Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 185 (2021) 302–310

Procedia Computer Science

www.elsevier.com/locate/procedia

Complex Adaptive Systems Conference Theme: Big Data, IoT, and AI for a Smarter Future
Malvern, Pennsylvania, June 16-18, 2021

Analysis of Patterns and Trends in COVID-19 Research

Christopher Dornick[a*], Amit Kumar[a*], Scott Seidenberger[a*], Elizabeth Seidle[a*†], Partha Mukherjee[a]

[a]*Pennsylvania State University, Great Valley, 30 E. Swedesford Road, Malvern, PA – 19355, USA*
**these authors contributed equally to this work*

Abstract

News and information surrounding the COVID-19 pandemic is ever-evolving and accumulating. Due to the global relevance and importance, it is critical to be able to parse through the available information in an efficient and reliant manner to gauge scientific progression and understandings surrounding COVID-19. In this research, abstracts from a corpus of scientific articles are evaluated using different Natural Language Processing (NLP) techniques, including Term Frequency-Inverse Document Frequency (TF-IDF), Latent Dirichlet Allocation (LDA), Bidirectional Encoder Representations from Transformers (BERT), and sentiment analysis, to better understand the breadth of extant literature. Results from the analyses show that in the very large corpus datasets, a large group of documents encompasses the overall or dominant general theme. However, the smaller clusters of documents reveal very precise and niche themes. Generalized COVID-19 is the dominant theme present in largest clusters. Smaller clusters include more specific terms (e.g., popular drugs, popular terms, key features/impacts related to COVID). With the resulting clusters, sentiment analysis was run to discover slight fluctuations over time depending on cluster with an overall relatively neutral sentiment. Overall, the precision of the BERT clusters distinguishes niche topics within the large corpus of literature and enables interesting and meaningful text analytics.

Keywords: TD-IDF; LDA; BERT; COVID-19; Topic-Modelling; Sentiment Analysis; Clustering; Text Mining; Text Analytics

†Corresponding author. Tel.: +1-814-933-7786.
E-mail address: ear5131@gmail.com

1877-0509
10.1016/j.procs.2021.05.032

1. Introduction

The COVID-19 pandemic has caused an unprecedented global stand-still. Consequently, COVID-19 is impacting a variety of areas spanning medicine, the environment, economics, social policy, and more, giving practically everyone a personal interest in the ongoing research and discoveries.

In the wake of the first global outbreak, the burgeoning scientific literature concerning COVID-19 is difficult to comprehend due to its volume and diversity. However, it is beneficial for people in various roles (e.g., policymakers, researchers, and business-owners) to understand the research trends with reference to COVID-19 for decision-making and planning.

In this case, a summative analysis of the extant literature about COVID-19 could be helpful to gauge progress, identify challenging areas, and possibly discover opportunities for innovation, strategies, or new research. An analysis in such a rapidly growing and changing area of research can help people appraise the status of the research and improve their general understanding for decision-making and planning. This study uses unsupervised NLP topic modelling and clustering techniques to uncover patterns and trends in COVID-19-related scholarly articles.

Specifically, this research will address:

- What are the underlying research topics?
- What is the sentiment of the overall research and the specific research topics?

2. Literature Review

Latent Dirichlet Allocation (LDA) models use a probabilistic method to cluster topics for documents and they perform well with long-length texts [1]. However, they require iterative hyperparameter tuning that can make the model performance vary. Additionally, LDA does not consider relations among the different underlying topics [1].

In information retrieval, term frequency–inverse document frequency (TF-IDF) is a probabilistic metric used to quantify word importance within a document or corpus [2, 3]. The TF-IDF is a multiple of the word frequency (TF) and the number of documents in the corpus that contain the word (proxy for word eliteness; IDF) [4]. The majority of recommendation systems behind text libraries use TF-IDF as their weighing scheme [5]

Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm that is uniquely suited to find clusters of arbitrary shapes. The algorithm is based on the density of data points, clustering closely packed points together [6]. HDBSCAN is a hierarchical conversion of DBSCAN used to extract clusters based on different density thresholds. [7]

Bidirectional Encoder Representations from Transformers (BERT) is the first deep, bidirectional pre-trained model on unlabeled text. It was developed by Google researchers in 2018 [8]. Traditional models are pre-trained unidirectional, left-to-right. The benefit of bidirectional is the model can better learn the relationship of words and detect the nuances of language. BERT generates numeric representations of its embeddings. The embeddings produced are a powerful data transformation tool that then provide much greater performance to subsequent algorithms that delivered clustering (e.g., HDBSCAN) and dominant topic extraction (e.g., LDA).

Transformer based, pre-trained models, have performed exceptionally well over the past few years in various NLP tasks [9]. DistilBERT is a version of the BERT framework that is lighter-weight (40% size reduced) but retains 97% of BERT performance to create easily interpretable topics and clusters in the documents [10]. Sentence-level transformers have proven to be successful in capturing whole document-level embeddings [11].

Prior COVID-19 text analysis has focused primarily on analyzing social media/public forum posts [12, 13]. In a similar study as our paper, Jelodar et al. conducted topic modeling and sentiment analysis related to COVID-19 reddit posts [14]. Our research differs because it is focused on analyzing scientific literature, one of the goals for the creation of the CORD-19 database [15].

3. Data Collection and Pre-Processing

3.1. Data Collection

The data are organized and made available by AI2, CZI, MSR, Georgetown, NIH and The White House for free on Kaggle [16, 17]. This dataset serves as a repository of scholarly articles about COVID-19, SARS-CoV-2, and other related coronaviruses. The publications' metadata are available in a downloadable csv file. As of September 15, 2020, the raw metadata file includes information about 253,545 articles.

3.2. Data Pre-Processing

The abstract text will be the source of text data for the analyses to provide additional and more nuanced information compared to just using the articles' titles. Additionally, articles in the data span a wide breadth of years (i.e., 1816 – 2021), with most articles being published in the past two years, and spanning 17 languages.

The following pre-processing steps were used to filter the raw dataset into the final dataset with 57,921 final articles, referred to as "BERT data" moving forward:

- Removed articles with publication year prior to 2019 to ensure high COVID-19 relevance.
- Removed articles without abstracts.
- Article text converted to lowercase and then removed articles with duplicate abstracts.
- Removed non-English articles.
- Removed articles without relevance to the pandemic. Articles without "corona", "sars" and/or "covid" in the title or abstract were excluded to avoid potentially irrelevant articles.
- Removed stop words, punctuation, special characters, URLs, and numbers from the abstracts for some of the subsequent analyses.

Additional text pre-processing steps used for the subsequent TF-IDF and LDA exploratory analyses include, referred to as "highly-processed data" moving forward:

- Tokenize, stem and lemmatize abstract text.
- Remove high content words to avoid biasing the topic clusters towards general subject matter. These words are ubiquitous not because they are meaningful, but rather they are the topic of our study. Words include: covid, sarscov, coronavirus, and virus.
- Remove verbs and adjectives using parts of speech (POS) tagging with the NLTK Python library.

4. Exploratory Data Analysis

Prior to performing the additional text cleaning descriptive text statistics were gathered about the articles' abstracts. Abstracts in the corpus of 57,921 articles were on average 198 words and on average 1164.3 characters. Abstracts were on average 8 sentences and words within abstracts were on average 5.9 characters long. TF-IDF and LDA are performed to evaluate the corpus for word importance and exploratory cluster topics. Results to the LDA will be compared to the results from the BERT analysis in the conclusion.

4.1. Term Frequency-Inverse Document Frequency (TF-IDF) Model

A TF-IDF analysis was used on the highly-processed data to identify the most relevant words and not necessarily the most frequent words within our data. TF-IDF was run on the highly-processed data. Each word's TF-IDF value within each abstract was summed and ranked in descending order, see Table 1.

Table 1. TF-IDF Top 25 Words in Abstracts.

Word (Rank)				
patient (1)	sever (6)	test (11)	result (16)	effect (21)
infect (2)	health (7)	care (12)	hospit (17)	may (22)
disea (3)	use (8)	respiratori (13)	treatment (18)	Include (23)
case (4)	studi (9)	risk (14)	model (19)	acut (24)
pandem (5)	clinic (10)	report (15)	data (20)	symptom (25)

Additional important words were identified by filtering further using TF-IDF modeling and varying min_df and max_df parameters (High frequency = 25% - 50% of articles, medium frequency = 10% - 25%, low frequency = 0.01% - 1% of articles). A TF-IDF analysis is also used to evaluate the final BERT clusters, as mentioned in the results section.

4.2. Latent Dirichlet Allocation (LDA) Model

Two exploratory LDA models were used on the highly-processed data. LDA is subject to parameter tuning (number of passes, number of latent topics, iterations, alpha, etc.). One LDA model evaluated and classified abstracts into 5 unique topics and the other LDA model evaluated and classified abstracts into 4 unique topics. Both models use 10 dominant or keywords, as reported in the table 2 below. The LDA using 5 topics seems to be potentially more useful because – apart from topic 1 which appears to perform like a residual miscellaneous category - the categories appear to have more distinct subject areas.

Table 2. LDA Models' Keywords

Topic	Keywords: 5 Topic Classifications	Keywords: 4 Topic Classifications
0	Health, system, world, survey, person, research, effect, state, also, may	Treatment, drug, effect, may, trial, cancer, also, system, could, role
1	Group, cancer, age, food, school, treatment, year, loss, news, or	Case, day, rate, age, number, death, year, time, may, period
2	Gene, site, two, drug, network, three, also, region, could, may	Group, gene, two, three, also, panel, agreement, brain, could, fusion
3	Treatment, effect, drug, may, trial, also, system, could, role, research	Health, system, world, research, effect, survey, also, may, person, time
4	Case, day, rate, number, death, age, time, year, period, may	

5. Methodology

5.1. BERT Embeddings for Clustering and Topic Generation

In our study the abstracts were processed with a sentence transformer that uses DistilBERT [10] as the underlying pre-trained language model. This transformer will tokenize and process a string of arbitrary length and transform it into a numeric vector of 768 elements. This vector is a contextual representation of the string that is processed through the pre-trained DistilBERT model. The base abstract was used as the string passed to the transformer, because the transformer and pretrained BERT model is trained to represent the words and sentences in context. Calculating the embeddings was computationally expensive, and performed on a cloud GPU cluster, where the embedding vector was then pushed to the MongoDB database for each document.

The next step is to employ the embeddings as a numeric representation of the abstracts. The first step in topic generation is to perform a dimensionality reduction on the data through Uniform Manifold Approximation and Projection (UMAP) [18]. It is an algorithm that is similar to t-SNE, but better for general non-linear dimensionality reduction. The 768-dimensional BERT embeddings (768 is the number of BERT vectors generated from data) are too high dimensional for efficient clustering. The UMAP reduced the 768-dimension space for each abstract to 10 dimensions. Figures 1 and 2 are 2-dimensional UMAPs of the data for visual representation.

Density-based HDBSCAN algorithm was used with the UMAP processed data to make the initial clusters. The minimum number of clusters to pass to the HDBSCAN is the main topic of model evaluation explored next in section 5.2. The application of cluster validity indicators to determine an adequate number of clusters was not considered because our goal was not to identify all the possible/optimum clusters in the whole dataset but only the top ones to identify some of the most important topics. The desired number of top clusters was identified by varying "min_cluster_size". The dimensionality of the data set had already been reduced to desired optimum number before clustering level also using UMAP. Finally, topic generation was accomplished through a class-based approach to TF-IDF similar to the approach discussed above. All documents assigned to a class will then be part of a "single document" that will have a TF-IDF performed on it to find the most important words associated with what makes that class distinct from others. The combination of the top ranked words will then be used to create the human-interpretable topic.

5.2. BERT Model Evaluation

Since this is an unsupervised task that is dealing with complex medical data, instead of numeric model performance metrics, we had two doctor of medicine students compare the outcomes of our LDA results with, and without BERT embeddings. Their domain knowledge was critical in understanding how the minimum cluster size for the HDBSCAN affected the output of the LDA by cluster on the BERT embedded abstract information. Using the macro view as the starting point, where the min cluster size was set at 100, the microscopic view had to have a min cluster size that produced both interpretable and useful clusters for specific analysis. As the number of clusters increases geometrically (refer Table 3) with a decreasing min cluster size, a few different options were presented for validation to the domain experts. These min cluster sizes were 60, 30, and 15.

The two raters were given the cluster results, along with the 4 dominant words extracted from each structure per the LDA. They were instructed to flag each cluster by whether they deemed the 4 dominant words for the cluster as "Interpretable", that is, there is a coherent relationship in the dominant topic words for the cluster. Next, if a cluster was flagged as "Interpretable", they had to determine if the cluster was "Useful". If a set of words was too specific or too vague where the rater wouldn't have been able to use those dominant words to understand the cluster relationship in a useful context, they would flag the cluster with a "0" for "Useful". The r Cohen's Kappa of Interpretability and Cohen's Kappa of Usefulness are calculated to evaluate performance on the cluster size that yielded the highest mean interpretability and usefulness.

The results from the qualitative validation by the domain experts are provided in Table 3.

Table 3. BERT Model Validation.

Min Cluster Size	Number of Clusters	Mean Interpretability	Cohen's Kappa Interpretability	Mean Usefulness	Cohen's Kappa Usefulness
60	39	0.641026	-	0.564103	-
30	64	0.90625	-	0.671875	-
15	147	0.918367	0.657	0.843537	0.357

The Interpretability inter-rater reliability score of 0.657 is "Substantial Agreement" [15]. The Usefulness inter-rater reliability score of 0.357 is "Fair Agreement." These results show that both raters scored the lower min cluster size to be both more interpretable and more useful. The mean useful score of 0.844 shows that although both raters agree in the magnitude of useful clusters, they had less agreement on specifically which clusters were useful. Still, the level of agreement is appropriate to assess the model valid to address the research questions.

5.3. Sentiment Analysis

Each abstract was analyzed for sentiment polarity using a rule-based method and assigned a continuous score between [-1.0, 1.0] using sentiment function from the TextBlob Python library. Negative numbers being negatively valence, positive numbers being positively valence, and 0 being neutral. This algorithm has very little computing expense which is advantageous when working with a large corpus. Additionally, because ratings are on a continuous

scale there is more nuance provided in the data than what would present if using a binary sentiment classification (e.g., positive or negative). This is particularly helpful when analyzing a more scientific and objective corpus, where sentiment may be more subdued.

6. Results

6.1. BERT Clusters

As discussed in section 5.1, two different UMAPs on the BERT embeddings were run. The first to capture the macroscopic topographical structure of the documents, where the dimensionality was reduced to two components and the local neighborhood size was 100 documents. The second UMAP was run to further isolate distinct clusters from the overall superstructure discovered by the first UMAP, to explore the nuanced differences in the densities of documents. The second UMAP reduced dimensionality to 5 and had a local neighborhood size of 50 documents.

The UMAP embeddings were then used as inputs for HDBSCAN clustering. The HDBSCAN algorithm is density based, stable, and requires little hyperparameter input. The only input is the minimum cluster size, which has been set to 15 documents.

The following figures, 1 and 2, show the cluster membership by color.

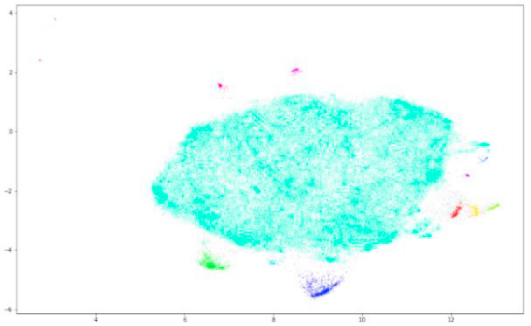

Fig. 1. Macroscopic topic density structure.

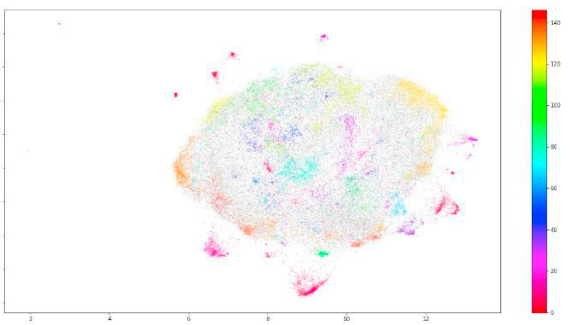

Fig. 2. Microscopic topic density structure.

The macroscopic topographical structure shows that most of the documents fall into a superstructure, with some outlying clusters. The microscopic topographical structure shows that within the main superstructure, there are distinct clusters of other topics. To extract the topics from the clusters, an LDA for each cluster was performed. The LDA algorithm used is the same LDA model used in the previous LDA model section. There will be a single dominant topic assigned to each cluster, with the 4 top words representing the dominant topic displayed (see Tables 5 and 6 below). Comparing Tables 4 and 5 versus Table 2, it is clearer to discern a relationship among the dominant words and, therefore, easier to qualitatively assign a theme to each cluster. This is supported by the Cohen's Kappa values discussed in section 5.2 and illustrates the power of BERT versus solely performing LDA on the highly processed dataset.

Table 4. Sample of large clusters from the macroscopic view.

Selected Cluster Number	Number of Documents	Dominant Topic Words	Additional Words
7	53114	sarscov, sever, case, health	respiratori, report, care, treatment, virus
4	1113	pregnanc, neonat, nur, report	matern, mother, delivery
9	739	health, region, number, epidem	outbreak, measur
13	257	clinic, type, gluco, mortal	hospit
11	238	bmi, sarscov, p, mortal	diabet, age, higher, outcome
10	186	eat, chang, system, secur	countri, lockdown, consumpt, insecur
2	181	control, dentist, treatment, oral	

Table 5. Sample of clusters from the microscopic view from within the main superstructure.

Selected Cluster Number	Number of Documents	Dominant Topic Words
119	145	distanc, lockdown, rate, interv
130	140	specimen, day, clinic, saliva
144	136	use, test, effect, result
4	118	model, estim, iranian, epidem
143	103	lesion, p, featur, case
30	98	worker, healthcar, sarscov, respond
21	97	phase, pulmonari, complic, may
135	94	group, v, care, hydroxychloroquin
75	280	pm, case, china, concentr
104	287	depress, fear, individu, peopl

6.2. Sentiment Analysis

The sentiment analysis was initially performed on the highly-processed abstract data. Most articles were relatively neutral sentiment. The relationship between sentiment ratings and time were tested using a coefficient of determination with R2 <0.01 indicating no association between the two variables.

Using the sentiment polarities from the highly-processed data and the BERT cluster labels for each article, the majority of articles are relatively neutral regardless of topic cluster and the average sentiment across clusters do not differ (see Figure 3). For all clusters, there are a handful of outliers with more extreme sentiment ratings, mimicking a leptokurtic distribution. While it may appear that some clusters are more extreme, the number of outliers is proportionate to the number of articles within the cluster. For example, cluster 7 encompasses 93% of the articles, while cluster 12 encompasses only 19 total articles. Similar distributions and patterns were found for the microscopic BERT clusters.

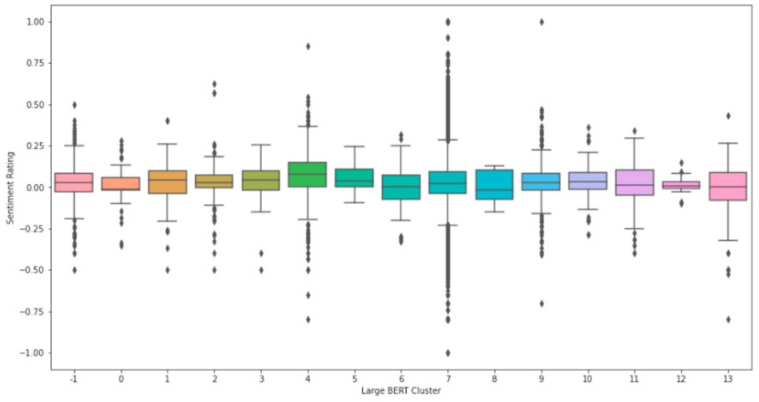

Fig. 3. Boxplot of sentiment ratings by Macroscopic BERT clusters.

No linear association or clear non-linear pattern between time and sentiment exists when looking at the corpus as a whole or when evaluating the corpus by BERT clusters. There are very few consistent patterns across clusters and trends do not appear to exist at this time. Perhaps as more data are collected across outbreak seasons, trends may start to appear. Despite no obvious patterns or trends, a few clusters experienced abrupt changes in average sentiment at specific points in time. This data could be used to investigate fluctuations in sentiment by topic areas for specific points in time. Data should then be interpreted based on what was happening within that timeframe.

7. Discussion

The ability of the BERT model to distinguish niche topics within the large corpus of literature allows for meaningful and potentially useful categorizations. By creating specific clusters, topic popularity and relevance can be measured. A sample of the top three inferred topics based on dominant words utilizing TF-IDF follows: Topic – “Key

Characteristics", Dominant Words – "lockdown", "fear", "depression", "ICU"; Topic – "Origin, Spread, Control", Dominant Words – "China", "test", "isolation"; Topic – "Drugs and Testing", Dominant Words – "remdesivir", "chloroquine", "RT-qPCR". Similarly, the topics can be further analyzed for sentiment and/or temporal changes. For example, the small BERT cluster 78, related to finance and economic related subjects, experiences a drastic change to a more positive sentiment during the month of August. When looking at articles from August for the small BERT cluster 78, words like 'gold' and 'monetarily' appear alongside the other frequent keywords like 'oil' and 'finance'. Indicating that perhaps this coincides with, or at least follows shortly after, the uptake in the market during that month [20].

COVID-19 is impacting a variety of areas spanning medicine, the environment, economics, social policy, and more; giving practically everyone a personal interest in the ongoing research and discoveries. The average person does not have the time or the domain knowledge to read through different scientific articles to get reliable information, especially when information is being discovered and disseminated so rapidly. The techniques used for understanding the COVID-19 research have the potential to reveal important challenge areas, the status of existing research, an understanding of research progress, and possible opportunities for innovation or new research.

8. Limitations and Future Research

The natural language processing algorithms utilized were computationally expensive to perform on our dataset. Therefore, increasing the MongoDB storage capacity was necessary to query the text data and to perform the analyses in the study. Despite the evaluation metrics validating the study, model parameters were not extensively optimized for the HDBSCAN, LDA, and TF-IDF during the tuning process. The framework is in place to easily adjust parameters and introduce updated research papers to the dataset in the future. As the focus of the research was primarily focused on text mining within a large corpus, latest real-time data was not prioritized for the analyses. The data used for the study was collected in early fall 2020 and COVID-19 research is fast evolving. That being stated, real-time data could be used to update the models developed in this study. As more time with COVID-19 lapses, NLP and other possible machine learning techniques can be used to evaluate the COVID-19 scientific literature for trends and patterns in sentiment, emotional classifications, topic areas, etc.

Increased processing and storage bandwidth, allowing for fine tuning and optimization of the parameters, and using real-time data are potential improvements to our study. Additionally, unused fields from the publications' metadata, such as academic journal title, could be included for deeper analyses.

9. Conclusion

This study examined the extant COVID-19 scholarly articles to organize the large corpus into different topics and perform sentiment analyses. While accurate unsupervised text mining and analyses involve multiple levels of data filtering and processing as well as appropriate combination of various models and hyperparameter tuning to derive final set of useful information, the BERT clusters exhibited more specific and helpful topic modelling than the LDA clusters based on the highly-processed dataset. This paper exhibits utility of BERT clusters for conducting text analytics.

Acknowledgements

We would like to thank Vikram Eswar, M.D. and Ryan Guzek, M.D. for their domain knowledge and assistance with model validation.

References

[1] Lee, S., Song, J., and Kim, Y. (2010) "An empirical comparison of four text mining methods. *Journal of Computer Information Systems*, **51(1)**, 1-10.

[2] Aizawa, A. (2003) "An information-theoretic perspective of tf–idf measures." *Information Processing & Management*, **39(1)**, 45-65.

[3] Zhang, W., Yoshida, T., and Tang, X. (2011) "A comparative study of TF* IDF, LSI and multi-words for text classification." *Expert Systems with Applications*, **38(3)**, 2758-2765.

[4] Qaiser, S., and Ali, R. (2018) "Text mining: use of TF-IDF to examine the relevance of words to documents." *International Journal of Computer Applications*, **181(1)**, 25-29.
[5] Beel, J., Langer, S., Genzmehr, M., Gipp, B., Breitinger, C., and Nürnberger, A. (2013) "Research paper recommender system evaluation: a quantitative literature survey." *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, 15-22.
[6] Ester, M., Kriegel, H. P., Sander, J., and Xu, X. (1996) "A density-based algorithm for discovering clusters in large spatial databases with noise." *Kdd*, vol. 96, no. 34, 226-231.
[7] Campello, R. J., Moulavi, D., Zimek, A., and Sander, J. (2015) "Hierarchical density estimates for data clustering, visualization, and outlier detection." *ACM Transactions on Knowledge Discovery from Data* (TKDD), **10(1)**, 1-51.
[8] Chang, M.W., Devlin, J., Lee, K., and Toutanova, K. (2019) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv.org. https://arxiv.org/abs/1810.04805v2.
[9] Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., and Huang, X. (2020) "Pre-trained models for natural language processing: A survey." arXiv preprint arXiv:2003.08271.
[10] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019) "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108.
[11] Cohan, A., Feldman, S., Beltagy, I., Downey, D., and Weld, D. S. (2020) "Specter: Document-level representation learning using citation-informed transformers." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2270-2282.
[12] L. Li, W. Zhao, W. Mao, C. Ye, X. Chu, and Z. Chu. (2020) "Data Mining of Weibo for Public Sentiment Evolution regarding COVID-19." *2020 IEEE International Symposium on Product Compliance Engineering-Asia* (ISPCE-CN), Chongqing, China, 1-6. doi: 10.1109/ISPCE-CN51288.2020.9321848.
[13] S. Verma, A. Paul, S. S. Kariyannavar, and R. Katarya. (2020) "Understanding the Applications of Natural Language Processing on COVID-19 Data." *2020 4th International Conference on Electronics, Communication and Aerospace Technology* (ICECA), Coimbatore, 1157-1162. doi: 10.1109/ICECA49313.2020.9297490.
[14] H. Jelodar, Y. Wang, R. Orji, and S. Huang. (2020) "Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach." *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, 2733-2742, Oct. 2020. doi: 10.1109/JBHI.2020.3001216.
[15] Lu Wang, L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Wilhelm, C., Xie, B., and Kohlmeier, S. (2020) "CORD-19: The Covid-19 Open Research Dataset." ArXiv, arXiv:2004.10706v2.
[16] AI2, CZI, MSR, Georgetown, NIH, and The White House (2020) "COVID-19 Open Research Data Challenge (CORD-19)." Retrieved September 15, 2020 from https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge?select=metadata.csv
[17] Allen AI. (2020) CORD19, https://github.com/allenai/cord19/blob/master/README.md
[18] Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W., Ng, L. G., and Newell, E. W. (2019) "Dimensionality reduction for visualizing single-cell data using UMAP." *Nature Biotechnology*, **37(1)**, 38-44.
[19] Landis, J. R., and Koch, G. G. (1977) "The measurement of observer agreement for categorical data." *Biometrics*, 159-174.
[20] The Market Intelligence Desk Team with Market Insite. (2020). *August 2020 Review and Outlook*. Retrieved November 11, 2020 from https://www.nasdaq.com/articles/august-2020-review-and-outlook-2020-09-02