Prompt, Condition, and Generate: Classification of unsupported claims with In-Context Learning

Anonymous ACL submission

Abstract

Unsupported and unfalsifiable claims we encounter in our daily lives can influence our view of the world. Characterizing, summarizing, and – more generally – making sense of such claims, however, can be challenging. In this work, we focus on fine-grained debate topics and formulate a new task of distilling, from such claims, a countable set of narratives. We present a crowdsourced dataset of 12 controversial topics, comprising more than 120k arguments, claims, and comments from heterogeneous sources, each annotated with a narrative label. We further investigate how large language models (LLMs) can be used to synthesise claims using In-Context Learning. We find that generated claims with supported evidence can be used to improve the performance of narrative classification models and, additionally, that the same model can infer the stance and aspect using a few training examples.

1 Introduction

While concise and clear arguments are often in short supply in online debates, such discussion still tends to follow particular motions (Levy et al., 2014), opinions (Li et al., 2020), human values (Kiesel et al., 2022), or narratives (Christensen et al., 2022). These debates generally consist of various components of arguments (claims, evidence, etc.) on a topic and are likely to have associated attributes like stance or aspect. For example, "Cloning humans for reproductive purposes is unethical and unacceptable, but creating cloned embryos solely for research – which involves destroying them anyway – is downright criminal," has a negative *stance* on the *topic* of cloning. The text conveys that its *aspect* is unacceptable because of the *evidence* that creating cloned embryos solely for research involves destroying them. Hence this text is an *argument*. In the absence of evidence, it would have been a *claim*, e.g., "Cloning humans for reproductive purposes is unethical."

Claims can be unverifiable or unfalsifiable for purposes of fact-checking in real world scenarios (Glockner et al., 2022). Hence, claims and arguments in online debates are frequently discarded during initial claim check-worthiness detection, or may be determined to have insufficient information to determine the veracity, and are therefore often not suitable for fact-checking pipeline (Augenstein, 2021). Instead of discarding the claims and arguments, we propose that one should instead identify the unsupported claim or *narrative*, e.g., "human cloning is wrong," with the aim of helping fact-checkers focus their efforts.

As noted in Section 2, there exists literature for generation of arguments or claims using large language models (LLMs). Yet, no work so far has studied how narratives from online debate portals (Christensen et al., 2022) relate to *argumentative* texts (Habernal and Gurevych, 2016) and how LLMs can help model the narratives. Defining what is meant by narratives and developing a suitable dataset (with both general & controversial claims) and suitable approaches for studying it is a critical first step towards building more general-purpose fact-checking systems for analyzing statements for which it is hard to find evidence. In this work, we close this gap by studying how to employ LLMs for generating argumentative text (Schiller et al., 2021) that follows a given narrative. We start by providing a narrow definition of a narrative, after which we formulate the task of narrative prediction using a new and diverse dataset for training and evaluation.

Fig. 1 illustrates our proposed steps in inferring important attributes using few-shot In Context Learning (ICL) and sampling of the subsequent arguments and claims that are used for downstream tasks such as narrative prediction. We note that while the explicit/implicit terminology is useful for painting a mental picture, the extraction or prediction of aspects in principle is not limited to text explicitly mentioning the attributes, and can be applied to predict attributes that are implicitly mentioned (by reading between the lines).

In summary, the contributions of this paper are:

1. *A specific definition* for narratives, along with an analysis of how this differs from arguments, claims, and motions;
2. *A new dataset and task*, consisting of online comments and tweets labelled for narrative prediction.
3. *A computational approach* that generates arguments/claims which are, in turn, used to generate synthetic tweets with a specified aspect and stance.
4. *A narrative classification approach* that summarizes all claims from a fine-grained debate into a list of unsupported claims using a language model.
5. *Empirical insights* into the impact and challenges of classifying tweets and generating new tweets consistent with particular narratives.

2 Related Works

Corpora of textual claims considering various controversial topics have often been used in the study of rhetoric and argumentation, including summarization (Stammbach and Ash, 2020), optimization (Skitalinskaya et al., 2022), identifying human values (Kiesel et al., 2022), robustness of arguments (Sofi et al., 2022), controllable text generation (Schiller et al., 2021), and studying what constitutes an argument (Trautmann et al., 2020).

Prior work on claim and argument summarization has been beneficial in different tasks and domains. In early works, summarization was used for explainable fact-checking (Stammbach and Ash, 2020; Mishra et al., 2020) and has recently been used to denoise tweets (Bhatnagar et al., 2022). In the latter case, they study textual arguments, but only as a summarization task as seen on social media (e.g., Twitter). However, neither of them focus on fine grained data. IBM-debater (Ein-Dor et al., 2020) and UKP-Corpus (Stab et al., 2018) involve mining of fine-grained data, but they deal only with arguments and not claims.

Our work aims to provide the best of both worlds. However, simply combining these approaches, i.e., summarization of textual arguments (e.g., tweets) for fine grained topic debates, will not be sufficient to find the narratives. Additionally, with real world tweets, abstractive summarization is still underdeveloped in the field of computational argumentation, as compared to summarization of plain text. Due to the data efficiency of prompt-based methods for tasks like abstractive summarization, binary classification, etc.(Chung et al., 2022; Sanh et al., 2022), we propose to explore these methods as they are related to our work.

Several fine-grained approaches have been explored in argument mining (Hansen and Hershcovich, 2022) (Trautmann et al., 2020; Schiller et al., 2021), however, they do not explicitly focus on narrow debates, e.g. "crypto currencies as a fiat currency," they instead treat broader controversial topics like "minimum wage."

In our work we create a new dataset, focusing on narrow debate topics, by relying on an argument mining annotation scheme based on (Hansen and Hershcovich, 2022), consisting of various categories of arguments found in online debates. Where (Hansen and Hershcovich, 2022) compare arguments in terms of categories (normative or factual arguments), we propose and study the new task of predicting controversial narratives from tweets. Perhaps most similar to our work is (Christensen et al., 2022), which proposed a human-in-the-loop-based model to cluster different unfalsifiable claims using crowdsourced triplets similarities.

Our approach also includes generating arguments and claims, augmenting existing data (details in Section 3). For analysis of their quality, we compare these with ground truth text using automatic metrics and human evaluations, and considers persuasiveness, grammatical correction, meaning preservation, and argument quality (Skitalinskaya et al., 2022; Habernal and Gurevych, 2016).

3 Task and Data

This section introduces our definition of a narrative, and a proposed task, and presents the data used for development and evaluation.

3.1 Narrative Definition

The theoretical underpinning of this paper hinges on a proposed relationship between the number of possible narratives (Def. 3.1) found in a fine-grained online debate. With this in mind we now define the *parrot hypothesis*.

The Parrot Hypothesis In a given social media debate, the thoughts and opinions contributed by commenters resolve to a countable set of distinct

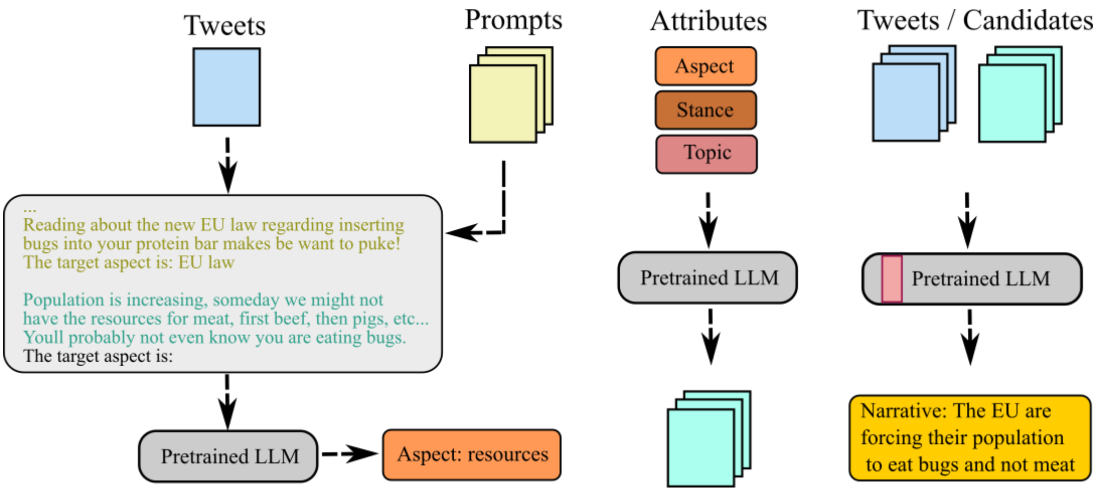

Figure 1: Predict, Condition, and Generate framework. Using In-Class Learning (ICL) with an LLM, we first predict the aspect and stance of a tweet, then we condition the LLM on these attributes to generate candidates. Finally, using the candidates and original tweets, we finetune the LLM for narrative prediction.

narratives. While users could, in principle, state their views in a concise, distilled manner, they often prefer to write embellished variants or personal takes that require reading between the lines.

At its core, the parrot[1] hypothesis seeks to limit the variations of statements to a countable amount of claims related to various topics that are frequently debated online. Additionally by narrowing the scope of a debate some of the coarse grained claims will become irrelevant instances. The result of using the hypothesis in action is that we can turn a fine-grained debate into a classification problem.

It may not be immediately obvious how the parrot hypothesis makes sense in argument mining. After all, shouldn't it be possible to generate an infinite amount of arguments and claims within an debate? We argue that when the scope of the debate narrows, the claims in the fine grained debate will be few in number, but distinct enough to be classified. It has been observed that online debates will have a number of arguments and claims emerge that the majority of users will back up their stances with, despite being worded differently. (Boltužić and Šnajder, 2015) Given the above clarification, we can now exhibit our primary definition.

Definition 3.1. *Narrative:* We use the term *narrative* to refer a shortly written unsupported claim, which has been reduced from its original argument discourse unit where its evidence type is not a survey or alternatively is an unfalsifiable or unverifiable claim.

[1]We use "parrot" in the sense of "parroting talking points," except that we don't assume the commenters are necessarily being fed talking points without their knowledge.

This differs from the concept of motion (Sofi et al., 2022; Ein-Dor et al., 2020; Levy et al., 2014), which is defined as a high level claim, but it is required to imply a clear positive or negative stance towards a topic, and often also contains an action that should be taken as a result. In contrast, a narrative does not need a clear stance nor an encouragement to take action.

3.2 Narrative Prediction

We approach the problem of narrative prediction on social media, by focusing on on tweets. We define the task of computational narrative prediction as follows.

Task Given a tweet t regarded as a statement by a participant in a debate, and a set of possible narratives $\mathcal{N}$, rewrite t into a narrative n such that:

- the narrative is written as an unsupported claim,
- only one narrative n can be selected for each tweet from $\mathcal{N}$, and
- n preserves the meaning of t as much as possible.

While we assume that t is already phrased such that it looks like a claim or an argument, the approaches proposed later in the paper are based on the likelihood of t "looking" like a claim or argument rather than basing it on evidence type for t before being used for classification.

Note that a tweet can contain multiple narratives, and it can follow a narrative explicitly or implicitly. In this case, the goal is to identify the correct explicitly stated narrative given a tweet.

3.3 Annotation scheme

To collect relevant data, we define an annotation scheme, which consists of a fine-grained topic, a sentence, and a narrative. We also showcase how other datasets following similar annotation schemes of (Schiller et al., 2021) can be of use in our study. Attributes from such datasets include the stance and aspect, which we find useful for generating sentences, akin to their CTRL model, to enhance performance of downstream tasks. As in (Schiller et al., 2021), we define an aspect as a continuous substring of an argument or claim which is a recurring subtopic that expresses the issue-specific key rationale for its conclusion, and define the stance to argue for or against the mentioned aspect that is not necessarily mentioned in the argument.

3.4 Dataset creation

We propose a new dataset called TN9, which includes selected topics from UKP-Corpus (Schiller et al., 2021; Stab et al., 2018) and (Hansen and Hershcovich, 2022).The topics and other key statistics about our dataset and its comparison with UKP-Corpus can be seen in Table 1.

	UKP-Corpus	TN9
Annotations	Aspect/Stance/Narrative	Narrative
Tweets (train/test)	30k/1.9k	90k/5.4k
Topics	Abortion, Cloning, Nuclear Energy	AGI, Attractiveness, Alternative Meat, Corporate culture, Crypto, Baby Formula, Influencer, Transport, Mental health
Source (Sentence)	Reddit	Twitter
Source (Labels)	mTurk	mTurk

Table 1: Summary of the datasets.

	Few-shot CoT
Shot	Tweet: [Tweet] Answer: Let's think step by step [Explanation] Therefore, the answer is [answer]
Tweet CoT Answer	Tweet: [Tweet] Answer: Let's think step by step <CoT> Therefore, the answer is* <answer>
	Direct Few-shot
Shot	Tweet: [Tweet] Aspect: [answer] ..
Tweet Answer	Tweet: [Tweet] Aspect: * <answer>

Table 2: Different prompt setups for few-shot Chain of Thought (CoT) and direct few-shot prediction of aspects and stances.

3.4.1 Scraping

We start with scraping relevant data from Twitter. We first execute a series of searches combining different keywords and sentences/phrases, highlighting different statements in a topic (as shown in the Appendix). We do this for 40 different keywords per topic from 2016-2022 and search for as many fields (e.g., images, links, and other metadata) as possible using the Twitter API.

3.4.2 Filtering and Data Cleaning

To ensure that we are working with arguments and claims, we next perform filtering steps. First, we remove duplicates but maintain identical sentences with different hashtags after removing retweets, quote tweets, links and videos, as well as mentions of users, token and media mentions. Second, we replace unreadable hexadecimal representations of unicode characters with their respective character, and encode the text with ascii characters. This results in 98,187 English tweets in total, around 11k tweets for each topic.

3.4.3 Dataset Annotation

Annotation is conducted using Amazon Mechanical Turk in 2 rounds. In round 1, we design a pretest to ensure that the workers know what constitutes an argument (Rinott et al., 2015; Trautmann et al., 2020) by showing examples or pure claims, either being unfalsifiable (Christensen et al., 2022), unverifiable(Petroni et al., 2022) or contain an evidence type that is not a study (Hansen and Hershcovich, 2022). Details of this task has been described in Appendix. After passing at least 4 out of 5 questions regarding classifying if a sentence was a claim (by choosing the claim type) or argument (by choosing evidence type), the workers could begin working on our HIT for next round.

In round 2, HIT asked workers a) if a given sentence is 1 out of $\sim$ 40 different claim or an argument b) annotate a given tweet with unsupported claims.

Geographic distribution Figure 2 presents statistics of the geographical distribution of the tweets. Most tweets don't have an associated geocode, and those that do can be either exact geocodes or simply mention the city that the user has registered. Like (Huang and Carley, 2019) only 2% of all tweets had available geotags and the tweets are found to be predominantly from the US, where the userbase is numerically the largest.

Topic	Sentence	Narrative
Crypto	you are promoting crypto which is a scam helping thieves and criminals you are also full of plastic parts and fillers profitable for the pharmaceutical and cosmetic industry	Influencers are scamming their fans using crypto
Formula	My congressman here voted NO on , lowering gas prices, NO on the baby formula bill, NO on contraception (?!!), and NO on other helpful bills. It is unbecoming to complain about economic hardship and then contribute to it.	People are reselling baby formula to other countries for higher prices
AGI	And on the other side, AGI will be the single greatest technology to alleviate human suffering in all of history.	AI will not replace humans but augment them

Table 3: Example labeled sentences and corresponding narratives.

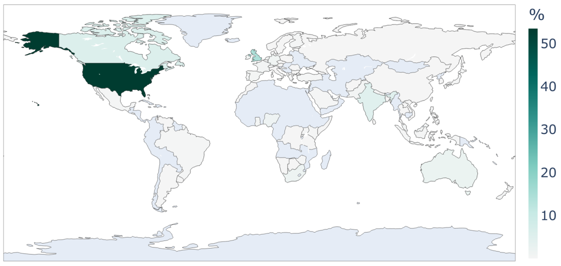

Figure 2: Visualization of the percentages of the number of tweets per country. Around 2% of the tweets had a specific geolocation.

4 Method

Given our new dataset, we can start modelling the narratives present in the tweets. To do this we both classify and summarise the tweets and along the way we ask the following questions:

Is it necessary to do parameter efficient finetuning for summarisation or could we simply do multi-class classification (MCC)? Will in context learning (ICL) using the real tweets improve performance and it be used to generate synthetic examples for a particular debate using the same LLM instead of scraping data? Could we benefit from utilizing argumentative attributes such as aspects and stances as done in (Schiller et al., 2021) for more fine-grained control of the generative process?

4.1 Prompt, Condition and Generate

Method: To predict narratives, we first investigate the effectiveness of a multi class classification setup using two methods: a) Classification *SFT_head*:using T0 encoder with fine tuned head where tweet t is feeded to the encoder and we predict 1 out of n narratives and b) Generation *SFT*: using a fine-tuned T0 model(Sanh et al., 2022) which uses the LoRA setup from (Liu et al., 2022) where t is given to the encoder and we decode the narrative n.

Additionally we investigate the effect of including generated synthetic tweets based on the stance and aspects using a generative model based on different LLMs during finetuning. As illustrated in Figure 1 we first prompt the language model for the aspects and stances. Next we condition the generation of a synthetic tweet (candidate) based on the predicted aspects and stances, finally we incorporate the candidates in our original pipeline for further finetuning of our model. We then compare these 3 methods, ICL, COT (Wei et al., 2023) and Cal (Zhao et al., 2021), with fully supervised BERT span predictors as baseline.

We follow (Liévin et al., 2023) in denoting x the target label (stance, aspect), y an input prompt and z the generated answer from an LLM denoted as p_θ. In the COT setting, sampling $\hat{z} \sim p_\theta(z|y)$ is a two-steps process (first generate the CoT, then extract the answer), otherwise multiple examples are given and the answer is extracted as pictured in Table 2. Using a sampling temperature τ, and $k-1$ examples $(x_1, y_1) \ldots, x_{k-1}, y_{k-1})$ we sample an answer from the generative LLMs as:

$$p_\theta(x|y) \approx \mathbb{1}\left[x \in \hat{z}_i\right], \quad \hat{z} \sim p_\theta(z|y) \qquad (1)$$

where $\mathbb{1}\left[x \in \hat{z}_i\right]$ takes value 1 if the target label (BIO tags for aspect and binary label for stance) x is identical to output $\hat{z}$, otherwise it decreases proportionally for each wrong tag. For the stance task it takes the value 1 when both answer x and completion $\hat{z}$ contain the same word (for/against) and is otherwise 0. We sample multiple completions using beam search to explore multiple possibilities.

One key observation is that we don't utilize verbalizers but instead restrict the possible decoding output only to the words considered in the sentence. This is because what we essentially wish to accomplish is few-shot span prediction and words from the vocabulary $\mathcal{V}$ could be our target.

few-shot ICL and COT: We study two classes of prompts: the *direct* prompt and few-shot CoT, as summarized in Table 2. The direct prompt directly generates an answer using a given prompt and previously seen examples with answers, similar to (Brown et al., 2020). The few-shot CoT framework is similar to (Wei et al., 2023) which provides a reasoning behind the given target labels before it predicts the answer, as seen in Table 2.

Calibration As noted in (Zhao et al., 2021), LLMs can be biased towards the training examples and the order of their occurrence. To mitigate this we estimate the bias towards each answer by feeding in a test input that is content-free, e.g., "N/A" and "". We then fit an affine transformation to "calibrate" the model's output probabilities to cause a uniform prediction for "N/A".

4.2 Evaluation

We predict narratives on 7548 test cases (629 per topic), with automatic and manual evaluation:

Automatic Evaluation As we finetune our model on generated (e.g. synthetic tweet) and real data (e.g. real world tweets), we compare them using several metrics like precision oriented BLEU (Papineni et al., 2002), recall oriented Rouge-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and finally chrF (Popović, 2015). To automatically quantify to what extent a candidate (synthetic tweets) contains the meaning of the original claim, we compute their semantic similarity in each case using the BERT-score(Zhang* et al., 2020).

For the TN9 dataset we have to infer the aspects and stances as they are not given as ground truth data, unlike the UKP-Corpus. Following (Schiller et al., 2021) using the UKP dataset we consider the F1, Acc, recall and precision metrics for the aspect prediction using BIO tags. For the stance prediction we perform binary classification, and consider the same metrics as for the aspects.

Manual Evaluation Before we fine-tune the narrative classification model, we focus on measuring the generative quality of model (which generates synthetic tweets as mentioned in Section 4.1) itself in a manual annotation study. We do this to ensure that the generate text is sensible to read for humans. For each generative model and topic we select 10 generated candidates and acquire 2 independent crowdworkers via MTurk at 18$/hour. In the first study, the annotators scored all generated candidates with respect to the four considered quality metrics: (1) argument quality (2) persuasiveness, (3) meaning preservation and (4) fluency. For assessing the argument quality we follow (Schiller et al., 2021; Skitalinskaya et al., 2022), for persuasiveness we follow (Habernal and Gurevych, 2016), and for quantifying to the quality of the generated candidate we use (Skitalinskaya et al., 2022), using these Likert scales:

- Argument Quality. 1 (notably worse than original), 2 (slightly worse), 3 (same as original), 4 (slightly improved), and 5 (notably improved)
- Persuasiveness. 1 (generated text less persuasive than original), 2 (equally persuasive), 3 (generated text is more persuasive) (choose one argument as being more persuasive or both as being equally persuasive.)
- Fluency. 1 (major errors, disfluent), 2 (minor errors), and 3 (fluent)
- Meaning Preservation. 1 (entirely different), 2 (substantial differences), 3 (moderate differences), 4 (minor differences), and 5 (identical)

Lastly we report the inter-annotator agreement (Cohen, 1960) and krippendorffs alpha (Krippendorff, 2004) between 2 annotators.

Setups	**UKP**	**TN9**
SFT_head	5.23 %	5.51%
SFT	5.75%	5.88%
SFT_T0-arg	7.58%	7.72%
SFT_T5F-arg	7.36%	7.64%
SFT_BLOOM-arg	6.67%	6.81%
SFT_CTRL-arg	7.31%	–

Table 4: Summary of the Narrative prediction F1 micro accuracy using SFT of a T0 model.We denote training with 600 additional sentences generated using attributes like stance and aspects and using different models with arg. We test against a classification setup where, SFT_head refers to encoder + head T0 model, with N outputs corresponding to N narratives per topic.

5 Experiments

5.1 Narrative prediction

Table 4 shows the average F1 micro accuracy between the decoded narrative text finetuned using our approach and the (tokenized) target narrative. We investigate whether adding additional synthetic tweets t_g to original dataset could improve the F1 accuracy. We experimented with generating 600 candidates (t_g) by providing a topic and using a LM (p_θ) to predict the stance t_s and aspect t_a from the original tweets t in the test set. Given the topic, t_s and t_a we generate t_g using p_θ and use them together as a dataset. This is used to fine tuning p_θ with their target narratives n to generate model predictions t_n. With this approach, we observe an increase in accuracy by 2 % depending on the model generating the data. We use the T5-flan-3B model (Chung et al., 2022), an API call to BLOOM-176B

Approach	BLEU	RouL	Meteor	BERT-score	chrf
ACNAG					
CTRLUKP	8.3	12.1	16.4	83.7	23.1
BLOOM	6.5	13.6	16.2	84.8	**31.06**
T5-flan	10.8	**20.6**	16.4	**90.5**	25.1
T0	**13.6**	20.3	**16.7**	90.2	25.2
TN9					
BLOOM	7.94	9.2	9.7	82.1	23.8
T5-flan	11.2	13.7	9.5	87.4	18.7
T0	12.3	13.1	9.2	87.8	18.9

Table 5: Automatic evaluation: Average performance of each model on 629 test cases per topic

Model	Persuasiveness	Fluency	Argument	Meaning
ACNAG				
CTRLUKP	2.1	2.3	3.6	3.4
BLOOM	1.9	**2.8**	**4.2**	4.1
T5-flan	2.2	1.8	3.2	3
T0	**2.6**	**2.8**	3.5	**4.5**
TN9				
BLOOM	2	**2.7**	3.4	**3.5**
T5-flan	**2.4**	2.3	**3.6**	3.3
T0	**2.4**	2.5	3.4	**3.5**

Table 6: Human/Manual evaluation: Average scores on 10 sentences generated on each topic using different methods: Persuasiveness (1-3), Fluency (1–3), Argument quality (1-5) and Meaning (1–5).

(Workshop et al., 2023) and the CTRL generative model from (Schiller et al., 2021). Predictions are shown in Table 9 alongside their target narrative.

5.2 Stance correctness

Table 8 shows stance prediction using standard ICL, COT, contextual calibration and a fully supervised BERT model(baseline) trained on a subset of data from (Schiller et al., 2021) (10k random examples per topic for all 8 topics). The results reveal using various LM (mentioned in Section 4) can outperform the baseline on at least two topics in UKP-Corpus with Cloning topic being an exception. We believe this is because the distribution of stances in this topic makes in highly polarized.

5.3 Aspect Prediction

Table 10 shows that T0 perform worse than our best baseline trained on 80k examples. Also, we find that our baseline provides a better model on average, with increase in performance using more data from other topics. Similarly, we find our baseline performing at a similar level to the official results reported in Table 3 in (Schiller et al., 2021). The performance of T0 has quite high variance, but has the advantage that it only requires a handful of examples (4) to compete with the other baselines.

In Figure 3, we visualise the performance of T0

Model	Persuasiveness	Fluency	Argument	Meaning
CTRLUKP	0.2/0.3	0.2/0.3	0.2/0.2	0.4/0.4
BLOOM	-0.1/-0.1	0.4/0.4	0.1/0.2	0.3/0.1
T5-flan	0.1/0.1	0.3/0.3	0.1/0.4	0.5/0.4
T0	0.3/0.3	0.4/0.3	0.2/0.3	0.5/0.3

Table 7: Annotator agreement (Cohens kappa and krippendorffs alpha) using 2 annotators across all topics.

few-shot prediction given subsets of $k \leq 4$ examples and baseline models. Increasing the number of samples yields better results. The variance of the predictions is rather large, reflecting that using the samples is not always beneficial to the model.

5.4 Automatic Evaluation

Table 5 shows the quantitative metrics between t_g and t. The relatively low scores of BLEU (6.5) and ROUGE-L (9.2) indicate that revisions take place, however due to the high BERT-score (90.5) the meaning is largely preserved. Additionally the METEOR and Rouge-L scores are similar to (Schiller et al., 2021) indicating similar generative behaviour. TN9 overall has lower scores indicating that it is harder for the model to generate a sentences similar to the tweets from a predicted stance and aspect. Table 7 shows T0 being preferred for generating meaningful and persuasive texts. This is important as we will use the data in a finetuning setup.

5.5 Human Evaluation

As shown in Table 6, Krippendorff's alpha agreement is generally low, being 0.24 on average, which are common in subjective tasks (Wachsmuth et al., 2017). The inter-annotator agreement (Cohen, 1960) varies from model and attribute but is on average .25, which can be interpreted as "fair" agreement (Landis and Koch, 1977). Table 6 shows that human annotators find text generated by T0, having a higher persuasiveness (2.6) and having similar meaning to the source text (4.5) than the other methods. However, candidates from BLOOM and CTRL-UKP have a higher argument quality (3.5 vs. 3.6 and 4.2) and are more fluently written.

6 Conclusion

In this paper we introduced a new definition of narratives and how to model these in fine grained debates with large language models. Our approach is based on parameter efficient fine tuning using controlled text generation using attributes predicted using a handful of examples. We show that claims generated using our approach are genuine and sen-

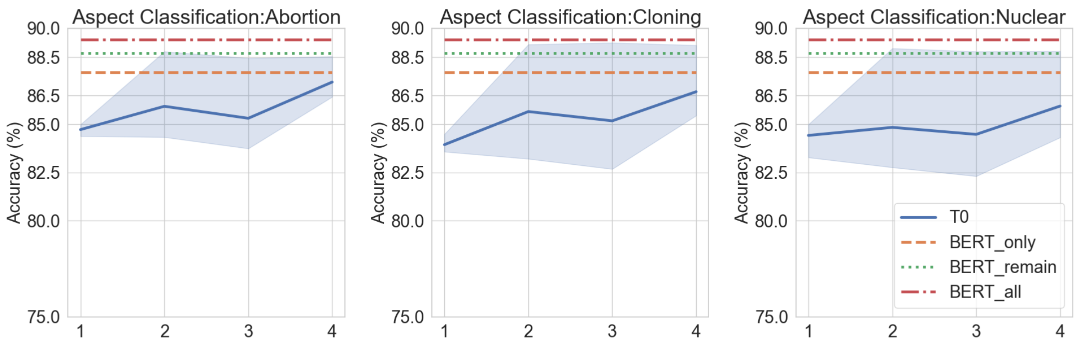

Figure 3: Aspect accuracy of few-shot ICL (`T0-3B`) on the Abortion, Cloning and the Nuclear Energy topic in the UKP dataset sampled with temperature $\tau \in \{0.7\}$. We report the average accuracy for a model using random subsets of $k' = 1 \ldots 4$ examples. We display the performances of the best finetuned baselines.

Topic	F1	Recall	Precision	Acc
Abortion$_{only}$	50.1	50.7	50.8	53.1
Abortion$_{icl}$	54.4	50.2	59.4	53.8
Abortion$_{cot}$	55.7	51.3	61.1	54.7
Abortion$_{cal}$	57.3	52.9	62.7	55.6
Cloning$_{only}$	75.5	75.5	75.6	75.8
Cloning$_{icl}$	59.3	50.3	72.4	54.9
Cloning$_{cal}$	62.4	53.4	75.1	56.7
Cloning$_{cot}$	60.6	51.58	73.8	55.9
Nuclear$_{only}$	37.1	50	29.5	58.9
Nuclear$_{icl}$	54.9	50.2	60.56	52.9
Nuclear$_{cot}$	57.4	51.3	61.1	53.6
Nuclear$_{cal}$	58.7	53.8	64.7	54.1
Abortion$_{remain}$	36.1	50	28.2	56.4
Cloning$_{remain}$	35.6	50	27.6	55.3
Nuclear$_{remain}$	37.1	50	29.5	58.9
Abortion$_{all}$	52.6	53.3	53.8	55.6
Cloning$_{all}$	77.1	76.8	77.6	77.6
Nuclear$_{all}$	37.1	50	29.5	58.9

Table 8: Average micro F1, recall and precision scores for stance prediction using binary classification (for=1,against=0). Top section uses a BERT$_{BASE}$ model with $_{only}$ indicating it is only trained on this topic. It is compared against the T0 model for few-shot ICL using just 4 random examples for prompting. The bottom section shows additional experiments where we show that a) $_{remain}$:training on the 5 remaining topics from (Stab et al., 2018) (i.e. excluding abortion, cloning & nuclear energy) before finetuning to a new topic and b) $_{all}$: trained on all 8 topics from (Schiller et al., 2021) don't perform as well as our chosen baseline i.e. $_{only}$.

Tweet t	**Model Prediction** t_n	**Target Narrative** n
Animals are not ingredients!	eating meat is murder	Eating meat is murder
Yall find hypermasculinity resulting in insecurities about the lack of a better body attractive? Lmaaoo	Hypermasculinity is problematic	Hypermasculinity in and of itself is the problem

Table 9: Sentences with predicted and target narrative.

Topic	F1	Recall	Precision	Acc
Abortion$_{only}$	68.5	67.7	69.4	87.7
Abortion$_{icl}$	66.9	66.5	67.2	87.1
Abortion$_{cot}$	67.2	66.7	67.7	87.3
Abortion$_{cal}$	68.2	67.8	68.7	87.8
Cloning$_{only}$	71.8	73.3	70.9	88.9
Cloning$_{icl}$	66.5	65.8	67.3	86.6
Cloning$_{cot}$	67.7	67.2	68.2	87.7
Cloning$_{cal}$	68.5	68.2	68.7	88.2
Nuclear$_{icl}$	66.1	65.4	66.7	86.3
Nuclear$_{only}$	73.1	73.5	72.8	89.9
Nuclear$_{cot}$	68.8	68.4	69.0	88.3
Nuclear$_{cal}$	68.4	68.2	68.5	88.1
Abortion$_{remain}$	71.6	72.1	71.2	88.7
Cloning$_{remain}$	74.9	75.12	74.93	90.5
Nuclear$_{remain}$	75.5	75.6	75.4	91
Abortion$_{all}$	72.9	72.3	73.8	89.4
Cloning$_{all}$	75.2	74.9	75.7	90.9
Nuclear$_{all}$	76.6	77.1	76.2	91.5

Table 10: Average micro F1, recall and precision scores for aspect prediction using IOB tags. The tags $_{only}$, $_{remain}$ and $_{all}$ indicate the same setup from table 8 using BERT$_{BASE}$ as baseline in the first section and few-shot ICL using the T0 model for few-shot ICL using 4 random examples for prompting.

sible in general. We fine-tune of model on our own dataset and the augmented UKP-corpus and outperform baseline approaches. In future work, we seek to examine multiple completions and ensambles similar to (Liévin et al., 2023) which enables to include examples of up to 100 examples for ICL, to reduce variance and outperform single-sample CoT methods using larger models (GPT-4, ChatGPT, LLama). Moreover, our approach considers each topic independently using a LLM but could be made to consider all simultaneously.

7 Limitations

Scaling to multiple topics For our approach, the prediction of narratives is topic specific and the number of models scales linearly with the with the topics. This is primarily because both the baseline method using a LM head cannot predict new classes and for the text2text approach it is theoretically possible to simply use one model, though initial experiments suggested a model per topic worked better. Instead of directly predicting the narratives, one could instead have ranked the list of narratives given a tweet. This gives us contextual information about the narratives, since they are written in text and not just as a class and provides a number of benefits including having one model for all topics but also new topics. Additionally it could also provide temporal evaluations by adding new emerging narratives to the list.

Scaling to more narratives The current approach requires a domain expert to writing down the particular narratives from the fine grained debate and does not model that there is a countable number of narratives within a specific domain. Finding the particular narratives is bottlenecked by knowing enough about the particular topic. Moreover, since it takes time to gather enough information about the different topics it makes it difficult to scale up to larger numbers of taxons.

Future work can explore automatic generation of the narratives given a list of tweets, and condense this list iteratively, and patch templates e.g., using pre-trained language models.

Directly modelling the initial argumentative text Finally, the approach we develop can operate on text that is claims or argument discourse units, but has no way of distinguishing between these or nonarguments. This precludes the model from being able to only predict a narrative if the text is indeed from the fine grained debate and can be tricked into providing narratives which the text doesn't follow.

References

Isabelle Augenstein. 2021. Towards explainable fact checking.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Varad Bhatnagar, Diptesh Kanojia, and Kameswari Chebrolu. 2022. Harnessing abstractive summarization for fact-checked claim detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2934–2945, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Filip Boltužić and Jan Šnajder. 2015. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115, Denver, CO. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Peter E. Christensen, Frederik Warburg, Menglin Jia, and Serge Belongie. 2022. Searching for structure in unfalsifiable claims. -.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2020. Corpus wide argument mining—a working solution. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7683–7691.

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing counter-evidence renders nlp fact-checking unrealistic for misinformation.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Marcus Hansen and Daniel Hershcovich. 2022. A dataset of sustainable diet arguments on Twitter. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 40–58, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Binxuan Huang and Kathleen M. Carley. 2019. A large-scale empirical study of geotagging behavior on twitter.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement of categorical data. *Biometrics, 33*, 33(1):159–174.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Jialu Li, Esin Durmus, and Claire Cardie. 2020. Exploring the role of argument structure in online debate persuasion. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8905–8912, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning.

Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2023. Can large language models reason about medical questions?

Rahul Mishra, Dhruv Gupta, and Markus Leippold. 2020. Generating fact checking summaries for web claims. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 81–90, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Fabio Petroni, Samuel Broscheit, Aleksandra Piktus, Patrick Lewis, Gautier Izacard, Lucas Hosseini, Jane Dwivedi-Yu, Maria Lomeli, Timo Schick, Pierre-Emmanuel Mazaré, Armand Joulin, Edouard Grave, and Sebastian Riedel. 2022. Improving wikipedia verifiability with ai.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online. Association for Computational Linguistics.

Gabriella Skitalinskaya, Maximilian Spliethöver, and Henning Wachsmuth. 2022. Claim optimization in computational argumentation.

Mehmet Sofi, Matteo Fortier, and Oana Cocarascu. 2022. A robustness evaluation framework for argument mining. In *Proceedings of the 9th Workshop on Argument Mining*, pages 171–180, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Dominik Stammbach and Elliott Ash. 2020. e-fever: Explanations and summaries for automated fact checking. In *Proceedings of the 2020 Truth and Trust Online (TTO 2020)*, pages 32 – 43, Arlington, VA. Hacks Hackers. Conference for Truth and Trust Online (TTO 2020) (virtual); Conference Location: online; Conference Date: October 16-17, 2020; Due to the Coronavirus (COVID-19) the conference was conducted virtually.

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schutze, and Iryna Gurevych. 2020. Fine-grained argument unit recognition and classification. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*. AAAI Press.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models.

A Implementation Details

Here we describe the implementation details for fine-tuning the 3B T0 model for narrative prediction, in addition to using different ICL strategies for stance and aspect prediction. For all downstream tasks, we use the same AdamW optimiser with linear learning rate decay and weight decay. Finetuning details such as number of epochs and learning rate is reported in Table 11

config	value
optimizer	AdamW
base learning rate	3e-4
weight decay	0.001
optimizer momentum	$\beta_1, \beta_2{=}0.98, 0.999$
adam epsilon	1e-8
batch size	8
learning rate schedule	linear decay
warmup steps percentage	10%
Number of epochs	10
batch size	8
maximum sequence length	128
maximum gradient norm	1
LoRA rank	4
LoRA init scale	0.01
LoRA layers	Self/Enc-Dec attention layers
LoRA scaling rank	1

Table 11: **Fine-tuning setting.**

Aspect prediction For our aspect prediction models we use the standard BERT model to predict a sequence of BIO tokens. We tokenise a given sentence using the `TreebankWordTokenizer` from the `nltk` package available for the Python programming language. For the ICL setup we force the T0 model to consider only the words in the given sentence by tokenizing the sentence and feeding it into `force_words_ids`, additionally we also force the decoding step to not include stop words in addition to special characters like " that appear in the sentence.

During decoding, we set the temperature τ = 0.7, top_p=0.9, number of beams equal to 5 to provide a variety of sentences following the same narrative.

Stance prediction For the Stance prediction we restrict decoding to one word only, and giving the model two choices *for* or *against* for the T0 model. For the baseline model we simply attach a LM head and do binary classification 0 =*for* and 1 =*against*.

Narrative prediction During finetuning we switch the standard T0 model out with T0-few with the LoRA setup and mainly keep the defaulthyperparameters but reduce the batch size to 4 and train a model for each topic for for 10 epochs. Each model takes around 3hrs to train on the 10k training sentences. In addition to this setup we also include sentences that we generated sentences using the topic, predicted stance and aspect using the CTRL-UKP model, T0-3B, T5-flan-3B and 175B Bloom model. The tweets we predict the stance and aspect is from the test set. Using these attributes we can generate similar sentences to the teat set to help enhance performance. We simply copy the target narratives as labels for the generated sentences and include them in the training dataset.

To give an example of the runtime for our code it takes 12 hours to complete 10 epochs for the T0-3B model using 1 TitanRTX-24GB and 1 Xeon E5-2620 v4 8c/16t - 2.1 GHz CPU, and 8 hrs using 1 A100-40GB and the same CPU for the T0-11B parameter model. We always have access to a minimum of 48GB of RAM but run our experiments using 64GB RAM.

B Search Query and Narrative Synonyms

Topic	**Search query**
AGI	
	AGI replace humans, AI replace humans, AGI technology
	AGI threat,AI threat,AGI beyond human intelligence,AGI rule the world
	AI useful,AI better than people,society help AI ,AI beats human
	human better AI, AGI achieve human ,AGI myth, human ethics AGI
	AI threat humanity, AI solves problem ,AI help climate, AI hype bad
	AI hype up, AI hype good, AGI future good, AI no common sense
	AI trust bad, AI superhuman, AI wants things that are absurd to humans
	AI billionaire control, AI just tool, AI wealth concentration
	AI wealth inequality, automation wealth inequality ,ai if statements
	ai uncontrollable, AI make me laugh, AI art steal
	AI mashup, AI demotivate, AI problems fix, AI fix our problems
	AI human nuance, AI data new oil, AI coming for job
Alternative meat	
	stop subsidizing meat,alternative meat fake, alternative meat unhealthy
	meat is murder, soy meat replacement, reduce meat consumption climate
	meat no sustainable, meat is unhealthy, food pyramid scheme
	subsidize green nutrition, increase production of meat
	exempt meat production from carbon taxes, carbon tax to food production
	invest Meat alternatives, Meat alternatives subsidized
	Plant based food subsidized,introduce meatless mondays
	Vegetarian vegan food encouraged,discourage vegan diet, subsidize fruits vegetables
	meat overconsumption, Plant based food encourage, Meat alternatives encourage
	plant based food sustainable, plant food is great, fresh organic food is good
	meat alternative food is good,red meat is bad, animals are not ingredients
	eat healthy food, raw food diet,flexitarian meat alternative
	big pharma alternative meat,alternative meat forced
	plant based food processed, plant based food remove meat
	animals eat meat humans too, eat plant save planet,eat
	meat save plant, meat ruining planet,alternative meat bugs
Attractiveness	
	spotlight effect, male gaze, male gaze exploiting, female gaze
	males observing attractive female, attractive hypermasculinity
	attractive masculinity,female sexual object, beauty standard money
	beauty standard protection, beauty standard shoe, beauty standard cloth
	beauty standard events, beauty standard desire, beauty standard objectify
	beauty standard stress,beauty standard stable, beauty standard fake
	beauty standard safe,sexual objectification patriarchy
	beauty standard disrespect,beauty standard gender role
	beauty standard escape, beauty standard equality, beauty standard dominance
	beauty standard ownership,beauty standard media, beauty standard unrealistic
	beauty standard transphobic, beauty standard harassed
	beauty standard academia,beauty standard university
	beauty standard cheating, beauty standard fetish, hygiene no beauty standard
	Toxic Masculinity attractive,attractive Bechdel test,enforcing stereotyp beauty

Table 12: Twitter search keywords

Table 12 - 14 lists the Twitter queries we used to retrieve the initial training data. Note that many of the words are either in their stem or shorted format in order to ensure a wider range of search results being returned. Per default twitter filter out sentences that does not contain tokens from the

Topic	Search query
corporate culture	
	corporate culture HR,company culture HR,work culture HR
	corporate culture toxic,company culture toxic,work culture toxic
	corporate culture unlawful,company culture unlawful,work culture unlawful
	company culture risk,corporate culture risk,work culture risk
	corporate culture speak up,company culture speak up,work culture speak up
	corporate culture abuse,company culture abuse,work culture abuse
	anti union company,don’t trust non profit,corporate culture no trust
	company culture no trust,work culture no trust
	corporate culture greed,company culture greed
	work culture greed,corporate culture millenialcompany culture millenial
	work culture millenial,their company do what they want
	corporate culture manager,company culture manager,work culture manager
	corporate culture stress,company culture stress,work culture stress
	corporate culture hard work,work pregnant,side hussle culture
	work culture loyalty,corporate culture loyalty,company culture loyalty
	work culture remote,corporate culture remote,company culture remote
	work culture ethic,corporate culture ethic,company culture ethic
	work culture family,corporate culture family,company culture family
	corporate culture cult,company culture cult,work culture cult
	corporate culture fun,work culture fun,company culture fun
	work culture perks,company culture perks,corporate culture perks
	job hop look bad,company dress code,corporate mass firing, quite quitting
	let it rot job,corporate culture disgusthing
	work culture disgusthing, company culture disgusthing
crypto	
	china crypto ban,china crypto mining,el salvador crypto legal
	crypto steal constitution,crypto banking the unbanked
	crypto financially free,crypto diversify asset,crypto people of color
	crypto trust technology not people,crypto access financial
	crypto bank failure,crypto better digital payment,crypto wealth builder
	crypto upwards mobility,crypto is an investment,crypto digital gold
	crypto short the bankers,crypto not democracy,crypto ruthless investors
	Bitcoin is a Platypus,crypto should be regulated,crypto needs rules
	crypto is a scam,crypto is for terrorists,crypto is for criminals
	crypto rich bailouts,crypto stock bubble,crypto unsustainable environment
	crypto ponzi scheme,crypto pump dump,crypto influencer ponzi
	crypto carbon tax,crypto great reset ,crypto own nothing happy
	crypto money laundering,crypto funding party
	crypto same as database, crypto is toxic
baby formula	
	baby formula scam,baby formula poison,baby formula breat milk
	baby formula inflammatory,baby formula infection,baby formula virus
	baby formula sustainable,baby formula weight
	baby formula replacement feeding,baby formula economic, baby formula hospital
	baby formula shame,baby formula guilt,baby formula husband feed
	formula feeding mental health,breastfeeding mastitis,breast milk propaganda
	breast milk infection,breast milk risk ,breast milk health
	breast milk is best,breast milk germ,baby formula propaganda
	breastfeeding guilt,breastfeeding negativity, breastfeeding anxiety
	breastfeeding public,breastfeed good citizen,breastfeeding shame
	breastfeeding sleeping,breastfeeding formula all nothing
	breastfeeding gender role,politically correct breastfeeding

Table 13: Twitter search keywords

query.

C Annotation Details

For our crowdsourcing of narrative annotations and human evaluation we use Amazon Mechanical Turk .Workers had to take a qualification test, have an acceptance rate of at least 80% based on 5 question, be located within the US, have successfully completed more than 1000 HITs before and have an approval rate of 98%. We paid 1 dollar per HIT for the dataset task which is to classify one tweet into one in roughly 40 narrative categories. Initially time spend on a HIT is much higher than when they complete their 25th hit as workers learn to memories the categories. For the human evaluation we get annotations from two crowdworkers and pay 2 dollars per HIT. Consent was obtained from the crowdworkers by including the warning for the pretest annotation: "By completing this test you will agree that subsequent HITs using this pretest as a prerequisite can be used for data collection in relation to research projects", similarly for

Topic	Search query
Influencer	
	Influencer real job, Content creator full-time job
	Influencer popular career choice, Career social media influencer
	Social media influencer pay, Social media influencer doesn’t pay well
	social media influencers real job, Social media influencers unemployed
	Job title influencer, quitting job influencer
	Influencer marketing is big money, Influencer marketing not authentic
	Social media influencer cute name employed, Self-employed influencer
	Social media youths employed, adults influencer jobs
	followers get a job,quitting jobs influencer jobs
	social media influencers jobs money, influencer deal hate
	social media job followers, social media influencers work hard
	hard work influencer, Influencer new job career
	popular career social media influencer, Influencer full time
	influencer job pays, Influencer boring job
	social media influencers getting paid, influencer easy money
	influencer no respect, social media influencer celebrity
	social media influencer waste of time
Mental health	
	Mental health-related sports ,checking mental health athletics
	Mental health for athletes important,Mental health concern athletes
	Mental health concern for student athletes
	Sport reduces stress depression, sports affect mental health
	sports healthy mind,Athletic mental health awareness
	recognize mental health athletic, Prioritize mental health sports schools
	sports, mental health ,Sports coach mental health
	Initiative mental health sports,Well being athletic
	Mental health identify issue sports, Sports support mental health
	Sports stigma mental health, Stigma mental health atheletics
	University atheltics metal health awareness
	Stigma challenges sports mental well being
	male dominated sport toxic, vulnerability weakness sport
	athletes no real problems, athletes trans problems
	sports mental health flu ,sports mental health kill
	sports mental health of money,Sport mental health brutal
	Sport drug mental health, Sport racism mental health
	athlete burnout young age,athlete burnout young
	athlete blame media,athlete work late
	sport alienation,Sports mental health religion
	Sport mental health religion
Transportation	
	public transportation good,public transportation work
	cheap public transportation, comfortable public transportation
	bus better than car, public transportation environment
	buses safer driving,trains better than flight
	train better climate,Climate Action Public Transport
	public transport safer, car culture climate, public transit affordable
	flights less time trains, trains more expensive
	buses carry more people,cars carry less people
	cycling decrease car traffic, cycling better air quality
	public transport less pollution, public transport less CO2
	public transportation personal space, public transportation germs
	public transportation disease, public transportation covid
	public transportation rural, buses middle class
	buses poor people,cars rich people ,car only wealthy
	less drivers safer streets,public transportation night unsafe
	public transportation night comfortable, tax car poor
	use bike dangerous,public transit not profitable
	public transport profitable ,highways profitable
	car centric bad, public transportation useless
	car give freedom independence, car give independence

Table 14: Twitter search keywords

we get consent from people whose data we are using though the Twitter Term of Services. The data collection procedure was approved by our internal ethics review board.

During our annotation of the narrative labels we discovered that the returned answers tend to be biased towards the top 10 first possible answers that could be selected in the HIT. To mitigate potential bias we manually went though the top 3 most frequent answer for each topic in the validation set and relabel the corresponding tweets.

D Crowdsourced Annotations

For this paper, gathering annotations has happened over three annotations rounds, each focusing on different sections of the paper.

D.1 Pretest

The first crowdsourcing task is that of a pretest, which is used to determine if workers are suitable for our main annotation mask. It is based on data from (Hansen and Hershcovich, 2022) and focuses on correctly classifying two different types of labels: Pro/con and evidence.

D.1.1 Pro/Con

Pro/con is a binary label. The tweet is annotated as (+1) for pro when a clear claim has a positive or supportive stance towards the topic. It is annotated as (-1) when it has a clearly antagonistic or attacking stance towards it the topic. We exclude data for which has no clear stance.

Instructions for annotators: Given a tweet your task is to annotate it with its stance in relation to the topic topic. The stance is either pro or con (for or against a topic). In this case select pro if you find that the tweet is supportive towards the topic, and con if it is hostile instead. Remember that a tweet with hostile remarks can still be supportive of the topic, as we want to find the stance towards the topic and not the tweet itself.

D.1.2 Evidence

Evidence as a label has 6 classes. The tweet is annotated using any of the labels: Normative, Study, Expert, Fact, Anecdotal or unrelated/no evidence. The description for each of these labels are taken from (Hansen and Hershcovich, 2022): Anecdotal refers to "a description of an episode(s), centred on individual(s) or clearly located in place and/or in time." Expert refers to a "testimony by a person, group, committee, organisation with some known expertise / authority on the topic. Study refers to "results of a quantitative analysis of data, given as numbers, or as conclusions" Fact refers to "A known piece of information about the world without a clear source for the information" Normative refers to "an added description for a belief about the world" No evidence refers to "the tweet does contain evidence, but it is not related to the topic, or it does not have any evidence."

Instructions for annotators: The task is to annotate a tweet with the type of evidence it contains. Evidence is a statement used to support or attack a topic or claim. Evidence can be present in combination with a claim, or it can also be self-contained if it is just stating facts or referencing studies related to the topic. If the evidence is unrelated to the discussed topic, it is marked as unrelated. If you feel that multiple types of evidence is present in the tweet, choose the one that you think best describes the main piece of evidence in the tweet. Remember that your task is to annotate the type of evidence that is in the tweet regardless of your views and if the evidence is true or not.

D.2 Narrative annotation

The main crowdsourcing task of this paper is essentially claim classification. Given a tweet the workers determine if the tweet is a claim or argument with an evidence type that is not a study (as taken from the definition of study in (Hansen and Hershcovich, 2022)). Then if the tweet is a claim then they should select the most similar claim from a list of options. If no option is suitable, they should select "No claim in list is similar to the tweet".

Instructions for annotators: The task here is to annotate a tweet given a list of claims that the tweet might be similar to. Of course, each tweet can be relevant for more than one claim, but it can also be irrelevant and should be annotated as such. Therefore, given that the topic is select the claim which you find the tweet most similar to (regardless of your views on the list of claims, the topic and the tweet itself). Remember that the surrounding context of a tweet can be missing, and that people may be sarcastic.

D.3 Human evaluation of generated claims

The last crowd sourcing campaign is the human evaluation in which we evaluate how well a generated claim compared against the original claim (It is generated from the predicted stance and aspect from the original claim). We follow primarily (Skitalinskaya et al., 2022) for definition of argument quality, meaning and fluency, but also (Schiller et al., 2021) for fluency and persuasiveness. These generated claims are then used for finetuning a LLM for improved narrative prediction.

Instructions for annotators: In this task, you will identify if a generated claim is similar to or has improved, without changing the overall meaning of the text. Each field contains a par of tweets, one being the original and the other a synthetic tweet that is trying to mimic it. Please rate each candidate along the following four perspectives: argument quality, fluency, meaning and persuasiveness.

Argument Quality has a scale from 1 to 5: 1 (notably worse than original), 2 (slightly worse), 3 (same as original), 4 (slightly improved), 5 (notably improved) Does the generated claim improve over the original claim? Things to look for include: specifying a fact, simplifying the sentence, adding clarity, adding additional information such as facts, adding, editing or removing links for external resources.

Meaning has a scale from 1 to 5: 1 (entirely different), 2 (substantial differences), 3 (moderate differences), 4 (minor differences), 5 (identical) Here we wish to measure if the generated claim have the same overall meaning as the original. Adding extra information that does change the objects or events described in the claim should not penalise the score.

Persuasiveness runs from 1 to 3. 1 (generated text less persuasive than original), 2 (equally persuasive), 3 (generated text is more persuasive) (choose one argument as being more persuasive or both as being equally persuasive.) Here we wish to measure if the generated claim is more useful in a debate about a certain topic than the original claim. Adding additional text that explains an event or fact more in depth should be rewarded.

Fluency runs from a scale form 1 to 3: 1 (major errors, disfluent), 2 (minor errors), 3 (fluent) Here we want you to to compare the generated sentence with the original one and ask if the sentence is written in fluent English and makes sense? You should consider rewarding the generated claim in case of improved grammar, spelling and punctuation of generated claim over the original claim.

E Narratives per topic

Topic	Narrative
Abortion	
	Abortion reduces crime
	Abortion should not be allowed
	Everyone has a right to life
	A fetus is a real persons
	Abortion is painful for the fetus
	Abortion is not murder
	Abortion reduces the value of human life
	Women that go through abortion face social stigma or guilt
	Supporting abortion is societal pressure
	Abortion gives mothers the option of giving birth to healthy children
	No abortion option for poor women is injustice
	A fetus is not a real persons
	Abortion is murder
	Planned children lead better lives
	Modern medicine makes abortion is less of a risk
	Women choose what to do with their bodies
	Couples that cant get kids want to adopt
	Do not have kids if you fear they will be born with defects
	Fathers have no say if the mother wants abortion
	Abortion is not painful for the fetus
	Removing abortion can put some pregnant woman at risk
	Women that abort have no dignity
	abortion leads to mental diseases
	Abortion encourages more sex
	Authority are against performing abortion
	anti-abortion is counterproductive
	Abortion is inhumane
	restricting abortion enforce traditional gender stereotypes
	women that have been raped should have right to abort
	Abort is morally wrong
	abortionists are in it for the money
	Fetuses should be protected
	Children who almost got aborted might feel rejected
	pro-life views makes no sense
	Parents must know if their child has an abortion
	No claim in the list is describing the tweet
AGI	
	AI is just hype
	AI art unlike human art does not have any value
	AI is just if else statements
	AI has no common sense
	Current AI is not superhuman
	AIs do not have empathy
	AI is bad because it is not as good as human
	AI can make you laugh
	AI will not replace humans but augment them
	AI is bad as it replaces artist
	AGI is just a myth
	AI is for the most part uncontrollable
	AI will create more problems than it solves
	You cannot trust AI
	AI will not take your job
	Data is important to make good AI
	AGI will rule the world
	We will get AGI sooner than excepted
	AI is a threat to humans
	AI will fix our problems
	AI cannot recreate human nuances
	AI will take your job
	AI will demotivate you from working
	AI is stealing from artist
	You cannot trust people who hype up AI
	AI will help us solve climate change
	AI will live up to its hype
	AI is already superhuman
	AI is just a tool
	AI is power hungry just like the billionaires who control it
	AI furthering the wealth inequality
	AI is a general purpose technology like electricity
	No claim in the list is describing the tweet

Table 15: First list of narratives

Topic	Narrative
Alternative meat	
	we could stop subsidising highly processed foods
	Investing in Meat alternatives is good and profitable
	meat production is not sustainable
	alternative meat is not viable for a healthy diet
	big pharma is behind the alternative meat
	animals eat meat so humans should too
	Eating meat is immoral
	alternative meat does have enough proteins
	plant based food is made to remove meat
	meat cause cancer and can be deadly
	plant based food are sustainable food
	red meat is bad
	We should subsidise Meat alternatives and Plant based food
	Being Vegetarian or vegan allows you to be healthy
	plant food and meat alternatives is great
	animals are not ingredients
	You do not need meat to hit the gym often
	transport of goods is more harmful to the planet than meat or plants
	alternative meat is a pyramid scheme
	alternative meat tastes bad
	alternative meat is unhealthy as a diet
	alternative meat is forced upon the consumer
	Plant based meals are highly processed and is not good
	We should eat plants to save planet
	alternative meat is fake
	Eating fewer plants and more meat will save the plant
	We should reduce meat consumption to protect the planet
	We should import a carbon tax to food production
	We should increase production of meat
	We should stop subsidising meat to allow for alternative meat
	Eating meat is murder
	exempt meat production from carbon taxes
	Being flexitarian allows you to get enough nutritions
	fresh organic food is good
	A vegan diet is unsustainable for the planet
	soy meat will not and cannot replace meat
	Eating bugs instead of meat will never be reality
	No claim in the list is describing the tweet
Attractiveness	
	Women and especially lesbians are exploiting the male gaze
	Forcing your standard of beauty on every women is trans phobic
	The male gaze encourages physical and sexual violence against women
	the lack of beauty standards warrants cheating
	beauty standards are pathetic and fake
	Academia is like any other industry where beauty standards play into how women are treated
	the female gaze and male gaze are distorted terms used on social media
	Just because people do not fit the beauty standard, does not mean that you can disrespect them.
	misogynists hate women that take back ownership of their bodies and reject beauty standards
	beauty standards serve to perpetuate a misogynistic society
	Hygienic actions like shaving is beyond beauty standards or gender roles
	Beauty standards are sexist
	We should be free from sexual objectification and beauty standards
	Corporations try to make you buy stuff though beauty standards
	beauty standards are unrealistic
	feminism and gender bending is enforcing stereotypical beauty standards
	beauty standards are toxic
	women who do not meet conventional beauty standards are not women
	beauty standards is nothing but a money making scheme
	beauty standards that cater to minorities are trained to be inclusive
	Women who don't fit societal beauty standards get catcalled and harassed
	social media attempts to hierarchize beauty to maintain dominance over others
	beauty standards are hard to escape
	Masculinity is not toxic but attractive
	Beauty standards are bad and stressful for young people
	No natural humans look like that
	Beauty is everywhere
	Beauty standards are racist
	Hypermasculinity in of itself is the problem
	No claim in the list is describing the tweet
Cloning	
	clones can perfect can give humans preferable qualities
	Cloning can save lives or cure humans
	Scientist that clone are acting unethically
	Couples without kids would rather use cloning than employ surrogates or IVF
	transplanting organs can be made easier and more successful by cloning
	Cloning can potentially create premature ageing
	Cloning in morally wrong
	cloning could provide childless couples with an enhanced or enlarged family
	cloning can be use to reduce risks
	Cloning will cause parents to customise their children
	Cloning is medicine; advances in cloning are advances in medicine
	People could keep on living due to cloning
	Cloning affects negatively to the reproductive processes
	Better cloning techniques offer higher chances of success with less moral hazards
	People that get cloned maintain their personality
	Cloning is playing God
	Cloning is for evil purposes
	Humans will become a product with cloning
	Cloning animals is cruel
	Cloning someone is not a safe thing to do
	Cloning is a natural
	Cloned people do not have souls
	Cloning will be accepted some day
	Cloning is akin to murder or manslaughter
	You lose a sense of individuality when cloned
	No claim in the list is describing the tweet

Table 16: Second list of narratives

Topic	Narrative
Corporate culture	
	non profit companies and for profit companies are equally untrustworthy
	Remote work makes it difficult to maintain company culture
	Jobs needs to be treated with respect
	Companies simply want drones that do not ask questions
	there is a lot of hustle culture for millennials
	Working remove does not mean that company culture are not important
	If you want a great company culture you should hire great people with a good work ethic
	Corporate culture is bad
	corporate culture is racist
	Companies can do what they want to do
	millennials do not want to work
	side effect of hustle culture is often a counterproductive narrowing of focus
	corporations are getting tax cuts while you are getting fired
	So many popular companies actually have a terrible culture and a bad work ethic
	Victims of abuse are ignored and silenced
	corporations that mass fire employees say that Nobody wants to work in the media or get record high profits
	mass firings are commonplace in corporate takeover
	Our company culture is good because we have fun
	corporate culture do not reward hard work
	job hopping looks bad on your resume
	Companies are anti union
	corporate culture so rotten to the core by greed
	Getting a stable job is getting harder over time
	A fun company culture does not care about your work life balance
	Some office cultures are like a cult and are not healthy for you
	loyalty to the company trumps everything
	Managers of a company cause nothing but trouble
	perks from a company are useless
	Young people refuse to enter or stay in the workforce
	Companies that make you feel like a family is good
	Companies that say they that you are a family is brainwashing you to comply
	Hard work gets you everywhere
	corporate culture promising generous pay and perks in order to mask workers disposability and exploitation
	loyalty to the company means absolutely nothing in this day and age
	You can make a remote workers feel part of the team without being in the same physical space
	An employee is a representative of a company and should respect the dress code and look professional
	work dress code is discriminatory
	Corporations that do mass firings are greedy
	Do not trust companies
	No claim in the list is describing the tweet
Crypto	
	People will sell you out but you can trust crypto
	crypto is used solely for pump and dump schemes
	crypto allows a complete transformation of the global economy though the great reset
	Influencer are scamming their fans using crypto
	crypto is ponzi scheme
	crypto is a scam
	crypto is for money laundering
	crypto is undiscriminatory to people of color
	crypto should be regulated
	crypto is legal way to pay
	crypto is hijacked by ruthless investors
	crypto is for terrorists
	crypto is toxic
	crypto is just a way to diversify your assets
	crypto can bank the unbanked
	crypto is for criminals
	crypto is a mean for the rich to get bailouts
	crypto is unsustainable for the environment
	crypto better digital payment than credit cards
	crypto is used to fund political parties
	crypto is yet another tech bubble ready to burst and fail
	Bitcoin is like a Platypus it is not real
	crypto is simply digital gold
	crypto allows one to become financially free
	crypto allows easy access to financial services
	crypto can help fighting greedy bankers
	crypto is an investment
	crypto is used to steal the constitution
	Crypto is not democratic
	crypto should be carbon taxed
	crypto is useful when banks are failing
	crypto mining should be banned
	blockchain is no different from a database
	crypto is a wealth builder
	WEF want you to own nothing and be happy, crypto avoids this problem
	crypto provides upwards mobility
	No claim in the list is describing the tweet

Table 17: Third list of narratives

Topic	Narrative
(Baby) Formula	
	breastfeeding is natural and is not politically correct
	women are pressured into breastfeeding and gets stressed
	baby formula is costly
	baby formula is killing babies
	It does not have to be all or nothing
	the political right vote against bills to make things more expensive
	breastfeeding and baby formula is risky
	baby formula ends up in foreign countries instead of the US where it should be
	Some people are allergic to breast milk and need formula
	People hate babies when they make abortion illegal and remove baby formula
	baby food industry is promoting propaganda
	breastfeeding can cause HIV
	baby formula is poisonous
	baby formula is good as fathers can feed their baby
	baby formula is best if you cannot breastfeed
	It is important to secure enough baby formula in an economic crisis
	breast milk is best
	The political left is out to remove babies
	People are reselling baby formula to other countries for higher prices
	Do what is best for you and the baby
	The baby formula shortage is one big scam
	Some people cannot tolerate formula and need milk
	breast milk has a lot of antibodies that can help the baby fight off infection
	people publicly shame women who breastfeed in public
	mental health is more important than breast feeding
	breastfeeding is healthier than baby formula
	baby formula can cause infection
	breastfeeding will lowers risk of breast cancer
	The political right have caused a baby formula shortage
	breast is best campaign causes anxiety in moms who cannot breastfeed
	No claim in the list is describing the tweet
Influencers	
	Influencers can be damaged by everything they say
	social media in the west is like opium for kids
	influencers just want to get rich quick
	influencer marketing are not authentic
	social media career is not sustainable
	influencers are creative
	influencers earn too much money
	social media people are toxic and rude
	an influencer is a social media celebrity
	Normal jobs are boring
	influencing is indeed hard work
	influencers understand the use cases of products and want to help
	you should not quit your job and become an influencer
	social media people are just plagiarising other people
	Becoming an influencer allows you to live the good life
	influencers do not know hard work
	influencers wants to be their own boss
	dealing with hate is part of a social media job
	the numbers of followers do not make you successful
	influencers are not respected
	influencer is not an adult job
	jobless youth are spending too much time on social media platforms
	If you have too few followers you should get a job
	influencers are wasting their time
	being an influencer is easy
	influencer is not a real job title
	No claim in the list is describing the tweet
Mental Health in sports	
	male dominated sports are toxic for women
	In sports people get into drugs when mental health declines
	sport help you alleviate stress
	The money made in sports should go to mental health organisations not admin staff
	team sport is a brutal business
	female athlete are not projected
	When athletes get in trouble the blame the media
	sports is like religion it is bad for your mental health
	In sports racism and mental health issues goes hand in hand
	athletes do not have any problems
	athletes are or must become hard workers
	athletes does not have real mental health problems
	Work late nights, don't take sick days
	Be tough, vulnerability is weakness
	Entering sports in an early age led to burnout
	getting help is stigmatising
	elite athletes have unfair genetic advantages
	sports athletes are manipulated
	mental health is not masculine
	athletes are only thriving professionally if they thrive personally
	mental health should not be treated like the flu
	trans people should not participate in male or female sports
	be ashamed if talking about mental health
	sports athletes are depressed
	you do as your told as an athlete
	alienation cause mental health issues in sports
	athletes should not worry because they have a lot of money
	talking about mental health is showing weakness
	athletes that speaks up about health issues are silenced
	No claim in the list is describing the tweet

Table 18: Fourth List of narratives

Topic	Narrative
Nuclear Energy	
	Nuclear reactor is easy to control
	Deciding what to do with regards to long term disposal of nuclear energy waste is difficult
	Nuclear energy will be available for use longer than oil for example
	Nuclear energy will contaminate the environment
	Nuclear energy is dangerous
	Nuclear energy can give us unlimited energy
	nuclear energy waste can be recycled
	Nuclear energy is good
	Using nuclear energy to solve problems that arise is logical
	Nuclear energy leads to more violence
	nuclear power produce carbon free energy
	nuclear energy is not efficient
	nuclear power is financially burdensome
	There is no significant risk with nuclear energy that cannot be said about other agents as well
	nuclear energy is dirty
	nuclear energy is not safe
	nuclear energy makes poor nations dependant on rich nations
	Every country can use nuclear energy unlike everything else
	Nuclear energy relies too heavily on subsidies
	There is not a good plan for storing or disposing of nuclear energy waste so we should use it
	Nuclear plants only produce electricity and cannot replace oil and gas
	Using nuclear power will lead to nuclear war.
	renewable energy is a more viable option than nuclear energy
	Nuclear energy are favoured by certain social structures like capitalism
	decentralised nuclear energy production is efficient
	Nuclear power is needed to stabilise climate change.
	Nuclear energy should not even be considered as an energy source
	Nuclear energy is much more harmful than beneficial
	Green energy will make nuclear energy obsolete
	Nuclear energy will increase the cancer in humans
	nuclear energy is not renewable energy
	Nuclear reactors are vulnerable to terrorist attack
	nuclear energy is more reliable than renewable energy sources like solar
	No claim in the list is describing the tweet
Transport	
	trains are better than flights
	public transportation is unsustainable for rural areas
	public transportation is useless car is better
	cars give you the freedom of Independence
	buses are safer than cars
	fewer drivers equals safer streets
	public transportation is comfortable
	public transport results in less pollution
	trains are better for the climate
	public transportation has no personal space
	It is important that public transit works
	cars are good when there is no good alternative
	flights are better than trains
	public transportation is for poor people
	public transportation is ridden with disease
	highways are not profitable
	public transit only works if affordable
	using bikes are dangerous
	public transportation is filled with germs
	cars are worse than buses as it carries less people
	buses are better than cars
	public transportation is good business
	taxing car and roads hurt the poor people
	cars are for rich people
	public transportation is unsafe at night
	car centric infrastructure is bad
	trains are too expensive
	No good transportation is a reflection of the government
	cycling will decrease car traffic
	public transit is not profitable
	public transportation is good
	No claim in the list is describing the tweet

Table 19: Fifth list of narratives