataset display similar sentiment under Covid-19 topics compared to other topics. However, words like “positive” and “crisis” show different sentiments in Covid-19 related news, so we substituted their VADER model scores with our custom scores.

In addition, we also picked out some novel words not covered by VADER but have high frequency and extreme sentiment scores, such as “Coronavirus” and “Pandemic,” and appended them to the VADER lexicon. The modifications made to the VADER lexicon are shown in Table 3.

Source	Token	Sentiment Score
VADER	crisis	0
	positive	-1
	great	0
	authority	-1
Novel	coronavirus	-1
	pandemic	0.5
	outbreak	-1.5
	virus	-1
	lockdown	-1
	trump	-1.5

Table 3: The existing words in VADER that are modified and the novel words added to the VADER lexicon

With the customized VADER model, we scored the 714 news articles previously labeled. Four aggregation methods were used to compute various VADER scores for an article:

1. VADER score of the entire article
2. Average VADER score of sentences
3. Average VADER score of paragraphs
4. Average VADER score of the first and last paragraph

Results showed that scoring the entire article with VADER produces a different result from an article’s average sentence or paragraph scores.

Because VADER scores are continuous values between -1 and 1, so they have to be mapped into the three discrete classes of negative, neutral, and positive before evaluating the F1 scores. For each aggregation method, we exhaustively enumerated over all possible boundary values for the mapping with 0.05 granularity and chose the boundary that yields the highest macro-F1 scores against the manual labels. We mapped the article’s sentiment scores to predicted labels and evaluated each aggregation method using the metrics described in section 3.2.

3.3.1 VADER Model Results

Using the updated VADER model, we made sentiment predictions on the 714 sentences we labeled, with the sentiment boundaries optimized as described above. Table 4 shows the evaluation results for different score aggregation methods and the corresponding score boundaries. As our experiment indicates, scoring all sentences in the article and taking their average produced the highest macro-F1 score and accuracy. For this aggregation method, the optimal boundaries for a neutral class are 0.00 and 0.15. Any article with a score less than 0.00 will be predicted as negative, and any article above 0.15 will be predicted as positive.

It is not surprising that the best performance is achieved by using average sentence scores. The evaluation datasets Hutto and Gilbert (2014) used to benchmark VADER against other baseline models consist of sentence-level snippets instead of longer paragraphs or articles. Therefore, VADER is intrinsically more suitable for sentence-level sentiment analysis. To be more specific, the sentiment intensity heuristics in VADER’s model decides that a few tokens with extreme intensities could steer an entire article’s sentiment. By scoring individual paragraphs or sentences and aggregating them, these tokens would have a lesser impact on the article-level sentiment score (Hutto and Gilbert, 2014).

We then made the same predictions on these 714 articles using the unmodified VADER model as well as the tonal indicator from GDELT. Table 5

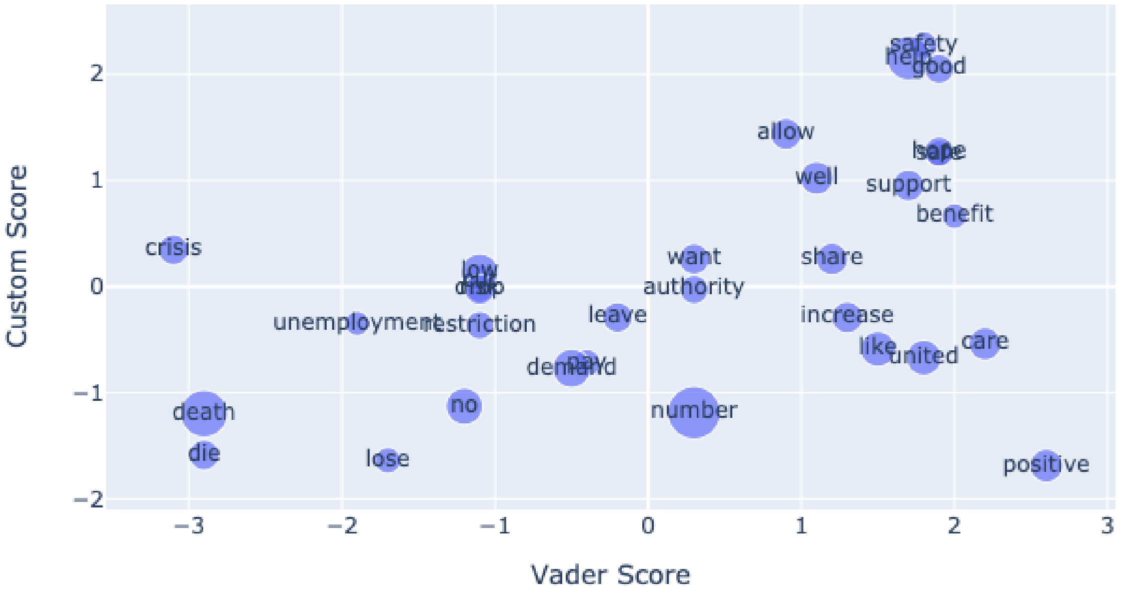

Figure 2: A comparison of our custom scores and VADER scores for words that appear more that 15 times in the sampled 714 new articles. The size of the blue circle indicates the word's frequency.

	Entire Article	Sentence Avg	Paragraph Avg	First & Last Paragraph Avg
Correlation	0.4850	**0.5504**	0.5438	0.1849
Accuracy	0.5924	**0.6064**	**0.6064**	0.4524
Macro-F1 Score	0.5852	**0.6103**	0.6048	0.4512
Negative F1 Score	0.7490	0.7609	**0.7666**	0.5799
Neutral F1 Score	0.2857	**0.3878**	0.3161	0.2367
Positive F1 Score	0.5096	0.4946	**0.5241**	0.3739
R^2	0.2272	**0.3092**	0.2901	0.2020
Neutral Boundaries	0.25-0.95	0.00-0.15	0.00-0.15	0.15-0.5

Table 4: The Spearman's Rho correlation, accuracy, Macro-F1 scores, F1 scores for all classes, and R^2 values, and optimal sentiment score boundaries for each aggregation method.

shows the F1 scores achieved by modified VADER, original VADER, and GDELT's in tonal indicators.

Both versions of the VADER model have a higher F1 score than the GDELT prediction, showing that a more robust rule-based model may be superior to GDELT's naive lexicon approach (Leetaru, n.d.). The results also indicates that our adjustments of VADER lexicon indeed improved the performance on Covid-19 news in terms of accuracy and weighted macro-F1 scores. Specifically, the modified VADER outperforms the vanilla one on negative and neutral classes, and under performs on positive classes. Due to the higher volume of negative and neutral news in our data, the modified VADER shows the best overall performance among the three methods.

3.4 Machine Learning Approach - BERT

The machine learning experiment for this project was running on a computer with a Tesla K80 GPU.

Since we only have 1553 manually labeled sentences which is a rather small training size for a BERT model, we used the modified VADER from 3.3 to enhance the training size to 19,000 and train the machine learning model. We then use it to predict sentiment classes on new sentences and compare to the human labels.

We also trained the model directly with the manually labeled sentences as a comparison (1400 sentences for training and 153 for evaluation).

3.4.1 BERT Model Results

We reported the evaluation results in Table 6.

The first column of Table 6 shows the results of the model trained on the modified VADER gener-

	Modified VADER	**Original VADER**	**GDELT Indicators**
Accuracy	**0.6064**	0.6050	0.6036
Macro F1	**0.6103**	0.6008	0.5942
Negative F1	**0.7609**	0.7490	0.7497
Neutral F1	**0.3878**	0.3434	0.3056
Positive F1	0.4946	**0.5209**	0.5263

Table 5: Comparison of accuracy, macro F1 score, negative, neutral, and positive F1 scores for the modified VADER, original VADER, and GDELT tonal indicators on the manually labelled 714 articles.

ated labels. Please notice that the model is evaluated on the dataset labeled by human. The result is not as good as the modified VADER from 3.3. This is not surprising because the machine learning model was trained on labels generated by VADER but evaluated on the human labels. The accuracy loss will pass from VADER to the machine learning model.

The second column of Table 6 reported the results of the model trained on the human-generated labels. Both the accuracy and the F1 scores are better than the modified VADER. However, we have concerns here as the model may be overfitting on the training data since there are only 1400 sentences for training.

3.5 Analysis

With the modified VADER as described in 3.3, we calculated the sentiment scores for all the 122,826 news and generated the daily average news sentiment scores and volumes. Since the original daily values show an apparent weekly pattern (e.g., negative emotions on weekdays and slightly natural emotions on weekends), we then applied a moving average of 7 days on both of the time series. We also calculated the daily sentiment using BERT described in 3.4, which showed a similar pattern as the modified VADER. For simplicity, we only present the former one here.

We compared the sentiment time series with S&P 500 Index in Figure 3. Two steep decreases exist in the sentiment curve around late February and early June. They correspond to the beginning of the massive spread of coronavirus in the U.S and the George Floyd event, respectively. A similar pattern occurs in the S&P 500 Index around the same periods with 2-3 weeks of lag. This indicates that the news sentiment can provide some leading signals to forecast the macro-economic indices.

4 Related Work

Many types of research have been done in different approaches for news sentiment analysis. Some of them are directly connected to the coronavirus disease. We will briefly discuss here some related work that has been done previously on sentiment analysis.

Lamsal (2020) introduced a way to collect Covid-19 related comments from Twitter on a daily basis. He also included active keywords and hashtags in the dataset.

Aguilar et al. (2020) created a newspaper-based sentiment indicator for Spain and used it to monitor Spanish economic activity in real-time. They specifically used their model to predict the current economic recession due to Covid-19.

Taj et al. (2019) explored lexicon-based approaches for sentiment analysis on news articles. They collected 2225 news from the BBC news website and compared the dictionary-based methods and corpus-based methods.

Aslam et al. (2020) analyzed the emotions and sentiments of 141,208 new related to coronavirus disease. They compared sentiment scores with time and reached the conclusion that people have a strong negative emotion to the coronavirus disease. They also include a list of high-frequency Covid-19 related positive and negative vocabulary in their article.

Oyebode et al. (2020) collected COVID-19-related comments from six social media platforms and focused on revealing critical phrases linked to people's emotions to the disease.

5 Conclusion

We explored the worldwide emotional changes in the Covid-19 pandemic using news data.

We have worked on both the lexicon-based approach and machine learning approach and compared the results. For the lexicon-based approach,

	Dataset Labeled by VADER	Dataset Labeled by Human
Accuracy	0.5329	0.6667
Negative F1	0.6213	0.7299
Neutral F1	0.5057	0.6429
Positive F1	0.4056	0.5490
Macro F1	0.5319	0.6608

Table 6: Comparison of accuracy, negative, neutral, positive, and macro F1 scores for the machine learning models that are trained on the datasets labeled by VADER and by human. Both models are evaluated on the human labeled dataset.

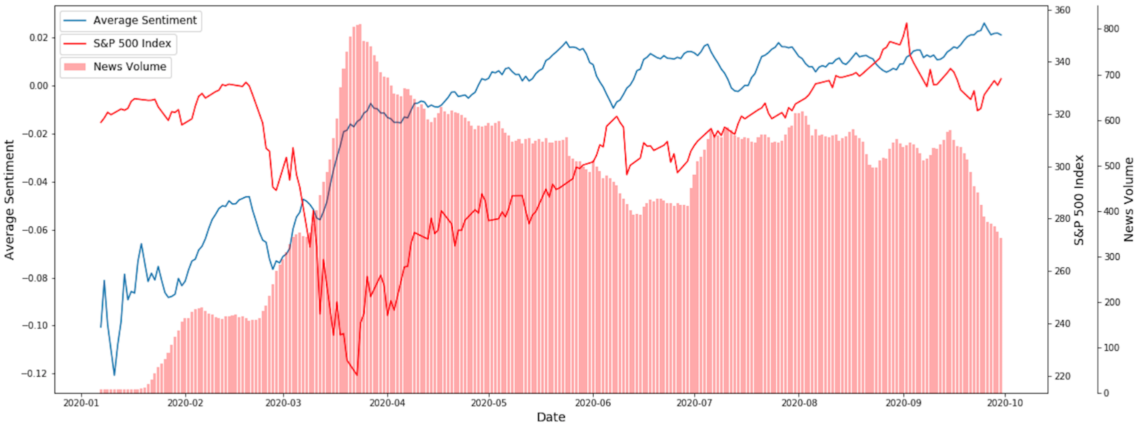

Figure 3: The time series of daily average news sentiment (blue line) and S&P 500 Index (red line) plotted together with the daily news volume (red bars), from January 1st, 2020 to September 30th, 2020.

our modified VADER model has a better performance than the original VADER model. Moreover, we found that averaging over sentences is the best granularity. For the machine learning approach, the model trained by the human-labeled news has better performance, but there may be a risk of overfitting.

We also compared the sentiment time series with the S&P 500 Index and identified some forecasting power.

For future work, we think our project can be improved in three directions:

1. The original VADER model was fine-tuned on social media materials (Hutto and Gilbert, 2014). We have shown that a simple fine-tuning can improve the performance on Covid-19 related news. We want to explore if the lexicon-based approach can be used in other news themes, such as politics, economics, or video games.

2. Our machine learning model trained on the human-labeled news has better performance than the others, but there is a chance of overfitting. We wonder if the model will still achieve high accuracy once we have collected enough manually labeled data for training.

3. Our current approach of fine-tuning a lexicon-based model and using it to generate training data for the machine learning model is very similar to few-shot learning (Wang et al., 2020). Increasing the accuracy of the lexicon model will definitely benefit the downstream machine learning model.

In the end, we are looking forward to the end of this Covid-19 pandemic in the near future when people's life can get back to normal.

References

Pablo Aguilar, Corinna Ghirelli, Matías Pacce, and Alberto Urtasun. 2020. Can news help measure economic sentiment? an application in covid-19 times. *SSRN Electronic Journal*.

M-Faheem Aslam, Tahir Awan, Jabir Hussain Syed, Aisha Kashif, and Mahwish Parveen. 2020. Sentiments and emotions evoked by news headlines of coronavirus disease (covid-19) outbreak. *Humanities and Social Sciences Communications*, 7.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014*.

Rabindra Lamsal. 2020. Coronavirus (covid-19) tweets dataset.

Kalev H. Leetaru. n.d. The gdelt project.

Manish Munikar, Sushil Shakya, and Aakash Shrestha. 2019. Fine-grained sentiment classification using BERT. *CoRR*, abs/1910.03474.

Daniel Myers and James W. McGuffee. 2015. Choosing scrapy. *J. Comput. Sci. Coll.*, 31(1):83–89.

Oladapo Oyebode, Chinenye Ndulue, Dinesh Mulchandani, Banuchitra Suruliraj, Ashfaq Adib, Fidelia Orji, Evangelos Milios, Stan Matwin, and Rita Orji. 2020. Covid-19 pandemic: Identifying key issues using social media and natural language processing.

Adam Hale Shapiro, Moritz Sudhof, and Daniel J. Wilson. 2020. Measuring news sentiment. *Journal of Econometrics*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

S. Taj, B. B. Shaikh, and A. Fatemah Meghji. 2019. Sentiment analysis of news articles: A lexicon based approach. In *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*.

Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. 53(3).

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *CoRR*, abs/1506.06724.