NLP Research Note

Henry Baker - 228755

1 Research Question

*What are the topics and narratives within climate change misinformation discourse on social media; how do they vary across platforms?**

> ***NB**: This was my stated research question. Given the challenges of developing an externally-valid model for inference and uninformative results, the task of this research note shifted towards misinformation modelling *per se*. A comparative approach across platforms was dropped entirely for reasons discussed.

Introduction

Climate change misinformation presents a growing challenge for social media users, undermining public understanding, eroding trust in scientific consensus, and delaying policy action. Peer-reviewed studies have shown that misinformation narratives exploit scientific uncertainties and amplify polarisation (van der Linden et al., 2017; Cook et al., 2019). As global temperatures risk exceeding the 2°C limit of the Paris Agreement (IPCC, 2021), developing reliable technical tools to classify and understand misinformation in online discourse is increasingly critical for policy intervention.

This research note aims to analyse climate change misinformation using natural language processing (NLP) - specifically to develop a climate change misinformation classifier and apply a topic modelling analysis to identify prevalent themes within climate misinformation discourse on social media. NLP affords advantages over qualitative methods for large-scale analyses, allowing for efficient processing of vast datasets and the uncovering of underlying macro patterns out of reach of manual approaches - this is the intention behind our analysis. However, our findings indicate that the generic NLP-based modelling approaches naively applied here (training binary misinformation classifiers on large n amounts of example text) struggled to account for the contextual variability and subtlety of misinformation, which by definition mimics truthful communication. Off the back of these preliminary results, more considered NLP approaches are discussed.

By highlighting these limitations, this research note points to the need for knowledge-grounded NLP methods, such as retrieval-augmented classifiers (RAC) and knowledge-enhanced models, which can incorporate external knowledge to address misinformation.

Data & Methods

Data Preparation (i): 'Model Development' Dataset (in script `1_data_processing`)

For developing (training, validating, and testing) our classifier models, we initially use the openly available *Twitter Disinformation* dataset on Hugging Face (Minassian, 2021) containing tweet texts labelled as either 'factual' or 'misinformation' ('factual' should better be understood as *'not misinformation'* or *information* and contains text clearly characterised by opinion and other modes of communication beyond assertion of facts in the scientific sense). We repartitioned the dataset into 92,394 examples for training and validation and 12,267 for testing. A class imbalance of 70:30 (factual-to-misinformation) was maintained; drawing from the literature, 30% was considered a reasonable upper bound to ensure sufficient exposure to positive examples for training rare-event classifiers (Shu et al., 2017).

We implemented two distinct data preparation workflows: one for the CNN and RNN family of models, another for the Hugging Face Transformer models. Given dataset size, all workflows—including data preparation, model development and inference—used batched processes for memory and efficiency. For preprocessing text data for the RNN family of models, we used a custom tokenization pipeline using SpaCy's `en_core_web_sm` lightweight model for word-based tokenization and vocabulary-based filtering (integrating a `CountVectorizer` to build the vocabulary). To reduce noise, we set a minimum document frequency threshold of three occurrences. Tokens were indexed by frequency to create a vocabulary mapping. The final vocabulary size was 74,703 and the token frequency represented a Zipf's Law-like distribution (see Appendix (i)), where a concentrated number of stop words and punctuation tokens dominate, and a long tail for lower-ranked tokens forming the general rest (with 'Trump' as the only substantively-interesting word-token within the top 30 most frequent). This highlights the need for a custom analyser (applied to both the test and inference data) for robustly handling stop words, out-of-vocabulary (OOV) and rare tokens (unseen in the train data). The tokenized outputs for all datasets were cached as pickle files, which are accessible via our GitHub repository.

For batching, DataLoaders were configured to handle integer-encoded sequences truncated to a maximum length of 300 tokens, padded as necessary. At this threshold, the average number of tokens per sequence was 233.76 in the training dataloader, and 228.37 in the test. For the vanilla RNN models, shorter sequence lengths were later experimented with to address exploding/vanishing gradient issues during training, balancing a narrowed context window against exploding/vanishing gradients. This trade off was non-constraining when applied to short form tweets, but was a major limitations when applying the vanilla RNN to longer sequences. We created two versions of the DataLoaders: one with sequence lengths metadata added (for RNNs) and one without (for CNN-based models). Pre-trained FastText embeddings (from Bojanowski et al, 2015; made available by Facebook AI Research (FAIR)) were mapped to the vocabulary to construct an embedding matrix (dimensions 77,217; 300) which served as input to the models for richer semantic representation.

For Transformer-based models, we employed a pre-trained BERT tokenizer (associated with the selected Transformer and checkpoint) to the text sequences, which were truncated to 512 tokens and padded. Hugging Face Dataset objects were created for both training and evaluation workflows, and dynamic embeddings were handled natively by the model's attention mechanism. The Transformers' dynamic attention-based embedding generation eliminated the need for static embedding matrices, simplifying this second workflow while also allowing for a wider context window, which provided a noticeable advantage only in the more challenging classification scenarios. During model tuning, we compared tokenized sequence lengths between workflows (with a maximum of 300 tokens for RNN/CNN workflows vs. 512 for Transformers). When working with longer-form textual sequences there were noticeable differences (see below), but for Twitter input data, performance improvement based on sequence length was negligible given that the average sequence length was already below 250 tokens. The only exception was the vanilla RNNs, which struggled even with lengths of 250 tokens.

Modelling (in script `2_model_dev`)

> **A note on modelling:**
>
> The modelling capacity discussed below exceeds the requirements for short-form Twitter analysis. Initially designed for a research note covering multiple social media platforms, time constraints and a possible category error in our approach (coverd in the discussion) led to a narrower focus on Twitter exclusively. Consequently, the model applied to Twitter data is somewhat of a 'sledgehammer to crack a nut', but offers a foundation for potential future comparative analysis.

We trained various model architectures with some limited hyperparameter tuning. Cached pickle files for the best-performing models are available on GitHub. Performance was evaluated using F1 scores (balancing precision and recall) and accuracy. Models were trained for 10 epochs, except Transformer models, which were trained for up to 30 epochs with early stopping after 3 epochs of no F1 improvement.

We implemented a basic Convolutional Neural Network (CNN) with a 1D convolutional layer, adaptive average pooling, and a final linear layer for binary classification. Some limited experiments with convolutional layers, filter sizes (64, 128, 256), and kernel sizes (3, 5, 7) aimed to capture n-gram contexts and feature hierarchies, and while initial improvements were observed, given the short form nature of the input text sequences *all* models demonstrated sufficiently capacity and expressivity with minimal tuning required. As such, further tuning provided diminishing returns which were marginal and often reversed when working across different model development datasets. For the vanilla Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) Networks, and Gated Recurrent Units (GRUs) we also experimented with larger hidden sizes (64, 128, 256), stochastic dropout in fully connected layers, bidirectionality, and shorter sequence lengths for the vanilla RNN. We also attempted pooled representations (e.g., max pooling, mean pooling) across time steps, although these yielded imperceptible advantages against the cell/unit structure of the LSTM/GRUs. Although vanilla RNNs struggled with longer sequence lengths, given the capacity of the LSTM/GRU implementations (especially with bidirectional configurations and additional dropout regularisation) and the GRU's particularly rapid convergence (due to fewer parameters), we took the GRU as our non-Transformer model of choice - with an F1 score of .94 - and didn't further optimise the others.

To fine-tune the Transformer model, we utilized a pretrained BERT base model (110m parameters; the uncased version) and its associated tokenizer, both obtained from Hugging Face. Using the default learning rate scheduler from Hugging Face's Trainer class, training usually continued for between 9-13 epochs, with ~.9 F1 score reached usually within 4 epochs, and the best performing model (by F1 score on the validation set) selected.

BERTweet:

BERTweet was initially chosen due to its pretraining on 845 million English tweets collected from 2012 to 2020, making it particularly adept at capturing the linguistic nuances, informal language, and stylistic features characteristic of tweets. Similarly, BERTweet tokenizer uses RoBERTa-style tokenization, optimized for tweets and handles Twitter-specific tokens better than standard BERT tokenizers (hashtags are treated as distinct tokens; it processes `@usernames` effectively; and emojis are retained as meaningful tokens).

However, the size of the model made it impractical to train given available compute resources and so we resorted to the lighter weight BERT base models.

The Transformer model consistently achieved the highest prediction scores across all the model development datasets we went through across this project. For the Twitter-specific model development, this advantage was negligible against the RNN family of models; it was only when developing models for longer-form social media text sequences that the Transformer's ability to handle long sequential data became significant. Despite the Transformer's relative out-performance, we use the GRU as our baseline model for downstream analysis given its comparative explainability - a worthwhile trade off for a negligible dip in in F1 score.

Data Preparation (ii): 'Inference' Dataset (in script `1_data_processing`)

We initially ran inference using the trained models on an unlabelled Kaggle dataset, *Climate Change Twitter* (Die9origephit, 2020). Basic RegEx was also applied to remove URLs and extraneous numbers, which had not been present in the training or validation sets to ensure a degree of consistency between training and inference. This was a research design choice as they may have contained useful information for classification. Parallel data preparation pipelines were conducted for both Transformer- and RNN-based models. The Transformer workflow used the pretrained tokenizer, while the RNN workflow employed the trained custom analyser which word tokenised the sequences and mapped FastText embeddings.

Applying the selected bidirectional GRU model to the inference dataset, 62.6% of inputs were predicted as 'misinformation' (the Transformer model was broadly equivalent). This contrasted sharply with the (loose) consensus around expected distribution in the literature (Shu et al, 2018; Vosoughi et al, 2018, Hassan et al, 2024), as well as the training data distribution. Upon manual inspection of the classified texts, the model's outputs appeared arbitrary, failing to capture substantive differences between the two classes. Selected examples of arbitrary classification can be found annotated in Appendix (ii).

This outcome provided an unpromising foundation for conducting a meaningful topic modelling exercise or making substantive claims about the content of climate misinformation on social media. To address this, before proceeding we decided to recalibrate the model by implementing primary attribution techniques to better understand its classification criteria. We employed Integrated Gradients (IG) (with a zero baseline) and Local Interpretable Model-Agnostic Explanations (LIME) to identify primary attributions, aggregating token-level scores to the word level. Both techniques revealed broadly similar patterns of model behaviour (see Appendix (iii) for a side-by-side comparison). As we had access to model gradients and focus on global model behaviour (rather than explanations of individual predictions) we prioritised IG attribution scores where the two diverged. Both methods revealed that the model failed to extract meaningful patterns from the data. Most concerningly, most climate-related terms carried large coefficients driving a misinformation classification, this was true of both the Transformer and the GRU model. See Appendix (iii).

We hypothesised these seemingly invalid and implausible inference results derived from a bias in the model development dataset, where climate terminology and related-themes were disproportionately associated with misinformation labels. A simple count confirmed that climate-related tweets in the model-development dataset were indeed overrepresented among misinformation labels: of the 10,267 climate-related examples from the 92,000 training examples, over 8,000 of them were associated with misinformation labels by our estimation. Furthermore, inspection of the dataset's Hugging Face data card revealed that it included data from a *Natural Hazards Twitter Dataset*, where disaster-related tweets had been somewhat arbitrarily re-labelled as 'misinformation' while non-disaster tweets were mapped as 'factual'. This likely propagated a bias through to the results. This represented a "garbage in, garbage out" (GIGO) issue, prompting us to adopt a more deliberate approach to data acquisition for model development.

Data Revisions & Augmentation (in scripts `1_data_processing; 3_inference`)

Through multiple iterations of data refinement and augmentation, followed by model retraining and inference, we developed a more robust classifier. Evaluating classifier performance at inference followed three stages of increasing granularity:

1. **Predicted class ratio:** We loosely expected <15% misinformation prevalence as a (generous) upper bound in a given sample of real-world data, loosely based on the literature.
2. **Analyst inspection:** A sample of predicted labels was reviewed against input text.
3. **Primary attribution analysis:** Attributions (Integrated Gradients, validated by LIME scores) were inspected for alignment with domain knowledge.

Given more time, Bayesian uncertainty estimation techniques (e.g., Monte Carlo Dropout, Ensemble methods) would have been preferable additions.

We hypothesised that no universal vocabulary, style or rhetorical pattern of misinformation exists, Instead exposing the model to substantively relevant examples of *climate-specific* misinformation would allow it to develop a representation of a body of *domain* knowledge with which to evaluate content for factuality or not. To do so, we filtered the model development dataset on a climate keyword pattern using a combination of RegEx-based exact matching, phrase matching (using spaCy's `PhraseMatcher`), and fuzzy matching (with `rapidfuzz` at a .8 similarity threshold) to maximise both precision and recall - see Appendix (iv). Filtering for climate-related keywords resulted in a dataset biased toward misinformation examples. We undersampled and augmented with additional twitter misinformation datasets - see references - which were also filtered for climate-relevance and similarly processed, to arrive back at a 70:30 factual-to-misinformation split of climate-relevant training examples.

Final Model Evaluation Metrics

After iterative data refinement we obtained the final model evaluation metrics on the test sets.

	CNN	**Vanilla RNN**	**LSTM**	**GRU**	**Transformer**
Tuning; Architectural Extensions	# Conv layers # Filters Kernel size	Hidden sizes Dropout Seq length	Stacking Bi-directional Dropout	Bi-directional Dropout	BERT Base (uncased) BERTweet
Highest Test F1 Score	.934	.724 (no convergence)	.936	.940	.985

Table 1: Optimal model implementations on final model development dataset - pickle files made available on GitHub.

It was surprising that the CNN out-performed some of the RNN-based models, likely due to the structure of short tweets in which local patterns or features are more informative than sequential dependencies. Consequently this advantage disappeared when training and evaluating on longer form news and non-Twitter social media content, where the CNN saw the largest deterioration, with an F1 score of .16, and an accuracy rate of .56 reflecting the underlying distribution of classes in the evaluate dataset rather than any model capacity to distinguish between them. The Transformer consistently outperformed all models in all contexts.

Also of note, training the vanilla RNN at a 300 token maximum - unlike subsequent LSTM and GRU implementations - induced instability in learning across epochs, characteristic of exploding/vanishing gradients issues. Shortened sequence lengths helped (150 tokens), but narrowed the context window. We could have further experimented with gradient clipping or an adaptive learning rate, but since the final LSTM/GRU implementations were more than sufficient for our purposes, we moved on.

A note on dataset augmentation and platform-specific limitations:

We further hypothesised that tweets were a limited text format for downstream analysis of climate misinformation, given their short sequences and the importance of URL-linked content and embedded data to constructing the meaning of the text.

Accordingly, we experimented with augmenting both the model-development and inference datasets to include other longer-form social media text. We further augmented the model-development dataset only with high-quality fake news classification datasets from diverse online media sources including both new and legacy media outlets. Although these training inputs included non-social media content (potentially inducing a difference in underlying distributions between development and inference), we hypothesised they would provide valuable learning for the classifier given the expanded textual format and volume of high-quality labelled training data available.

Two composite datasets (for development and inference) were created, with text processing and class- and source-balancing applied to ensure consistency. Evaluations of the model on the composite dataset had a deleterious effect on evaluation scores (a .66 F1 score for the best-performing GRU; .71 for the Transformer).

Without more considered data acquisition and processing (and perhaps further modelling work), a platform-agnostic misinformation classifier remains out of reach for this project. Instead we settled on a Twitter-specific classifier (using exclusively Twitter data in both the model development and inference datasets). This was a practical decision taken in light of time constraints. However, it limits the external validity of the results to non-Twitter platforms, and comes with the caveat that text sequences taken from tweets remain a limited format for downstream analysis of misinformation dynamics - something which became readily apparent.

Final Inference Metrics

For downstream inference, data refinements reduced the misinformation classification rate to 8.1% for the best-performing GRU, and 27.2% for the Transformer (see Appendix (v)). This aligned more closely with our (reasonably informed) prior expectations, and a manual inspection of the labels suggested a closer fit (see Appendix (vi) for annotated inspection). Words with the highest IG coefficients still remained uninterpretable / arbitrary, with multiple stop words and tokens carrying greater weight than domain-relevant examples - see Appendix (vi). Similarly, there remained surprising (and likely wrong) classifications. Further work is needed to improve the robustness of the models to out-of-distribution inference challenges, and our subsequent analysis lacks validity as a result.

A final inspection of the composite model development data found remaining problematic examples of non-Twitter data that had been included in supposed Twitter-only datasets, alongside questionable ground truths (based on our qualitative assessment). This is disappointing given the lengthy consideration we had given the data work elements of this research, but highlights the importance of high quality training data, and we expect the remaining issues necessarily lie here. If further time was available we would direct still further attention towards the data acquisition part of the pipeline.

Analysis of Final Results (in script `4_analysis`)

A note on the validity of this analysis:

The models lack robustness for inference on out of the model development dataset. By narrowing the scope to Twitter, we addressed external validity partially, but internal validity on Twitter-only data remains suboptimal. Despite these limitations, we proceed with the analysis here as it aligns with the research note's stated purpose. However, due to the models' unsatisfactory performance at inference, this analysis should be considered preliminary, and the subsequent discussion will address the broader implications of these non-robust findings.

Inference was conducted using the top-performing RNN model (a bidirectional GRU with dropout) and a fine-tuned base BERT Transformer model. As noted, the GRU's positive classification rate on the inference dataset was 8.1% (1,535 cases), compared to the Transformer's 27.2% (5,186 cases). The discrepancy in positive classification rates between the two models is not immediately clear and requires further investigation.

While we can inspect the GRU's misinformation classifications with attribution techniques (Integrated Gradients and LIME), to identify the top global features driving its predictions (see GRU attribution print outs in the `4_analysis` script), interpreting the Transformer model is more challenging due to its use of dynamic embeddings, which complicates global model interpretation (see Transfromer attribution print outs in the `4_analysis` script). A like-for-like comparison is thus challenging. To robustly interpret the Transformer's attributions would require aggregating them across examples to detect broader patterns, which introduces further noise and is impractical for this study. Instead, we manually review the results to conclude that the GRU provides more 'reasonable' classifications - see Appendix (vi) for annotated outputs of the first 10 text inputs and predicted labels). This conclusion, aligns with the parallel ML principle of Occam's Razor, leading us to take the GRU's classifications as the basis for further interpretation.

Topic Modelling

The topic modelling results (see Appendix (vii)) confirm limitations in the model's classifications. Several topics are heavily influenced by platform-specific artifacts, such as "https", "replying", "tweet", "thread" and other non-substantive terms, which undermine their coherence and interpretability (topics 1, 2 and 3) - further analysis would need to handle this platform-specific noise. Topics with terms like "climate", "change", "hoax" and "policy" show some alignment with expected climate discussions but remain fragmented and lack depth in addressing misinformation narratives. As is often the case in topic modelling the first topic is too general for meaningful interpretation; unfortunately that remains the case through both topics 2 and 3 as well). Other topics focus on temporal patterns (the months of topic 3 perhaps reflect the timing of discussions rather than contributing directly to understanding misinformation narratives) or political themes, such as Australian politics (topic 5's "government", "australia", "pacific", "election", "auspol" and "action"), but fail to provide actionable insights into the dynamics of climate misinformation. Topic 4 seems to bridge climate and public health narratives ("covid," "global," "health"), offering a potentially interesting intersection for further analysis, but it is then diluted by the inclusion of substantively irrelevant terms, limiting interpretability.

Similarly, visualising these keywords as a network or plotting them using input-level embedding coordinates (see Appendix (vii)) did not yield much additional analytical value. Taking a moment to unpack/critique the network analysis: the highly connected nodes which act as pivots for multiple topics ("climate", "change", "tweet", "global") are uninformative; more informative subtopic words ("health", "care", "young", or the Australian politics topic) are represented as isolated nodes, suggesting the public health/Australian politics angles are less integrated within the broader topics and remain if not fringe, then non-dominant issues.

Given the unforeseen time-sinks faced with the modelling and data acquisition, we leave further parsing of results for later iterations on this pipeline, once higher quality model training datasets have been implemented, producing more robust climate misinformation classifiers.

Discussion

The task of modelling disinformation in online social media text proved to be more challenging than anticipated, raising pertinent questions as to (i) the suitability of NLP for detecting disinformation and (ii) the feasibility of developing a context-agnostic model for misinformation detection. These challenges have policy-relevant implications for how social media companies, regulators, and researchers address the issue.

(i) Suitability of NLP For Modelling Disinformation

Our preliminary results suggest there is no universal pattern—whether syntactic, stylistic, or referential—that reliably characterises misinformation. Fundamentally, misinformation is designed to mimic truthful communication, deceptively conveying untruths in a way that mirrors legitimate information. This makes detecting misinformation as a reified concept problematic. Attempting to build a generic classifier to identify all misinformation, as originally pursued in this research, represents a categorical fallacy.

Instead, we suggest that both platform and domain-specific contexts must be foregrounded when designing misinformation classifiers. This does not preclude the use of NLP in misinformation classification (especially given the extremely high F1 scores achieved on the test sets, indicating that there are patterns and structure that can reliably be leveraged for learning) but highlights the need for a refined, contextually-driven approach. Rather than assuming the existence of an irreducible, context-independent signal of misinformation, a more effective framing is as a fact-checking task. Textual inputs should be evaluated against a predefined body of knowledge to assess their veracity.

In this vein, we propose that high-performing classifiers require substantively relevant examples to build representations of the body of knowledge they draw upon. While our results do not conclusively support this hypothesis, neither do they definitively refute it, requiring us to accept the null within the scope of this study. Future efforts might explore retrieval-augmented classifiers (RAC), knowledge-enhanced neural networks, memory-augmented networks, or large language models (LLMs) with emergent capabilities resembling real-world knowledge or experience.

(ii) Context Is All You Need

Our analysis underscores the critical role of both platform and domain context. Evidence for this includes the strong performance metrics achieved when models were trained and evaluated on platform- and domain-specific misinformation. However, these results deteriorated sharply when cross-platform cross-domain learning and inference were attempted (or even inference across different datasets taken from the same platform). It remains an open question whether platform or domain generalisation poses the greater challenge, but next steps would be to identify the necessary acquisition (and preprocessing) pipeline to achieve valid inference results even within the same platform and domain. We put significant effort into this part of the pipeline but still find problematic content and ground truths in the finally-assembled model training dataset. In an environment of readily available open datasets on platforms such as Kaggle and Hugging Face, duly considered data acquisition, verification and processing are paramount.

Domain-context: Initial results suggested that domain context might not be the primary obstacle, as models performed well on test datasets containing diverse domain-variant examples. On the other hand, applying these trained models to unlabelled, climate-specific Twitter data at inference yielded mixed results. In contrast, refining the training dataset to include only climate-specific examples improved the validity of inference labels, indicating that domain context does in fact play a significant role. That this improvement derived from narrowing the training dataset also contradicts an early hypothesis we had formed: that poor inference results were due to an overly-narrow training vocabulary (taken from the training set) and a subsequent high proportion of out-of-vocabulary (OOV) tokens when applied to the inference dataset. An analysis of tokenisation revealed that OOV tokens were not excessively numerous, suggesting the issue lay elsewhere. Further investigation is needed to fully understand these dynamics.

Platform-context: Similarly, expanding the training and evaluation datasets to include multiple platform sources degraded model performance immediately on the test set, indicating that the models leveraged platform-specific structure in the data for their predictions.

This research note does not make the claim that a generic context-agnostic classifier is unachievable. Many potential areas for improvement remain unexplored in this study - especially at initial points in the pipeline (from data acquisition to tokenisation and preprocessing). Nevertheless context-sensitive solutions like RACs, knowledge-enhanced neural networks, and memory-augmented networks show promise. These approaches raise critical questions about the nature of "knowledge"—whose knowledge is represented, who determines its validity, and under what criteria. These questions inevitably extend beyond technical considerations, implicating social, moral, and epistemological challenges.

References

Literature

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135–146. https://doi.org/10.1162/tacl_a_00051

Cook, J., Nuccitelli, D., Green, S. A., Richardson, M., Winkler, B., Painting, R., ... & Skuce, A. (2013). Quantifying the consensus on anthropogenic global warming in the scientific literature. *Environmental Research Letters, 8*(2), 024024.

Cook, J., Oreskes, N., Doran, P. T., Anderegg, W. R. L., Verheggen, B., Maibach, E. W., ... & Green, S. A. (2019). Consensus on consensus: A synthesis of consensus estimates on human-caused global warming. *Environmental Research Letters, 11*(4), 048002. https://doi.org/10.1088/1748-9326/11/4/048002

Hassan, I., Musa, R. M., Latiff Azmi, M. N., Razali Abdullah, M., & Yusoff, S. Z. (2024). Analysis of climate change disinformation across types, agents and media platforms. *Information Development, 40*(3), 504–516.

Intergovernmental Panel on Climate Change (IPCC). (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press. https://www.ipcc.ch/report/ar6/wg1/

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter, 19*(1), 22–36.

Treen, K. M. D. I., Williams, H. T., & O'Neill, S. J. (2020). Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change, 11*(5), e665.

van der Linden, S., Leiserowitz, A., Feinberg, G. D., & Maibach, E. W. (2017). Inoculating against misinformation. *Science, 358*(6367), 1141–1142. https://doi.org/10.1126/science.aar4533

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151.

Datasets

Minassian, R. (2021). *Twitter misinformation* [Dataset]. Hugging Face. https://huggingface.co/datasets/rouper
misinformation

Die9origephit. (2020). *Climate change tweets* [Dataset]. Kaggle. https://www.kaggle.com/datasets/die9origeph
change-tweets

Konradb. (2020). *Reddit lies tweets* [Dataset]. Kaggle. https://www.kaggle.com/datasets/konradb/reddit-lies-tweets

Shahane, S. (2020). *Fake news classification* [Dataset]. Kaggle. https://www.kaggle.com/datasets/saurabhshaha
news-classification

Depak, M. (2020). *FakeNewsNet* [Dataset]. Kaggle. https://www.kaggle.com/datasets/mdepak/fakenewsnet

Thedevastator. (2020). *Social media sentiment and climate change* [Dataset]. Kaggle. https://www.kaggle.com/ media-sentiment-and-climate-change

Reddit Climate Change. (n.d.). *Reddit climate change* [Dataset]. https://www.kaggle.com/datasets/pavellexyr/ reddit-climate-change-dataset

Models

Hugging Face. (n.d.). BERTweet. Retrieved November 30, 2024, from https://huggingface.co/docs/transformers

Appendix

GitHub repo (for scripts & pickle files): https://github.com/henrycgbaker/nlp_research_note

Running `2_model_dev` will call preceding & auxillary scripts automatically.

(i) Vocabulary

```
Vocab size:  77215

Vocab first 20 idx printed:  [(’<pad>’, 0), (’<unk>’, 1), (’the’, 2), (’,’, 3), (’.’,
4), (’to’, 5), (’of’, 6), (’and’, 7), (’a’, 8), (’in’, 9), (’ ’, 10), (’that’, 11),
(’-’, 12), (’on’, 13), (’for’, 14), (’is’, 15), (’s’, 16), (’it’, 17), (’said’, 18),
(’he’, 19)]
```

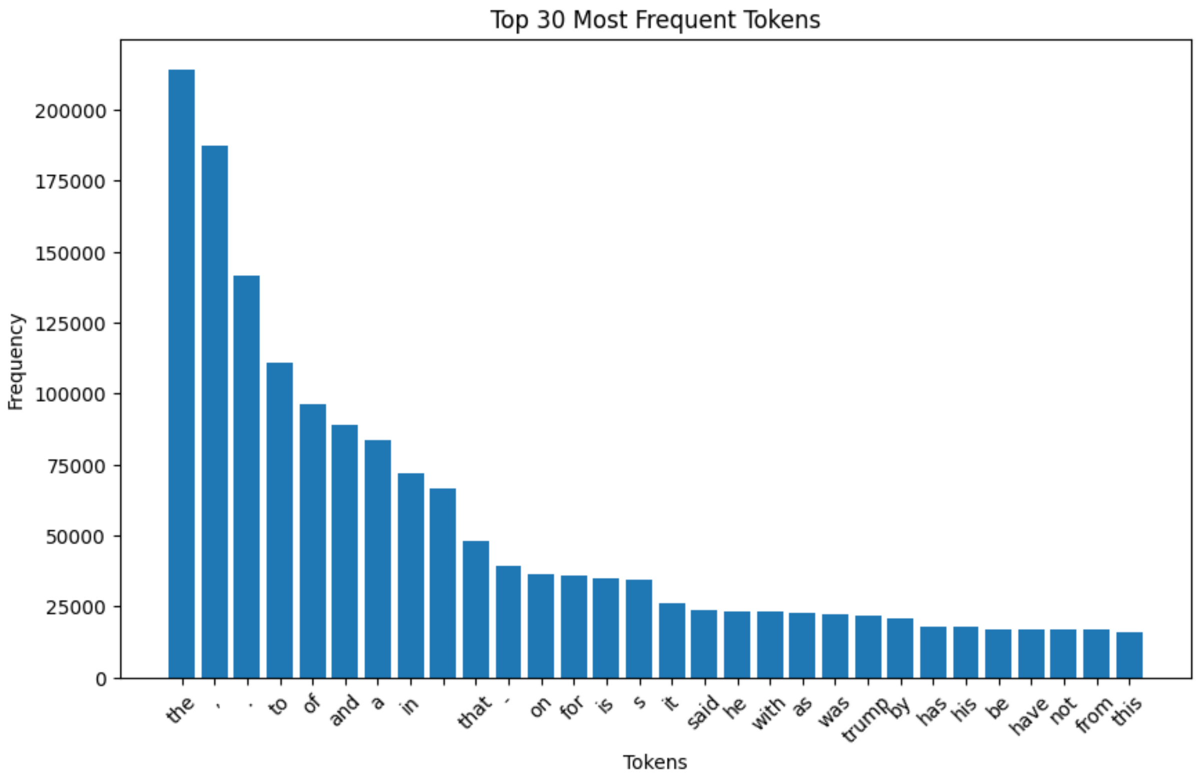


Figure 1: Vocabulary: top 30

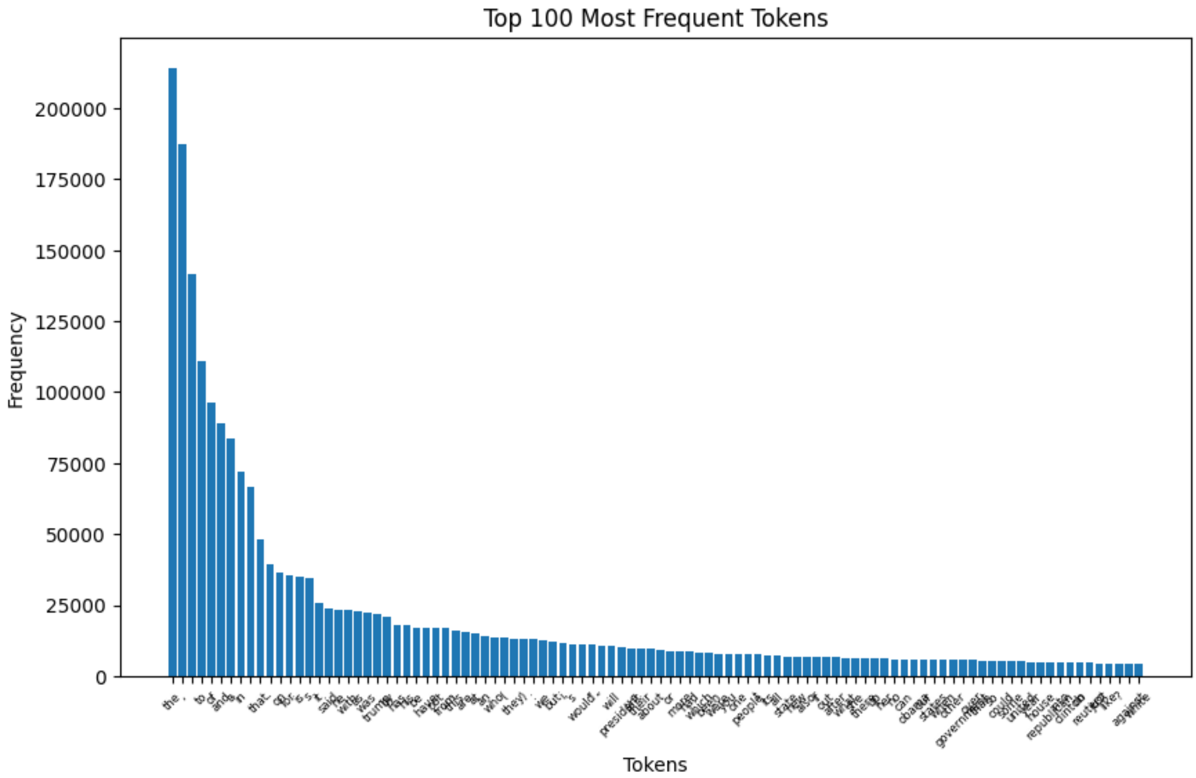


Figure 2: Vocabulary: top 100

(ii) Sample of inference positive classifications from (pre data refinement) modelling

NB: this small selection of sequence-labels were selected as they reflect the obvious mis- inappropriate-classifications occurring when running inference before data refinement took place.

Text	Predicted Label
A new study suggests sharks will need to adapt, move or die as climate change could soon render their nurseries uninhabitable.	`misinformation`
Coastal lowland tiles are ones that can flood with climate change. This thread has some pictures that show examples: https://forums.civfanatics.com/threads/climate-change-quirks.653220/	`misinformation`
The devastating threat of climate change is on our doorstep.	`misinformation`
The Hidden Truth about Global Warming and the Paris Climate Accord | The Climate Change Biz Boom	`misinformation`
"These investments are critical to building a stronger economy in the First State that is more resilient to threats like climate change. That's why"	`misinformation`
"Have we let climate change go too far?? (We have)"	`misinformation`
"Individual active transportation efforts [are] a good example to help pave the way for world leaders to be held accountable for their climate change, & public health, commitments" Check out the commentary by JPAH editorial board member, Tom Kane	`misinformation`
Labor leader Anthony Albanese has gained an edge over Prime Minister Scott Morrison on climate change policy ahead of the next election, with 37 per cent of voters backing Labor's target to cut carbon emissions |	`misinformation`

Table 2: Sample of inference classifications from (pre data refinement) model

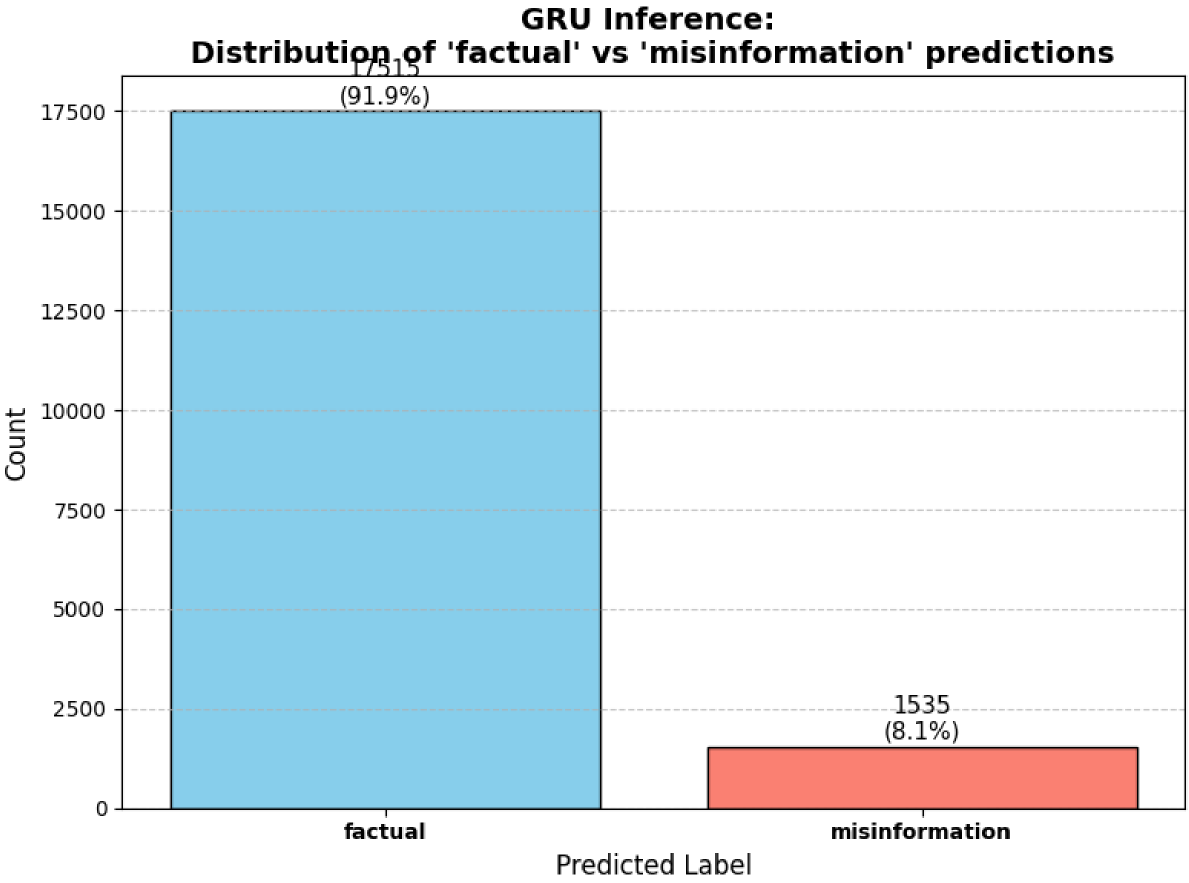
(iii) Integrated Gradients (IG) vs Locally Interpretable Model-Agnostic Explanations (LIME) analysis of the preliminary model's behaviour

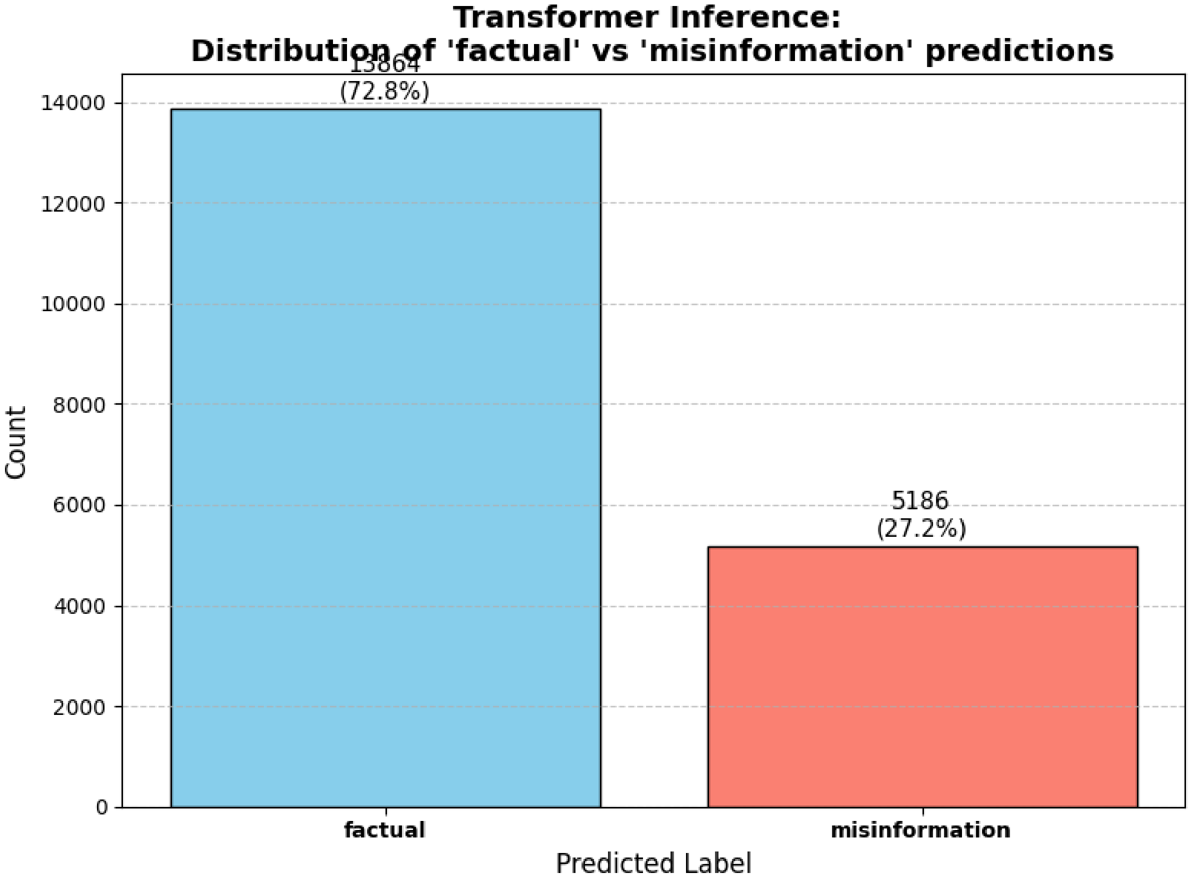
We worked to include output here, but could not get it to render; please refer to script `4_analysis` for the printed results.

(iv) Climate Keyword Pattern

```
keywords = [ "climate", "sustainability", "global warming", "carbon", "greenhouse",
"emissions", "renewable", "biodiversity", "ecology", "sustainable", "fossil fuels",
"energy transition", "carbon footprint", "net zero", "solar power", "wind energy",
"climate crisis", "carbon neutrality", "deforestation", "environmental", "pollution"
```

(v) Final Results: Inference Classification rates

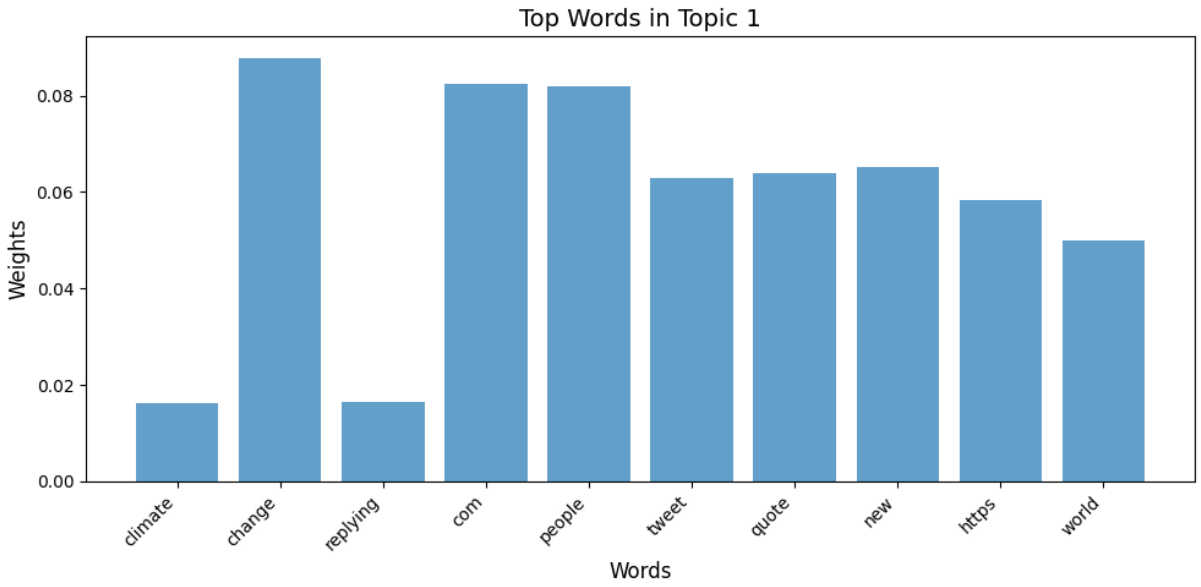


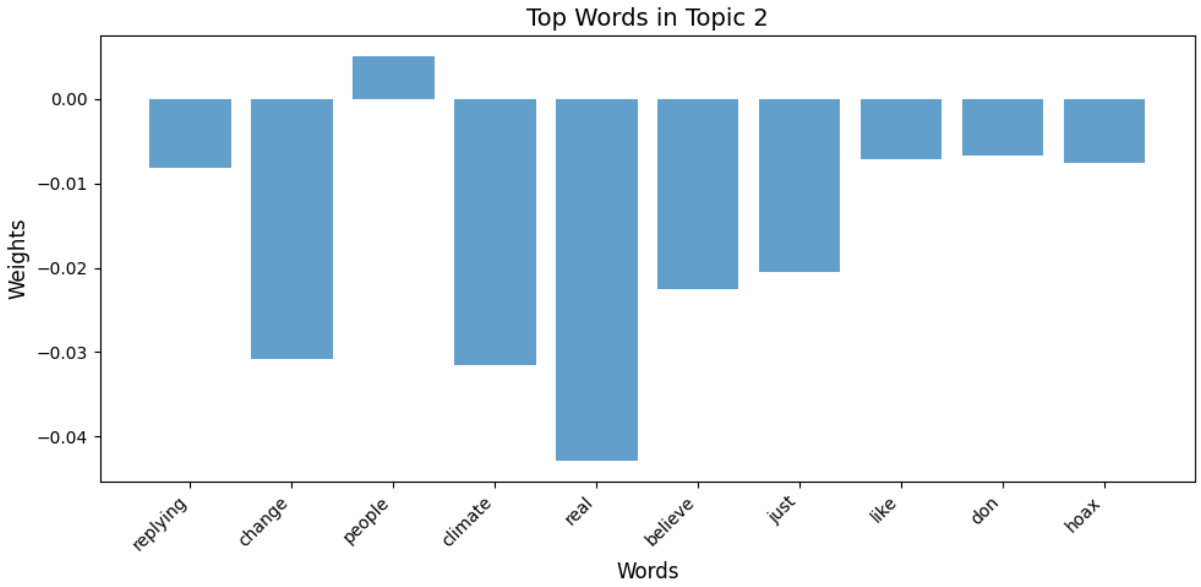

Transformer Inference:
Distribution of 'factual' vs 'misinformation' predictions
13864
(72.8%)
5186
(27.2%)
Count
0
2000
4000
6000
8000
10000
12000
14000
factual
misinformation
Predicted Label

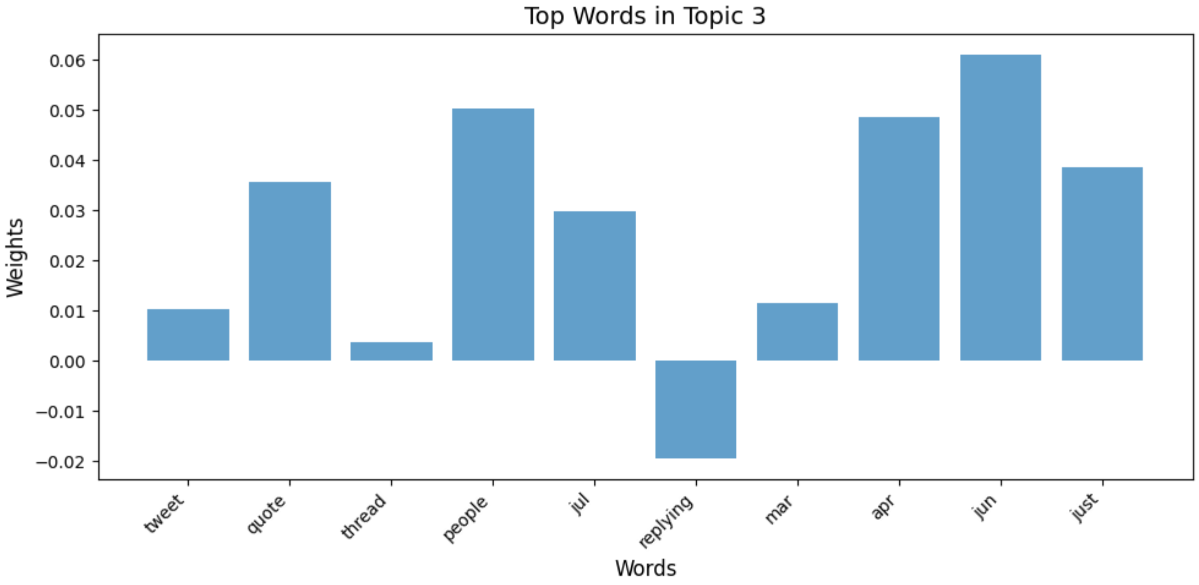
(vi) Sample of inference classifications from final modelling (first 10 by index)

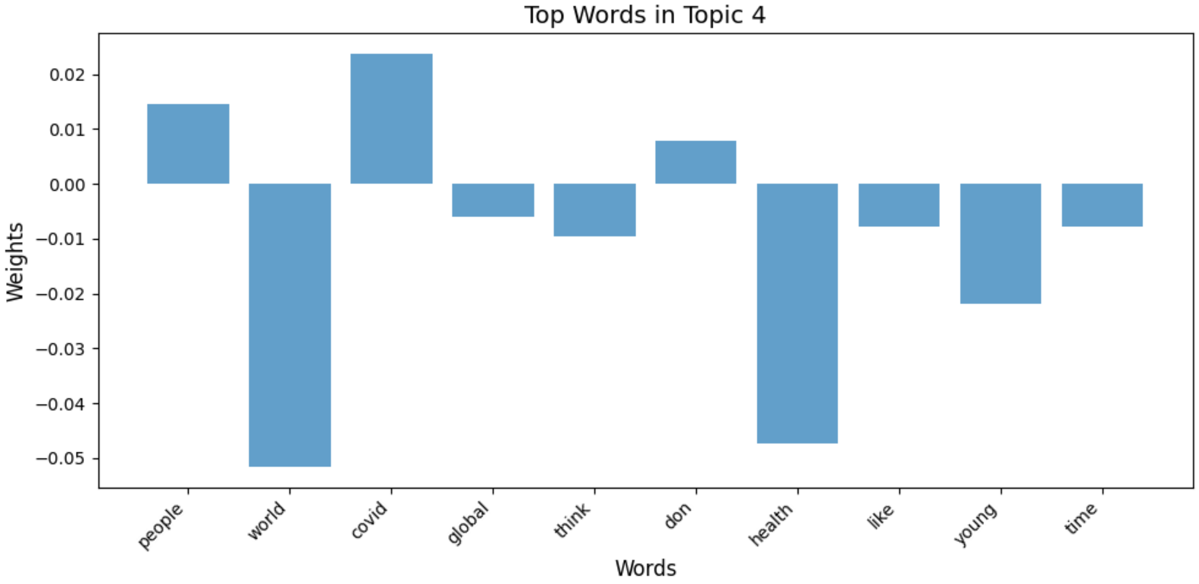
Text	Predicted Label (Transformer)	Predicted Label (GRU)	Analyst Comment
The only solution I've ever heard the Left propose for climate change is more taxes, more control and less freedom. They have one playbook and it fails every single time.	misinfo	factual	Good fit by GRU; Transformer's classification likely driven by partisan rhetoric - as this is neither factual nor misinformation, rather an opinion. For cases like these, misinfo or not is an unhelpful binary.
Climate change doesn't cause volcanic eruptions.	misinfo	factual	Good fit by GRU.
Vaccinated tennis ball boy collapses in the tennis court due to climate change.	factual	factual	Difficult to classify - Tweets are a suboptimal format for identifying misinformation.
North America has experienced an average winter, with temperatures and snowfall totals in line with historical trends. Do not be fooled. This phenomenon is known as "Asymptomatic Climate Change".	factual	factual	Good fit by both.
They're gonna do the same with Climate Change when it starts to get really bad. Quote Tweet rec @joeywreck · Jan 17 They really want you to fucking live with covid.	factual	factual	Difficult to classify - Tweets are a suboptimal format for identifying misinformation.
HELLO AMERICA, Who would have ever thought the World could be taken; by fearing Climate Change, the Common Cold, and the FLU?	factual	factual	Most likely misinformation.
fucking hell this weather makes me really fucking anxious bc climate change is only getting worse and worse and my god we'll all be dead within a couple of years	factual	factual	The 'couple of years' is hyperbole, but could be misread as misinformation. Still, both models seem to focus on anything promoting climate change as real is to be trusted. Hard to classify.
Great to finally have this important UNESCO/SCOR publication finished outlining techniques and approaches to tackling harmful algal bloom research in the context of climate change. Thanks to all the authors for excellent chapters. https://oceanexpert.org/document/29762	misinfo	factual	A prominent example of assertion of scientific fact - GRU a good fit, Transformer misclassified badly.
Climate change is one of the world's most pressing problems. RedCodeForHumanity So, it makes no sense when mainstream media publishes an article about ClimateChange with a paywall. This is profit over people and planet. MSMFail Should	factual	factual	Both classify well.

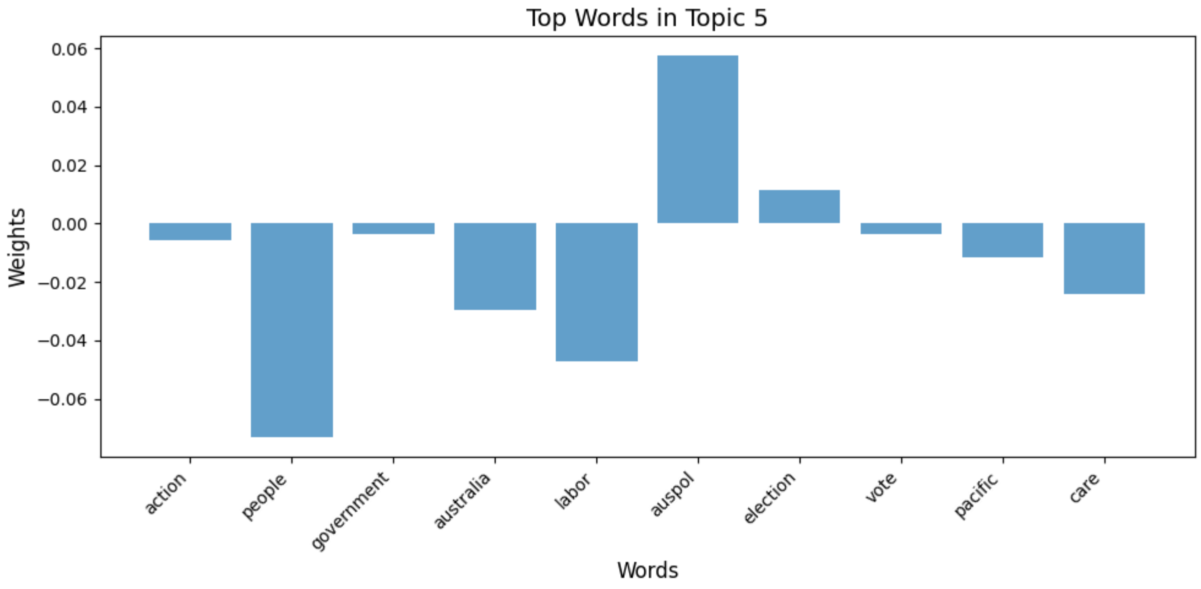
(vii) Topic Modelling

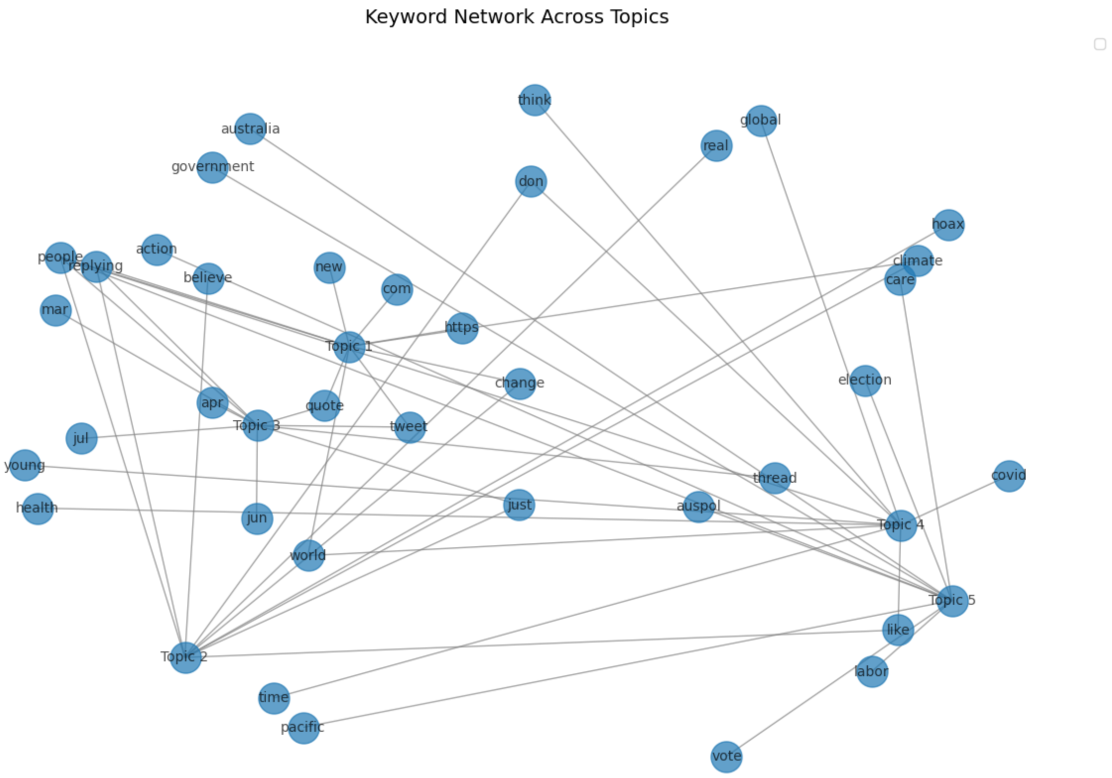


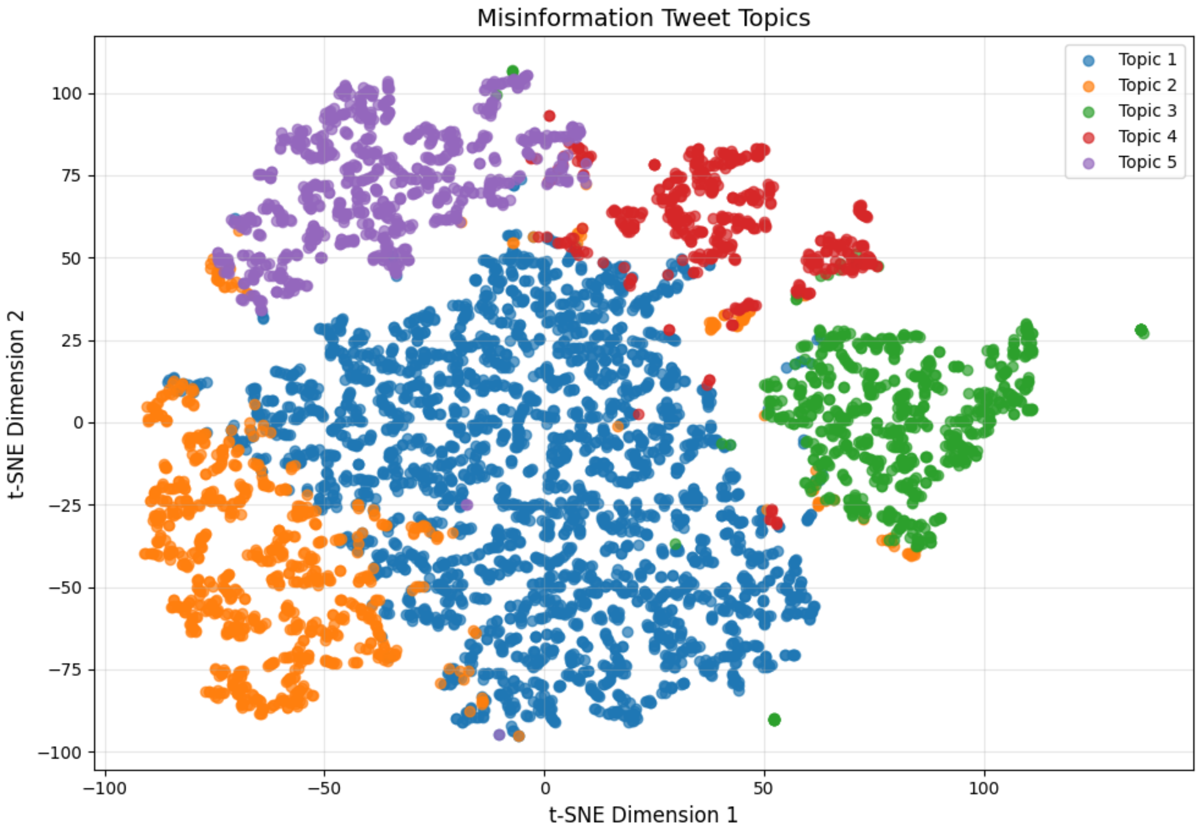



Top Words in Topic 3
Weights
0.06
0.05
0.04
0.03
0.02
0.01
0.00
−0.01
−0.02
tweet
quote
thread
people
jul
replying
mar
apr
jun
just
Words


Top Words in Topic 4
Weights
0.02
0.01
0.00
−0.01
−0.02
−0.03
−0.04
−0.05
people
world
covid
global
think
don
health
like
young
time
Words


Top Words in Topic 5
Weights
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
action
people
government
australia
labor
auspol
election
vote
pacific
care
Words


Keyword Network Across Topics
think
global
australia
real
government
don
hoax
action
people
replying
new
climate
believe
care
com
mar
https
Topic 1
election
change
apr
quote
Topic 3
tweet
jul
covid
young
thread
just
auspol
health
jun
Topic 4
world
Topic 5
like
Topic 2
labor
time
pacific
vote


Misinformation Tweet Topics
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
t-SNE Dimension 2
t-SNE Dimension 1
100
75
50
25
0
−25
−50
−75
−100
−100
−50
0
50
100