Table 9: Comparison of models for multi-label classification.

Model	Macro-F1	Weighted-F1	Jaccard	Accuracy
Only predicts labels				
LLDA	0.1188	0.2134	0.1543	0.0675
RoBERTa	0.5837	0.6993	0.6713	0.6010
CT-BERT	0.5860	0.7144	0.6884	0.6087
HateXplain	**0.6007**	**0.7153**	**0.6928**	**0.6294**
Predicts labels and explanations				
CAML	0.3449	0.5039	0.4365	0.3675
ExPred	0.5870	0.6924	0.6652	0.5715
Multi-Task	0.5775	0.7029	0.6761	0.5937
Multi-Task (w GRU)	**0.5991**	**0.7104**	**0.6855**	**0.6037**

We have also made publicly available the implementations of the benchmark methods tried on the dataset (as described in the next section) as well as some helper scripts (such as scripts for loading the dataset into appropriate Numpy arrays).

4 TASKS ON THE DATASET

In this section we shall discuss three tasks that can be directly applied to the dataset – (1) multi-label classification, (2) explainable classification, and (3) tweet summarization. We also apply several state-of-the-art methods for benchmarking each of the three tasks.

4.1 Multi-label classification

In the standard multi-label classification task, each data point (tweet in our case) has to be assigned to one or more classes (anti-vaccine concerns in our case). This is in contrast with the multi-class (or single-label) classification in which a data point has to be assigned with exactly one (out of many) classes.

Data splitting: For creating a benchmark, we split the tweet-set into a train, a validation and a test set in the ratio of 70-10-20. Since the distribution of classes is skewed (as seen in Table 3), a random split would cause a class imbalance. Thus we decided to split the data using the iterative stratification method [34, 39] that aims to balance the classes in a multi-label setting. We used the sk-multilearn library [40] to perform this stratification.

Metrics for Evaluation: Given the set of predicted and gold standard labels, we have used 4 different metrics to estimate the performance of the models. First we calculate the F1-score for all the 12 classes separately and find the (i) Macro-average, and (ii) Weighted-average (where the weights of the classes are proportional to the class frequencies). We also calculate the Jaccard similarity between the predicted label-set and the gold standard label-set over each tweet, and average the Jaccard similarity values over all tweets. Finally we also calculate the subset accuracy – for a particular tweet, a predicted set of labels is considered a match only if it *exactly* matches with the set of gold standard labels. All metrics were calculated using standard functions from the Scikit-Learn package [28].

Benchmarking methods: We experimented with representative methods from different families of multi-label classifiers. We try out a topic modelling method (Labeled LDA) since these are what have been used by prior works for understanding anti-vaccine opinions (as described in Section 2.1). We tried some basic transformer based classifiers, namely RoBERTa-Large (known to perform very well for several NLP tasks) and CT-BERT-v2 [22] which is a BERT-Large model pretrained on millions of COVID-19 tweets (known to give good classification performance on COVID-related tweets [29]). Next we modified the HateXplain [17] model to work for multi-label scenario by joining the explanations for different labels.

We also tried a few models which incorporate the explanations, either to predict only the labels or to predict the labels and generate explanations jointly. In this family of models, we tried the CAML [21] model which uses CNN and Attention mechanism and can be used to get explanations for all the labels. We also use a simplified version of the recently-developed ExPred [46] model, and modify it for a multi-label setting. This model generates explanations along with a label prediction auxiliary task and then predicts the labels again. The "*Multi-task*" model contains a shared CT-BERT encoder and two decoders which separately predict the labels and generate the explanations. We also try a variation of the Multi-Task model where we pass the token embeddings through a GRU first before generating the explanations.

For each of the above-stated methods, we used a batch size of 16 and a maximum sequence length of 128. For HateXplain in particular we obtained best results with a batch size of 2. The models were trained and validated on corresponding datasets for 20 epochs. The model that yielded the best Macro-F1 score on the validation dataset was saved, and this model was used to calculate the performance metrics on the test dataset.

Performances: The results of different methods are given in Table 9. It is seen that CT-BERT outperforms RoBERTa slightly due to the domain pretraining. In contrast, the LLDA model vastly underperforms, highlighting the limitations of the previous works which tried to use topic modelling for extracting concerns about vaccines. CAML does not perform too well because it uses word2vec embeddings which are known to not be as good as BERT embeddings. The HateXplain model seems to perform the best achieving a Macro-F1 score of 0.6007, while the Multi-Task model with GRU performs second best with a Macro-F1 score of 0.5991, while also generating explanations. In terms of class-wise F1 score, all the models seem to perform badly on the 'None' and 'Conspiracy' classes which contain some confusing arguments. The 'Country' and 'Religious' classes are also very sparse, which lead to models under-performing on these classes. Details are omitted for brevity.

4.2 Explainable classification

The next task is that of generating explanations along with the predicted labels. The explanation generation part is a sequence-labelling task, where for a given tweet text as input, each token in the text is to be marked if it is part of the explanation or not. This is to be repeated for each of the predicted labels for the tweet. Thus, for a given tweet, the output of the task will be (i) the predicted labels, and (ii) a list of binary vectors, one for each predicted label (where each vector is of the same length as the number of tokens in the tweet). In these binary vectors, an element being 1 implies that the corresponding token in the tweet-text is part of the explanation for the corresponding label.

Table 10: Comparison of models for explanation generation.

	Explanation metrics			*Tuple metrics*	
Model	**IOU-F1**	**Jaccard**	**Token F1**	**M-F1**	**Acc**
CAML	0.0774	0.1899	0.2886	0.0351	0.0498
ExPred	0.0560	0.1582	0.2487	0.0249	0.0156
Multi-Task	0.3476	0.3336	0.4154	0.2704	0.2854
Multi-Task (with GRU)	0.3015	0.2920	0.3451	0.2355	0.2724

Table 11: Comparison of models for Summarization task.

Model	**ROUGE-1 F1**	**ROUGE-2 F1**	**ROUGE-L F1**
Extractive models			
LexRank [7]	21.48	2.29	15.73
PacSum [47]	19.14	2.09	12.69
COWTS [33]	31.54	4.52	21.07
Abstractive models			
BART [15]	28.60	5.04	19.19
Pegasus [45]	25.20	5.16	20.00
T5 [31]	27.73	5.34	19.38

Metrics for evaluation: For evaluating the explanations generated by different methods, we use some of the hard selection metrics defined by DeYoung et al. [4]. Among these, the first metric IOU-F1 (Intersection-Over-Union F1) is useful for detecting partial matches. The IOU is the size of the overlap between two sets of tokens divided by the size of the union of the two sets of tokens (also known as the Jaccard coefficient). For a particular tweet and label, a predicted explanation is counted as a match if and only if the IOU with the ground truth explanation is ≥ 0.5 [4]. These matches are then used to calculate the multi-label F1 score as described previously in Section 4.1. Additionally, the token-level F1 score and Jaccard coefficient is also calculated and averaged over all data points. We collectively refer to these metrics as *Explanation metrics*.

Separately, we also *evaluate the tuples of predicted labels and explanations together*. We consider a (label, explanation) tuple to be a match if and only if the predicted label is present in the gold standard set of labels and the predicted explanation overlaps at least 50% with the corresponding gold standard explanation (IOU ≥ 0.5). The Macro-F1 and the Accuracy is then calculated over these matches as before. We refer to these metrics as *Tuple metrics*.

Benchmarking methods: For this task, we use the same models which generate explanations (CAML, ExPred, Multi-Task and Multi-Task w GRU), as described in the previous subsection on multi-label classification. We also use the same data splits along with the experimental settings. Hence we are omitting these details here for brevity. It must however be noted that for the 'CAML' model, the explanations were extracted from the Attention layer weights, by taking the top-8 tokens with the highest weights. For the rest of the models, the explanations were predicted as a sequence-labelling task, with those tokens being generated that have the sigmoid of the logits ≥ 0.5 (similar to a multi-label prediction).

Performances: The metric scores on the test set for the explanation generation task are given in Table 10. The two variations of the Multi-task models seem to perform the best on these tasks, with the standard model achieving an IOU-F1 score of 0.3476 and the one with a GRU achieving 0.3015. Especially for the tuple metrics, these models comprehensively outperform the CAML and ExPred models. Similar to the multi-class classification scores, the IOU-F1 scores are low for the 'Country' and 'Religious' (sparse) classes.

4.3 Summarization

Multi-document summarization (a special case of which is tweet summarization) is a classical IR task that aims to produce a short summary of a large set of documents, that contains as much information content as possible while reducing redundant information. Summarization algorithms come in two flavours – i) Extractive, which select a few tweets out of all the tweets, and ii) Abstractive, which generates words to create a coherent summary like humans do. We have tried a few methods of each type to generate class-wise summaries and get some benchmark scores.

Metrics for Evaluation: For evaluation of the summaries, we use the popular ROUGE metrics. Specifically, we report the ROUGE-1 F1-score (that considers unigram matches between the gold standard summaries and an algorithm-generated summary), ROUGE-2 F1-score (that considers bigram matches) and the ROUGE-L F1-score (that considers longest sequence matches).

Benchmarking methods: We have employed a few popular summarization algorithms to benchmark our summarization dataset. Among extractive methods, we have used the graph-based LexRank summarizer [7], PacSum [47] that defines its own centrality measures, and COWTS [33] which is an Integer Linear Programming based summarizer designed for disaster-related tweets. Among abstractive methods, we have employed different *pretrained* transformer based encoder-decoder models such as T5 [31], BART [15], and Pegasus [45], that differ in their pre-training strategies.

For each class, we used the extractive models to generate summaries of around 20 tweets each, from the document consisting of all labeled tweets belonging to that class. Similarly, the abstractive models were used to generate class-wise summaries of around 250 words each. It is to be noted here that the pre-trained abstractive models have a limitation on the size of input documents that they can summarize. Hence, for each class, we first split the corresponding tweet dataset into almost equal-sized chunks such that the length of each chunk is less than or equal to the maximum permissible length. We obtain smaller summaries from each chunk and concatenate these to form a longer summary. The final summary of length around 250 words was obtained by greedily selecting the top-ranked sentences (from the longer summary) based on their TF-IDF scores w.r.t. all tweets in the corresponding class.

Performances: The ROUGE scores on the summarization task are given in Table 11. Among the extractive models, COWTS performs the best achieving a ROUGE-2 F1 score of 4.52 and a ROUGE-L F1 score of 21.07. The performances of the different abstractive models are similar. All models achieve especially low ROUGE-2 F1 scores, and it is a potential research direction to improve summarization performance over this dataset.

4.4 Summary of the section

We tried different state-of-the-art models on three tasks to establish benchmark results on the CAVES dataset. For the classification task the highest Macro-F1 score achieved was 0.6007, which is a moderate score considering it is a multi-label setting. Explanations generated were of decent quality with the best IOU-F1 of 0.3476, though the performance for some other models seem to be quite low. For the summarization task, the ROUGE-2 scores achieved were low. This may be due to various reasons such as the specific vocabulary used in the anti-vax text may be unique, repetition of similar concepts in multiple tweets, etc. These results suggest that the CAVES dataset and associated tasks pose interesting challenges, and more specific models need to be developed to tackle the tasks.

5 OTHER POTENTIAL USES OF THE DATASET

In this section, we highlight some potential applications of our proposed dataset, other than the ones we have discussed so far.

Distribution of concerns over time: Our dataset can be used to study the change in the distribution of different anti-vaccine concerns over time (and for training models to track such changes when applied on large scale Twitter data). In Figure 3, we show the month-wise frequency distribution of the largest 6 classes in our labeled set of tweets (in terms of the number of tweets in a class). We observe some interesting peaks in the figure which can be mapped to certain real-world events (e.g., as reported in the AJMC articles on COVID vaccine developments throughout 2020 [37] and 2021 [38])–

- There is a spike in the *Pharma* class around September 2020 which could be explained by two events – Pfizer expanding phase 3 trials of its vaccine, and Astrazeneca trial halting due to complications faced by a patient.
- There is a spike in the *Unnecessary* class in October 2020, which may be due to the FDA's approval of Remdesivir as a COVID-19 drug (which made many people feel that vaccines are unnecessary).
- The *Rushed* class has a spike in November 2020 likely due to Pfizer and AstraZeneca reporting completion of their trials.
- The *Side-effects* class has the peak in April-May of 2021, likely due to some adverse reactions to the Johnson & Johnson vaccine being reported at the end of April.
- The spike in the *Ineffective* class in July 2021 is likely owing to the reports of the Pfizer Vaccine not being effective against the delta variant of COVID-19.

It is to be noted here that the labeled set of tweets in the CAVES dataset might not be fully representative of the actual temporal distribution of tweets, due to inadvertent biases that might have crept in as part of the various steps we took for selecting the tweets. Hence, one should be careful in drawing strong trends or temporal conclusions just by analyzing the labeled dataset.

Generating highlights for explainable summarization: The benchmark methods we have used in Section 4.3 are standard summarization models that work only with the tweet texts. Our dataset can facilitate designing of models that utilize the explanations to improve the summarization task. Similar to Wang et al. [43], a select-then-generate framework could be designed that first highlights the reasoning spans as explanations and then generates a summary while focusing on the highlighted spans. Such as approach can not only improve the interpretability of extractive summarization models, but can also provide suitable explanations behind generation of particular phrases in case of abstractive models.

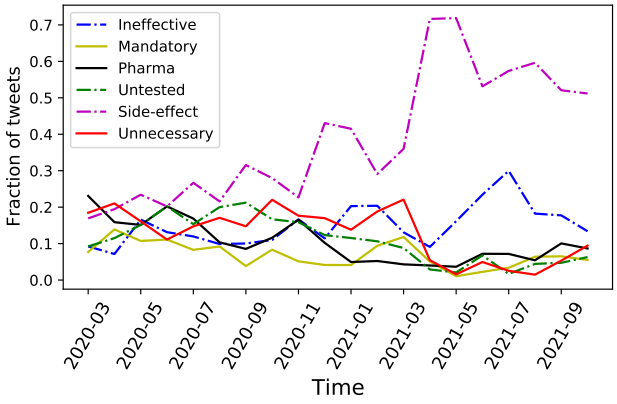

Figure 3: Frequency distribution of tweets corresponding to different concerns over time.

Conspiracy detection: Our dataset contains a set of tweets related to conspiracies around COVID-19 vaccines, as reported in Table 3. The dataset can thus be used to benchmark automatic COVID-19 conspiracy theory detection models such as [35]. Our dataset can further facilitate the design and evaluation of multi-task models that not only detect conspiracy-related tweets but also generate explanations for the same.

6 CONCLUSION

We have built a dataset of tweets that is important from a societal standpoint as it identifies concerns that people have towards vaccines, as well as facilitates explainable classification in a multi-label setting. The dataset also contains summaries of different classes and hence can be used to develop or test summarization algorithms. We have provided some benchmark results on the three different primary tasks, and discussed some other potential retrieval tasks.

The benchmark results point towards the need for improved, customized models for addressing the tasks. For example, apart from the tweet texts, the models can potentially incorporate additional (meta) information from the tweets or the users who posted the tweets to improve scores. Given the timely importance of the CAVES dataset, we believe it will instill enough interest within the community in the near future, to develop better methods for the proposed tasks.

ACKNOWLEDGMENTS

The project is partially supported by research grants from Accenture Corporation and DRDO, Government of India (through the research project titled "Claim Detection and Verification using Deep NLP: an Indian Perspective"). S. Poddar is also supported by the Prime Minister's Research Fellowship (PMRF) from the Ministry of Education, Government of India.

REFERENCES

[1] Erika Bonnevie, Allison Gallegos-Jeffrey, Jaclyn Goldbarg, Brian Byrd, and Joseph Smyser. 2021. Quantifying the rise of vaccine opposition on Twitter during the COVID-19 pandemic. *Journal of communication in healthcare* 14, 1 (2021), 12–19.

[2] Ziqiang Cao, Chengyao Chen, Wenjie Li, Sujian Li, Furu Wei, and Ming Zhou. 2016. Tgsum: Build tweet guided multi-document summarization dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[3] Liviu-Adrian Cotfas, Camelia Delcea, Ioan Roxin, Corina Ioanăş, Dana Simona Gherai, and Federico Tajariol. 2021. The longest month: Analyzing covid-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement. *IEEE Access* 9 (2021), 33203–33223.

[4] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4443–4458.

[5] Kuldeep Dhama, Khan Sharun, Ruchi Tiwari, Manish Dhawan, Talha Bin Emran, Ali A Rabaan, and Saad Alhumaid. 2021. COVID-19 vaccine hesitancy–reasons and solutions to achieve a successful global vaccination campaign to tackle the ongoing pandemic. *Human Vaccines & Immunotherapeutics* 17, 10 (2021), 3495–3499.

[6] Soumi Dutta, Vibhash Chandra, Kanav Mehra, Asit Kumar Das, Tanmoy Chakraborty, and Saptarshi Ghosh. 2018. Ensemble Algorithms for Microblog Summarization. *IEEE Intelligent Systems* 33, 3 (2018), 4–14.

[7] Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research* 22 (2004), 457–479.

[8] Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics* 9 (2021), 391–409.

[9] Alexander R Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749* (2019).

[10] Keith Gunaratne, Eric A Coomes, and Hourmazd Haghbayan. 2019. Temporal trends in anti-vaccine discourse on Twitter. *Vaccine* 37, 35 (2019), 4867–4871.

[11] Ruifang He, Liangliang Zhao, and Huanyu Liu. 2020. TWEETSUM: Event oriented social summarization dataset. In *Proceedings of the 28th International Conference on Computational Linguistics*. 5731–5736.

[12] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.

[13] Neil F Johnson, Nicolas Velásquez, Nicholas Johnson Restrepo, Rhys Leahy, Nicholas Gabriel, Sara El Oud, Minzhang Zheng, Pedro Manrique, Stefan Wuchty, and Yonatan Lupu. 2020. The online competition between pro-and anti-vaccination views. *Nature* 582, 7811 (2020), 230–233.

[14] David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research* 5, Apr (2004), 361–397.

[15] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.

[16] Irene Li, Tianxiao Li, Yixin Li, Ruihai Dong, and Toyotaro Suzumura. 2021. Heterogeneous Graph Neural Networks for Multi-label Text Classification. *arXiv preprint arXiv:2103.14620* (2021).

[17] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14867–14875.

[18] Richard McCreadie, Cody Buntain, and Ian Soboroff. 2019. Trec incident streams: Finding actionable information on social media. (2019).

[19] Tanushree Mitra, Scott Counts, and James W Pennebaker. 2016. Understanding anti-vaccination attitudes in social media. In *Tenth International AAAI Conference on Web and Social Media*.

[20] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*. 1–17.

[21] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of NAACL-HLT*. 1101–1111.

[22] Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503* (2020).

[23] Martin M Müller and Marcel Salathé. 2019. Crowdbreaks: tracking health trends using public social media data and crowdsourcing. *Frontiers in public health* 7 (2019), 81.

[24] Minh-Tien Nguyen, Dac Viet Lai, Huy Tien Nguyen, and Minh Le Nguyen. 2018. Tsix: a human-involved-creation dataset for tweet summarization. In *Proc. International Conference on Language Resources and Evaluation (LREC)*.

[25] Tasmiah Nuzhath, Samia Tasnim, Rahul Kumar Sanjwal, Nusrat Fahmida Trisha, Mariya Rahman, SM Farabi Mahmud, Arif Arman, Susmita Chakraborty, and Md Mahbub Hossain. 2020. COVID-19 vaccination hesitancy, misinformation and conspiracy theories on social media: A content analysis of Twitter data. (2020).

[26] Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. (2006).

[27] Elise Paul, Andrew Steptoe, and Daisy Fancourt. 2021. Attitudes towards vaccines and intention to vaccinate against COVID-19: Implications for public health communications. *The Lancet Regional Health-Europe* 1 (2021), 100012.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[29] Soham Poddar, Mainack Mondal, Janardan Misra, Niloy Ganguly, and Saptarshi Ghosh. 2022. Winds of Change: Impact of COVID-19 on Vaccine-related Opinions of Twitter users. In *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM'22)*.

[30] SV Praveen, Rajesh Ittamalla, and Gerard Deepak. 2021. Analyzing the attitude of Indian citizens towards COVID-19 vaccine–A text analytics study. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 15, 2 (2021), 595–599.

[31] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

[32] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[33] Koustav Rudra, Subham Ghosh, Niloy Ganguly, Pawan Goyal, and Saptarshi Ghosh. 2015. Extracting situational information from microblogs during disaster events: a classification-summarization approach. In *Proceedings of the 24th ACM international on conference on information and knowledge management*. 583–592.

[34] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 145–158.

[35] Shadi Shahsavari, Pavan Holur, Tianyi Wang, Timothy R Tangherlini, and Vwani Roychowdhury. 2020. Conspiracy in the time of corona: Automatic detection of emerging COVID-19 conspiracy theories in social media and the news. *Journal of computational social science* 3, 2 (2020), 279–317.

[36] Kalyani Sonawane, Catherine L Troisi, and Ashish A Deshmukh. 2021. COVID-19 vaccination in the UK: Addressing vaccine hesitancy. *The Lancet Regional Health–Europe* 1 (2021).

[37] AJMC Staff. 2021. A Timeline of COVID-19 Developments in 2020. *AJMC* (2021). https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020

[38] AJMC Staff. 2021. A Timeline of COVID-19 Vaccine Developments in 2021. *AJMC* (2021). https://www.ajmc.com/view/a-timeline-of-covid-19-vaccine-developments-in-2021

[39] Piotr Szymański and Tomasz Kajdanowicz. 2017. A network perspective on stratification of multi-label data. In *First International Workshop on Learning with Imbalanced Domains: Theory and Applications*. PMLR, 22–35.

[40] Piotr Szymański and Tomasz Kajdanowicz. 2017. A scikit-based Python environment for performing multi-label classification. *arXiv preprint arXiv:1702.01460* (2017).

[41] Ke Tao, Fabian Abel, Claudia Hauff, Geert-Jan Houben, and Ujwal Gadiraju. 2013. Groundhog day: near-duplicate detection on twitter. In *Proceedings of the 22nd international conference on World Wide Web*. 1273–1284.

[42] Gianmarco Troiano and Alessandra Nardi. 2021. Vaccine hesitancy in the era of COVID-19. *Public health* 194 (2021), 245–251.

[43] Haonan Wang, Yang Gao, Yu Bai, Mirella Lapata, and Heyan Huang. 2021. Exploring Explainable Selection to Control Abstractive Summarization. In *Proc. AAAI Conference on Artificial Intelligence*. 13933–13941.

[44] Xiaoyi Yuan, Ross J Schuchard, and Andrew T Crooks. 2019. Examining emergent communities and social bots within the polarized online vaccination debate in Twitter. *Social media+ society* 5, 3 (2019), 2056305119865465.

[45] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*. PMLR, 11328–11339.

[46] Zijian Zhang, Koustav Rudra, and Avishek Anand. 2021. Explain and predict, and then predict again. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 418–426.

[47] Hao Zheng and Mirella Lapata. 2019. Sentence Centrality Revisited for Unsupervised Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 6236–6247.