# Curating corpora with classifiers: a case study of clean energy sentiment online

Michael V. Arnold,[1,*] Peter Sheridan Dodds,[1,2] and Christopher M. Danforth[1,3]

[1] *Computational Story Lab, Vermont Complex Systems Center, MassMutual Center of Excellence for Complex Systems and Data Science, Vermont Advanced Computing Core, University of Vermont, Burlington, VT, USA*
[2] *Department of Computer Science, University of Vermont, Burlington, VT, USA*
[3] *Department of Mathematics & Statistics, University of Vermont, Burlington, VT, USA*

(Dated: February 7, 2023)

The ubiquity of social media posts containing broad public opinion offers an alternative data source to complement some shortcomings of traditional surveys. While surveys collect representative samples and achieve relatively high accuracy, they are both expensive to run and lag public opinion by days or weeks, which could be overcome with a real-time data stream and fast analysis pipeline. One challenge in this pipeline we seek to address is selecting the best corpus of relevant documents for analysis. Querying with keywords alone often includes irrelevant documents that are not easily disambiguated with bag-of-words NLP methods. We explore methods of corpus curation to filter irrelevant tweets using transformer-based sentence embedding models, fine-tuned for our binary classification task on hand-labeled tweets, and achieving F1 scores of up to 97%. The low cost and high performance of fine-tuning such a model suggests it should be widely adopted as a pre-processing step for social media datasets with uncertain corpus boundaries.

## I. INTRODUCTION

☐ Add scintillating introduction.
☐ Computational social science runs on twitter data. 7.7M google scholar results for twitter, with over 140k in just 2022. How many of these studies query for a hashtag or keyword to build a corpus? How many would benefit from training a classifier to filter out irrelevant tweets?

The wide-spread availability of social media data has resulted in an explosion of social science studies as researchers adjust from data scarcity to abundance in the internet age [1, 2].
☐ Analysis of social media data promises to supplement traditional polling methods by allowing for rapid, near real-time measurements of public opinion, and allowing for historical studies of public language. Polling remains the gold standard for measuring public opinion where precision matters, [3]
☐ Question

When researchers characterize online discourse around a specific topic, a few approaches are available. Each comes with trade-offs, both in researchers' time as well as the resulting precision and recall of the corpus.

Querying for relevant hashtags, which signal an author's intent to attach their post to broader topic, can match with relevant posts with high precision, but often low recall. Hashtag based queries have been used by researchers to construct focused corpora of tweets ranging from sports and music [4, 5], to natural disasters, political activism, and protests [6–13].

Alternatively, researchers can query for posts with an expansive set of keywords to increase recall at the expense of precision. Expanding this set of keywords can be done with by experts with domain expertise, algorithmically, or a combination. Broad expert crafted keyword lists have been used to study social responses to the COVID-19 pandemic [14–17]☐ Add more examples here?. Other researchers have used algorithmically generated lists of keywords, by comparing the distribution of words in a corpus of interest to a reference corpus and selecting words with high rank-divergence contributions [18–21]. But continued expansion beyond the most relevant keyword necessarily reduces precision. Researchers can further refine the set of relevant keywords to balance precision and recall, or add complexity to their queries with negation or Boolean operators to require multiple keywords.

While some social media datasets can be sufficiently curated with simple heuristics or rules-based classifiers, others could benefit from an alternative paradigm. We argue for a two step pre-processing pipeline that combines broad, high recall keyword queries with fine-tuned, transformer-based classifiers to increase precision. This approach can trade the labor costs associated with building rules-based filters, for the cost of labeling social media data for few-shot learning [22], while still achieving high precision.

Since the introduction of Word2Vec in 2013 and GloVe in 2014, the natural language processing community has had access to high quality, global word embeddings [23, 24]. These embeddings are trained vector representations of words from given a corpus of text, enabling word comparisons with distance metrics. However, global embeddings will average the representations of words, making them unsuitable for document classification where key terms have multiple meanings. The development of attention ☐ More about transformers ☐ specifically about sentence embeddings, including different ones we tried. While transformer based language models like BERT provide state of the art performance

* mvarnold@uvm.edu

on natural language processing tasks, they are computationally expensive to run. Creating sentence embeddings with Siamese, transformer networks, promises large improvements in computational efficiency for sentence classification, while maintaining state-of-the art performance [25].
☐ text classification for longer texts has been less successful [26]. Text classification with a large number of classes is also challenging [26][27]
☐ more of classifying tweets [28] [29]

Today, sophisticated, pre-trained language models are readily accessible to researchers through [30], and can be easily fine-tuned with a limited amount of labeled data, so [22, 31]
☐ more on renewable energy specifically [32, 33]

Although we focus on the topic of renewable energy, we hope our methods are broadly applicable to any text-based social media dataset.

## II. METHODS AND DATA

We explore the performance of contextual sentence embedding based text classifiers for corpus curation through a selection of case studies.

### A. Description of data sets

In this study we examine ambient tweet datasets, collections of tweets that are anchored by a single keyword or set of keywords. From Twitter's Decahose API, a 10% sample of all tweets, we begin by selecting tweets containing valid user-provided locations in the United States. From this selection, we query for tweets that both contain keywords of choice and are classified as English by FastText [34]. We define the results of this query as the unfiltered ambient corpus.

We chose three keywords related to non-fossil fuel energy generating technologies, `'Wind'`, `'Solar'`, and `'Nuclear'`. Over the study period from 2016 to 2022, these keywords matched 3.43M, 1.39M, and 1.29M tweets in our subsample, respectively.

While the terms of service with Twitter do not allow us to publish raw tweets, we provide relevant tweet IDs for rehydration. ☐ Add tweet_id extractor to git repo.

### B. Relevance classification

For classification we hand-labeled a random sample of 1000 matching tweets for each keyword as either 'relevant' (R) or 'not relevant' (NR) to energy production. We've made of tweet IDs and corresponding labels available for both the training data as well as predicted labels for the full dataset. ☐ add this

We trained three models for comparison, two contextual sentence embeddings, `all-mpnet-base-v2` and `all-MiniLM-L6-v2`, and one non-contextual model, created by averaging `GloVe` word embeddings. We've listed the performance of the models in Tabel I.
☐ Add some more model F1 scores, maybe a for L6, and for Glove if you can do it with Hugging Face, else go edit

| | Wind | Solar | Nuclear |
|---|---|---|---|
| % Relevant | 4.7% | 43.7% | 16.0% |
| F1 | 0.90 | 0.95 | 0.86 |

TABLE I.

### C. Corpus Comparison

☐ Corpus comparison

### D.

☐ Compare performance of contextual sentence embeddings to glove?
☐ Expand on fine-tuning and classification procedure.

## III. RESULTS

☐ Explain what we found
☐ Solar tweets were nearly evenly split with 47% of the corpus being relevant and 53% being not relevant by volume of words.
☐ plot of relevant vs non-relevant tweets in the embedding space. ☐ sentiment plot

### A. Solar energy case study

Of the three case studies we found the R `'solar'` tweets corpus evolved most relative to the corresponding NR corpus. Looking at the sentiment timeseries in 3, we see little difference between the ambient sentiment of the R and NR corpora prior to 2019, before NR ambient sentiment, shown in red, sharply falls as the R corpus appears to remain on trend. For the standard deviation of ambient sentiment, which measures how broad is the distribution of sentiment scores for each LabMT word in the ambient corpus, we also observe a dramatic increase in 2019.

We contend that the process of selecting relevant social media documents to include in a corpus is just as important as the NLP measurement tools used to quantify sentiment. The difference in resulting sentiment measurement, between what would have been measured without a classifier (the R + NR corpus in purple) and the improved measurement after filtering with a classifier (the R corpus in blue) is stark. Looking at only the combined R

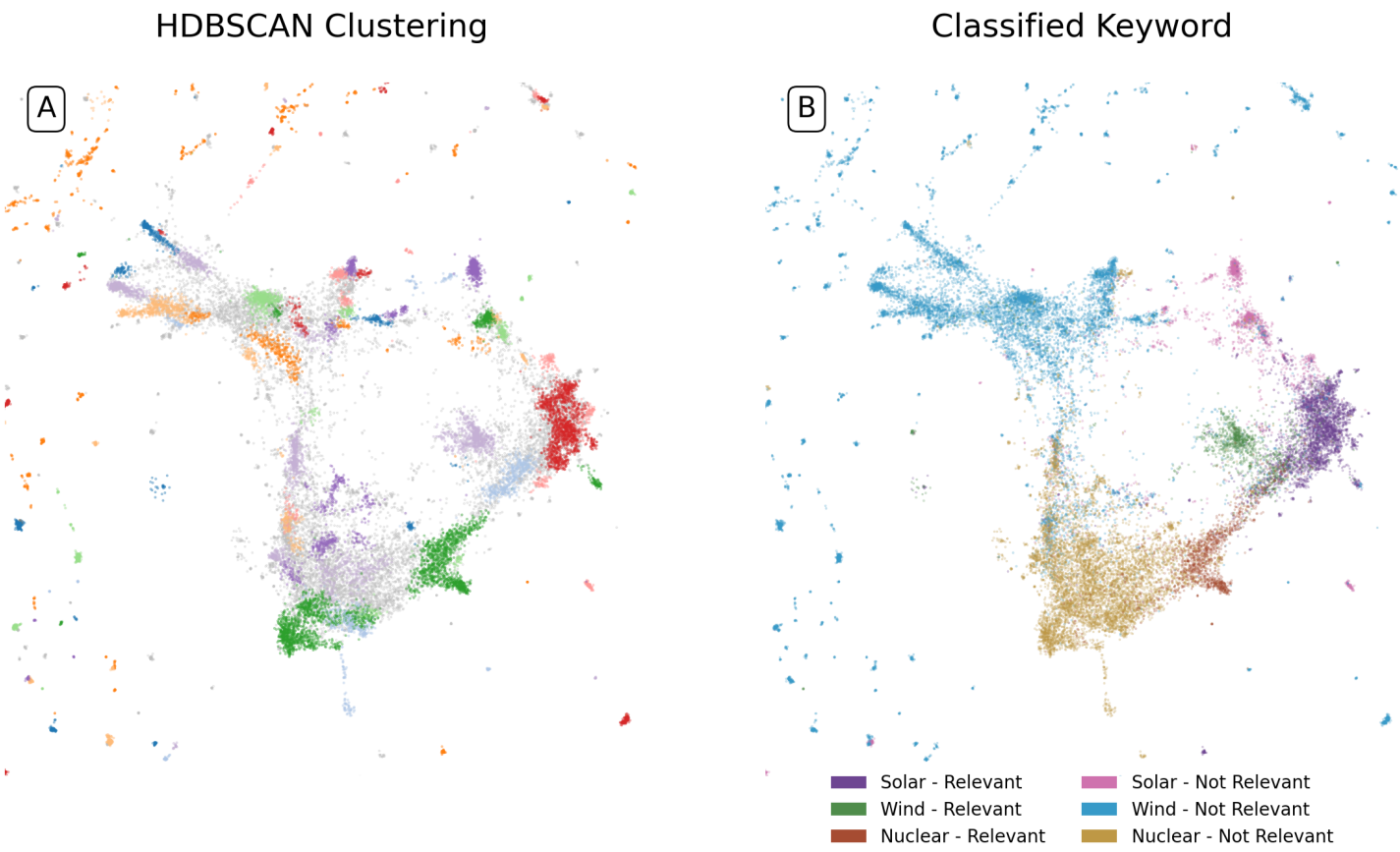


FIG. 1. **Embedded tweet distribution plot** for the combined datasets. Using a pre-trained model for semantically meaningful sentence embeddings, `all-mpnet-base-v2`, we plot the distribution of tweets within this semantic space. In both plots points are tweets projected into 2D using UMAP for dimensionality reduction. In panel A, we perform density based, hierarchical clustering using HDBSCAN and color by cluster. In panel B, we color by both the keyword used to query and the classification as relevant or not relevant to the topic of clean energy. Relevant tweets containing the keywords '`wind`', '`solar`', and, to a lesser extent, '`nuclear`' are relatively close together on the right in the embeddings, while not relevant tweets are more dispersed.

+ NR measurement, researchers could incorrectly conclude that language surrounding '`solar`' has decreased dramatically since 2019.

Focusing on only the R '`solar`' sentiment timeseries, it is clear that there was no dramatic drop in sentiment around '`solar`', and this language remains more positive relative to Enlish language tweets in general. ☐ sentiment timeseries
☐ sentiment shifts ☐ allotax

### B. Wind Energy Case Study

☐ The unclassified '`wind`' tweets corpus had the lowest proportion of revelant tweets; only 5% of the human labeled subset was related to clean energy. The n-gram '`wind`' is used in many different contexts besides energy generation, from the weather to figurative uses like 'second wind' and the phrase 'wind up.'

We found the ambient sentiment of the R '`wind`' corpus has been slightly more positive than average language use on Twitter.

### C. Nuclear Energy Case Study

☐

## IV. CONCLUDING REMARKS

☐ Bring it home.

☐ limitations, not a representative sample, etc While our contribution

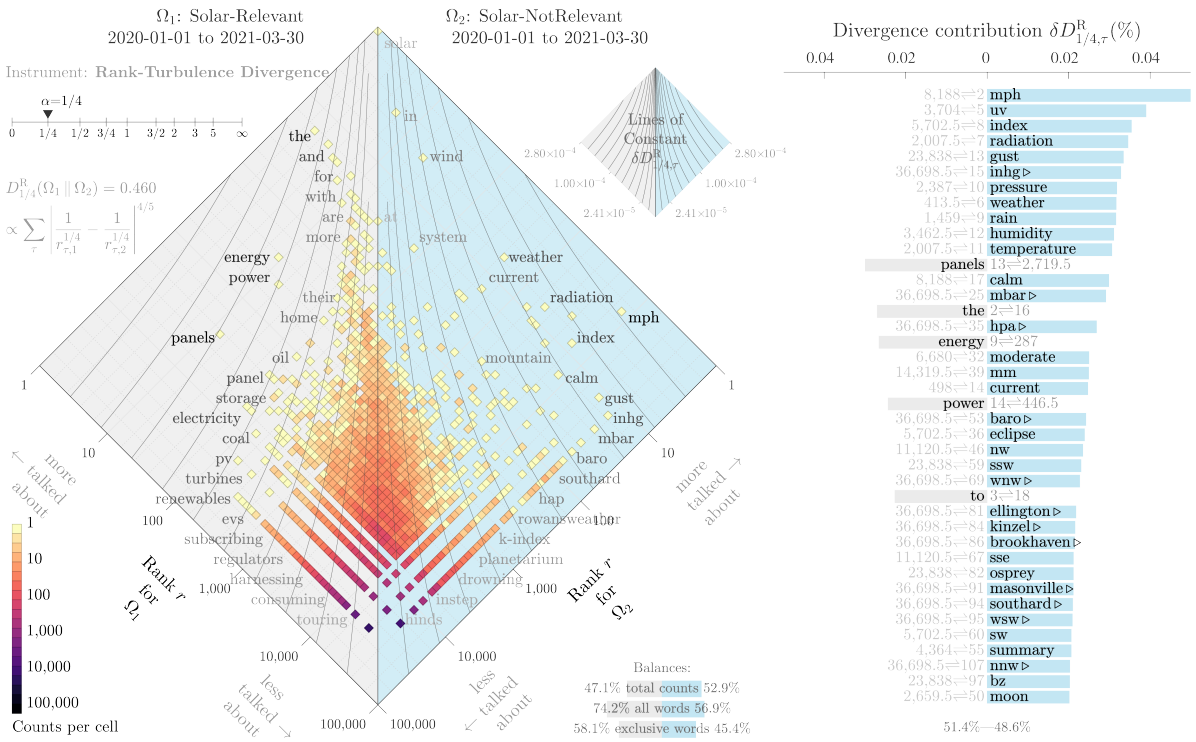


FIG. 2. **Allotaxonograph comparing the rank divergence of words classified as relevant to solar energy discourse to those containing the keyword 'solar' but classified as not relevant.** On the main 2D rank-rank histogram panel, words appearing on the right have a higher rank in the "relevant" subset than in "not relevant", while phrases on the left appeared more frequently in the "not relevant" tweets. The panel on the right shows the words which contribute most to the rank divergence between each corpus. We observe that many words associated with weather bots, such as "mph," "uv," and "pressure," are more frequently used in non-relevant posts, while words like "panels," "energy," and "power," used more in tweets relevant to solar energy. Notably, commonly used function words, such as "the," "and," and "are," are off-center in the rank-rank histogram, a further indication that many of the "not relevant" tweets are from automated accounts publishing weather data rather than using conversational English. The balance of the words in these two subsets is noted in the bottom right corner of the histogram, showing the percentage of total counts, all words, and exclusive words. For this example the two subsets are nearly balanced, indicating that the filtered corpus contains less than 50% of word tokens from the raw query. See Dodds *et al.* [18] for a full description of the allotaxonometric instrument.

## ACKNOWLEDGMENTS

We are grateful for . . .

[1] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Computational social science. Science, 323(5915):721–723, 2009.

[2] D. Lazer, E. Hargittai, D. Freelon, S. Gonzalez-Bailon, K. Munger, K. Ognyanova, and J. Radford. Meaningful measures of human society in the twenty-first century. Nature, 595(7866):189–196, 2021.

[3] E. M. Cody, A. J. Reagan, L. Mitchell, P. S. Dodds, and C. M. Danforth. Climate change sentiment on twitter: An unsolicited public opinion poll. PloS one, 10(8):e0136092, 2015.

[4] M. Blaszka, L. M. Burch, E. L. Frederick, G. Clavio, and P. Walsh. # worldseries: An empirical examination of a twitter hashtag during a major sporting event. International Journal of Sport Communication, 5(4):435–453, 2012.

[5] S. C. Choi, X. V. Meza, and H. W. Park. South korean culture goes latin america: Social network analysis of kpop tweets in mexico. International Journal of Contents, 10(1):36–42, 2014.

[6] Z. C. Steinert-Threlkeld, D. Mocanu, A. Vespignani, and J. Fowler. Online social networks and offline protest. EPJ Data Science, 4(1):1–9, 2015.

[7] G. Lotan, E. Graeff, M. Ananny, D. Gaffney, I. Pearce, et al. The arab spring— the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. International journal of communication,

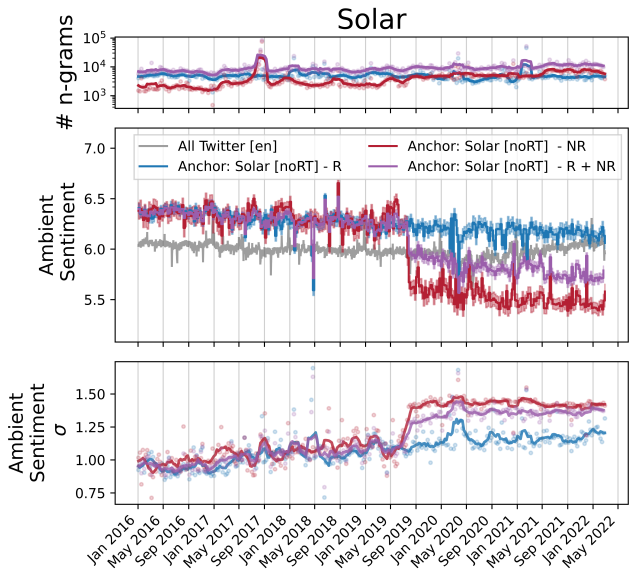


FIG. 3. **Ambient sentiment time series comparison for relevant (R), non-relevant (NR), and combined tweet corpora, all containing the keyword 'solar'.** In the top panel we show the number of tokens with LabMT [35] sentiment scores in each corpus on each day. 'Relevant' tweets, in blue, have more scored tokens early on, but the number tokens in 'not relevant' tweets increase in relative proportion over time. The center panel shows the average sentiment for each corpus, including measurement of English language tweets as a whole in gray for comparison. Before 2019, the measured sentiment for both corpora are comparable, but later the mean sentiment of 'non-relevant' tweets drops. In the bottom panel we plot the standard deviation of the sentiment measurement, which captures a broader distribution of sentiment scores for 'non-relevant' tweets. Without classification filtering, the ambient sentiment measurement would been misleading, appearing as though the sentiment contained in tweets containing the word 'solar' has dramatically dropped in 2019, when in fact sentiment has only modestly declined.

5:31, 2011.

[8] D. Freelon, C. D. McIlwain, and M. Clark. Beyond the hashtags:# ferguson,# blacklivesmatter, and the online struggle for offline justice. Center for Media & Social Impact, American University, Forthcoming, 2016.

[9] S. J. Jackson, M. Bailey, and B. F. Welles. #HashtagActivism: Networks of race and gender justice. Mit Press, 2020.

[10] R. J. Gallagher, E. Stowell, A. G. Parker, and B. Foucault Welles. Reclaiming stigmatized narratives: The networked disclosure landscape of# metoo. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW):1–30, 2019.

[11] R. J. Gallagher, A. J. Reagan, C. M. Danforth, and P. S. Dodds. Divergent discourse between protests and counter-protests:# blacklivesmatter and# alllivesmatter. PloS one, 13(4):e0195644, 2018.

[12] Y. Gorodnichenko, T. Pham, and O. Talavera. Social media, sentiment and public opinions: Evidence from# brexit and# uselection. European Economic Review, 136:103772, 2021.

[13] M. V. Arnold, D. R. Dewhurst, T. Alshaabi, J. R. Minot, J. L. Adams, C. M. Danforth, and P. S. Dodds. Hurricanes and hashtags: Characterizing online collective attention for natural disasters. PLoS one, 16(5):e0251762, 2021.

[14] S. Shugars, A. Gitomer, S. McCabe, R. J. Gallagher, K. Joseph, N. Grinberg, L. Doroshenko, B. F. Welles, and D. Lazer. Pandemics, protests, and publics: Demographic activity and engagement on twitter in 2020. Journal of Quantitative Description: Digital Media, 1, 2021.

[15] E. Chen, K. Lerman, E. Ferrara, et al. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. JMIR public health and surveillance, 6(2):e19273, 2020.

[16] J. Green, J. Edgerton, D. Naftel, K. Shoub, and S. J. Cranmer. Elusive consensus: Polarization in elite communication on the covid-19 pandemic. Science advances, 6(28):eabc2717, 2020.

[17] Covid-19 stream, May 2020.

[18] P. S. Dodds, J. R. Minot, M. V. Arnold, T. Alshaabi, J. L. Adams, D. R. Dewhurst, T. J. Gray, M. R. Frank, A. J. Reagan, and C. M. Danforth. Allotaxonometry and rank-turbulence divergence: A universal instrument for comparing complex systems. arXiv preprint arXiv:2002.09770, 2020.

[19] T. Alshaabi, M. V. Arnold, J. R. Minot, J. L. Adams, D. R. Dewhurst, A. J. Reagan, R. Muhamad, C. M. Danforth, and P. S. Dodds. How the world's collective attention is being paid to a pandemic: Covid-19 related n-gram time series for 24 languages on twitter. Plos one, 16(1):e0244476, 2021.

[20] J. Minot, M. Trujillo, S. Rosenblatt, G. De Anda-Jáuregui, E. Moog, A. M. Roth, B. P. Samson, and L. Hébert-Dufresne. Distinguishing in-groups and onlookers by language use. In Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pages 157–171, 2022.

[21] A. M. Stupinski, T. Alshaabi, M. V. Arnold, J. L. Adams, J. R. Minot, M. Price, P. S. Dodds, and C. M. Danforth. Quantifying changes in the language used around mental health on twitter over 10 years: Observational study. JMIR mental health, 9(3):e33685, 2022.

[22] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys (csur), 53(3):1–34, 2020.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[24] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014.

[25] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 11 2019.

[26] S. Gao, M. Alawad, M. T. Young, J. Gounley, N. Schaefferkoetter, H. J. Yoon, X.-C. Wu, E. B. Durbin, J. Doherty, A. Stroup, et al. Limitations of transformers on clinical text classification. IEEE journal of biomedical and health informatics, 25(9):3596–3607, 2021.

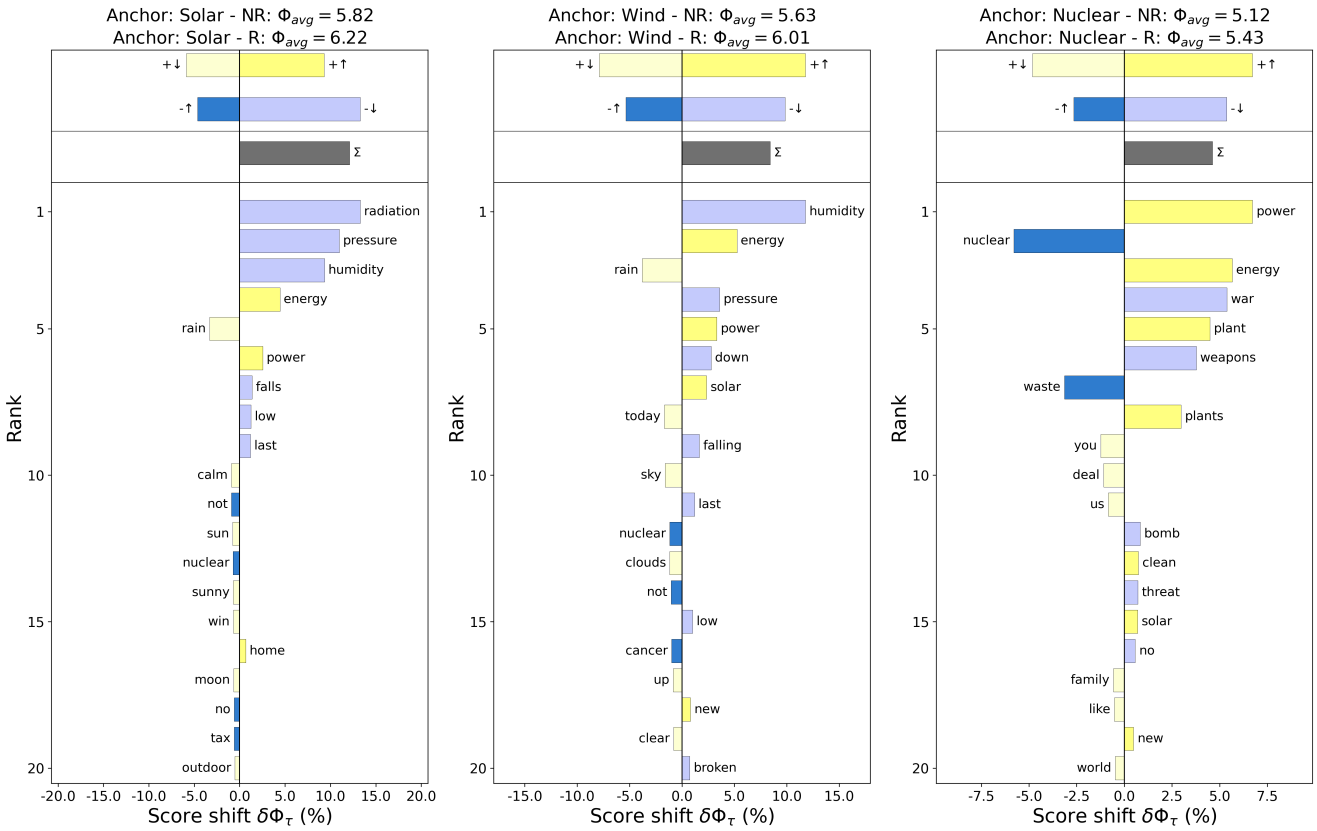


FIG. 4. **Sentiment shift plots comparing the classified 'relevant' (R) and 'non-relevant' (NR) tweet corpora for tweets containing the keywords 'solar', 'wind', and 'nuclear'.** We show the 20 top contributing words to the difference in LabMT sentiment between the corpora. 'Relevant' tweets, those related to clean energy are happier on average for all keywords when compared to 'not relevant' tweets. Sad words that are less common in 'relevant' '**solar**' tweets are 'radiation', 'pressure', and 'humidity', which largely refer to the weather. Happy words like 'energy' and 'power' are more common in 'relevant' tweets compared to tweets not relevant to solar energy. For '**wind**' sad terms like 'humidity' and 'pressure' are less common in 'relevant' tweets, while happy terms like 'energy', 'power', and 'solar' are more common in tweets relevant to wind as a renewable energy source. For '**nuclear**', relevant tweets are on average more positive due to sad words like 'war', 'weapons', and 'bomb' are less common in relevant tweets, while happy words like 'power' and 'energy' are more common. Some sad words like 'nuclear' and 'waste' do occur more frequently in relevant tweets.


[27] W.-C. Chang, H.-F. Yu, K. Zhong, Y. Yang, and I. S. Dhillon. Taming pretrained transformers for extreme multi-label text classification. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pages 3163–3171, 2020.

[28] D. Antypas, A. Ushio, J. Camacho-Collados, L. Neves, V. Silva, and F. Barbieri. Twitter topic classification. arXiv preprint arXiv:2209.09824, 2022.

[29] D. Quercia, H. Askham, and J. Crowcroft. Tweetlda: supervised topic classification and link prediction in twitter. In Proceedings of the 4th Annual ACM Web Science Conference, pages 247–250, 2012.

[30] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pages 38–45, 2020.

[31] L. Yan, Y. Zheng, and J. Cao. Few-shot learning for short text classification. Multimedia Tools and Applications, 77(22):29799–29810, 2018.

[32] J. Kim, D. Jeong, D. Choi, and E. Park. Exploring public perceptions of renewable energy: Evidence from a word network model in social network services. Energy Strategy Reviews, 32:100552, 2020.

[33] A. Jain and V. Jain. Sentiment classification of twitter data belonging to renewable energy using machine learning. Journal of information and optimization sciences, 40(2):521–533, 2019.

[34] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 427–431. Association for Computational Linguistics, April 2017.

[35] P. S. Dodds, E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, et al. Human language reveals a universal positivity bias. Proceedings of the national academy of sciences, 112(8):2389–2394, 2015.