# A Dataset of Sustainable Diet Arguments on Twitter

**Marcus Astrup Hansen** **Daniel Hershcovich**
Department of Computer Science
University of Copenhagen
dh@di.ku.dk

## Abstract

Sustainable development requires a significant change in our dietary habits. Argument mining can help achieve this goal by both affecting and helping understand people's behavior. We design an annotation scheme for argument mining from online discourse around sustainable diets, including novel evidence types specific to this domain. Using Twitter as a source, we crowdsource a dataset of 597 tweets annotated in relation to 5 topics. We benchmark a variety of NLP models on this dataset, demonstrating strong performance in some sub-tasks, while highlighting remaining challenges.

## 1 Introduction

In Natural Language Processing (NLP), impact on climate change is usually only framed in the context of efficiency Strubell et al. (2019); Schwartz et al. (2020); Puvis de Chavannes et al. (2021). While efficiency improvements are welcome, we risk greenwashing NLP and further neglecting the field's potential to positively impact climate change. Hershcovich et al. (2022) proposed to strive towards *net positive* climate impact of NLP by developing beneficial applications (see §3 for related work in this direction).

In IBM's Project Debater (Slonim et al., 2021), a large team of researchers created a system capable of autonomously debating a human in a structured environment. While the system could not convince many people to switch positions, it helped to educate people about certain topics. This can be regarded as a first step towards behavioral change (Boström, 2020; Lockie, 2022).

In this paper we propose to apply debating technology to promote behavioral change that benefits the environment and climate: namely, mining arguments that can convince people to undergo a shift to a more climate-friendly diet (see §2). Our focus is on extracting and labeling argumentative structures used in online social media—specifically, Twitter—and compiling them into a domain-specific English dataset for green nutrition. Our annotation focuses on subjective and anecdotal evidence, shifting away from traditional argument mining methods where more strict explicit evidence is preferred. This shift is motivated by sociological research that shows that anecdotal stories are more persuasive in changing people's opinion (Petty et al., 1981; Hidey et al., 2017). Finally, we train and benchmark baseline models on the dataset, showing promising results but also identifying important challenges.[1]

## 2 Sustainable Diets

To successfully transform our societies to become more sustainable, we need to focus on improving the sustainability of our diets. The EAT-Lancet report (Willett et al., 2019) has marked this as a shift away from excessive consumption of animal protein-heavy diets. However, unfortunately, such diets are generally quite prevalent in many developed countries. The science behind the benefits of such a transition is quite well established (Prag and Henriksen, 2021), but there is still a lack of incentives to change habitual behaviors for people participating. To change such incentives and habits requires actions from all aspects of society, including individual consumers. Loorbach (2009) argues that the social transition of our diets to become more sustainable requires us to continuously monitor and evaluate processes across all the societal facets to help solve issues and update practices.

Therefore to successfully transform our society to consume a sustainable diet for a successful green transition, we must change the social and cultural conditions and traditions around green nutrition. However, Graça et al. (2019) shows that there is solid evidence that established dietary preferences

[1]The dataset and models can be found in https://github.com/danielhers/sustainable-diet-arguments-twitter.



*Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 40 - 58

are hard to change for large consumer segments due to negative taste perceptions and lack of knowledge and skills about healthy and green foods. Here, we address this challenge by aiming to collect arguments covering various aspects, beyond the obvious ones about health and climate. Regardless of which aspects are more convincing, the end result will benefit the climate—our rationale is that the end will justify the means.

## 3 Related work

**Positive environmental impact.** Machine learning and related fields have a substantial potential to help address climate change (Kaack et al., 2022). Some of the potential paths where NLP can be used for a positive impact include helping people understand their carbon footprint, facilitating behavior change towards more sustainable practices, informing policy and supporting education and financial regulation (Rolnick et al., 2019). Cross-disciplinary research with social science can help improve the understanding of large-scale discourse spread over multiple channels regarding climate change (Stede and Patz, 2021). Successful examples include compliance verification of corporate reports: Bingler et al. (2022) examined annual corporate reports and found many engage in "cheap talk" (greenwashing), e.g., lacking specificity in climate goals and activities. Biamby et al. (2022) created a dataset for and detected images with misleading captions on Twitter for several topics, including climate change. These efforts allow for better policy shaping and steering of the online discourse around climate change, which we hope to achieve with our work too.

**Project Debater.** As part of the Debater project (Slonim et al., 2021), Ein-Dor et al. (2019) created an end to end argument mining system where topics are used to mine for arguments in a very large corpus of English newspaper articles and Wikipedia articles. Toledo-Ronen et al. (2020) subsequently automatically translated the argument corpus to five languages, projecting the labels from English. They additionally collected and annotated crowdsourced arguments in these languages natively, annotating argument quality and evidence. They used a large group of annotators with rigid guidelines, resulting in high quality multilingual arguments. We use a similar framework and methodology, but use Twitter as a corpus and focus on English only in this paper.

**Argument mining from social media.** Early work on argumentation mining from Twitter found it is a feasible but challenging task, due to unique linguistic properties (register, domain, noisy data) and differences with respect to established argumentation theories and phenomena, e.g,. the need to distinguish opinions from facts (Habernal and Gurevych, 2017; Dusmanu et al., 2017). More recently, Schaefer and Stede (2020) created a dataset of 300 German tweets containing the word "Klima" (climate), annotated for three labels: argumentative, claim and evidence. They experimented with different models for classifying tweets, using an argument mining pipeline (Schaefer, 2021) that first filters out irrelevant tweets, then extracts ADUs (argument discourse units, namely claims or evidence), classifies the tweets as either supporting or attacking a claim and builds a graph of ranked arguments. They stressed the importance of argument quality prediction as part of the pipeline. Our approach is similar to Schaefer and Stede (2020)'s, but we leave argument quality to future work. As examples for alternative approaches, Schaefer and Stede (2021) annotated 3244 German Facebook comments on a political talk show's page from February 2019. They classified toxic, engaging and fact-claiming comments, focus mainly on the latter due to their relation to evidence detection for argument mining. Cheema et al. (2022) created a multimodal argument mining dataset with the focus on verifiable claims, manually annotating 3000 tweets for three topics covering COVID-19, climate change and technology. They found that identifying check-worthy claims was subjective for both students and experts, and that pre-trained models yield the best performance for both modality types. Wojatzki and Zesch (2016) created a dataset of argumentative tweets for the topic of atheism, using stance as a proxy for implicit arguments. They allowed annotators to mark text as lacking context or being ironic, and asked them to annotate the stance of arguments towards the topic. They then used this measure as the signal for implicit arguments. For explicit arguments, the annotators could only annotate stance towards targets if they had textual evidence.

## 4 Annotation Scheme

We define an argument mining annotation scheme based on previous work (Aharoni et al., 2014; Ein-Dor et al., 2019; Slonim et al., 2021; Schaefer and Stede, 2020), consisting of Topics and the annota-

tion labels Argumentative, Claim, Evidence, Evidence type and Pro/Con.

**Topics.** To be useful for debates and analysis, arguments are mined with respect to a *topic*—"a short, usually controversial statement that defines the subject of interest" (Aharoni et al., 2014). Topics need to be short, clear, dividing, and relevant to our central theme of sustainable nutrition. We also wish for the topics not to be too specific—for high coverage, we choose broad and simple topics:[2]

| | |
|---|---|
| **T1** | *We should reduce the consumption of meat* |
| **T2** | *Plant-based food should be encouraged* |
| **T3** | *Meat alternatives should be encouraged* |
| **T4** | *Vegan and vegetarian diets should be encouraged* |
| **T5** | *We should pursue policies that promote sustainable foods* |

**Argumentative.** Argumentative is the label that denotes if a tweet is argumentative for any topic. This means the tweet contains argumentative structures such as claims or evidence while having a clear stance on some topic. We define arguments broadly, including those that do not refer to the topic explicitly but whose stance toward it is only implied. Indeed, Wojatzki and Zesch (2016) achieved a similar result by using stance detection as a proxy. If an argument is not clear in its stance, i.e., it is neutral or unrelated, it is not be considered argumentative.

**Claim.** A claim is a standpoint towards the topic being discussed (Schaefer and Stede, 2020). We expand upon this definition by allowing the standpoint to indirectly acknowledge the topic discussed, which is *implicit argumentation*, or explicitly when directly acknowledging the discussed topic. If a claim is not related to the discussed topic, it is not considered a claim. A claim should further be able to exist in a self-contained manner, not relying on external references to fully convey the claim and stance it takes towards the topic. Therefore, it should be able to fully articulate the entire claim without the need for external reference. A tweet referencing others' stance towards the topic is not considered a claim.

**Evidence.** Evidence is a statement that explains a stance towards the topic. It can be stated in combination with a claim, or it can be self-contained if it is just stating a fact or referencing studies related to the topic. Therefore a tweet does not have to co-occur with a claim to contain evidence relevant to the topic, and as such, evidence is not dependent on a claim when annotating (see §5). A tweet can still contain claims with supporting evidence as part of its text. If evidence is unrelated to the discussed topic, it is not considered evidence.

**Evidence type.** Evidence is labeled as one of the following types. The first three types are from Rinott et al. (2015), while we propose the last two based on preliminary exploration of our data:

1. **Anecdotal.** A description of an episode(s), centered on individual(s) or clearly located in place and/or in time.
2. **Expert.** Testimony by a person, group, committee, an organization with some known expertise/authority on the topic.
3. **Study.** Results of a quantitative analysis of data, given as numbers, or as conclusions.
4. **Fact.** A known piece of information without a clear source, regardless of whether it is a *true* fact or not. See example in Figure 1a.
5. **Normative.** Description of a belief or value the author holds. See example in Figure 1b.

See Table 1 for examples from the dataset. If its type is unclear, a tweet should not be considered evidence, and might be a claim instead. If neither is clear, the tweet itself might lack context and should not be considered argumentative.

**Pro/Con.** The stance of a tweet towards a topic depends on a claim or evidence being present in the tweet. Moreover, if there is no clear stance, the tweet should not be considered argumentative.

## 5 Dataset

Here we describe the procedure for collecting and annotating our dataset of tweets containing arguments related to the topics described in §4.

**Scraping.** The corpus used for annotation is a collection of tweets scraped from Twitter using tweepy[3] by iteratively creating queries by a combination of keywords[4] and n-grams from an initial set of topics. For each query, we scrape a maximum of 1000 tweets. We remove retweets, quote tweets, links and videos, as well as tweets with less than three words, resulting in 31840 English tweets in total. User mentions are replaced with

[2] Note that T5 is more complex and specific. It covers a specific type of tweets that we noticed during early annotation work, discussing sustainable food policy.

[3] https://github.com/tweepy/tweepy

[4] See Appendix D for a listing of the queries.

(a) Humans should not eat animals (Claim) as we don't need meat to fulfill our nutritional needs. (Evidence: Fact)

(b) It is morally wrong to eat and cause animals pain to fulfill our nutritional needs. (Evidence: Normative)

Figure 1: Simplified examples of arguments for the topic T1 *(We should reduce the consumption of meat)*. In (a) the evidence type is Fact, since no source is given. In (b) it is Normative, as it describes a belief but is more elaborate than a claim. Note that the level of granularity in our dataset is a whole tweet rather than spans within a tweet. Here, spans are indicated to illustrate which part of the tweet suggests that it should have a particular label.

| Evidence type | Example | Topic(s) | Pro/Con |
|---|---|---|---|
| Anecdotal | *We are on the green bean diet here, too! I love them. Mom hasn't tried broccoli* 😱 | T1, T2, T3, T4, T5 | Pro |
| Expert | *Many fruit & veg (which contain natural acid) don't trigger flare ups- The list is long and varied (obs this may not apply to you) but after a little digging I found some doctors do reccomend a plant based diet to ease the inflammation. Going meatless is even recommended by ICA* | T2, T4 | Pro |
| Study | *According to a 2022 study, eating an optimal #diet of whole grains, legumes, fish, fruits, vegetables and nuts can improve life expectancy by how many years?* | T2, T4 | Pro |
| Fact | *Hey eco-friendlies! The well known high-carbon company McDonalds produces 1.5 MILLION tonnes of food packaging alone* 🤢*! Fun fact carbon footprints are important! Tune in for more behind closed door stats!* | T5 | Pro |
| Normative | *The dangerous of this thing is that our vegan extremists will start interfering in this..* | T4 | Con |
| Unrelated/no evidence | *Give your children healthy food to avoid the dad bod haha* | | |

Table 1: Examples from the dataset of tweets containing evidence for each evidence type, the topics for which they were annotated as evidence and their pro/con annotation for each of the topics.

<*MENTION*>. Hashtags and emojis are kept as they contain relevant information.

**Relevance-based filtering.** Upon initial inspection, we find that most tweets are not relevant to any of our topics, despite matching our queries. Therefore, before sampling data for annotation, we use an information retrieval system to extract the most relevant tweets in relation to our topics. We use a neural ranking model trained for semantic search (Reimers and Gurevych, 2019) that was trained to score the relevance of an answer to a question. We deem this a decent proxy for our retrieval system as we want to find tweets that take a stance and explain their claims and evidence in the context of a topic. We elaborate more on this model in §6.

**Sampling.** When generating our dataset, we sample 250 tweets at random from the full unfiltered corpus, and combine this set with 347 random tweets filtered by the semantic model.

**Annotation.** Annotation is conducted using Amazon Mechanical Turk[5] in rounds as described in Figure 2. Five workers annotated each instance. The annotations guidelines are given in Appendix A. First, tweets are annotated as for whether they are Argumentative regardless of a topic. Second, annotators are presented with an Argumentative tweet as well as our list of topics, and are asked to select the topics for which the tweet is a Claim. This ensures that annotators judge the topics relative to each other and are thus more consistent (despite their conceptual overlap) than if each tweet/topic pair were annotated separately. Separately, annotators are presented with an Argumentative tweet as well as one topic, and are asked to select the Evidence Type of the tweet with respect to the topic (or indicate that it is not Evidence). This is done to facilitate the annotation of the heterogeneous Evidence label. The binary label is then derived from this annotation by collapsing all types as positive.

[5] https://www.mturk.com

| Topic | Tweets | Arg | ADUs | Claims | Evidence | Claims with evidence | Pro | Con |
|---|---|---|---|---|---|---|---|---|
| T1 | 597 | 387 | 118 | 63 | 89 | 34 | 77 | 37 |
| T2 | 597 | 387 | 130 | 92 | 85 | 47 | 89 | 38 |
| T3 | 597 | 387 | 85 | 42 | 63 | 20 | 58 | 27 |
| T4 | 597 | 387 | 156 | 106 | 112 | 62 | 99 | 54 |
| T5 | 597 | 387 | 140 | 60 | 113 | 33 | 96 | 37 |
| Full set | 2985 | 1935 | 629 | 363 | 462 | 196 | 419 | 193 |

Table 2: Statistics for the different topics and the overall full set. Arg: Argumentative. ADU: argument discourse units (Claim or Evidence). Labels are based on majority voting among annotators. Of the ADUs, we see more Evidence than Claims. Pro/con labels are rather unbalanced, with a bias towards positive stance.

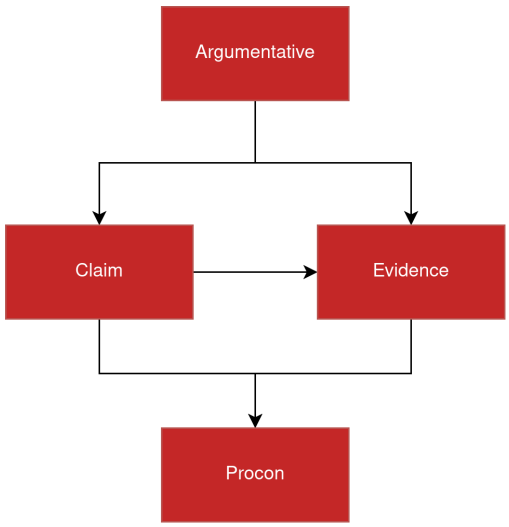


Figure 2: Dependencies between annotation rounds.

Finally, Claims and Evidence are presented along with one topic at a time, and annotators are asked to indicate whether they support or contest the topic.

**Inter-annotator agreement.** We calculate the average inter-annotator agreement for our crowd-sourced data using Cohen's kappa. The resulting scores are 0.49 for Argumentative, 0.47 for Claim, 0.15 for Evidence (including type) and 0.63 for Pro/Con. The low agreement for Evidence is likely due to the multi-class label being harder to agree upon than a binary label.

**Statistics.** Table 2 presents statistics of the labeled dataset. Most annotators labeled a substantial amount of tweets as Argumentative. However, only a minority actually contained ADUs (Claims/Evidence). This discrepancy can be attributed to the Argumentative label being decoupled from the topic itself: an Argumentative tweet might only be relevant for another topic, either within our set of five topics or for a different topic altogether.

Like Cheng et al. (2022), we find substantially more Evidence than Claims, though their Evidence depends on Claims. Evidence seems to generally be more prevalent than Claims in online discourse. This can also result from our annotation procedure, where Claims require identifying relevant topics, and Evidence requires identifying the type. On the other hand, the broad types of Evidence we allow and the fact that they are not dependent on Claims allows for more Evidence than in other datasets.

The fact that Pro/Con labels are biased towards positive stance could be due to online discourse being more prevalent for the Pro side rather than other domains. The annotators' preconceived notions might have played a role in them being more inclined to select Pro in situations where they could have been uncertain due to the topic's definitions.

**Topic overlap.** In Figure 3, we see how much each topic's tweets overlap with other topics as a percentage of their combined number of tweets, where they both have either evidence or claim. We see that all topics have roughly 20% of their tweets overlapping with another topic. This is not surprising as the topics are all very similar, and tweets can easily be relevant for more than one topic at a time.

**Evidence types.** Figure 4 shows the distribution of Evidence types in the dataset. Most Evidence is Normative or Anecdotal, reflecting online discourse being less strict, which lends itself to using weaker types of Evidence to explain one's stance.

## 6 Experiments

To evaluate the ability of existing models to mine arguments according to our scheme, we conduct a series of experiments with various approaches.


Figure 3: Percentage of claims (above) and evidence (below) overlapping between topics: T1=meat, T2=plant, T3=alternative, T4=vegan, T5=policy.

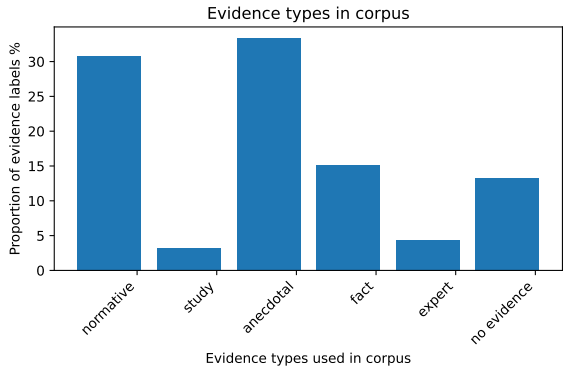


Figure 4: Distribution of Evidence types in the dataset. Note that "no evidence" is considered a type due to the combined annotation procedure, where Evidence is annotated immediately along with its type (or as "no evidence" when no type is applicable).

### 6.1 Information Retrieval

We experiment with information retrieval baselines, rating how likely a document is to be relevant for a query:

BM-25 (Trotman et al., 2014) is a standard retrieval model based on TF-IDF scores of exact token matches, used in many systems and should give a good benchmark for the difficulty of retrieving claims and evidence just from topic queries. It returns an unbounded positive score, which we cut off at 1.

`multi-qa-MiniLM-L6-cos-v1`[6] is a sentence transformer (Reimers and Gurevych, 2019) based on MiniLM (Wang et al., 2020), which is a distilled version of `UniLM v2` (Bao et al., 2020), which was pre-trained on 160GB text corpora from English Wikipedia, BookCorpus, OpenWebText, CC-News and Stories. `multi-qa-MiniLM-L6-cos-v1` was fine-tuned on the concatenation of multiple question answering (QA) dataset, totalling about 215M instances. This is the same semantic search model we used in §5 for filtering tweets, and therefore this experiment should give us a good idea of how well our models perform compared to a model that has had an impact on the selection previously. The model returns a score between 0 and 1. We use 0.5 as the cut-off for classification.

The retrieval models are unsupervised, and consider neither argumentativeness, which is independent of the topic, nor pro/con (stance classification). However, they serve as a baseline for claim and evidence detection, as those tasks have a retrieval aspect. We use the tweet as a document and the topic as a query, scoring their relevance and using the resulting scores from the models for classification.

### 6.2 IBM Debater

IBM Debater offers implementations for various argument mining components (Slonim et al., 2021), and provides an API,[7] which we use as a baseline representing existing argument mining models. It has been trained on a different type of data from different domains and with stricter annotation guide-

[6] `https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1`

[7] `https://early-access-program.debater.res.ibm.com`

lines. We evaluate their "zero-shot transfer" to our dataset, without any further training.

### 6.3 Fine-tuned RoBERTa

Pretrained language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have been used successfully on similar datasets (Cheng et al., 2022; Schaefer and Stede, 2021).[8] We fine-tune and evaluate `cardiffnlp/twitter-roberta-base`,[9], which was trained on a dataset containing 58M tweets (Barbieri et al., 2020), specifically to handle user identifier tokens and emojis.[10] For claim, evidence, and pro/con, the topic plays an essential role in the classification. To encode tweet-topic pairs, we combine the tweet and topic using a separator token (`[SEP]`).

Our dataset contains probabilities for each label according to the distribution over the different annotators. We use cross-entropy with raw probabilities rather than rounding the labels, and fine-tune the RoBERTa encoder as part of the training. The hyperparameters used are: learning rate 5e-5, batch size 5, weight decay 0.05 and the `adamw` optimizer.

### 6.4 XGBoost + RoBERTa

Schaefer and Stede (2020) used XGBoost (Chen and Guestrin, 2016) in combination with BERT on a similar dataset to ours. We evaluate this model on our dataset across all label targets. We train the XGBoost classifier on top of frozen contextualized embeddings from RoBERTa (again, `cardiffnlp/twitter-roberta-base`), since XGBoost is not a neural model and does not support backpropagating gradients to fine-tune the underlying encoder. All labels in this experiment are determined by majority vote and take the values $\{0, 1\}$ except for pro/con, which takes the values $\{-1, 1\}$.

Here we have two ways of embedding the topic: the first approach is to embed the tweet only on its own, which is done for the argumentative task as it is not dependent on the topic. The other method is to combine the tweet and topic using a separator token (`[SEP]`), which is used for the other tasks.

We run a grid search with three-fold cross-validation for each task, over three hyperparameters: learning-rate $\in \{0.01, 0.03, 0.06\}$, max-depth $\in \{1, 3, 5, 6, 7, 8, 9, 10\}$ and number of estimators $\in \{1, 2, 5, 7, 10, 15, 20, 25, 30, 40, 60, 80, 100\}$. We select the best combination based on macro F1 score.

[8] See Appendix E for experimental replication of previous results.

[9] https://huggingface.co/cardiffnlp/twitter-roberta-base

[10] The model's tokenizer does not have `<MENTION>` as a token. Instead, it recognizes `@user`, so we replace all our `<MENTION>` tokens with `@user` in preprocessing.

| Model | Macro F1 | Macro P | Macro R |
|---|---|---|---|
| Majority Class | 0.39 | 0.32 | 0.50 |
| Random Class | 0.49 | 0.50 | 0.50 |
| Fine-tuned RoBERTa | 0.51 | 0.51 | 0.51 |
| RoBERTa + XGBoost | **0.67** | **0.69** | **0.67** |

Table 3: Results from models evaluated on the **argumentative** task. P and R are precision and recall, with their attached averaging type. Highlighted are the best performing model for their task and averaging type.

| Model | Macro F1 | Macro P | Macro R |
|---|---|---|---|
| Majority Class | 0.45 | 0.41 | 0.50 |
| Random Class | 0.50 | 0.50 | 0.50 |
| BM25 | 0.50 | 0.61 | 0.55 |
| multi-qa-MiniLM | **0.67** | **0.73** | **0.65** |
| IBM-API | 0.57 | 0.57 | 0.57 |
| Fine-tuned RoBERTa | 0.48 | 0.50 | 0.47 |
| RoBERTa + XGBoost | 0.51 | 0.59 | 0.53 |

Table 4: Results on the **claim** task.

### 6.5 Experimental Setup

To classify for argumentative tweets, we only use the tweets and disregard the topics. We subsequently only use argumentative tweets (according to the human annotation) when experimenting with detecting claims and evidence. Pro/con classification is only evaluated on for tweets containing evidence or claims (according to the human annotation). We perform 3-fold cross-validation with maximum 15 epochs, using early stopping based on validation macro F1 evaluated every 20 batches with patience set to 5.

We also report results for Majority Class and Random Class baselines, which respectively select the most common label for each task (based on the training set), and a random class with uniform probability.

| Model | Macro | | |
|---|---|---|---|
| | F1 | P | R |
| Majority Class | 0.44 | 0.39 | 0.50 |
| Random Class | 0.49 | 0.50 | 0.50 |
| BM25 | 0.52 | 0.60 | 0.56 |
| multi-qa-MiniLM | **0.64** | **0.66** | **0.63** |
| IBM-API | 0.46 | 0.51 | 0.57 |
| Fine-tuned RoBERTa | 0.48 | 0.49 | 0.48 |
| RoBERTa + XGBoost | 0.47 | 0.60 | 0.51 |

Table 5: Results on the **evidence** task.

| Model | Macro | | |
|---|---|---|---|
| | F1 | P | R |
| Majority Class | 0.40 | 0.34 | 0.50 |
| Random Class | 0.52 | 0.53 | 0.54 |
| IBM-API | **0.59** | **0.60** | **0.59** |
| Fine-tuned RoBERTa | 0.45 | 0.48 | 0.45 |
| RoBERTa + XGBoost | 0.53 | 0.53 | 0.54 |

Table 6: Results on the **pro/con** task.

## 7 Results

The results are shown in Table 3 for argumentative, Table 4 for claim, Table 5 for evidence and Table 6 for pro/con. In Table 3 we see XGBoost performs well on all metrics for the argumentative task. The fine-tuned RoBERTa does not perform well on the argumentative task, underperforming both the random and majority baselines. In Table 4 we see that `multi-qa-MiniLM-L6-cos-v1` outperforms all other models with a large margin for claims. BM25 only matches when there is an overlap in vocabulary between tweet and topic, which `multi-qa-MiniLM-L6-cos-v1` does not require. Table 5 shows similar results for evidence, where `multi-qa-MiniLM-L6-cos-v1` and BM25 outperform all other models. One interesting result is the relatively large dip in performance for the IBM-API for evidence with respect to claims, suggesting the change in annotation style for evidence has a significant impact compared to previous works. On the pro/con task (Table 6), both the IBM-API and RoBERTa + XGBoost outperform the baselines in all metrics, but not the fine-tuned RoBERTa. The IBM-API has the best performance in this case, by a large margin.

**Input encoding for XGBoost.** The different methods of combining topics and tweets for XGBoost (see §6) have a relatively small impact on performance. The concatenation method outperforms the `[SEP]` method in pro/con and claim, and therefore we only report results using it in the tables.

**XGBoost vs. fine-tuning.** Overall, XGBoost performs well compared to the fine-tuned RoBERTa. This could be due to training issues or a lack of data: we only have about 600 unique tweets, with only a fraction of them being annotated as containing claims and evidence, causing issues of sparsity and dataset imbalance. This could be mitigated by using a different training approach or annotating more examples in the future.

**Success of retrieval models.** The retrieval models perform well in the claim and evidence tasks, where `multi-qa-MiniLM-L6-cos-v1` performs the best overall. Of course, this result should be interpreted with great skepticism, as it is likely due to the filtering process we did early in our dataset compilation (§5) and should not be discounted as it has added some bias to the data. However, it also shows that the filtering process did have a decent impact on scoping in on tweets most likely to contain argumentative structures. Therefore, the `multi-qa-MiniLM-L6-cos-v1` results could be interpreted as the proportion of retrieved tweets containing argumentative structures. The BM25 model performs well with its precision scores for the binary average, which makes sense as it requires a vocabulary overlap between the tweet and topic. Due to a relatively low overlap between the tasks for the tweet and their topics (see Appendix C), BM25 only needs one token to overlap for it to mark it as relevant and therefore will retrieve quite a few false-positive tweets on average. Nevertheless, this could also be because each topic only has a few keywords, making them good queries. Overall, the retrieval models make a good baseline for future evidence and claim tasks experiments.

**IBM Debater.** The IBM-API models also perform well for the pro/con and claim task. However, surprisingly, the model performs poorly on the evidence task. This could be due to a shift in the task definition, since we added two new types of evidence: normative and fact. They account for nearly half of all the annotated evidence. However, anecdotal evidence is based on the IBM Debater definitions and is the most frequent type of evidence, so the issue might be one of several. First,

the semantic structures in tweets are hard for the IBM models to adapt to, causing them to miss most evidence. Another reason could be that annotators have overused anecdotal evidence where it should have been labeled as normative or fact or not as evidence. Overall, the IBM models have performed exceptionally well, considering they have never seen data of this type when compared to other baselines.

## 8 Discussion and Limitations

While we frame our dataset around sustainable diets, it is, in fact, focused on plant-based diets. Many other aspects are relevant for sustainability, including production, geographical location, genetic modification, transportation, water consumption, land preservation and health. We leave these issues to future work.

The topics used in this paper are simple by design. They are all quite similar, which might cause some correlation issues when training models. For instance, T1 (discussing meat consumption) has a significant overlap with T3 (discussing meat alternatives) of 22% for claim and 26% for evidence. However, it can also indicate the presence of other topics that are similar to both. For instance, when arguing for reducing meat, people might use animal welfare as evidence. Therefore, topic exploration and expansion could be done further to improve the spectrum of topics in the dataset and explore how relationships between topics are made and related in debates.

The dataset is a starting point for training argument mining models. It is balanced in the distribution of the claims and evidence across the topics, with a minor overlap between topics of roughly 20%. Our annotation guidelines are robust enough to be used for crowdsourced and expert annotation. The low agreement in the crowdsourced annotations for evidence may be improved by better guidelines or a different annotation methodology, but they may simply be a reflection of inherent subjectivity. This will be investigated in future work.

One issue with this dataset is its relative lack of context for many of the tweets due to them referencing outside tweets or responding to other users in a discussion. There is good potential here to utilize this external context for further argument mining or further improve the detection of claims and evidence in the primary tweet. This could initially be done by annotating the current tweets as debate fragments if large parts are out of context. Here a debate fragment tweet would refer to a tweet in a larger debate with other users and could then be used for future extraction and more expansive mining of ADUs.

| **Fine-tuned RoBERTa** | |
|---|---|
| **Information** | **Unit** |
| 1. Is the resulting model publicly available? | No |
| 2. How much time does the training of the final model take? | 105 Seconds |
| 3. How much time did all model experiments take (incl. hyperparameter search)? | 4228 seconds |
| 4. What was the energy consumption (GPU/CPU)? | 333 Watt |
| 5. At which geo location were the computations performed? | Denmark |
| 6. How much CO2eq was emitted to train the final model? | 0.975g |
| 7. How much CO2eq was emitted for all experiments? | 39g |

Table 7: Proposed climate performance model card for our fine-tuned RoBERTa model experiments.

One major difference between previous work and ours is data size: our dataset contains only 597 unique tweets annotated for 5 topics, while Schaefer and Stede (2021) annotated 3244 Facebook comments and Cheng et al. (2022) annotated nearly 70k sentences. Future experiments on a larger dataset may result in a different conclusion with respect to the relative performance of the models.

## 9 Conclusion

We defined an annotation scheme for an argument mining task tailored for social media with a focus on argumentation for sustainable nutrition. We proposed two new types of Evidence: Normative and Fact. With this scheme we scraped and annotated a dataset containing 597 tweets for five different topics, resulting in a dataset of 2985 annotated tweet-topic pairs. XGBoost is a strong starting point for argument mining, and IBM Project Debater API is a robust zero-shot model for argumentation tasks.

## 10 Broader Impact

Our dataset and models were designed with the intention to have positive impact on the environment by promoting sustainable consumer practices: by mining for convincing arguments of various aspects related to sustainable diets, downstream applications can improve marketing of sustainable products. Implementation of the resulting technol-

| RoBERTa embeddings + XGBoost | |
|---|---|
| **Information** | **Unit** |
| 1. Is the resulting model publicly available? | No |
| 2. How much time does the training of the final model take? | 57 Seconds |
| 3. How much time did all model experiments take (incl. hyperparameter search)? | 456 seconds |
| 4. What was the energy consumption (GPU/CPU)? | 28 Watt |
| 5. At which geo location were the computations performed? | Denmark |
| 6. How much CO2eq was emitted to train the final model? | 0.08g |
| 7. How much CO2eq was emitted for all experiments? | 3.5g |

Table 8: Proposed climate performance model card for our RoBERTa + XGBoost model experiments.

ogy will enable more effective communication campaigns to increase adherence with dietary guidelines. Furthermore, by identifying diverse arguments, our work can contribute to ethnographic research on public opinions towards sustainable diets, and help shape public policy. Promoting responsible behaviour is an important gap, as food marketing is already driven by business incentives. However, the risk of manipulative *dual use* must be considered. Future applications of this work must involve AI ethics experts and be complemented by explainability methods and fact verification to guarantee reliability of generated claims and ensure alignment with expected values.

Negative impact on the environment as a result of the development and any potential deployment of the models must be taken into account as well. Tables 7 and 8 contain the climate performance model card for the fine-tuned RoBERTa and RoBERTa + XGBoost models, according to the guidelines defined by Hershcovich et al. (2022).

### 10.1 Data Statement

The following is our data statement following Bender and Friedman (2018):

*A. CURATION RATIONALE*

In order to have a potential net positive impact on promoting sustainable diets in the future, a dataset with a focus on dietary discussions was needed. Twitter was deemed an excellent source for this information and as such scraping of 31840 tweets was done in combination with relevance filtering. This has resulted in 597 tweets that has been annotated for 4 different tasks, each done for 5 different topics in relation to discussions around diets.

*B. LANGUAGE VARIETY*

The tweets in this dataset where scraped in April 2022 with the Twitter API.[11] The set of English tweets was scraped without information of regional variety, it is only known that they are written in English. But certain tweets make specific wordings from which it can be inferred they are from the US (en-US) or India (en-IN). More regions are most likely also represented in the dataset, but specifics are unknown.

*C. SPEAKER DEMOGRAPHIC*

The authors of the tweets demographics were not collected. The tweets originate from 597 unique users.

*D. ANNOTATOR DEMOGRAPHIC*

The data was annotated by a crowd of annotators procured from Amazon Mechanical Turk. The specific region used was the US East Coast. There is no demographic information available from Amazon Mechanical Turk users beyond the requirements set for workers to be allowed to work on HITs—in the case of this dataset the only requirement is a masters qualification. Assuming we have an even distribution of the known demographics on Amazon Mechanical Turk, we would have a slightly skewed split between genders with 57% identifying as female. The age distribution is towards the younger ages with 29.7% being between 18-29 and 36.8% 30-39 and the majority identifying as white 79.9%.[12]

*E. SPEECH SITUATION*

The tweets can contain a maximum of 280 characters and are written in a spontaneous and asynchronous format. The tweets were collected with a focus on diet, but parts of the tweets also cover climate, sustainability, animal welfare and policy as side effects of our scraping methods and the topics used for relevance filtering. The majority of the tweets are in response to other Twitter users' tweets, so the intended audience would be one of the two opposing sides in a debate around one of the 5 topics in this paper.

*F. TEXT CHARACTERISTICS*

The tweets are only in raw text format as we filtered out any tweets containing URLs, images and other non textual modalities. Many of the tweets

[11] https://developer.twitter.com/en/docs/twitter-api

[12] More information on the demographics on Amazon Mechanical Turk can be found in https://www.cloudresearch.com/resources/blog/who-uses-amazon-mturk-2020-demographics/.

contain references to other users or users' tweets in a conversation format. Therefore, some tweets' context is limited without added work to include the references. There are also emojis and hashtags present in a large section of the tweets.

*G. RECORDING QUALITY*: N/A

*H. OTHER*: N/A

*I. PROVENANCE APPENDIX*: N/A

## References


Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, Baltimore, Maryland. Association for Computational Linguistics.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International Conference on Machine Learning*, pages 642–652. PMLR.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Giscard Biamby, Grace Luo, Trevor Darrell, and Anna Rohrbach. 2022. Twitter-COMMs: Detecting climate, COVID, and military multimodal misinformation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1530–1549, Seattle, United States. Association for Computational Linguistics.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2022. Cheap talk in corporate climate commitments: The role of active institutional ownership, signaling, materiality, and sentiment. (22-01).

Magnus Boström. 2020. The social life of mass and excess consumption. *Environmental Sociology*, 6(3):268–278.

Gullal S. Cheema, Sherzod Hakimov, Abdul Sittar, Eric Müller-Budack, Christian Otto, and Ralph Ewerth. 2022. Mm-claims: A dataset for multimodal claim detection in social media.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Liying Cheng, Lidong Bing, Ruidan He, Qian Yu, Yan Zhang, and Luo Si. 2022. Iam: A comprehensive and large-scale dataset for integrated argument mining tasks.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017. Argument mining on Twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322, Copenhagen, Denmark. Association for Computational Linguistics.

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. 2019. Corpus wide argument mining - a working solution. *CoRR*, abs/1911.10763.

João Graça, Cristina A. Godinho, and Monica Truninger. 2019. Reducing meat consumption and following plant-based diets: Current evidence and future directions to inform integrated transitions. *Trends in Food Science & Technology*, 91:380–390.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Daniel Hershcovich, Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. Towards climate awareness in NLP research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Lynn H Kaack, Priya L Donti, Emma Strubell, George Kamiya, Felix Creutzig, and David Rolnick. 2022. Aligning artificial intelligence with climate change mitigation. *Nature Climate Change*, pages 1–10.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Stewart Lockie. 2022. Mainstreaming climate change sociology.

Derk Loorbach. 2009. Transition management for sustainable development: A prescriptive, complexity□based governance framework. *Governance*, 23:161 – 183.

Richard Petty, John Cacioppo, and Rachel Goldman. 1981. Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology*, 41:847–855.

Adam A. Prag and Christian B. Henriksen. 2021. Correction: Prag, a.a.; henriksen, c.b. transition from animal-based to plant-based food production to reduce greenhouse gas emissions from agriculture—the case of denmark. sustainability 2020, 12, 8228. *Sustainability*, 13(2).

Lucas Høyberg Puvis de Chavannes, Mads Guldborg Kjeldgaard Kongsbak, Timmie Rantzau, and Leon Derczynski. 2021. Hyperparameter power impact in transformer language model training. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 96–118, Virtual. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. 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.

David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Körding, Carla P. Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer T. Chayes, and Yoshua Bengio. 2019. Tackling climate change with machine learning. *CoRR*, abs/1906.05433.

Robin Schaefer. 2021. Building an argument mining pipeline for tweets. in online handbook of argumentation for ai (ohaai) volume 2, 2021. *CoRR*, abs/2106.10832.

Robin Schaefer and Manfred Stede. 2020. Annotation and detection of arguments in tweets. In *Proceedings of the 7th Workshop on Argument Mining*, pages 53–58, Online. Association for Computational Linguistics.

Robin Schaefer and Manfred Stede. 2021. UPAppliedCL at GermEval 2021: Identifying fact-claiming and engaging Facebook comments using transformers. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 13–18, Duesseldorf, Germany. Association for Computational Linguistics.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green AI. *Communications of the ACM (CACM)*, 63(12):54–63.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, Ron Hoory, Yufang Hou, Shay Hummel, Michal Jacovi, Charles Jochim, Yoav Kantor, Yoav Katz, David Konopnicki, Zvi Kons, Lili Kotlerman, Dalia Krieger, Dan Lahav, Tamar Lavee, Ran Levy, Naftali Liberman, Yosi Mass, Amir Menczel, Shachar Mirkin, Guy Moshkowich, Shila Ofek-Koifman, Matan Orbach, Ella Rabinovich, Ruty Rinott, Slava Shechtman, Dafna Sheinwald, Eyal Shnarch, Ilya Shnayderman, Aya Soffer, Artem Spector, Benjamin Sznajder, Assaf Toledo, Orith Toledo-Ronen, Elad Venezian, and Ranit Aharonov. 2021. An autonomous debating system. *Nature*, 591(7850):379–384. Publisher Copyright: © 2021, The Author(s), under exclusive licence to Springer Nature Limited.

Manfred Stede and Ronny Patz. 2021. The climate change debate and natural language processing. pages 8–18.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Orith Toledo-Ronen, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim. 2020. Multilingual argument mining: Datasets and analysis. *CoRR*, abs/2010.06432.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to bm25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium*, ADCS '14, page 58–65, New York, NY, USA. Association for Computing Machinery.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Walter Willett, Johan Rockström, Brent Loken, Marco Springmann, Tim Lang, Sonja Vermeulen, Tara Garnett, David Tilman, Fabrice Declerck, Amanda Wood, Malin Jonell, Line Gordon, Jessica Fanzo, Corinna Hawkes, Rami Zurayk, Juan Rivera, Wim Vries, Lindiwe Sibanda, and Christopher Murray. 2019. Food in the anthropocene: the eat–lancet commission on healthy diets from sustainable food systems. *The Lancet*, 393.

Michael Wojatzki and Torsten Zesch. 2016. Stance-based argument mining – modeling implicit argumentation using stance.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

## A Crowdsourced Annotations

Each annotator was paid according to 15$ an hour of work. From our experience with annotation, we could complete roughly 100 total tweets + topics worth of annotation work in 45 mins for all four labels. Rounding it up to 60 mins and annotating for one label at a time, we calculated a pay of 0.045$ for each tweet + topic pair for each label.

For this paper, gathering annotations has happened over four annotations rounds, each focusing on one of the four primary labels we use in this paper. Five different annotators were recruited to calculate a majority for each annotated label. Each round helped bootstrap the data needed for annotation of the next round. For instance, we did not want to annotate non-argumentative data for claims or evidence as most previous annotators have already deemed it non-argumentative and would therefore be a waste of annotation resources. Instead, we would first retrieve annotations for the argumentative tweets. Then ask a new set of annotators to annotate for claims or evidence on the argumentative tweets. Due to evidence requiring its type annotated we also use the results from claims annotation round to help narrow the combination of topics and tweets used for evidence annotation. Pro/Con was also dependent on either claim or evidence being found in a tweet-topic pair, so was the last step in the annotation process.

Due to the subjective nature of annotating for this paper, we did not want to dismiss workers' work. Despite clear instructions, different people will consider claims relevant while others will not consider them relevant. Instead, we would actively moderate the resulting annotations and block any annotator creating low-quality annotations during annotation. We did this by first pre-annotating a small set and then calculating an overlap with annotators. If the overlap were small, we would block them from continuing. However, some annotations were slow to gather and would take multiple days. This resulted in us having to reopen hits that were partially annotated. Therefore if any annotator had already completed a set of hits, we would block them from redoing that set of hits. However, this method was imperfect, so that the same annotator might have double annotated some tweets.

We did test out an alternative method for part of the claim annotations where we would have a short test that would qualify annotators for the more extensive annotation set if they performed well. However, this method took much more time for annotations to be collected and was therefore dropped. It was also discovered during postprocessing of the hits that some annotations had less than five annotators, and others had more. This was only for a minority of hits, and it is believed that duplication's of a few tweets in the early corpus were the reason. This was fixed for later annotation rounds but should be noted as it might impact later results.

### A.1 Argumentative

Argumentative was the first label to be crowdsourced, we only gave annotators two options, "argumentative" and "not argumentative". Argumentative gets labeled as 0 for non-argumentative and 1 for argumentative.

**Instructions for annotators:** The task here is to annotate tweets if they are stated in an argumentative manner. Argumentative is a broad concept but

essentially means that the tweet either contains evidence or claims that would be relevant for a debate about some topic.

### A.2 Claim

We changed the annotation task from annotating for implicit/explicit claims for a specific topic for claim annotation. Instead, we asked annotators to select one or more topics where the claim would be relevant. They were asked to label a tweet relevant for one of the topics described earlier or mark it as irrelevant for all, or not containing a claim. The former option would be used to detect unrelated argumentative tweets. It is labeled as 0 for not containing a claim relevant to the topic and 1 for containing one that is.

This change was made for a few reasons. First, it reduced the number of hits needed 5-fold from 1935 hits needing to be made to only 387 hits. It also ensured that we had all tweets evaluated for all topics. Lastly, asking them to select the most relevant topics should give a more precise estimation of relatedness to a topic.

The downsides of this approach were that people were much more likely to select only one topic to be relevant rather than selecting two or more, even if a tweet was relevant.

The data for claims took three rounds of hit generation. Therefore this data might have some duplicate annotation work done.

**Instructions for annotators:** The task here is to annotate a tweet in relation to a set of topics. Here the tweet can contain a claim that might be relevant to any one of the topics. Of course, each tweet can be relevant for more than one of the topics, but it can also not be relevant for any one of the topics and should be annotated as such. Therefore, select the topics in which you find the tweet contains a claim relevant to an argument in a debate or communication campaign about the topic (regardless of your views on the claim and the topic and whether you would use it).

A claim is a standpoint toward a topic being discussed either directly or indirectly. The claim should be able to clearly be identified in a tweet on its own without relying on an assumption from the reader. This is an important issue for response tweets as the user might implicitly support a claim relevant to the topic or add a claim to a stance on the topic. Therefore, such tweets should not be annotated as containing a claim. The claim should also clearly have a positive or negative stance toward the discussed topic. Implicit claims are different from explicit ones as they lack the syntactic connection to the topic. This means they omit parts of the discussed topic or have no direct connection to it; instead, they indirectly express a stance towards it. An example of this could be a tweet, "Gardening has been great for my family and me! Can't wait to collect the bounties of this year's harvest," which contains an implicit claim with a clear stance toward T2 and T4. Suppose the tweet contains a claim clearly discussing a different topic unrelated to any of the other topics. It should then be labeled with the "unrelated or no claim label" If the tweet does not contain a claim at all, then it should also be marked with the "unrelated or no claim label."

### A.3 Evidence

Annotating evidence was done differently from claims. Since evidence is very nuanced and has many different types, we did not want to simplify annotating evidence the same way claims were simplified. This risked annotators relying too much on their own interpretation of what evidence over time. Therefore we wanted them to select what type of evidence was in a tweet concerning a topic. So each tweet needed its type of evidence annotated for every topic, but this would explode the number of annotations needed as explained with claims. Therefore, we decided to limit a tweet to the topics where claims were found relevant by just one annotator. This limits the amount of annotation work to the most likely relevant tweet-topic pairs while not limiting future annotation work to expand evidence annotation for topics where claims were not detected.

Therefore annotators are prompted to annotate a tweet-topic pair for any of the labels "Normative", "Study", "Expert", "Fact", "Anecdotal" or "Unrelated or no evidence". The Evidence label is labeled as 1 for containing relevant evidence and 0 for not.

The main downside to this annotation methodology is that it increases the likelihood of people annotating evidence as relevant to a topic since they might be more focused on its type regardless of relevance and instructions. However, with this method, we get a much more nuanced picture of the evidence contained within tweets which could be used for future modeling.

We considered an alternative method where annotators would first annotate for evidence types and

| Labels | Guidance |
|---|---|
| Argumentative | Select this if the tweet is making a clear self-contained claim. A claim is self-contained if the statement is clearly taking a stance towards some topic. Claims can be reactions towards a topic, like showing excitement or disgust towards a topic. The tweet is also argumentative if it contains evidence of some sort. Evidence can be citing a study, referencing an expert, or stating facts or beliefs. They don't necessarily have to be true. |
| Not Argumentative | A tweet is not argumentative if it is not clearly stating a self-contained claim. This could be because the stance of the claim is not clear, or the tweet does not clearly articulate a claim. Questions and irony or humor are automatically not argumentative and should be labeled as such. |

Table 9: Guidance for the individual labels

then annotate for relevance. However, this method was dropped as it would require an extra round of annotations, and it is hard to annotate evidence type without a clear topic to measure it after. For example, one tweet might contain anecdotal evidence for one topic but fact evidence for another.

**Instructions for annotators:** The task here is to annotate tweets related to a topic where you have to annotate what kind of evidence a tweet 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. There exist different types of evidence, and if a tweet contains any evidence, it should have the kind of evidence annotated. If more than one type of evidence exists in the tweet, choose the type you think best describes main piece of evidence in the tweet that is relevant for the topic. Be aware that the same tweet can show up multiple times and that each time it might have to be annotated differently for its evidence depending on the topic. Some tweets include various types of evidence where parts of the evidence are only relevant for one topic but not another. Therefore one tweet might have normative evidence for one topic but expert evidence for another and no evidence for a third. Remember, your goal is to annotate what type of evidence is in the tweet and if the evidence could be used in debate/argument or public communication both for or against the specified topic. Regardless of your views on the topic and whether the evidence is true or not.

### A.4 Pro/Con

Pro/con was the last label to be annotated. It gets 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. If there is no clear stance, the tweet's label for pro/con is set to $0$ and it should be reevaluated as a relevant tweet.

Due to its dependence on claim and evidence being present and relevant, we selected a subset of annotations if the majority thought there was either claim or evidence and the claim and evidence were relevant. This can accidentally remove some relevant tweets for annotation, but future work could annotate them.

To force people to choose the stance a tweet has for a topic, we removed the neutral option in annotation, so people have to annotate for pro or con. We believe that this should be fine due to the previous annotations, as the tweets left should have a clear stance on the topics they were relevant for.

**Instructions for annotators:** The task here is to annotate a tweet's stance in relation to a topic. The stance can be either one of pro or con. Here pro is a positive or supportive stance towards the topic, whereas con is a negative or hostile stance towards the topic. It is very important that you remember that it is the stance towards the topic and not the stance in the tweet itself.

## B Annotation Examples

### B.1 Processing annotations

After gathering crowdsourced annotations, we have a list of individual user annotations we have to merge. We do not want to merge the annotations

| Evidence type | Guidance |
|---|---|
| Anecdotal | A description of an episode(s), centered on individual(s) or clearly located in place and/or in time. |
| Expert | Testimony by a person, group, committee, organization with some known expertise / authority on the topic. |
| Study | Results of a quantitative analysis of data, given as numbers, or as conclusions |
| Fact | A known piece of information about the world without a clear source for the information |
| Normative | An added description for a belief about the world |
| Unrelated or no evidence | The tweet does contain evidence, but it is not related to the topic, or it does not have any evidence. |

Table 10: Evidence type annotator guidance.

| Tweet & Topic | A | C | E | PC | Comments |
|---|---|---|---|---|---|
| *Lol - and the wash post is the PR firm and Whole Foods is the official food supplier* | 0 | 0 | 0 | 0 | This tweet answers with a joke or irony towards another unknown tweet and is therefore not argumentative. |
| Topic: T5 (*We should pursue policies that promote sustainable foods*). Tweet: *It would also be nice if our government could begin subsidizing more sustainable options (like plant based meat) vs things like beef but... i digress* | 1 | 1 | Normative | 1 | Here the claim is that plant-based options should be actively pursued explicitly by policy and implicitly through the encouragement of alternatives and reduction in meat. It uses normative evidence to support its claim. |
| Topic: T2 (*Plant based food should be encouraged*). Tweet: *Green taxes go into subsidizing development and production of green energy solutions. If we were on 100% renewables, our electricity prices would not have needed to go up. We need to move into self-sufficient green energy as soon as possible* | 1 | 0 | 0 | 0 | This tweet contains both claims and examples of normative evidence but is unrelated to the topic and should therefore be annotated as unrelated. |
| Topic: T1 (*We should reduce the consumption of meat*). Tweet: *Yes but to be fair: we can expect a massive increase in meat and dairy consumption in emerging countries that will severely limit the impact of whatever we do.* | 1 | 1 | Normative | -1 | This tweet contains a belief that emerging countries will remove any progress we make and is therefore taking an opposing stance towards the topic. |

Table 11: Example annotations. A: Argumentative. C: Claim. E: Evidence (type). PC: Pro/con.

into binary labels as this throws away any uncertainty from the annotators. We, therefore, want instead to merge into a probability spectrum that defines the overall confidence of the annotators. Of course, each label does this slightly differently due to their unique annotation strategies.

For the argumentative label, we calculate the probability by summing the number of annotators believing the tweet to be argumentative. Then divide the sum by the number of annotators.

For claim, we sum each topic added as relevant for a tweet and divide that by the number of annotators. We also calculate the unrelated probability for the claim in the same way.

For evidence, we sum each type of evidence and use the max probability for evidence. We also save the evidence type distribution and the unrelated probability.

Lastly, for Pro/Con, we sum the number of pro labels and con labels, divide by the number of annotators, and select the label with the highest probability. Since con has to be a value of between -1 and 0, we have to flip its probability if it is the max likelihood.

This gives us the probability of a tweet being argumentative. We can then set the cutoff point for the argumentative tweets at 0.5 for the majority and use them for new annotations or modeling. We can also use the probabilities themselves for modeling.

When using the resulting data, one can extract binary labels by rounding to the nearest integer.

| | |
|---|---|
| Average tweet token count | 29.67 |
| Average claim token count | 31.63 |
| Average evidence token count | 34.59 |
| Average topic tweet vocab share | 2.8% |
| Average claim, topic tweet vocab share | 6.9% |
| Average evidence, topic tweet vocab share | 4.9% |
| Average claim tweet to tweet vocab overlap | 5.6% |
| Average evidence tweet to tweet vocab overlap | 5% |

Table 12: Overall tweet statistics for tweet token count for each type and percentage of vocab sharing between tweet and topic, and tweet to tweet.

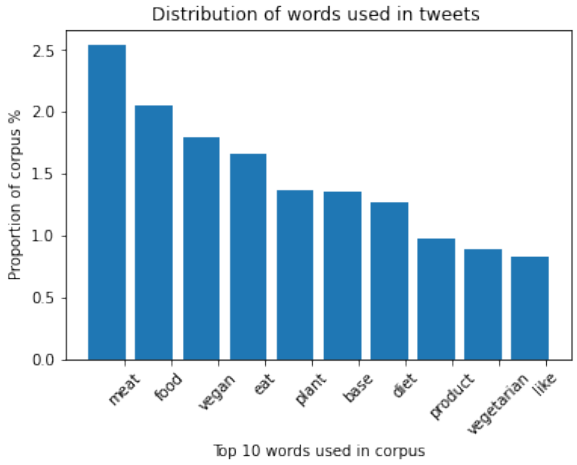


Figure 5: Top 10 words used corpus after stemming and removing stopwords from tweets

## C Statistics and Analysis

In Table 12 we have some general statistics regarding tweets and topics textual information. We see that claims and evidence have slightly more words than the average tweet. On the other hand, we see minimal vocabulary sharing between tweets and topics. This is probably because topics are quite short, while tweets are, on average, much longer. We see a more significant share of vocabulary for tweets containing claims and evidence in relation to their topics. However, tweets do not seem to share a large percentage of their vocabulary with each other, which shows the general difficulty for claim and evidence detection.

In Figure 5, we see the top 10 most used words in the corpus after having filtered out stopwords and stemmed the rest. Again, we see a general overlap with keywords from our topics, such as vegan, meat, and plant. Interestingly, ”plant” and ”base” almost occur the same amount, indicating a substantial usage of plant-based in tweets.

## D Tweet Retrieval Queries for Corpus Creation

English keywords: ”healthy food”, ”food”, ”green food”, ”veganism”, ”vegetable”, ”good recipe”, ”climate friendly recipe”, ”climate friendly diet”, ”healthy recipe”, ”sustainable diet”, ”green diet”, ”diet with vegetable”, ”vegetables are healthy”, ”fruit and vegetable”, ”fruit”, ”vegetarian”, ”vegan”, ”good vegan recipe”, ”good vegetarian recipe”, ”organic”, ”plant food is great”, ”fresh and organic is good”, ”varied and balanced diet”, ”beans”, ”sustainable meat”, ”legumes”, ”whole grains”, ”local farmers market”, ”plant based”, ”meat alternative”, ”plant based diet”, ”green food is really good”, ”animals are not ingredients”, ”eat healthy food”, ”raw food diet”, ”whole foods”, ”flexitarian”, ”raw foodism”, ”rawism”.

## E Experiment Replications

We tried to replicate some of the work of others to explore potential methods from which we would use for this paper. The two specific papers that are used for inspiration are both made by Schaefer and Stede (2020, 2021) .

### E.1 Fact-claiming & Engaging Comments

In Schaefer and Stede (2021) the data is 3244 German Facebook comments on a political talk show’s page from February 2019. The paper aims to classify toxic comments, engaging comments, and fact-claiming comments. They focus mainly on the fact-claiming comments due to its related nature to argument mining for evidence detection. They propose three models and two baseline models. The two baselines used are unigrams + SVM and Linguistic Features + XGBoost Chen and Guestrin (2016). The models they propose are:

- Fine-tuned BERT Embeddings + Transformer
- BERT Embeddings + Transformer
- BERT Embeddings + XGBoost

They don’t detail the implementation of the extra transformer layer on top of BERT. We assume this is a single layer added on top, followed by a liner classification layer. For the rest of the models, none of the hyperparameters are described for any of the

models. Instead, they explain that they used a development set for hyper parameter tuning for the models. This development set was created from 12.5% of the given training data. Another 12.5% was taken for a test set used to give them preliminary results. For the final evaluation they where given a new dataset of 944 unlabeled comments which where drawn from discussions of different show to avoid topical bias.

To replicate the results of Schaefer and Stede (2021), we use huggingface, Wolf et al. (2019), framework to fine-tune 2 BERT models of bert-base-german-cased[13], each focused on either subtask one or subtask two. The model is fine-tuned for 75% of the training set for one epoch. The optimizer used is Adam, with a learning rate of 5e-5 and no weight decay. The rest of the hyperparameters are left to the default setup of the TrainingArguments for huggingface's models. The models are trained on a binary classification task, which is done by loading in the BERT model as an AutoModelForSequenceClassification with two labels and fine-tuning it. Results from our replication and the original paper can be found in Table 13.
Our attempt at replicating the results are successful as we manage to get similar scores as reported (Schaefer and Stede, 2021) and exceeding them slightly in certain areas. Our results could probably be improved if we used some hyper-parameter search with the left over 25% of the training data. This experiment shows the advantage of using large language models as the base for further model experimentation.

### E.2 Climate Tweets

Schaefer and Stede (2020) focus on creating a new Twitter-based dataset. The dataset contains 300 labeled German tweets containing the word "klima" (climate). The tweets where annotated for three labels, those being argument, claim and evidence. Part of the paper then explores a modeling approach to evaluate the viability of this dataset on a set of models. They use XGBoost as their primary model, with the main difference being the features it is trained on for the different models. The features used are:

- Bigrams
- Pretrained BERT Embeddings
- Uni & Bigrams
- Linguistic & Twitter Features

Unfortunately, they don't report the hyperparameters used by any of the models in the paper. They train each model to do binary classification for one of three targets: argumentative, claim detection, and evidence detection. They report their results with F1 macro weighted, precision and recall. To replicate the results of Schaefer and Stede (2020) we use a similar setup as explained. We use flair as the framework Akbik et al. (2019) to generate Pretrained BERT Embeddings (Akbik et al., 2018) using bert-base-german-cased. We then use an XGBoost model that is trained on the embeddings [14]. Finally, we use grid search to optimize the hyperparameters over the dataset by doing three-fold cross-validation. The final hyperparameters used are 15 estimators with a max depth of 1 and a learning rate of 0.01. The rest are the default values used by XGBRFClassifier. When generating the results, we use 10 fold cross-validation as described in the paper. The data contains labeled tweets from two different annotations hence fourth expert 1 and expert 2, in their paper they don't describe which of these labels they use or if they combined them somehow, therefore we did the experiment with both set of annotations. Their annotations don't agree and their Cohen's Kappa inter annotator agreements are 0.53 for argumentative, 0.55 for claim and 0.44 for evidence. Results from our replication and the original paper can be found in Table 14.
Due to them not being allowed to share their raw tweets we had to fetch the original tweets from their id, which results in a loss of tweets due to the original being deleted. We therefore only had 212 tweets vs the original 300 for our model to train and evaluate on. We did check if any major imbalances had occurred compared to the original dataset, and found no major changes in the balance of the tweets. We therefore where training our models under similar conditions to the original authors with the only difference being size of data. This difference might have impacted the result's in our replication process, but as we get very similar results compared to the original paper, this impact is probably minimal. We see that for evidence we have a large difference in the results, which should be expected as this is where the annotators disagree the most in their labeling, with expert 2's annotations being the easiest

[13] https://huggingface.co/dbmdz/bert-base-german-cased

[14] https://github.com/dmlc/xgboost

| Approach | Subtask (ST) 2 | | | Subtask (ST) 3 | | |
|---|---|---|---|---|---|---|
| | F1 | Precision | Recall | F1 | Precision | Recall |
| Unigram SVM (ST 2)/LR (ST 3) | **0.671** | 0.665 | **0.688** | 0.654 | 0.667 | **0.688** |
| Linguistic Features XGBoost (ST 2)/RF (ST 3) | 0.670 | **0.681** | 0.664 | **0.693** | **0.710** | 0.685 |
| BERT Emb (FT) Transformer | **0.689** | 0.708 | **0.672** | 0.736 | 0.740 | 0.732 |
| BERT Emb Transformer | 0.669 | 0.701 | 0.640 | 0.722 | **0.758** | 0.690 |
| BERT Emb XGBoost | 0.669 | 0.685 | 0.654 | 0.717 | 0.736 | 0.698 |
| BERT (FT) Classification (Replication attempt) | 0.681 | **0.717** | 0.648 | **0.745** | 0.752 | **0.737** |

Table 13: Evaluation results from Schaefer and Stede (2021) and our replication. Emb: Embeddings. FT: fine-tuned.

| Features | Preproc | F | P | R |
|---|---|---|---|---|
| **Argumentative** | | | | |
| Bigrams | 1, p, s | 0.8 | 0.75 | 0.86 |
| BERT | p | 0.82 | 0.8 | 0.86 |
| Ours (expert 1) | | 0.83 | **0.89** | **0.98** |
| Ours (expert 2) | | **0.84** | **0.89** | **0.98** |
| **Claim** | | | | |
| Uni- & Bigrams | 1, p | 0.79 | 0.78 | 0.82 |
| BERT | p | **0.82** | 0.8 | 0.85 |
| Ours (expert 1) | | 0.80 | **0.87** | 0.97 |
| Ours (expert 2) | | **0.82** | **0.87** | **0.98** |
| **Evidence** | | | | |
| Uni- & Bigrams | 1, p | **0.67** | 0.68 | 0.68 |
| BERT | p, s | 0.59 | 0.59 | 0.62 |
| Ours (expert 1) | | 0.61 | 0.66 | 0.43 |
| Ours (expert 2) | | **0.67** | **0.69** | **0.78** |

Table 14: 10-fold cross validation results from Schaefer and Stede (2020) and our replication. F: weighted F1 score. P: weighted precision. R: weighted recall. l: lowercase. p: punctuation. s: stopword.

for the model to learn.