Inferring Climate Change Stances from Multimodal Tweets

Nan Bai
Wageningen University & Research
Wageningen, the Netherlands
Delft University of Technology
Delft, the Netherlands
n.bai@tudelft.nl

Ricardo da Silva Torres
Wageningen University & Research
Wageningen, the Netherlands
ricardo.dasilvatorres@wur.nl

Anna Fensel
Wageningen University & Research
Wageningen, the Netherlands
anna.fensel@wur.nl

Tamara Metze
Delft University of Technology
Delft, the Netherlands
t.a.p.metze@tudelft.nl

Art Dewulf
Wageningen University & Research
Wageningen, the Netherlands
art.dewulf@wur.nl

ABSTRACT

Climate change is a heated discussion topic in public arenas such as social media. Both texts and visuals play key roles in the debate, as they can complement, contradict, or reinforce each other in nuanced ways. It is therefore urgently needed to study the messages as multimodal objects to better understand the polarized debate about climate change impacts and policies. Multimodal representation models such as CLIP are known to be able to transfer knowledge across domains and modalities, enabling the investigation of textual and visual semantics together. Yet they are not directly able to distinguish the nuances between supporting and sceptic climate change stances. This paper explores a simple but effective strategy combining modality fusion and domain-knowledge enhancing to prepare CLIP-based models with knowledge of climate change stances. A multimodal Dutch Twitter dataset is collected and experimented with the proposed strategy, which increased the macro-average F1 score across stances from 51% to 86%. The outcomes can be applied in both data science and public policy studies, to better analyse how the combined use of texts and visuals generates meanings during debates, in the context of climate change and beyond.

CCS CONCEPTS

• **Applied computing** → **Sociology**; • **Information systems** → **Clustering and classification**; *Social networks.*

KEYWORDS

Multimodal Embeddings; Transfer Learning; User-Generated Content; Climate Change Claims; Sea-Level Rise; Public Policy

ACM Reference Format:
Nan Bai, Ricardo da Silva Torres, Anna Fensel, Tamara Metze, and Art Dewulf. 2024. Inferring Climate Change Stances from Multimodal Tweets. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3626772.3657950

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657950

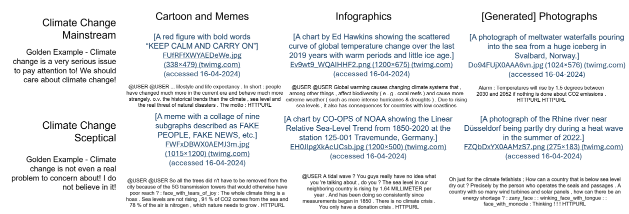

Figure 1: Examples from the collected multimodal dataset of climate change stances. The texts are normalized and translated into English. The visuals are verbally described while the original images can be accessed with their twimg URLs.

1 INTRODUCTION

As a heated public-policy topic in recent decades, climate change triggers debates in public arenas including social media [25, 36]. A wide range of stakeholders including scientists, news media, policymakers, and action groups, etc. actively express their views on this politically controversial issue, trying to make their voices heard and have an influence on the agenda [41]. Understanding people's stances towards climate change, i.e., whether they support or deny it, is crucial for monitoring and analyzing the dynamics in such debates. Literature in the emerging field of Climate NLP (Natural Language Processing) has been using verbal information to retrieve the existence of environmental claims [11, 35, 42], summarize the main topics covered during such debates [3, 8, 13–15, 40], detect the sentiment polarity [13–15, 38, 39, 43] and potential aggressiveness of discussants [14, 15], and infer the stances of users towards climate change [9, 14–16, 22, 38–40], possibly with a hierarchical reasoning chain of contrarian and sceptical claims [7, 26]. Whereas these studies focused on texts, visuals are equally important throughout the debate of policy controversies, containing complementary and conflicting messages [1, 17, 27, 31, 32]. Visuals, together with texts (Figure 1) can serve different framing functions, such as sense-making, emotion-triggering, value-portraying, etc., leading to various interpretations during public debates [31, 32].

Multimodal representation models with Contrastive Language-Image Pre-Training (CLIP), which projects textual and visual objects into the same high-dimensional vector space, are shown to be effective in fusing and generalizing the semantics within both

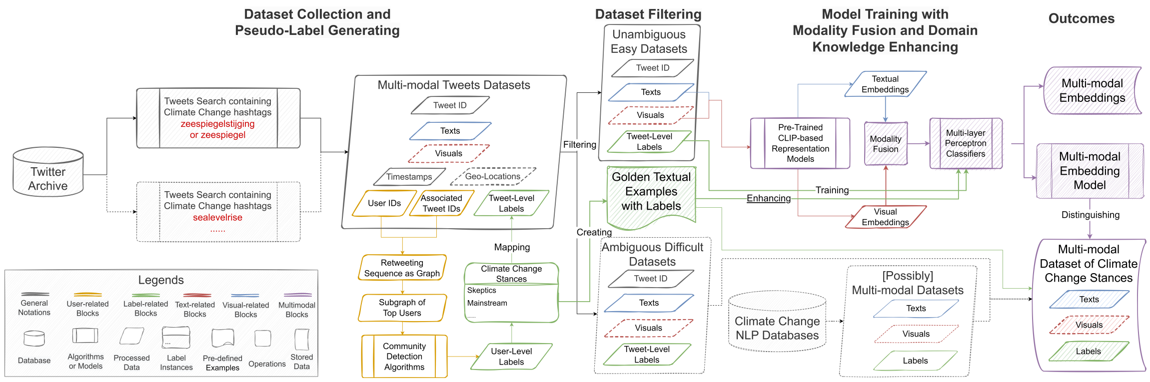

Figure 2: The workflow proposed in this paper. Dashed components are either optional (visuals and geo-locations) or not yet implemented in the presented version of the paper (ambiguous difficult datasets and other climate change databases).

modalities [28, 34]. The resulting multimodal embeddings (often by directly averaging the uni-modal embeddings) are also consistently used in downstream tasks, such as multimodal topic modelling, semantic search, zero- or few-shot learning, etc [18, 29]. However, the CLIP models were trained on a general corpus. They cannot necessarily distinguish the nuances between supports and sceptics towards a particular topic (e.g., climate change) *per se*, failing in zero-shot stance classification, as will be shown in Section 2.3.

This paper explores the capability of CLIP-based models to infer the climate change stances (CCS) of multimodal tweets, providing CCS-aware multimodal embeddings for downstream tasks. A workflow is proposed to collect multimodal datasets from Twitter, generate pseudo-labels about supporting or sceptical CCS, and classify the tweets using multimodal features with a simple strategy combining modality fusion and domain-specific enhancing. The classification performance generally increased with the experimented strategy, and visual-text pairs of similar natures were detected that reflected both sides of the debate, as shown in Figure 1. The new CCS dataset and multimodal embeddings prepare for future studies in data science and policy science to better analyze and understand the debate dynamics using both texts and visuals. Codes and the processed dataset are available at this repository.

2 METHODOLOGY

2.1 Problem Overview

Figure 2 provides an overview of the proposed framework, reflecting the aim of this paper to obtain both multimodal CCS datasets from Twitter and multimodal embeddings with CCS knowledge.

The workflow starts with collecting and processing multimodal datasets (left part of Figure 2). The datasets get user-level and tweet-level pseudo-labels about CCS or other labels of interest based on community detection algorithms. The labelled datasets are further filtered to only keep an *unambiguous* and *easy* subset to train classifiers for pragmatic reasons (middle part of Figure 2). Both textual and visual embeddings are extracted from variants of CLIP-based representation models. Different strategies of modality fusion [2, 34] are experimented with to fuse the embeddings, before feeding them into an additional Multi-layer Perceptron (MLP) classifier trained with domain-specific golden-label enhancing to infer whether a sample supports or denies climate change (right part of Figure 2). The end products are highlighted in the rightmost part of Figure 2. Note that part of the proposed workflow has not yet been implemented, i.e., inferring stances with ambiguous difficult subsets and integrating other climate change datasets.

2.2 Dataset Collection and Pre-processing

A multimodal Dutch dataset of texts and visuals concerning sea-level rise (zeespiegelstijging) with all available tweets containing relevant keywords was collected using Twitter API v2. Let i be the index of a generic sample of the dataset, then its raw data could be denoted as a tuple $\mathfrak{d}_i := (\mathfrak{I}_i, \mathcal{S}_i^{\mathrm{R}}, \mathfrak{u}_i, O_i, \mathfrak{t}_i, \mathfrak{l}_i)$, $\mathfrak{d}_i \in \{\mathfrak{d}_i\}_{i=1,2,\ldots,K_0}$, where $K_0 = 220,494$ is the initial size of the collected dataset related to climate change issues. $\mathfrak{I}_i$ is a three-dimensional image tensor representing one of 7410 unique visuals, where $\mathfrak{I}_i = \emptyset$ is also allowed when no visuals are attached in a tweet [1, 23]. Among the collected tweets, 57,038 (25.9%) are with non-empty visual features. $\mathcal{S}_i^{\mathrm{R}}$ is a raw paragraph within the tweet, which is first normalised into $\mathcal{S}_i^{\mathrm{NL}}$ by transforming repeated mentionings into '@USER' tokens, changing internet links into 'HTTPURL' tokens, and de-emojizing the emojis into verbal descriptions. Since many multimodal embedding models are trained with English corpus, the normalised Dutch texts are then translated into English $\mathcal{S}_i^{\mathrm{EN}}$ with Google Translator API from the Deep Translator library. $\mathfrak{u}_i$ is a user ID among 54,005 unique users. O_i is a set of user IDs that are either retweeted or mentioned in the tweet text, of which the posting user $\mathfrak{u}_i$ is also an instance. $\mathfrak{t}_i$ and $\mathfrak{l}_i$ represent the timestamp and the geo-location of the tweet if the information is available, where empty values are also allowed. In the collected dataset, $\mathfrak{t}_i$ ranges from April 26, 2007, to January 1, 2023, and only 1225 tweets are originally geo-tagged.

A retweeting sequence within the top-1000 users is formalized as a social network. With a community detection algorithm based on Clauset-Newman-Moore greedy modularity maximization [6, 19], the top users can be divided into sub-communities representing discourse coalitions [25, 31, 44]. In the context of this study, two prominent communities emerged within the top users. The hashtags extensively used by both communities were reviewed *post hoc* by domain experts, distinguishing them as typical users who support the mainstream view on anthropogenic climate change issues, and users who are sceptical about the existence or anthropogenic nature of climate change, respectively, i.e., typical in one of the CCS. This results in a user-level pseudo-label $y_j^{\mathrm{UL}} \in \{-1, 1, \varnothing\}$, where -1 labels the user as sceptical towards climate change (578 detected); 1 means that the user supports the mainstream (404 detected); and $\varnothing$ marks the unlabelled users. This user-level label is mapped to tweet-level $y_i^{\mathrm{TL}} \in \{-1, 1, \varnothing\}$ by labelling tweets that are merely associated with one type of users holding either supporting or sceptical stances in O_i. This mapping process ensures a collection of **unambiguous** examples to avoid confusion during training. Furthermore, for two unambiguous tweets $\mathfrak{d}_i, \mathfrak{d}_{i'}$ containing exactly the same pair of texts and visuals, it is possible that the pair of information is used by opposite parties, i.e., $y_i^{\mathrm{TL}} y_{i'}^{\mathrm{TL}} = -1$. This is not uncommon in social media debates, as the same message can be used by various parties differently, possibly also containing contradictory meanings depending on the context. However, this will make the later training unstable. Filtering out those cases (279 examples) can result in an **easy** sub-dataset that only contains straightforward examples.

By keeping the non-redundant, unambiguous, and easy examples, a final dataset $\mathfrak{D} := \{\mathfrak{d}_i\}_{i=1,2,\ldots,K_1}$ is obtained labelled with $y_i \in \{-1, 1\}$, where $K_1 = 49,316$ (22.3% of K_0). Among the tweets, 29,306 (59.4% of K_1) are climate change mainstream, and 20,010 (40.6% of K_1) are sceptics. The dataset is further randomly divided into training set, validation set, and test set with a proportion of 70/15/15.

2.3 Modality Fusion and Stance Classification

Pilot studies of this research showed that only by averaging the textual and visual embeddings, original CLIP-based multimodal embeddings confused the CCS and failed to give correct predictions in zero-shot classifications even when the stances are obvious for humans. Therefore, this study explores strategies of modality fusion and domain-knowledge enhancing to further improve the ability of classifiers to distinguish supporting and sceptical stances.

Let $\mathbf{f}_{\mathrm{CLIP}}$ denote the CLIP-based models with parameters $\mathbf{\Phi}_{\mathrm{CLIP}}$, then the textual embeddings $\boldsymbol{x}_i^{\mathrm{TEX}} = \mathbf{f}_{\mathrm{CLIP}}(\mathcal{S}_i^{\mathrm{EN}} | \mathbf{\Phi}_{\mathrm{CLIP}})$ and the visual embeddings $\boldsymbol{x}_i^{\mathrm{VIS}} = \mathbf{f}_{\mathrm{CLIP}}(\mathfrak{I}_i | \mathbf{\Phi}_{\mathrm{CLIP}})$ of data $\mathfrak{d}_i$ would be vectors with the same dimensionality $\mathbb{R}^{d \times 1}$. When $\mathfrak{I}_i = \varnothing$, $\mathfrak{I}_i$ is universally replaced with a white image of the same size, as $\boldsymbol{x}_\varnothing^{\mathrm{VIS}}$. Let $\mathbf{g}_{\mathrm{F}}$ denote modality fusion operations [2, 34], then the initial multimodal embedding $\boldsymbol{x}_i^{\mathrm{MULT}} = \mathbf{g}_{\mathrm{F}}\left(\boldsymbol{x}_i^{\mathrm{TEX}}, \boldsymbol{x}_i^{\mathrm{VIS}}\right)$ would be a vector with possibly a different dimensionality $\mathbb{R}^{d_0 \times 1}$. Let $\mathbf{f}_{\mathrm{MLP}}^{(t)}$ denote the first t layers of a τ-layer MLP classifier with parameters $\mathbf{\Phi}_{\mathrm{MLP}}$, the intermediate vectors can be written as $\boldsymbol{z}_i^{(t)} = \mathbf{f}_{\mathrm{MLP}}^{(t)}(\boldsymbol{x}_i^{\mathrm{MULT}} | \mathbf{\Phi}_{\mathrm{MLP}})$, $\boldsymbol{z}_i^{(t)} \in \mathbb{R}^{d_t \times 1}$, where $\boldsymbol{z}_i^{(0)} := \boldsymbol{x}_i^{\mathrm{MULT}}$, $\boldsymbol{z}_i^{(\tau)} \in \mathbb{R}^{2 \times 1}$, meaning that the τ_{th} layer generates the final 2-dimensional stance classification results.

As domain knowledge, sentences $\{\mathcal{S}_k^{\mathrm{G}}\}$ clearly referring to either side of CCS are prepared as **golden** examples, given arbitrary labels $y_k^{\mathrm{G}} \in \{-1, 1\}$. In this initial stage of exploration, one sentence per stance was used as golden examples, both of which can also be seen in the left side of Figure 1. For each epoch of training, in addition to the conventional classification loss with the loss function ℓ, an optional **enhancing** step also optimizes for the golden examples:

$$\mathcal{L}_{\mathrm{G}} = \sum_k \ell(\mathbf{f}_{\mathrm{MLP}}^{(\tau)}(\mathbf{g}_{\mathrm{F}}(\mathbf{f}_{\mathrm{CLIP}}(\mathcal{S}_k^{\mathrm{G}} | \mathbf{\Phi}_{\mathrm{CLIP}}), \boldsymbol{x}_\varnothing^{\mathrm{VIS}}) | \mathbf{\Phi}_{\mathrm{MLP}}), y_k^{\mathrm{G}}), \quad (1)$$

hypothetically pushing the intermediate vectors of different stance labels farther from each other. After rounds of training, the vectors $\boldsymbol{z}_i^{(t)}$ (especially $\boldsymbol{z}_i^{(\tau-1)}$) effectively become the new multimodal embeddings of data point $\mathfrak{d}_i$, and the chained models of $\mathbf{f}_{\mathrm{MLP}}^{(\tau-1)}$, $\mathbf{g}_{\mathrm{F}}$, and the original $\mathbf{f}_{\mathrm{CLIP}}$ become the new embedding model.

2.4 Experiments

Experiments were set up to evaluate the proposed strategy of combining modality fusion with golden-example enhancing, with both the trained classifier and the intermediate layers as embeddings. Without loss of generality, four variants of the CLIP checkpoints have been experimented with as the baselines: clip-ViT-B-32 (B-32), clip-ViT-B-16 (B-16), clip-ViT-L-14 (L-14), and the distilled multilingual clip-ViT-B-32, all implemented from the Sentence Transformer Python library [28–30]. In addition to the translated English sentences, the normalised Dutch sentences were also tested with the multilingual model (ML-EN and ML-NL, respectively).

Five versions of modality fusion $\mathbf{g}_{\mathrm{F}}$ were implemented: only keeping textual embeddings $\boldsymbol{x}_i^{\mathrm{TEX}}$, only keeping visual embeddings $\boldsymbol{x}_i^{\mathrm{VIS}}$, averaging $\boldsymbol{x}_i^{\mathrm{TEX}}$ and $\boldsymbol{x}_i^{\mathrm{VIS}}$, concatenating them, and merging them with a complex function similar to previous study [34]:

$$\mathbf{g}_{\mathrm{F\_complex}}(\boldsymbol{v}_1, \boldsymbol{v}_2) = [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_1 + \boldsymbol{v}_2, \boldsymbol{v}_1 - \boldsymbol{v}_2, \boldsymbol{v}_1 \odot \boldsymbol{v}_2]\,. \quad (2)$$

For each model variant, all versions of the modality fusion strategy were paired with the optional enhancing operation mentioned in Equation (1), resulting in 10 runs of experiments. 3-layer MLP models with the same hyper-parameter configuration were trained on mini-batches for 200 epochs, where cross-entropy was used as the loss function ℓ and early-stopping was implemented with the overall accuracy on the validation set. The models were eventually evaluated with the accuracy and the macro-average F1 scores of three sub-cases: examples that are truly multimodal ($\mathfrak{I}_i \neq \varnothing$), examples that only have textual information ($\mathfrak{I}_i = \varnothing$), and examples only containing visuals not previously seen in training.

Furthermore, the multimodal embeddings $\boldsymbol{z}_i^{(t)}$ (especially $\boldsymbol{z}_i^{(\tau-1)}$) from the intermediate layers of the trained MLP were used to compare the cosine similarity of those computed with the golden examples. The best-performing embeddings were eventually consulted to extract the closest multimodal examples from the dataset $\mathcal{D}$ that best align with each statement in the golden examples.

3 RESULTS AND DISCUSSIONS

3.1 Classification Outcomes

Merging the results with all CLIP variants on both validation and test sets, the ranges of the macro-average F1 scores are plotted in Figure 3. Fusing the multimodal embeddings generally increased the classification performance compared to any single modality, and the complex fusion strategy mentioned in Equation (2) was generally the most effective one. This is, however, not the case for multimodal examples with unseen visual images. Even though all text-visual pairs $(\mathcal{S}_i^{\mathrm{EN}}, \mathfrak{I}_i)$ are unique, it is possible that some visuals in the validation and test sets were previously paired with other sentences in the training set. The models may have remembered the associations of visuals with the labels and thus over-fitted on those visuals. This observation invites further investigation.

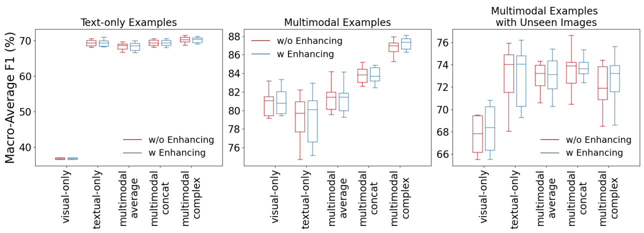

Figure 3: The macro-average F1 score of trained models on validation and test sets on the three sub-cases. The two parallel boxes show the results with or without the optional enhancing step of golden examples.

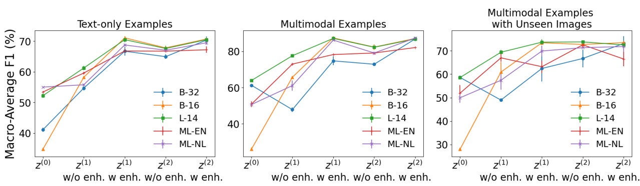

Figure 4: The macro-average F1 score based on cosine similarity of embedding layers on the three sub-cases. All embeddings are computed with complex fusion and enhancing. Error-bars show the range in validation and test sets.

Despite having the risk of overfitting the models with the additional enhancing step with golden textual examples, the performance also either increased or stayed unharmed. Instead, the enhancing solved a problem observed and mentioned in Section 2.3: 15 out of 25 trained models without enhancing still predicted wrong labels for the golden examples (e.g., the B-32 using multimodal-complex fusion strategy predicted both sentences as 'sceptical', and the one using text-only predicted both as 'mainstream'). The loss $\mathcal{L}_G$ dropped from 1.11±0.22 to 0.77±0.41 by adding the text-based enhancing, reversing the partially wrong predictions of 10 trained models except for the ones using visual embeddings only.

Similar effects with enhancing can be observed in Figure 4, where the best-performing models with complex multimodal fusion were evaluated. Regardless of the models being used, the macro-average F1 score of multimodal examples (middle sub-figure) significantly increased from $50.5\% \pm 13.4\%$ in the initial layer $z^{(0)}$ to $86.0\% \pm 2.0\%$ in the second-last layer $z^{(2)}$ with enhancing. Similarly, the accuracy increased from $61.6\% \pm 14.9\%$ to $87.6\% \pm 2.2\%$. Later intermediate layers generally performed better than earlier layers. Among the same layer, adding enhancing would significantly increase the general macro-average F1 score from $66.8\% \pm 9.0\%$ to $74.5\% \pm 7.9\%$, and the accuracy from $72.8\% \pm 6.7\%$ to $76.2\% \pm 8.1\%$, despite the discrepancy of ML-EN. The findings suggest that with both complex fusion in Equation (2) and text-based enhancing in Equation (1), multimodal embeddings acquainted with CCS knowledge could be obtained.

3.2 Discussion

Figure 1 has illustrated a few typical multimodal data samples from the collected dataset that were detected as semantically most similar to both golden examples, computed with the new embeddings generated from B-16 and L-14 variants of CLIP-based models (i.e., the best models shown in Figure 4). The new embeddings managed to catch the nuances of two stances and match them with correct and reasonable multimodal pairs of texts and visuals. Interestingly, from both sides of the argument, visuals of similar natures are being used, such as cartoons and memes, data visualization including maps and infographics, as well as real-world and/or imaginary photographs. Specifically, climate change sceptics also use scientific data to justify and strengthen their beliefs against the mainstream. This observation is consistent with previous studies discussing the use of visuals during policy controversies, going beyond climate change [17, 24, 31, 32]. Different types of visuals have similar functions of supporting the main stances towards the argument (here climate change) from a different angle. A more concrete and nuanced understanding of the roles of visuals and multimodal interactions can be obtained by thoroughly examining the universal use of different types of visuals and multimodal pairs throughout the debate, possibly augmented with spatiotemporal contexts [1, 8]. The next steps of this study will explore the behaviour of trained models and continue the analyses of multimodal persuasion strategies on the examples that are not necessarily easy and unambiguous, thus completing the workflow proposed in Figure 2. The same workflow could also collect datasets in other languages and countries containing more diverse keywords concerning climate change.

The proposed framework did not include human annotators. Rather, pseudo-labels for multimodal pairs were approximated based on community detection algorithms. Further human-in-the-loop evaluations and augmentations could increase the reliability of the dataset. The robustness of trained models need to be evaluated on other unseen existing text-only climate change datasets [4, 7, 14, 22, 35, 38, 40, 42, 43]. Integrating them as additional golden examples could further enhance the ability of trained embedding models and generate finer-grained datasets [7, 26, 40]. Broader model searching with hyper-parameter tuning and ensemble learning combining multiple trained models can potentially increase the performance and generalizability of the proposed approach [12]. To combat the issue of possible over-fitting mentioned in Section 3.1, strategies such as data augmentation, [domain-specific] regularization, and additional search with model architecture could help [26]. Extra penalties can be given to ambiguous examples during training. In follow-up studies, the obtained CCS-aware multimodal embeddings can be further extended to improve domain-specific multimodal topic modelling [8, 18, 25] and to assist multimodal framing analysis [4, 10, 31, 33, 37]. Moreover, the proposed framework can be experimented with other multimodal representation models other than CLIP, exploring the possibility of collaborating with the most recent advances in large language models [5, 20, 21, 45].

4 CONCLUSIONS

This paper explores the capability of CLIP-based models to infer climate change stances (CCS) from multimodal tweets. The proposed workflow yielded both a multimodal dataset about CCS and an embedding model to obtain CCS-aware vector representations. Both outcomes are rare in literature but provide potentials in data science and public policy research. It prepares for a systematic understanding of discourse coalitions during public debates and policy controversies in the context of climate change and beyond.

REFERENCES

[1] Nan Bai, Pirouz Nourian, Renqian Luo, and Ana Pereira Roders. 2022. Heri-graphs: a dataset creation framework for multi-modal machine learning on graphs of heritage values and attributes with social media. *ISPRS International Journal of Geo-Information* 11, 9 (2022), 469.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[3] Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2021. Automatic classification of neutralization techniques in the narrative of climate change scepticism. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 2167–2175.

[4] Dallas Card, Amber Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China, 438–444.

[5] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv:2310.09478

[6] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. 2004. Finding community structure in very large networks. *Physical review E* 70, 6 (2004), 066111.

[7] Travis G Coan, Constantine Boussalis, John Cook, and Mirjam O Nanko. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific reports* 11, 1 (2021), 22320.

[8] Biraj Dahal, Sathish AP Kumar, and Zhenlong Li. 2019. Topic modeling and sentiment analysis of global climate change tweets. *Social network analysis and mining* 9 (2019), 1–20.

[9] L Derczynski, K Bontcheva, M Liakata, R Procter, GWS Hoi, and A Zubiaga. 2017. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 69–76.

[10] Art Dewulf and René Bouwen. 2012. Issue framing in conversations for change: Discursive interaction strategies for "doing differences". *The Journal of Applied Behavioral Science* 48, 2 (2012), 168–193.

[11] Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. arXiv:2012.00614

[12] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. 2020. A survey on ensemble learning. *Frontiers of Computer Science* 14 (2020), 241–258.

[13] Cuc Duong, Qian Liu, Rui Mao, and Erik Cambria. 2022. Saving earth one tweet at a time through the lens of artificial intelligence. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Padua, Italy, 1–9.

[14] Dimitrios Effrosynidis, Alexandros I Karasakalidis, Georgios Sylaios, and Avi Arampatzis. 2022. The climate change Twitter dataset. *Expert Systems with Applications* 204 (2022), 117541.

[15] Dimitrios Effrosynidis, Georgios Sylaios, and Avi Arampatzis. 2022. Exploring climate change on Twitter using seven aspects: Stance, sentiment, aggressiveness, temperature, gender, topics, and disasters. *Plos one* 17, 9 (2022), e0274213.

[16] Max Falkenberg, Alessandro Galeazzi, Maddalena Torricelli, Niccolò Di Marco, Francesca Larosa, Madalina Sas, Amin Mekacher, Warren Pearce, Fabiana Zollo, Walter Quattrociocchi, et al. 2022. Growing polarization around climate change on social media. *Nature Climate Change* 12, 12 (2022), 1114–1121.

[17] Efrat Gommeh, Huub Dijstelbloem, and Tamara Metze. 2021. Visual discourse coalitions: visualization and discourse formation in controversies over shale gas development. *Journal of Environmental Policy & Planning* 23, 3 (2021), 363–380.

[18] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794

[19] Aric Hagberg, Pieter Swart, and Daniel S Chult. 2008. *Exploring network structure, dynamics, and function using NetworkX*. Technical Report. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

[20] Markus Leippold, Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, et al. 2024. Automated Fact-Checking of Climate Change Claims with Large Language Models. arXiv:2401.12566

[21] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023. Mmbench: Is your multi-modal model an all-around player? arXiv:2307.06281

[22] Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting Stance in Media On Global Warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3296–3315.

[23] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, New Orleans, LA, USA, 18177–18186.

[24] Tamara Metze. 2020. Visualization in environmental policy and planning: A systematic review and research agenda. *Journal of Environmental Policy & Planning* 22, 5 (2020), 745–760.

[25] Warren Pearce, Kim Holmberg, Iina Hellsten, and Brigitte Nerlich. 2014. Climate change on Twitter: Topics, communities and conversations about the 2013 IPCC Working Group 1 report. *PloS one* 9, 4 (2014), e94785.

[26] Jakub Piskorski, Nikolaos Nikolaidis, Nicolas Stefanovitch, Bonka Kotseva, Irene Vianini, Sopho Kharazi, and Jens P Linge. 2022. Exploring Data Augmentation for Classification of Climate Change Denial: Preliminary Study. In *Proceedings of the Text2Story'22 Workshop*. Elsevier, Stavanger, Norwway, 97–109.

[27] Elaine Teixeira Rabello, Efrat Gommeh, Andrea Benedetti, Gabriel Valerio-Ureña, and Tamara Metze. 2022. Mapping online visuals of shale gas controversy: a digital methods approach. *Information, Communication & Society* 25, 15 (2022), 2264–2281.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, Online, 8748–8763.

[29] Nils Reimers and Iryna Gurevych. 2019. 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, Hong Kong, China, 3982–3992.

[30] Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. arXiv:2004.09813

[31] Eduardo Rojas-Padilla, Tamara Metze, and Art Dewulf. 2024. Cliquepolitik: Multimodal online discourse coalitions on CRISPR-Cas genome editing technology. *Review of Policy Research* n/a, n/a (2024). https://doi.org/10.1111/ropr.12590 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ropr.12590

[32] Eduardo Rojas-Padilla, Tamara Metze, and Katrien Termeer. 2022. Seeing the visual: A literature review on why and how policy scholars would do well to study influential visualizations. *Policy Studies Yearbook* 12, 1 (2022), 103–136.

[33] Shamik Roy and Dan Goldwasser. 2020. Weakly supervised learning of nuanced frames for analyzing polarization in news media. arXiv:2009.09609

[34] Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. 2022. CLIP Models are Few-Shot Learners: Empirical Studies on VQA and Visual Entailment. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 6088–6100.

[35] Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Toronto, Canada, 1051–1066.

[36] Manfred Stede and Ronny Patz. 2021. The climate change debate and natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*. Association for Computational Linguistics, Online, 8–18.

[37] Tim M Stevens, Noelle Aarts, and Art Dewulf. 2020. Using emotions to frame issues and identities in conflict: farmer movements on social media. *Negotiation and Conflict Management Research* (2020). https://doi.org/10.1111/ncmr.12177

[38] Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. 2023. A multi-task model for sentiment aided stance detection of climate change tweets. In *Proceedings of the international AAAI conference on web and social media*, Vol. 17. AAAI Press, Limassol, Cyprus, 854–865.

[39] Apoorva Upadhyaya, Marco Fisichella, and Wolfgang Nejdl. 2023. Towards sentiment and Temporal Aided Stance Detection of climate change tweets. *Information Processing & Management* 60, 4 (2023), 103325.

[40] Roopal Vaid, Kartikey Pant, and Manish Shrivastava. 2022. Towards fine-grained classification of climate change related social media text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, Dublin, Ireland, 434–443.

[41] Christel W van Eck, Bob C Mulder, and Art Dewulf. 2020. Online climate change polarization: Interactional framing analysis of climate change blog comments. *Science Communication* 42, 4 (2020), 454–480.

[42] Francesco S Varini, Jordan Boyd-Graber, Massimiliano Ciaramita, and Markus Leippold. 2021. ClimaText: A dataset for climate change topic detection. arXiv:2012.00483

[43] Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. Climatebert: A pretrained language model for climate-related text. arXiv:2110.12010

[44] Hywel TP Williams, James R McMurray, Tim Kurz, and F Hugo Lambert. 2015. Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global environmental change* 32 (2015), 126–138.

[45] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. arXiv:2311.04257