# BERTweet's TACO Fiesta: Contrasting Flavors On The Path Of Inference And Information-Driven Argument Mining On Twitter

**Marc Feger**
HeiCAD - Heine Center for Artificial Intelligence and Data Science
Düsseldorf Germany
marc.feger@hhu.de

**Stefan Dietze**
GESIS - Leibniz Institute for the Social Sciences
Cologne Germany
stefan.dietze@gesis.org

## Abstract

Argument mining, dealing with the classification of text based on inference and information, denotes a challenging analytical task in the rich context of Twitter (now 𝕏), a key platform for online discourse and exchange. Thereby, Twitter offers a diverse repository of short messages bearing on both of these elements. For text classification, transformer approaches, particularly BERT, offer state-of-the-art solutions. Our study delves into optimizing the embeddings of the understudied BERTweet transformer for argument mining on Twitter and broader generalization across topics. We explore the impact of pre-classification fine-tuning by aligning similar manifestations of inference and information while contrasting dissimilar instances. Using the TACO dataset, our approach augments tweets for optimizing BERTweet in a Siamese network, strongly improving classification and cross-topic generalization compared to standard methods. Overall, we contribute the transformer WRAPresentations and classifier WRAP, scoring 86.62% F1 for inference detection, 86.30% for information recognition, and 75.29% across four combinations of these elements, to enhance inference and information-driven argument mining on Twitter.

## 1 Introduction

Twitter (now 𝕏) is a global hub for opinions, news, and information and serves as a primary data source for research, which had already recognized the value of its user-generated content prior to its transition to 𝕏 (Kwak et al., 2010; Boyd et al., 2010).

Argument mining describes the process of classifying texts by assessing their written content in terms of information and inference elements to identify arguments (Palau and Moens, 2009; Peldszus and Stede, 2013; Lawrence and Reed, 2019).

In the intersection of traditional machine learning and natural language processing, pre-trained transformers like BERT (Devlin et al., 2019) and its specialized variants, such as BERTweet (Nguyen et al., 2020), have set state-of-the-art classification standards (Houlsby et al., 2019; Sun et al., 2019). During fine-tuning, transformers create universal text representations providing contextual features for a soft-max classifier, meaning additional layers on top of the pre-trained model that are jointly optimized for downstream tasks (Devlin et al., 2019).

Thereby, the field of argument mining has also witnessed the benefits of transformer models like BERT for cross-topic classification (Bhatti et al., 2021; Thorn Jakobsen et al., 2021) and argument similarity (Reimers and Gurevych, 2019; Reimers et al., 2019; Thakur et al., 2021).

Besides the common methods of adjusting the in-task performance through parameter tweaks (Lan et al., 2019; You et al., 2019) or incorporating augmentations (Feng et al., 2021; Thakur et al., 2021), multi-task learning is recommended as an additional fine-tuning strategy (Sun et al., 2019; Stab et al., 2018). Thereby, multi-task learning denotes a prior phase of fine-tuning representations on auxiliary tasks such as clustering or semantic similarity before proceeding to the actual classification step and is argued to effectively reduce a model's sensitivity to spurious correlations (Liu et al., 2019; Tu et al., 2020), which in turn is key to cross-topic argument mining (Thorn Jakobsen et al., 2021).

We believe that acquiring robust and meaningful representations, in the sense of perceiving the constituent elements of arguments, prior to classification is particularly useful for the nuanced task of argument mining when applied to diverse topics.

Generalizability in terms of cross-topic classification is crucial for practical argument mining in realistic scenarios, both in general research (Daxenberger et al., 2017; Stab et al., 2018) and specifically on Twitter (Schaefer and Stede, 2021), necessitating models to focus on argument components while avoiding reliance on spurious correlations like topic words (Thorn Jakobsen et al., 2021).

*Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2256–2266
June 16-21, 2024 ©2024 Association for Computational Linguistics

In this paper, we pioneer the optimization of the understudied transformer BERTweet for argument mining on Twitter. Thereby, we refine its representations of tweets within the embedding space by specializing BERTweet to better encode inference and information across diverse topics.

Utilizing the TACO dataset (Feger and Dietze, 2024), offering the first strong baseline evaluations of BERTweet for argument mining on Twitter, we optimize the model's representation layers in a multi-task approach by accentuating the contrast between inference and information while centering similar manifestations before the actual classification step, for which we assume proximity to imply shared class signals (van Engelen and Hoos, 2020).

We achieve this by configuring a Siamese BERTweet network using SBERT (Reimers and Gurevych, 2019). Applying contrastive loss (Hadsell et al., 2006) and text augmentation techniques (Wei and Zou, 2019), this network teaches BERTweet to cluster tweet embeddings according to their respective roles in argument mining, that is, to generally encode the presence or absence of both inference and information in those representations used for classification. Hence, we aim for classifications driven by the argument constituting elements, steering clear of spurious correlations.

Utilizing BERTweet's enhanced embeddings, it excels in both closed and cross-topic argument mining on Twitter, outperforming several standard methods (Schaefer and Stede, 2021) in this domain.

Towards inference and information-driven argument mining on Twitter, we contribute:[1]

- A pre-classification fine-tuning approach for BERTweet, enhancing its capacity to represent information and inference for closed and cross-topic argument mining on Twitter.
- An augmentation strategy to reduce spurious entity and topic signals while increasing sentence variability in tweets.
- WRAPresentations[2], an enhanced BERTweet embedding model driven by inference and information, obtained through contrastive optimization on augmented TACO tweets.
- WRAP[3], our tweet argument classifier leveraging WRAPresentations for argument mining across diverse topics on Twitter.

## 2 Twitter Arguments from Conversations

Our primary dataset[4], TACO (Feger and Dietze, 2024), encompasses 1,734 tweets from 200 entire conversations spanning six topics: #Abortion (25.9%), #Brexit (29.0%), #GOT (11.0%), #LOTRROP (12.1%), #SquidGame (12.7%), and #TwitterTakeover (9.3%). So far, it stands as the sole publicly available labeled tweet dataset tailored for inference and information extraction, strategically addressing reply-patterns inherent to their emerging conversational contexts during annotation.

Annotations were conducted by six experts according to the Cambridge Dictionary definitions, differentiating *inference* as *a guess that you make or an opinion that you form based on the information that you have* and *information* as *facts or details about a person, company, product, etc*. With a robust agreement of 0.718 Krippendorff's $\alpha$, four classes emerged of these elements: *Reason* (inference and information), *Statement* (inference without information), *Notification* (information without inference), and *None* (neither element).

Table 1 details the class distribution of TACO.

| **Reason** | **Statement** | **Notification** | **None** |
|---|---|---|---|
| 581 (33.50%) | 284 (16.38%) | 500 (28.84%) | 369 (21.28%) |

Table 1: The class distribution of tweets in TACO.

On TACO, Vanilla BERTweet serves as the best performing baseline, excelling with 74.45% F1 for Reason, 56.66% F1 for Statement, 78.30% F1 for Notification, and 80.56% F1 for None after fine-tuning on these classes (Feger and Dietze, 2024).

## 3 Inference and Information-Driven Representations for Mining Arguments

In text classification, transformers like BERTweet use the final hidden state of the first token $[CLS]$ as the sequence representation. Classification involves a soft-max classifier added as an extension after the final representation layer, determining the label assignment for a tweet $t$ by evaluating the probability of each possible label $y$ as:

$$p(y|h) = softmax(\hat{W}h) \qquad (1)$$

where, $\hat{W}$ signifies the task-specific weights of the classification head, and $h$ represents the final representation of $t$ obtained with the transformer. Achieved through pooling an entire sequence representation via $[CLS]$, $h$ is expressed as

[1] github.com/TomatenMarc/TACO-Fiesta

[2] huggingface.co/TomatenMarc/WRAPresentations

[3] huggingface.co/TomatenMarc/WRAP

[4] github.com/TomatenMarc/TACO

$G_W(t) = h$, where the transformer is considered an independent function $G_W(t)$ with its distinct weights $W$, taking $t$ as input. For the specific classification task, both $\hat{W}$ and $W$ are jointly fine-tuned by maximizing the log-probability of the correct label, where $h$ implicitly undergoes optimization.

For optimizing class assignments on TACO, we emphasize the impact of specializing $h$ for encoding inference and information before classification.

Hence, we consider the pre-classification specialization of an embedding $h$ as a contrastive problem of semantic similarity, where tweets with similar expressions of the text dimensions inference and information are brought closer together, while those lacking in similarity are positioned farther apart.

### 3.1 Embedding Inference and Information

We measure the semantic similarity between two tweet representations, denoted as $h_1$ and $h_2$, using cosine distance:

$$D(h_1, h_2) = 1 - cos(h_1, h_2) \in [0, 2] \qquad (2)$$

a standard metric (Mikolov et al., 2013; Kim, 2014; Tai et al., 2015; Chen and He, 2020) for assessing text vector similarity. $D(h_1, h_2)$ reflects complete equivalence at 0, orthogonality at 1, and absolute dissimilarity at 2. Mainly defined by the cosine similarity $cos(h_1, h_2) \in [-1, 1]$, where $-1$ represents complete dissimilarity, 1 indicates equivalence, and values closer to 0 suggest orthogonality, this distance is length-independent and primarily influenced by the angle between two embeddings.

Building on this circumstance, we assume that the actual representation $h$ of a tweet can be normalized and lies on the $n$-sphere:

$$S(n) = \{h \in \mathbb{R}^{n+1} : \|h\| = 1\} \qquad (3)$$

Transferred to the Cartesian nature of arguments $h = \langle information, inference \rangle$, we consider their representations to live on the unit sphere $h \in S(1)$ (Wang and Isola, 2020; Khosla et al., 2020; Chen and He, 2020). In $h$, 1 signifies full presence, and $-1$ implies total absence of a component. Consequently, an ideal class center on the unit sphere heads towards the pole $\langle 1, 1 \rangle$ for Reason, $\langle -1, 1 \rangle$ for Statement, $\langle 1, -1 \rangle$ for Notification, and $\langle -1, -1 \rangle$ for None. A breakdown of this is shown in the upper part of Figure 1, acknowledging the realistic expectation that the actual embeddings may differ from the ideals while the objective is to get them closer to them.

### 3.2 Contrastive Siamese Network

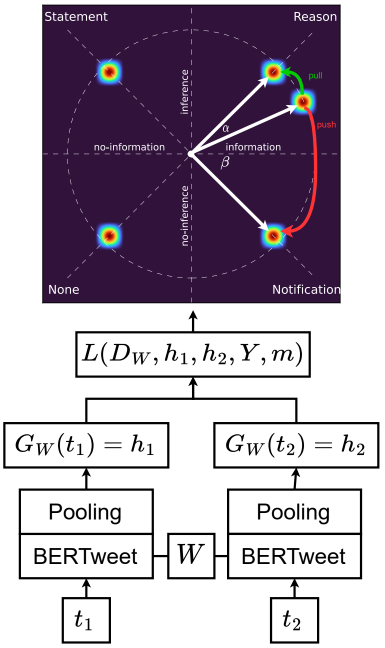


Figure 1: Visualization of the employed Siamese BERTweet architecture, with parameterized cosine distance $D_W(h_1, h_2)$ and contrastive loss $L(D_W, h_1, h_2, Y, m)$. Atop this architecture, the Cartesian embedding space for an argument representation $h = \langle information, inference \rangle$ is presented as target.

To address semantic similarity, a prevalent strategy involves enhancing representations through learning a metric (Chopra et al., 2005; Xing et al., 2002; Hadsell et al., 2006). Precisely, metric learning entails the implicit acquisition of a metric $D_W(h_1, h_2)$ parameterized by the weights $W$ of the representation model $G_W$ (Chopra et al., 2005). We seek to find $W$ such that the target metric:

$$D_W(t_1, t_2) = 1 - cos(G_W(t_1), G_W(t_2)) \qquad (4)$$

is smaller if $t_1, t_2$ are semantically similar, and higher if not.

By utilizing the identical embedding function $G_W(t)$ (BERTweet) with shared weights $W$ to learn the metric, our architecture is referred to as a Siamese network (Bromley et al., 1993; Chopra et al., 2005). Similar and dissimilar tweet pairs are provided as input to this network. To update the weights and optimize the network's performance, a loss function is applied on top of this architecture.

To attain the goal of increasing the differentiation between similar and dissimilar pairs, it is suggested to employ the contrastive loss (Chopra et al., 2005; Hadsell et al., 2006):

$$
\begin{aligned}
L(D_W, & h_1, h_2, Y, m) = \\
& (Y)\frac{1}{2}D_W(h_1, h_2)^2 + \qquad (5) \\
& (1-Y)\frac{1}{2}\{max(0, m - D_W(h_1, h_2))\}^2
\end{aligned}
$$

where, $h_1, h_2$ are two representations ($G_W(t_i) = h_i$) of different tweets $t_1, t_2$ to be optimized given $D_W(h_1, h_2)$ as metric. $Y$ denotes the binary label indicating if $t_1, t_2$ are similar ($Y = 1$) or contrasting ($Y = 0$). Furthermore, a margin value $m > 0$ is introduced as the minimal distance between two contrasting tweets.

When establishing $m$, our objective was to set $D_W(h_1, h_2)$ in a way that maximizes contrast between dissimilar pairs while avoiding overestimation of their true distance. Focusing on $D_W(h_1, h_2) \in [0, 1]$, representing positive similarity, we selected $m = 0.5$. This choice intuitively represents the minimum threshold for high similarity, yielding optimal results in our study.

With $m = 0.5$ we ensure that even if a representation closely matches an ideal center but is labeled as dissimilar, the optimized representation pushes $60°$ away and into an adjacent quadrant.

### 3.3 Augmentation of TACO

In the initial phase of processing TACO data, we generated a unique copy for each tweet through augmentation, denoted as A-TACO. Employing EDA (Easy Data Augmentation) techniques (Wei and Zou, 2019) of (1) synonym replacement, random (2) insertion, (3) swap, and (4) deletion, this procedure segregates our total ground truth into A-TACO, for optimization the embedding space of BERTweet prior to classification, and TACO, designated for fine-tuning and evaluating classifiers.

Maintaining independence between optimization and evaluation data is crucial to avoid further spurious correlations (Thorn Jakobsen et al., 2021) and ensure that the data includes essential class signals, thus enabling broad generalization across varying sentence structures and cross-topic evaluations.

Following technique (1), we utilized spaCy[5] to automatically identify as many entities and preselected keywords related to the six topics in the TACO dataset as possible. Subsequently, we replaced these words with the $[MASK]$ token, a placeholder commonly used by BERT-like models, including BERTweet, for predicting missing words. Particularly, we utilized BERTweet as a fill-mask model to generate new tokens for those masked in the input sequence (Kumar et al., 2020).

In order to increase the variability of word choice and sentence structure while minimizing semantic changes, the techniques (2-4) were applied to 10-90% of all words. Optimal coherence, with an average cosine distance of $\sim 0.08$ between the $[CLS]$ tokens of tweets and augmentations, is seen at a replacement rate of 10%, maintaining semantic consistency with entity and topic words being almost entirely changed or removed. Again, step (1) was applied to avoid reintroducing topic words. Refer to Table 2 for an augmentation example.[6]

| | |
|---|---|
| TACO | **Elon Musk** ready with 'Plan B' **if Twitter** rejects his offer **Read @USER Story \| HTTPURL #ElonMusk #ElonMuskTwitter #TwitterTakeover HTTPURL** |
| A-TACO | **Wenger** ready with 'Plan B' **as Wenger** rejects his offer **- HTTPURL via @USER** |

Table 2: An augmented Notification reminiscent of a general blog comment after replacing entities (Elon Musk and Twitter are changed to Wenger), deleting topic or entity references, including hashtags, and rewording the tweet while retaining its original substance.

## 4 Experimental Setup

This section outlines the protocols used for evaluating and optimizing BERTweet's embedding space with A-TACO and follow-up classification on TACO. We select macro F1 scores[7] for evaluation in response to the imbalanced distribution across TACO's four classes, guaranteeing an equitable analysis and underscoring a model's adeptness at managing heterogeneous data distributions. In our subsequent classification analysis, we also present the micro F1 scores[7] for each tweet class. Beyond this, we consider Recall to account for the generalizability of a model to unknown topics after fine-tuning in the pre-classification phase.

### 4.1 Models

In our approach, it is important to differentiate between the pre-classification fine-tuning for specializing embeddings and their subsequent fine-tuning tailored for mining arguments on TACO. In this context, we compare different ablations of our fine-tuning pipeline for embeddings before and upon classification, comparing their prediction strength with various common baseline models.

[5] spacy.io

[6] For more examples, see: README.md

[7] Precision and Recall for experiments are in the repository.

For both tasks, we utilize the Vanilla BERTweet model[8], with 12 transformer blocks and 12 self-attention heads processing sequences of up to 128 tokens, consistent with the best performing model reported for TACO (Feger and Dietze, 2024).

The first embedding model derived from Vanilla BERTweet, enhanced as described in Section 3 by applying contrastive loss within the Siamese network utilizing A-TACO to improve the cosine distance $D_W(t_1, t_2)$ for similar or dissimilar tweets, is referred to as WRAPresentations. For comparison, we introduce a second derivative, Augmented BERTweet, which undergoes pre-classification fine-tuning using the same tweets of A-TACO as WRAPresentations but directly optimizes $p(y|h)$ with standard cross-entropy loss.

Both these strategies aim to improve the representation $G_W(t) = h$ of any tweet $t$ used for subsequent classification $p(y|h)$ on TACO by incorporating augmented tweets of A-TACO and adjusting the internal weights $W$ in different ways to better encode argument components for each model $G_W$.

For classification on TACO, we utilize TF-IDF representations, where word frequency is widely recognized as a feature in strong baselines for argument mining on Twitter, which are Support Vector Machine (SVM) (Addawood and Bashir, 2016), Logistic Regression (LR) (Bosc et al., 2016; Dusmanu et al., 2017), and Random Forest (RF) (Dusmanu et al., 2017). These models go beyond considering individual words by incorporating tweet-related features like emoji, URL, and hashtag frequencies. Despite this, their potential for cross-topic generalizability remains unexplored.

For each classifier, we evaluate the average class length for classification to examine linguistic feature acquisition.

### 4.2 Pre-Classification Fine-Tuning

To enhance BERTweet's embeddings, we chose TACO's golden tweets with flawless annotation agreement, accounting for 70.3% of all tweets, with class distribution remaining largely consistent.

For the final evaluation, we employed the original golden tweets for #Abortion but augmentations of golden tweets for the remaining five topics during fine-tuning. #Abortion was deemed as holdout topic due to its highest dissimilarity when compared to the remaining topics, posing a greater classification challenge (Thorn Jakobsen et al., 2021). This provided initial insights into cross-topic generalization and the efficacy of fine-tuning with augmentations and predicting given real tweets. Pairs were formed for all tweet combinations, denoting tweets of the same class as similar $Y = 1$ and those of different classes as dissimilar $Y = 0$, yielding more dissimilar than similar pairs.

For the final validation set, 86,142 pairs were generated. The optimization data, divided into fine-tuning and test sets with a stratified $60/40$ ratio, yielded 307,470 and 136,530 candidate pairs, respectively. To ensure a balance between similar and dissimilar pairs, we chose the largest possible set such that both similar and dissimilar pairs are equally represented (Bromley et al., 1993; Chopra et al., 2005) while maintaining all tweets of the respective splits.

In total, 162,064 pairs were obtained for fine-tuning, 71,812 for testing, and 53,560 for final validation of the enhanced BERTweet representations prior to classification.

For all transformer models, we performed fine-tuning over $5$ epochs using an A100 GPU with 40 GB of memory, a batch size of 32, and a learning rate of $4e^{-5}$, which proved to be optimal for all models. The Siamese BERTweet network is implemented using SBERT (Reimers and Gurevych, 2019) as depicted in the lower part of Figure 1.

Additionally, we applied different fine-tuning strategies for WRAPresentations using both $[CLS]$ pooling, later used for classification, and $[MEAN]$ pooling, recommended for better sentence embeddings (Reimers and Gurevych, 2019).

### 4.3 Argument Mining on TACO

We evaluate the practicality of BERTweet's specialized embeddings on TACO, given the three argument mining tasks of (1) inference detection, (2) information recognition, and (3) classification of all four tweet classes, with a concurrent aim for cross-topic generalization.

For task (3), we trained a feed-forward neural network with two linear layers on top of each embedding model, undergoing $5$ additional fine-tuning epochs with the best performing parameters having a learning rate of $4e^{-5}$ and batch size of $8$, corresponding to the best model and parameters reported for TACO (Feger and Dietze, 2024). Again, we used a single A100 GPU with 40 GB of memory. Thereby, the results for tasks (1) and (2) are aggregations specific to class elements of task (3) predictions, focusing on inference or information.

[8] huggingface.co/vinai/bertweet-base

Extending our ablation strategy, classifiers were evaluated in two different setups to investigate the general effects of fine-tuning embeddings prior to classification and their subsequent adaptability to actual class signals (Peters et al., 2019).

In the first setup (Frozen), freezing embeddings allowed us to assess the benefits attributable to pre-classification fine-tuning. In the second setup (Dynamic), embeddings underwent further fine-tuning during classification head optimization, where we assessed their adaptability to task-specific learning. Success in both setups signifies a model's ability to represent argument components prior to classification and to adapt these fine-tuned representations to the specific classes of inferences and information.

We employed a 6-fold shuffled cross-validation, maintaining consistent splits for all classifiers across the six topics of TACO, to establish an upper-bound (Thorn Jakobsen et al., 2021). This closed-topic validation was then compared with cross-topic validation, where each of the six topics served as a unique testing set, and the remaining five topics were utilized for fine-tuning (Bosc et al., 2016; Daxenberger et al., 2017; Stab et al., 2018). Lower performance is expected in cross-topic validation, as classifiers are exposed to unseen topics.

# 5 Results

In this section, each model is investigated with respect to the actual tweets of TACO. First, we assess the embeddings of each transformation model in terms of their baseline notion of argument components and in terms of the four tweet classes, focusing on the structural differences of their representations. Second, we evaluate the different models in both closed and cross-topic classifications to determine their applicability to, and generalizability across, topics.

## 5.1 Results: Pre-Classification Fine-Tuning

| Model | P | R | F1 |
|---|---|---|---|
| Vanilla BERTweet-$[CLS]$ | 50.00 | 100.00 | 66.67 |
| Augmented BERTweet-$[CLS]$ | 65.69 | 86.66 | **74.73** |
| WRAPresentations-$[CLS]$ | **66.00** | 84.32 | 74.04 |
| WRAPresentations-$[MEAN]$ | 63.05 | **88.91** | 73.78 |

Table 3: Evaluation of within-class similarity and between-class separability of all transformer models using $[CLS]$ tokens as used during classification. These models were fine-tuned with A-TACO and evaluated on the TACO holdout topic #Abortion. Suffixes indicate pooling methods for optimizing the embedding spaces.

After pre-classification fine-tuning to enhance semantic similarity, we evaluate the optimized embedding models for classifying tweet pairs as similar or dissimilar given $D_W(t_1, t_2)$.

All fine-tuning strategies outperformed Vanilla BERTweet in terms of F1, compare Table 3.

We excluded WRAPresentations with $[CLS]$ pooling for follow-up classification due to the absence of discernible benefits in F1 compared to Augmented BERTweet and WRAPresentations using $[MEAN]$ pooling for pre-classification fine-tuning, each showing higher Recall scores.

Hence, we will refer to WRAPresentations-$[MEAN]$ as WRAPresentations.

In comparing Augmented BERTweet and WRAPresentations, both models show similar overall performance in terms of F1, but diverge in their emphasis on Precision and Recall. The results suggest that contrastive fine-tuning of representations is not inherently superior to directly optimizing $p(y|h)$ with augmented tweets. However, this strategy enhances Recall, with further distinctions expected in downstream task evaluations.

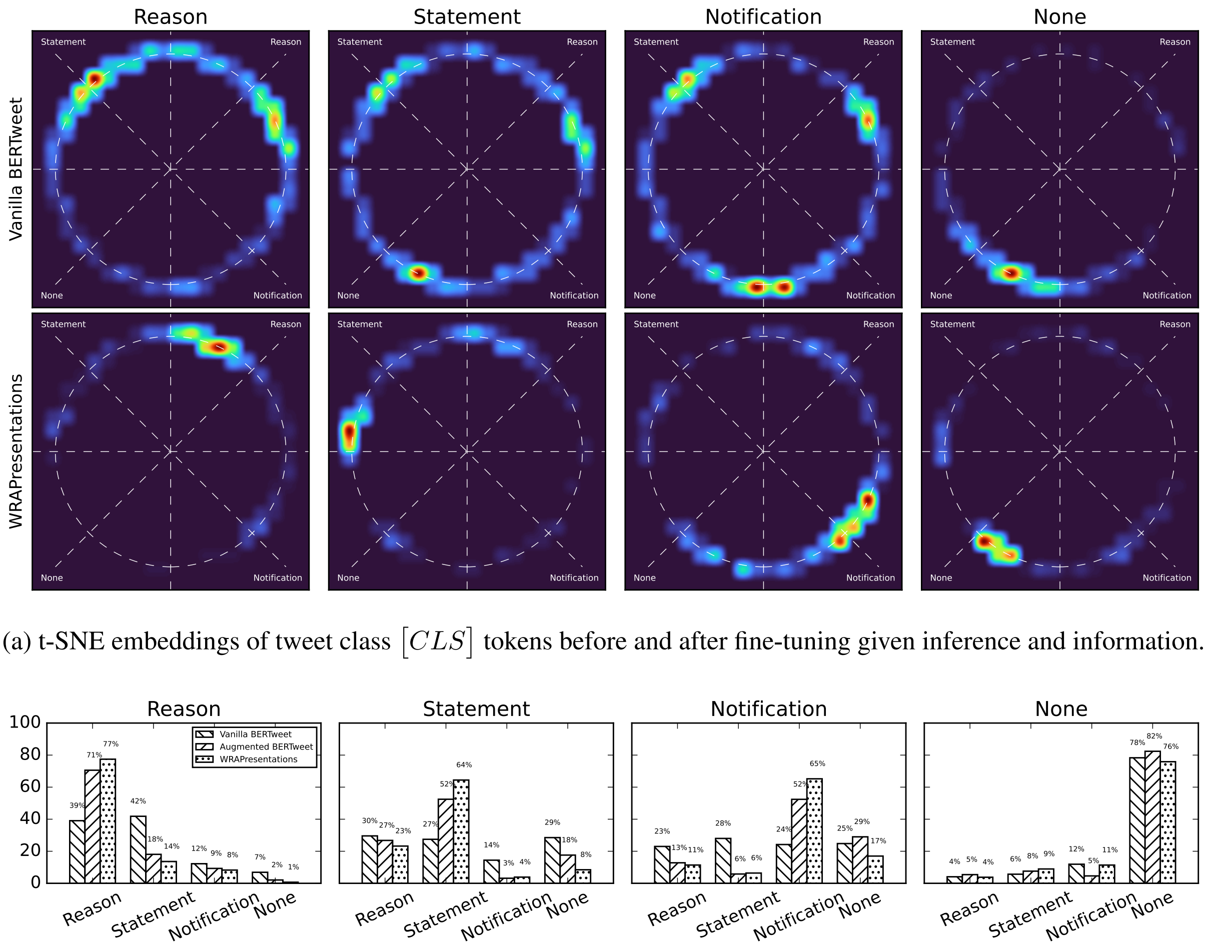
Nonetheless, we assume that the enhanced Recall at this stage is already a first indicator for later generalizations of classifications across topics. Additionally, we confirmed the effectiveness of pre-classification fine-tuning with A-TACO when applied to real tweets from an unseen topic.

Furthermore, we visually explored BERTweet's embedding space before and after fine-tuning, utilizing $[CLS]$ representations of all original tweets in TACO, as depicted in Figure 2(a).

Applying t-SNE for dimensional reduction (van der Maaten and Hinton, 2008; Jawahar et al., 2019), comparing Vanilla BERTweet with WRAPresentations showed enhanced class quadrant density, compare Figure 2(a), suggesting an improvement of class semantics given inference and information for a majority of tweets. Similar patterns, albeit at smaller numbers, are observed for Augmented BERTweet, see Figure 2(b).

Numerically, WRAPresentations improved tweet order by 38% for Reason, 37% for Statement, and 41% for Notification over Vanilla BERTweet. Despite a -2% decrease in the None class quadrant, None remains predominant, refer to Figure 2(b).

Augmented BERTweet closely matches WRAPresentations in representing tweets, excelling by 6% for None but lagging behind by -6% for Reason, -12% for Statement and -13% for Notification.



(a) t-SNE embeddings of tweet class $[CLS]$ tokens before and after fine-tuning given inference and information.

(b) Distribution of classes within the projected quadrants of the expected $\langle information, inference \rangle$ space.

Figure 2: Investigation on the impact of BERTweet's fine-tuning for the transfer of class semantics onto the expected $\langle information, inference \rangle$ space in terms of the $[CLS]$ tokens for tweet classification. Considering the classes, (a) highlights the tightening of tweet embeddings towards their respective ideal class poles. Considering the distribution of tweets, (b) emphasizes that each expected quadrant corresponds to the anticipated majority class.

| | Inference | | Information | | Multi-Class | |
|---|---|---|---|---|---|---|
| Model | Frozen | Dynamic | Frozen | Dynamic | Frozen | Dynamic |
| Closed-Topic (6-fold) Validation | | | | | | |
| **Length** | 62.34 | | 71.47 | | 38.26 | |
| **RF + TF-IDF** | 76.12 | | 80.56 | | 55.65 | |
| **Vanilla BERTweet** | 73.12 | 84.54 | 66.49 | 83.55 | 42.87 | 71.05 |
| **Augmented BERTweet** | 84.49 | **86.68** | 79.22 | 84.57 | 67.07 | 73.80 |
| **WRAPresentations** | **86.88** | 86.62 | **81.54** | **86.30** | **71.07** | **75.29** |
| Cross-Topic (6-fold) Validation | | | | | | |
| **Length** | 61.99 | | 71.55 | | 38.17 | |
| **RF + TF-IDF** | 73.93 | | 80.16 | | 53.29 | |
| **Vanilla BERTweet** | 70.28 | 83.15 | 66.15 | 82.22 | 39.00 | 68.12 |
| **Augmented BERTweet** | 84.20 | 84.25 | 79.38 | 83.31 | 66.41 | 69.99 |
| **WRAPresentations** | **86.83** | **86.27** | **81.54** | **84.90** | **70.93** | **73.54** |

Table 4: Macro F1 scores of each classifier for inference and information detection, and all four classes.

| | Reason | | Statement | | Notification | | None | |
|---|---|---|---|---|---|---|---|---|
| Model | Frozen | Dynamic | Frozen | Dynamic | Frozen | Dynamic | Frozen | Dynamic |
| Closed-Topic (6-fold) Validation | | | | | | | | |
| **Length** | 61.68 | | 20.19 | | 14.47 | | 56.72 | |
| **RF + TF-IDF** | 69.35 | | 17.30 | | 63.35 | | 72.62 | |
| **Vanilla BERTweet** | 66.05 | 74.98 | *00.00* | 53.99 | 43.80 | 77.62 | 61.63 | 77.62 |
| **Augmented BERTweet** | 74.50 | 76.82 | 49.53 | 58.37 | 70.95 | **80.28** | 73.29 | 79.71 |
| **WRAPresentations** | **77.34** | **78.14** | **58.66** | **60.96** | **72.61** | 79.36 | **75.67** | **82.72** |
| Cross-Topic (6-fold) Validation | | | | | | | | |
| **Length** | 61.78 | | 19.32 | | 14.49 | | 57.09 | |
| **RF + TF-IDF** | 68.61 | | 13.33 | | 62.75 | | 68.46 | |
| **Vanilla BERTweet** | 63.57 | 73.15 | *00.00* | 47.40 | 35.79 | 74.92 | 56.64 | 77.01 |
| **Augmented BERTweet** | 75.18 | 75.10 | 46.34 | 51.74 | 71.61 | 75.71 | 72.50 | 77.42 |
| **WRAPresentations** | **77.13** | **77.05** | **57.62** | **58.33** | **73.05** | **78.45** | **75.91** | **80.33** |

Table 5: Micro F1 scores for classifiers identifying the four classes in inference and information detection.

### 5.2 Results: Classification and Generalization

For simplicity, we present the outcomes of the RF classifier as best performing baseline and the average class length as minimal-performance indicator.

When turning to the closed-topic validation, WRAPresentations outperforms all classifiers except task (1), where dynamic embeddings in Augmented BERTweet exhibit performance nearly equivalent, as demonstrated in the upper half of Table 4. Quantitatively, WRAPresentations yields 86.88% F1 for task (1), 81.54% F1 for task (2), and 71.07% F1 for task (3) when frozen. Dynamically optimizing embeddings, WRAPresentations achieves 86.62% F1 for task (1), 86.30% F1 for task (2), and 75.29% F1 for task (3).

Shifting our attention to the more demanding task of cross-topic validation, assessing a classifier's ability to generalize to unseen topics, WRAPresentations demonstrates superior performance over all evaluations, thereby achieving 86.83% F1 for task (1), 81.54% F1 for task (2), and 70.93% F1 for task (3) when frozen. With dynamically adjusted embeddings, it achieves 86.27% F1 for task (1), 84.90% F1 for task (2), and 73.54% F1 for task (3), compare lower half of Table 4.

Further, WRAPresentations clearly improved performance for Statement, the least common and most difficult class to predict when comparing the remaining classifiers. Thereby, all other classifiers perform below or slightly above chance agreement for closed-topic validation and generalization across topics for this class, where Vanilla BERTweet even achieved 00.00% F1 when frozen, showcasing the necessity for adjusting classifiers and embeddings to specific classes, see Table 5.

## 6 Discussion

WRAPresentations consistently outperforms all models, except for a marginal -0.06% F1 decrease compared to Augmented BERTweet with dynamic representations for task (1) of closed-topic evaluation, while totally excelling across topics.

Augmented BERTweet performs stronger in detecting instances without inference, as demonstrated by the substantial 9.33% F1 increase for the Notification class with dynamic embeddings, see upper half of Table 5. Considering that tasks (1) and (2) are aggregations derived from the results of task (3), WRAPresentations enhances the overall performance of task (3) for achieving the best results, prioritizing an improvement in task (2) while incurring a slight decrease in task (1).

This effect emerges as further refinements for additional classification improvements can partially overwrite the enriched representations of inference and information in tweets, exposing unconsidered class signals during optimization of the head.

However, examining WRAPresentations' frozen states, superior in closed and cross-topic validation, underscores the advantages of our pre-classification fine-tuning focused on semantic similarity in tweets for enhanced classification strength, see Table 4, 5.

Supported by these cross-validated results, it appears that WRAPresentations can establish robust inference and information-driven representations for tweet classification, owing to our multi-task approach for systematically contrasting the argument-constituting elements in its embedding space, demonstrating adaptability and generalizability for all three argument mining tasks on Twitter, including the difficult Statement identification.

## 7 Conclusion and Ongoing Work

Our pre-classification multi-task fine-tuning approach considerably improves the specification of embeddings of BERTweet to encode diverse manifestations of inference and information, especially supporting the classification of tweets in TACO.

Enhanced by contrastive learning of semantic similarity, BERTweet's optimized embeddings excel a diverse range of argument mining approaches for Twitter, showcasing superior adaptability to class signals and cross-topic generalization.

In this regard, we can successfully contribute WRAPresentations, a contrastively optimized embedding model, and the advanced classification model WRAP for inference and information-driven argument mining across diverse topics on Twitter.

We also provide grounds for assuming that the augmentation of tweets constitutes a valuable asset within this domain of research.

Given our successful pre-classification fine-tuning with augmented tweets showing strong impact towards original tweets, we pose the two broader questions for argument mining regarding: (1) the necessity of using tweets for detecting arguments on Twitter, requiring investigation of whether tweet-like instances from other domains alone are sufficient, and (2) whether WRAPresentations or our contrastive learning approach can be transferred to build strong classifiers for domains other than Twitter.

## Limitations

For our work, we report the following limitations:

The field of argument mining on Twitter is subject to Twitter's strict data regulations, which allow only the publication of tweet identifiers but not their text. The costly Twitter API, offering only 1,500 free queries per month, complicates research reproducibility and risks data loss from deleted tweets when fetched by their identifiers. For this study, we used the TACO dataset from our previous study, which gave us full access to the data. Access to the source dataset can be granted on request for non-harmful research purposes, subject to appropriate and mandatory data protection agreements.

## Acknowledgments

We are very grateful for the attentive and constructive feedback from our anonymous reviewers. Your willingness to share your expertise and provide constructive feedback is deeply appreciated.

## References

Aseel Addawood and Masooda Bashir. 2016. "what is your evidence?" a study of controversial topics on social media. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 1–11, Berlin, Germany. Association for Computational Linguistics.

Muhammad Mahad Afzal Bhatti, Ahsan Suheer Ahmad, and Joonsuk Park. 2021. Argument mining on Twitter: A case study on the planned parenthood debate. In *Proceedings of the 8th Workshop on Argument Mining*, pages 1–11, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Bosc, Elena Cabrio, and Serena Villata. 2016. DART: a dataset of arguments and their relations on Twitter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1258–1263, Portorož, Slovenia. European Language Resources Association (ELRA).

Danah Boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *2010 43rd Hawaii International Conference on System Sciences*, pages 1–10.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann.

Xinlei Chen and Kaiming He. 2020. Exploring simple siamese representation learning. *CoRR*, abs/2011.10566.

S. Chopra, R. Hadsell, and Y. LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Marc Feger and Stefan Dietze. 2024. TACO – Twitter Arguments from COnversations. *Preprint*, arXiv:2404.00406.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. *CoRR*, abs/1902.00751.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *CoRR*, abs/2004.11362.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, WWW '10, page 591–600, New York, NY, USA. Association for Computing Machinery.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, page 98–107, New York, NY, USA. Association for Computing Machinery.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Informatics Nat. Intell.*, 7:1–31.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, Florence, Italy. 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.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

Robin Schaefer and Manfred Stede. 2021. Argument mining on twitter: A survey. *it - Information Technology*, 63(1):45–58.

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.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? *CoRR*, abs/1905.05583.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Terne Sasha Thorn Jakobsen, Maria Barrett, and Anders Søgaard. 2021. Spurious correlations in cross-topic argument mining. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 263–277, Online. Association for Computational Linguistics.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Jesper E. van Engelen and Holger H. Hoos. 2020. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume abs/2005.10242.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Eric Xing, Michael Jordan, Stuart J Russell, and Andrew Ng. 2002. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

Yang You, Jing Li, Jonathan Hseu, Xiaodan Song, James Demmel, and Cho-Jui Hsieh. 2019. Reducing BERT pre-training time from 3 days to 76 minutes. *CoRR*, abs/1904.00962.