9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

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

Social Media Corpus Curation, Natural Language Processing, Text Classification, Sentence Embeddings, Sentiment Analysis

Extended Abstract

Recent advances in text classification have enabled an alternative model to curate social media corpora for computational social scientists. For the past two decades, researchers have largely handcrafted keywords lists to query relevant social media posts [1]. Others have attempted improve on the data collection process to automate the keyword selection process [2] to boost recall, or add negation keywords in an attempt to increase precision. However, the process of creating a text data set can often remain opaque; with keywords chosen by subject matter experts there is little transparency into the trade-offs of including or excluding individual terms, or how that effects the resulting corpus.

We argue that corpus curation is well-framed as a text classification problem and that contextual sentence embeddings are suitable features for this task [3]. With few-shot learning, leveraging pre-trained transformer based language models, researchers can now label a few hundred messages, a task of no more than two hours, and scale up to classify millions of tweets on a laptop.

We demonstrate our approach with a set of case studies examining online sentiment surrounding clean energy technologies on Twitter. The collections of tweets that are anchored by a single keyword or set of keywords. From Twitter's Decahose API, a random 10% sample of all public tweets, we select tweets containing a user-provided locations [4]. We extracted these locations from a free text location field in each user's bio, if the text matched a valid `'city, state'` string in the United States [5, 6]. From this selection, we query for tweets that both contain keywords of choice and are classified as English by FastText [7]. We define the results of this query as the unfiltered ambient corpus.

We choose 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.

After manually labeling 1000 tweets as relevant or not relevant to clean energy for each subset, we train a text classifier using a pre-trained sentence embedding model, `all-mpnet-base-v2`. We achieve F1 scores of up to 0.95 on our binary classification task. Our classifier detects an influx of bots would have a dramatic effect on sentiment measurements taken on an unfiltered corpus defined by broad list of keywords, but filters them to recover a mostly stable sentiment signal over our study period.

We believe our method of setting corpus boundaries is more transparent than defining the boundaries with expert selected keyword lists. Instead, researchers using this paradigm would share their labeled training data, so reviewers could inspect if the heuristics described by authors matches the labels. 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.

References

[1] Sarah Shugars, Adina Gitomer, Stefan McCabe, Ryan J Gallagher, Kenneth Joseph, Nir Grinberg, Larissa Doroshenko, Brooke Foucault Welles, and David Lazer. Pandemics, protests, and publics: Demographic activity and engagement on twitter in 2020. *Journal of Quantitative Description: Digital Media*, 1, 2021.

[2] Joao P Carvalho, Hugo Rosa, Gaspar Brogueira, and Fernando Batista. Misnis: An intelligent platform for twitter topic mining. *Expert Systems with Applications*, 89:374–388, 2017.

[3] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[4] Decahose api.

[5] Tyler J Gray, Andrew J Reagan, Peter Sheridan Dodds, and Christopher M Danforth. English verb regularization in books and tweets. *PloS one*, 13(12):e0209651, 2018.

[6] Kelsey Linnell, Michael Arnold, Thayer Alshaabi, Thomas McAndrew, Jeanie Lim, Peter Sheridan Dodds, and Christopher M Danforth. The sleep loss insult of spring daylight savings in the us is observable in twitter activity. *Journal of big data*, 8:1–17, 2021.

[7] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas 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.

[8] Peter Sheridan Dodds, Eric M Clark, Suma Desu, Morgan R Frank, Andrew J Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M Kloumann, James P Bagrow, et al. Human language reveals a universal positivity bias. *Proceedings of the national academy of sciences*, 112(8):2389–2394, 2015.

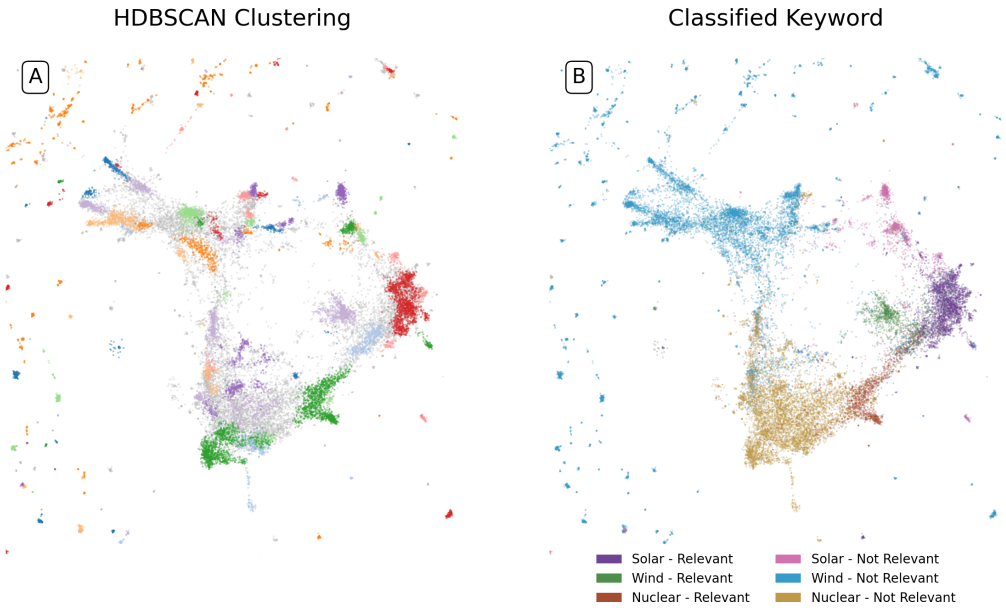

Figure 1: Tweets containing the keywords `'solar'`, `'wind'`, and `'nuclear'` placed in a semantically meaningful embedding space. In panel A, we perform unsupervised clustering with HDBSCAN and color tweets by the resulting cluster. In panel B, we show the results of our text classifier, which infers if tweets are relevant or not relevant to the topic of clean energy. We color by keyword and classification result. Dimensionality reduction into 2D is performed with UMAP. Locally, tweets that are closer together in the reduced embedding tend to be more semantically similar, but global positions and the coordinates are not meaningful. Tweets classified as relevant to clean energy are tightly grouped on the right side.

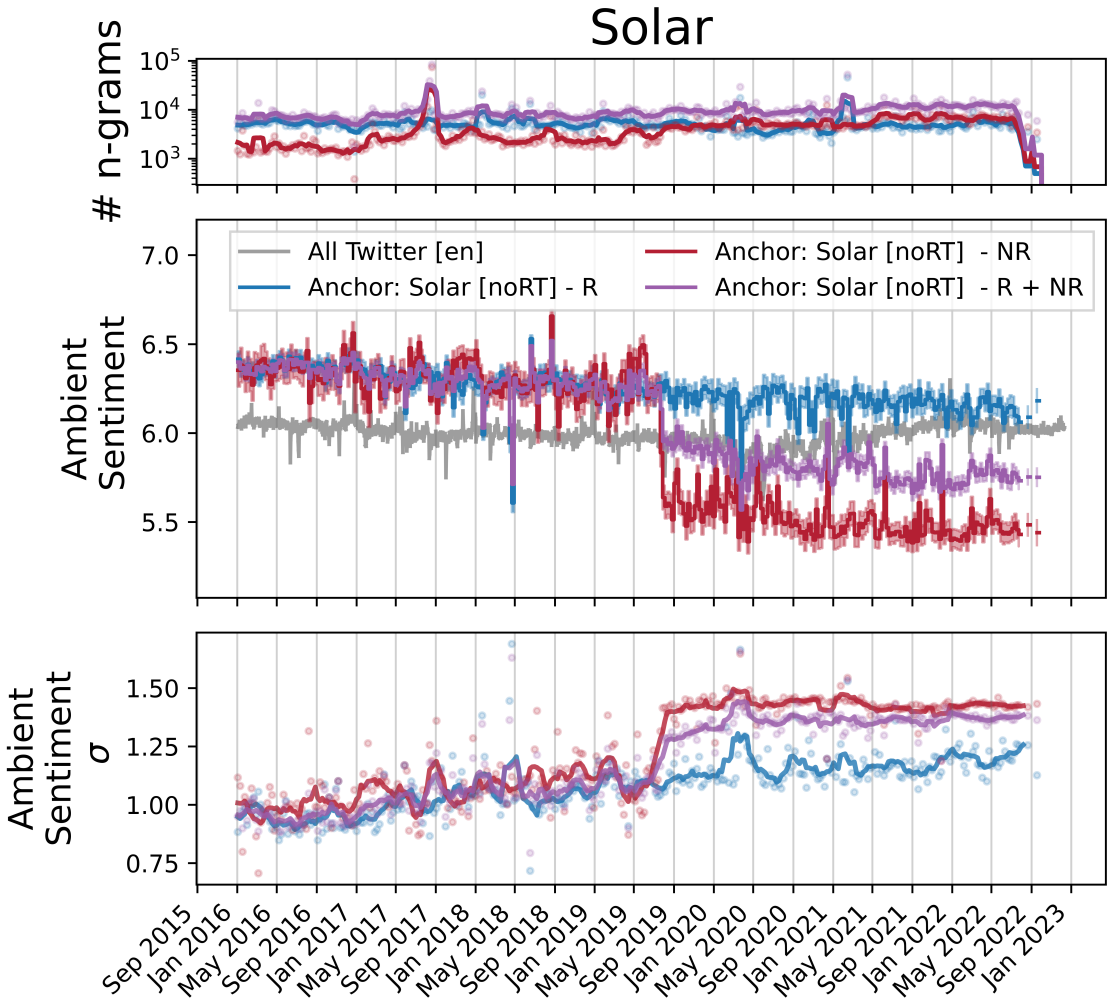

Figure 2: **Ambient sentiment time series comparison for relevant (R), not-relevant (NR), and combined tweet corpora, containing the keyword `solar`.** In the top panel we show the number of tokens with LabMT [8] sentiment scores in each corpus on each day. 'Relevant' tweets, in blue, have more scored 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 a 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.

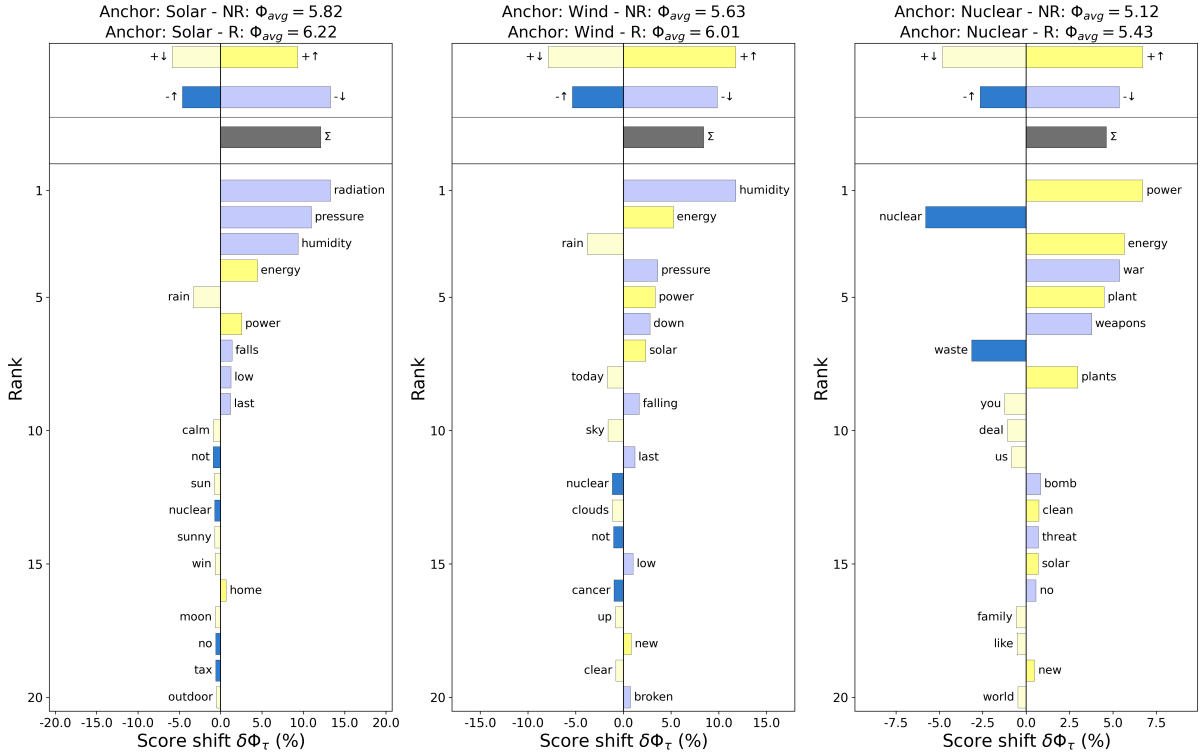

Figure 3: **Sentiment shift plots comparing the classified 'relevant' (R) and 'not-relevant' (NR) tweet corpora for tweets containing the keywords `solar`, `wind`, and `nuclear`.** We show the 20 top contributing words to the difference in LabMT sentiment between the corpora. 'Relevant' tweets, those 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' being 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.