Argument Mining on Twitter: A Case Study on the Planned Parenthood Debate

Muhammad Mahad Afzal Bhatti*, **Ahsan Suheer Ahmad***, **and Joonsuk Park***†
*Department of Math and Computer Science, University of Richmond
†NAVER AI Lab
{mahad.bhatti, ahsansuheer.ahmad}@richmond.edu
park@joonsuk.org

Abstract

Twitter is a popular platform to share opinions and claims, which may be accompanied by the underlying rationale. Such information can be invaluable to policy makers, marketers and social scientists, to name a few. However, the effort to mine arguments on Twitter has been limited, mainly because a tweet is typically too short to contain an argument—both a claim *and* a premise. In this paper, we propose a novel problem formulation to mine arguments from Twitter: We formulate argument mining on Twitter as a text classification task to identify tweets that serve as premises for a hashtag that represents a claim of interest. To demonstrate the efficacy of this formulation, we mine arguments for and against funding Planned Parenthood expressed in tweets. We first present a new dataset of 24,100 tweets containing hashtag #StandWithPP or #DefundPP, manually labeled as SUPPORT WITH REASON, SUPPORT WITHOUT REASON, and NO EXPLICIT SUPPORT. We then train classifiers to determine the types of tweets, achieving the best performance of 71% F_1. Our results manifest claim-specific keywords as the most informative features, which in turn reveal prominent arguments for and against funding Planned Parenthood.

1 Introduction

The goal of argument mining is to automatically extract arguments—typically defined as consisting of both a claim and at least one premise supporting the claim—from text in various domains. By analyzing the argumentative structure, we can not only identify claims, but also gain a deeper understanding of the evidence and reasons behind the claims (Rahwan et al., 2009; Mochales and Moens, 2011a; Peldszus and Stede, 2013; Lippi and Torroni, 2015; Budzynska and Villata, 2016; Lawrence and Reed, 2020). And the domains for argument mining have quickly expanded to include less formally written text on social media (Wyner et al., 2012; Goudas et al., 2014; Park and Cardie, 2014; Morio and Fujita, 2018; Chakrabarty et al., 2019).

Yet, the effort to mine arguments from Twitter has been limited due to a rather obvious reason—tweets are often too short to contain an entire argument, i.e., a claim and a premise (Dusmanu et al., 2017). For this reason, existing approaches to argument mining on Twitter typically focus on identifying claims, evidence, or either, but not both at the same time (Addawood and Bashir, 2016; Bosc et al., 2016a; Dusmanu et al., 2017; Wührl and Klinger, 2021). This is not ideal, since the underlying rationale can be as important as the claim.

To mine full arguments from Twitter, we propose a novel problem formulation based on an observation—some trendy hashtags serve as placeholders for claims, which may or may not be supported by the contents of the tweets containing them. In the case of the Planned Parenthood debate, the two opposing sides use hashtags #StandWithPP and #DefundPP to specify their respective claims; only a subset of the tweets serve as premises supporting the given claim. For instance, Example 3 in Table 1 can be interpreted as an argument in which the premise "#AllLivesMatter even the unborn." (= *all live matter, even the unborn*) supports the claim to "#DefundPP" (= *Planned Parenthood should not be funded by the government.*). In contrast, Example 6 cannot be considered a premise for the claim, as it does not provide a reason to "#StandWithPP"(= *Planned Parenthood should continue to be supported by the government*). In the case of Example 10, it is not even clear whether or not the user supports the claim #StandWithPP represents. Thus, the tweet cannot be considered a premise for the claim. (While both Examples 6 and 10 are not considered premises, distinguishing the two can be useful for compiling a quantitative summary, e.g. the number of tweets showing support—the former should count as a supporting tweet, unlike the latter.)

Proceedings of The 8th Workshop on Argument Mining, pages 1–11
Punta Cana, Dominican Republic, November 10–11, 2021.

2-class	3-class	#	Example Tweet
PREMISE	SUPPORT WITH REASON	1	I **#StandWithPP**. Routine healthcare shouldn't exist just for the rich.
		2	*@user* helps men too! **#StandWithPP**
		3	#AllLivesMatter even the unborn. **#DefundPP** #DefundPlannedParenthood
		4	God Has A Plan For Every Life #PraytoEndAbortion **#DefundPP** #DefendLife #innocent
NON-PREMISE	SUPPORT WITHOUT REASON	5	**#StandWithPP** now and forever.
		6	I wish everyone shouldn't hate on Planned Parenthood so fucking much **#StandWithPP**
		7	#YouHadMeAt I'll do everything in my power to ***#DefundPP***
		8	Tell your Senators to Defund Planned Parenthood SIGN & RT **#DefundPP** *[url]*
	NO EXPLICIT SUPPORT	9	Legal Troubles Continue for Group Attacking Planned Parenthood **#StandWithPP** *[url]*
		10	#SextaComMFSDV **#StandWithPP** Citizen Khan Grow your Twitter followers *[url]*
		11	*@user* Paid staffers? **#defundpp**
		12	Dont listen to the Daily Bugle . Spider-Man is a force for good .**#StandWithPP** #PeterParker

Table 1: Example Tweets for Each Class (2-class and 3-class setup). The user can show support with reason (SUPPORT WITH REASON), show support without providing a reason (SUPPORT WITHOUT REASON) or make their stance unclear through an irrelevant or overall confusing tweet (NO EXPLICIT SUPPORT).

Henceforth, we call a hashtag representing a claim a *claim-hashtag*, and a tweet serving as a premise a *premise-tweet*. From an argument mining perspective, the claim is already known for tweets containing a claim-hashtag, i.e., the claim represented by the claim-hashtag. And such tweets can be easily retrieved using the Twitter API or simple text matching. Thus, the main challenge is in determining whether a given tweet is a premise-tweet. In other words, we formulate argument mining on Twitter as a text classification task to identify premise-tweets for claim-hashtags of interest.

To demonstrate the efficacy of the proposed formulation, we mine arguments for and against funding Planned Parenthood expressed on Twitter: We first present a new dataset of 24,100 tweets containing hashtag #StandWithPP or #DefundPP, each manually labeled as SUPPORT WITH REASON, SUPPORT WITHOUT REASON, or NO EXPLICIT SUPPORT. We then train several classifiers and test them on 30% of the dataset held-out in advance. We find that fine-tuned BERT performs the best, achieving 71% F_1. We also show that claim-specific words serve as the most important features for this task, which in turn reveal important arguments for and against funding Planned Parenthood.

Why Planned Parenthood? The Planned Parenthood debate is multi-faceted, involving issues like the personhood of fetuses, women's rights, and health services to people of various socioeconomic status. A major benefit of automatically extracting arguments from Twitter is that it provides an easy access to arguments people have made. This is especially helpful for complex topics like Planned Parenthood, where unique but noteworthy arguments can be lost in the midst of others. From a practical perspective, each side of the Planned Parenthood debate has a dominantly used hashtag, allowing us to target two specific hashtags and gain a holistic view of the debate.

Our main contributions are threefold:

- We propose a novel problem formulation for mining full arguments—both a claim and a premise—on Twitter.
- We present a newly annotated dataset consisting of 24,100 tweets[1], which is 10 to 80 times bigger than existing datasets for mining arguments from Twitter.
- We identify prominent arguments for and against funding Planned Parenthood expressed on Twitter by analyzing the most informative features.

2 Related Work

2.1 Argument Mining

Argument mining has been used in various domains over the years. These include text written by professionals—such as legal documents (Moens et al., 2007; Wyner et al., 2010; Mochales and Moens, 2011b) and newspaper articles (Reed et al., 2008)—as well as student essays (Stab and Gurevych, 2014; Wachsmuth et al., 2016) and online user comments and reviews (Wyner et al., 2012; Goudas et al., 2014; Park and Cardie, 2014). In addition, researchers have tackled dialogues (Budzynska et al., 2014), political debates (Lippi and Torroni, 2016), clinical trials (Mayer et al., 2018), peer reviews (Hua et al., 2019) and news blogs (Basile et al., 2016). While

[1]Tweet IDs and labels are available at joonsuk.org

this is a diverse set of domains, they share a common trait that documents are long enough to contain full arguments, often multiple of them in a single document. Thus, argument mining involves identifying argumentative spans of text, determining argumentative units—e.g. premise and claim—within the arguments, and recognizing the argumentative structure connecting the units. However, tweets are typically too short to contain full arguments, preventing the use of standard argument mining approaches (Dusmanu et al., 2017).

There has been some pioneering work on mining arguments from tweets, as summarized by Schaefer and Stede (2021). To get around the issue of tweets being too short to contain an entire argument, researchers typically seek to identify argumentative tweets—tweets that contain an argumentative unit, e.g. claim *or* premise (Bosc et al., 2016a,b; Dusmanu et al., 2017; Wührl and Klinger, 2021). For instance, Bosc et al. (2016a,b) distinguish argumentative tweets from non-argumentative ones. For tweets containing a claim, they further distinguish opinion from factual tweets. For tweets containing evidence, they seek to identify the source. Addawood and Bashir (2016); Addawood et al. (2017) also identify argumentative tweets, which are further broken down into six different types, such as *expert opinion* and *blog*. Schaefer and Stede (2020) present several task formulations, where the closest one to ours is identifying evidence tweets (for a claim expressed in what they call a *context tweet* or a reply tweet).

Our work, however, specifically targets tweets containing both a claim (in the form of a hashtag) and a premise. This enables the full argument to be reconstructed for each argumentative tweet. In addition, our newly annotated dataset is significantly larger than the datasets used in previous tweet argument mining research, more than 10 to 80 times the size depending on the task (Dusmanu et al., 2017; Schaefer and Stede, 2020). This will enhance the reliability of the experiment results and analyses.

2.2 Planned Parenthood

Planned Parenthood is a non-profit organization that provides reproductive health services in the US and abroad.[2] Whether or not the US government should continue to fund Planned Parenthood has been the subject of ongoing debate, mainly due to the controversial practice of abortion (Halva-Neubauer and Zeigler, 2010; Devi, 2015; Silver and Kapadia, 2017); researchers have argued over the legality and subsequent funding for abortion (Primrose, 2012; Wharton et al., 2006). Supporters of Planned Parenthood have presented several arguments, including that it provides other medical services (Silver and Kapadia, 2017; Stevenson et al., 2016; House and Goldsmith, 1972). Those against Planned Parenthood also have expressed their position, mostly arguing against the practice of abortion (Halva-Neubauer and Zeigler, 2010; Ziegler, 2012; Devi, 2015).

The general public has also been voicing their opinions through various social media platforms such as Twitter. While Twitter provides a convenient means to express opinions, gathering such opinions for analysis is not as straight forward. This is unfortunate, as many arguments with compelling reasons and evidence are present in tweets; they are not used to further the discussion surrounding Planned Parenthood in a productive manner. Our work is a step toward addressing this issue by enhancing the efficiency of communication.

3 Data

#StandWithPP and #DefundPP represent the two opposing sides on the issue of the US federal government funding Planned Parenthood, or of Planned Parenthood itself. The claims represented by the hashtags can be stated as follows:

- **#StandWithPP:** Planned Parenthood should continue to receive federal funding.
- **#DefundPP:** Planned Parenthood should not receive federal funding.

Tweets containing either of these hashtags were collected over a span of two months. Prior to preprocessing, there were a total of 20,314 and 12,470 tweets containing #StandWithPP and #DefundPP, respectively.

3.1 Preprocessing

As part of the preprocessing, we first removed duplicate and otherwise uninformative tweets that can be easily identified[3]: tweets by the seven most frequently tweeting users (these are mostly auto-generated spams with repetitive content); tweets

[2] https://www.plannedparenthood.org/

[3] These were all NO EXPLICIT SUPPORT tweets, technically, but we removed them from the dataset, as they can be easily identified by pattern matching, without training a classifier.

Claim-Hashtag	PREMISE	NON-PREMISE		Total
	SUPPORT WITH REASON	SUPPORT WITHOUT REASON	NO EXPLICIT SUPPORT	
#StandWithPP (Training)	4,432 (35.9%)	2,940(23.8%)	4,962 (40.3%)	12,334
#StandWithPP (Test)	1,852 (35.0%)	1,316 (24.9%)	2,118 (40.1%)	5,286
#DefundPP (Training)	2,000 (44.1%)	486 (10.7%)	2,050 (45.2%)	4,536
#DefundPP (Test)	861 (44.3%)	193 (9.9%)	890 (45.8%)	1,944
Total	9,145 (37.9%)	4,935 (20.5%)	10,020(41.6%)	24,100

Table 2: Distribution of Classes in the Dataset. 30% of the tweets for each hashtag were randomly put in the held-out test set.

with a URL and two or more special character (this is a noticeable pattern for tweets in our dataset simply sharing URLs to news sites with random special characters to catch people's attention); tweets with fewer than 4 tokens; and tweets in which @-mentions, URLs, or hashtags make up more than 35% of the tokens.

The filtering process reduced the number of tweets to 16,870 and 7,230 for #StandWithPP and #DefundPP, respectively. Then, all @-mentions were masked to protect the users' privacy. Any URLs were also masked, as our goal is to recognize premises in the body of tweets.

3.2 Annotation

The dataset was then annotated using the Amazon Mechanical Turk[4] service. The annotators were asked to classify each tweet as one of the three possible classes:

- **SUPPORT WITH REASON:** The user supports the claim represented by the claim-hashtag and presents a reason, regardless of the validity and strength.
- **SUPPORT WITHOUT REASON:** The user supports the claim represented by the claim-hashtag, but does not provide a reason.
- **NO EXPLICIT SUPPORT:** All other tweets. Typically, the user has a neutral or unclear stance toward the claim represented by the claim-hashtag, such as news tweets. In some cases, the user uses a claim-hashtag to present a counter-argument to people supporting the claim, rather than to show support.

We ran a pilot study in which annotators were asked to annotate tweets for which we had the gold standard labels. Out of 100 annotators who participated, we identified 32 reliable annotators to annotate the dataset.

Then, each tweet was annotated by two annotators, where disagreements were resolved by an adjudicator. We observed a reasonable agreement, Krippendorff's α of 0.79. A common source of disagreement was incomplete information, e.g. "8 Unbelievably Heartbreaking Quotes From Women Who Aborted Their Own Babies | *[URL]*: *[URL]* #DefundPP." Depending on the quotes presented in the URL, this tweet can be for or against Planned Parenthood: What is heartbreaking could be abortion itself or the process of abortion due to the lack of access to adequate health services. (Given the presence of the hashtag #DefundPP, it is likely that the quote, and in turn this tweet, is against abortion and Planned Parenthood. However, the annotators were asked not to assume the presence of a hashtag as a sign of support, as it is not always true.)

Table 2 summarizes the resulting dataset. Note that this is after removing obvious spam tweets during preprocessing as described above. Thus, the percentage of NO EXPLICIT SUPPORT is higher in reality. Also, there is a noticeable difference between #StandWithPP and #DefundPP tweets in terms of the class distribution; a significantly smaller portion of the latter are SUPPORT WITHOUT REASON tweets. We suspect that this is because changing the status quo requires more convincing arguments. Thus, people arguing to defund Planned Parenthood are more likely to support their claim with a reason or evidence.

4 Premise-Tweet Identification

Argument mining consists of several subtasks, such as identifying argumentative spans of text, determining argumentative units—e.g. premise and claim—within the arguments, and recognizing the argumentative structure connecting the units. In this work, however, the claim is easily identifiable, as we assume that it takes the form of a hashtag, i.e., claim-hashtag, that is known in advance. Thus, the core of our approach to mining arguments on Twitter is deciding whether or not a given tweet is a premise-tweet for a given claim-hashtag. To tackle the task, we train fine-tuned BERT, CNN,

[4] http://www.mturk.com

Model	#StandWithPP				#DefundPP			
	Prec	Rec	F_1	Acc	Prec	Rec	F_1	Acc
Baseline approaches adopted from Schaefer and Stede (2020)								
- XGBoost with UNIGRAMS	.682	.676	.669	.676	.665	.682	.667	.682
- XGBoost with UNIGRAMS + BIGRAMS	.697	.686	.679	.686	.671	.686	.671	.686
- XGBoost with BERT Word Embedding	.542	.543	.528	.543	.534	.549	.532	.549
CNN with GloVe Word Embedding (CommonCrawl)	.675	.661	.650	.661	.607	.669	.634	.669
CNN with GloVe Word Embedding (Twitter)	.696	.689	.689	.689	.678	.685	.656	.685
Fine-tuned DistilBERT	.680	.675	.670	.675	.643	.683	.656	.683
Fine-tuned BERT	**.714**	**.714**	**.713**	**.714**	**.719**	**.728**	**.718**	**.728**

Table 3: Experiment Results for 3-Class Classification (SUPPORT WITH REASON vs SUPPORT WITHOUT REASON vs NO EXPLICIT SUPPORT). The experiments were conducted independently for #StandWithPP and #DefundPP tweets. Also, each entry in the table is the weighted average of the measures computed with respect to the classes.

and XGBoost classifiers as detailed in this section.

Note that we are also interested in distinguishing non-premise-tweets that support the claim (SUPPORT WITHOUT REASON) from those that do not (NO EXPLICIT SUPPORT); This is because the sheer number of tweets supporting a claim can be used to generate a statistical summary of people's support for the claim. Thus, we formulate argument mining on Twitter as a classification task with three classes: SUPPORT WITH REASON, SUPPORT WITHOUT REASON, and NO EXPLICIT SUPPORT.

4.1 Fine-tuned BERT

Given the successful use of fine-tuned BERT on various text classification tasks (Croce et al., 2020; Tian et al., 2020), we fine-tune a pre-trained BERT to premise-tweet classification task using our training set. For the experiments, we fine-tune the 'bert-base-uncased' pre-trained model (Wolf et al., 2020), which consists of 12 BERT attention layers, 768 hidden nodes, and 12 attention heads, with a total of 110M parameters. Using the BERT Tokenizer (Wolf et al., 2020), each tweet is represented by token, segment, and position embedding. Lastly, in order to classify tweets, the model is augmented with a fully-connected classification layer with ReLu activation on top of the pooled output from BERT. An AdamW optimizer is used for regularization (Loshchilov and Hutter, 2019).

In addition to BERT, we also test the efficacy of DistilBERT, which is a much simpler and faster model that can match the performance of BERT in some cases (Sanh et al., 2019).

4.2 Convolutional Neural Network (CNN)

While BERT's attention mechanism is shown to be effective for capturing both short and long distance relations between words in documents, a simple CNN may suffice given the brevity of tweets. Thus, we also experiment with CNN. Following the framework presented by Kim (2014), each tweet is represented as an $n \cdot k$ matrix, where n is the length of the tweet and k is the dimensionality of the word vectors. For word representation, we employ two versions of the GloVe word embedding (Pennington et al., 2014): A 200-d version trained on tweets, since we are working with tweets; and a 300-d version trained on Common Crawl, since a higher-dimensional embedding may be more effective. For both, we limit the size of the vocabulary to a million tokens.

4.3 eXtreme Gradient Boosting (XGBoost)

XGBoost is an extension to Gradient Boosting that has shown to be effective in several classification tasks (Stein et al., 2019; Qi, 2020). Schaefer and Stede (2020) show that using XGBoost to classify evidence tweets has promising results. Given the similarity of one of their setups to ours, we use as baselines the 3 variations they employed: XGBoost with UNIGRAMS, XGBoost with UNIGRAMS + BIGRAMS, and XGBoost with BERT word embeddings. The booster we use is a gradient boosting tree, with a standard max depth of 6 for a tree. The algorithm minimizes the multi-class log loss function, and applies a variation of softmax to get the predicted output probabilities.

5 Experiments

5.1 Setup

For each claim-hashtag in our dataset—#StandWithPP and #DefundPP—the classifiers were trained and tested on the respective training and held-out test sets (See Table 2). For optimizing hyper-parameters, 5-fold cross validation was done on the training set. The dropout rate was $p = 0.5$ for CNN and $p = 0.1$ for BERT. The learning rate was $lr = 0.001$ for CNN and $lr = 2e^{-}5$ for BERT.

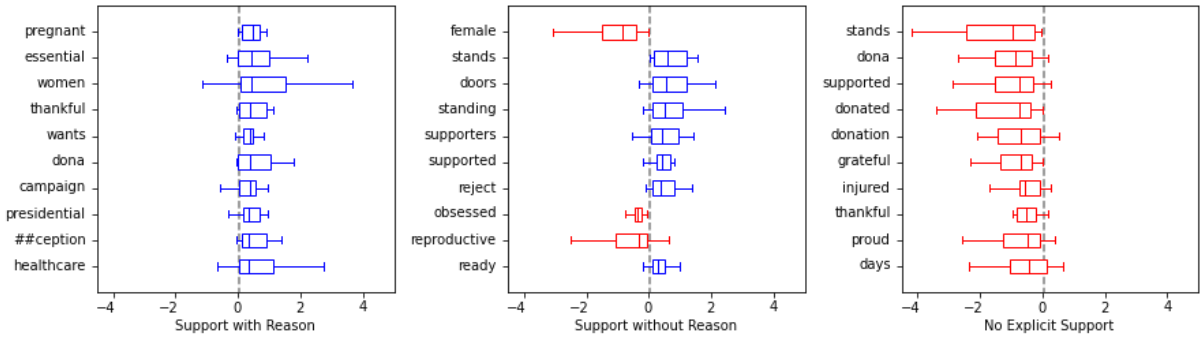

Figure 1: Words that had the biggest influence on the classification decision for BERT fine-tuned on **#StandWithPP** tweets, sorted by the median absolute SHAP value. Positive values are colored blue, and negative, red.

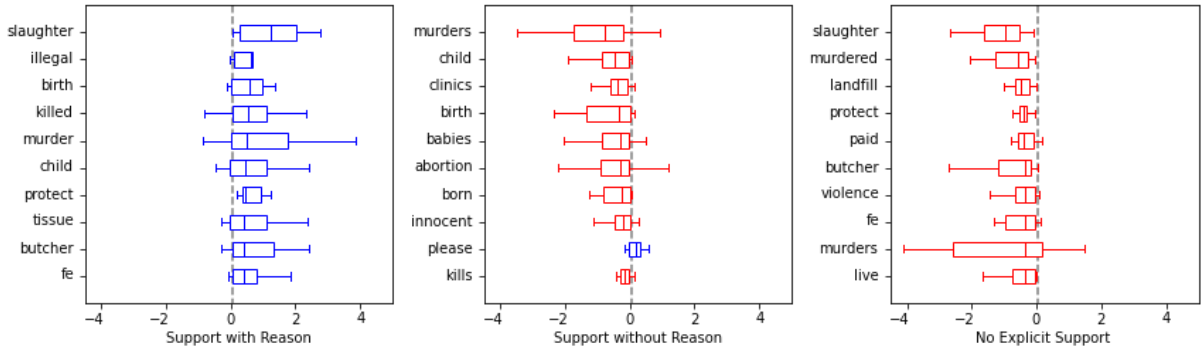

Figure 2: Words that had the biggest influence on the classification decision for BERT fine-tuned on **#DefundPP** tweets, sorted by the median absolute SHAP value. Positive values are colored blue, and negative, red.

Batch size was $b = 50$ for CNN and $b = 32$ for BERT. The number of epochs was 15 for CNN and 4 for BERT. For the XGBoost baselines, we used the same setup from Schaefer and Stede (2020), but the models were trained and tested on our training and test sets.

5.2 Results & Analysis

The experiment results are summarized in Table 3. Fine-tuned BERT outperformed the rest across the board. This is not surprising given the strong state-of-the-art performances of transformer-based models across various NLP tasks. DistilBERT, a much smaller and faster version of BERT, is noticeably worse than BERT, but is comparable to the CNNs. CNN with GloVe-Twitter performs slightly better than CNN with GloVe-Common Crawl; we suspect that the word embedding trained on tweets is more effective, since our dataset is also a collection of tweets. For the XGBoost baselines, using ngrams proved to be more effective than using BERT word embeddings. This is consistent with the results from Schaefer and Stede (2020), though the datasets are different, and thus a direct comparison cannot be made. We suspect that the straightforward mapping between dimensions and words in ngrams is better suited for XGBoost than multiple dimensions collectively representing a word in a word embedding; this is because XGBoost is a decision tree based approach that learns to weigh each feature (dimension) differently.

Figures 1 and 2 are the median SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) values of the top features for fine-tuned BERT, the best performing model. The SHAP value for a feature with respect to a class indicates the level of influence the given feature had in classifying a tweet as the given class. Here, the influence can be either positive or negative, indicated by the sign of the SHAP value. And the bigger the magnitude, the heavier the influence it had on the classification decision. The median for each word is calculated across all occurrences of the word in the test set. Note that words that occur fewer than 10 times in the test set were excluded from the plot.

Similar patterns are exhibited for both claim-hashtags. For SUPPORT WITH REASON, the words with large absolute SHAP values tend to be keywords for prominent arguments for the given claim.

#	Actual	Predicted	Tweet
1	S+R	S+R	[CLS] grateful to see judge po ##sner standing up for women ' s rights and calling out ridiculous restrictions on abortion providers . # stand ##with ##pp [SEP] [PAD] [PAD]
2	S+R	S+R	[CLS] for 100 years the brave people of planned parent ##hood have provided health care to women . here ' s to 100 more . # stand ##with ##pp [SEP]
3	S+R	S+R	[CLS] just when i think [UNK] can ' t get any lower they t ##wee ##t this , how about all those babies you butcher ##ed # def ##und [SEP]
4	S+R	S+R	[CLS] pp needs to be def ##und ##ed to the extent that un ##born baby slaughter can no longer be tolerated in america . # def ##und ##pp # [SEP]
5	S-R	S-R	[CLS] i just supported national day of solidarity on [UNK] / / [UNK] t _ ur ##l # stand ##with ##pp # investigate ##cl ##ini ##c ##vio ##lence [SEP] [PAD]
6	S-R	S-R	[CLS] please pray today that barack obama ' s veto to stop def ##und ##ing planned parent ##hood is thwarted . the u . s . senate voted in [SEP]
7	NES	NES	[CLS] 26 hours since a terrorist attacked a co [UNK] 3 civilians killed , 8 injured ##0 go ##p pre ##z hopeful ##s have mentioned the attack # stand [SEP]
8	NES	S+R	[CLS] so far state & amp ; fed investigation on the women ' s heath care provider turned up no evidence of wrong ##do ##ing # stand ##with ##pp [SEP]
9	S-R	S+R	[CLS] " your words and actions hurt women and create a climate of di ##sr ##es ##pe ##ct that makes violence possible . " # stand ##with ##pp [UNK] [SEP]

Table 4: Example tweets and classifications by fine-tuned BERT. Tokens are highlighted in blue if they have positive attribution scores with respect to the predicted class, and red if negative. The darker the color, the higher the absolute value of the score. The class names are abbreviated as follows: SUPPORT WITH REASON (S+R), SUPPORT WITHOUT REASON (S-R), and NO EXPLICIT SUPPORT (NES)

.

In the case of #StandWithPP, the words “women" and “healthcare" rank high; they typically appear in tweets that emphasize women’s rights or the need for healthcare in general as reasons to support Planned Parenthood (See Examples 1 and 2 in Table 4). In the case of #DefundPP, words that emphasize babies and framing abortion as murder rank high (Examples 3 and 4).

For NO EXPLICIT SUPPORT, most of the words with large absolute SHAP values have negative values, meaning the existence of these words was taken as a sign that the given tweet is not NO EXPLICIT SUPPORT. In other words, lacking strong characteristics of the other classes is the characteristic of NO EXPLICIT SUPPORT(Example 7). This is partially due to our having removed spam tweets with obvious patterns during preprocessing. Otherwise, those patterns may have had positive SHAP values of large magnitudes.

For SUPPORT WITHOUT REASON, however, the two claim-hashtags exhibit some differences. For #StandWithPP, words that appear in clear statements of support, e.g. “support" and “stands [with Planned Parenthood]," have positive influence on classifying a tweet as SUPPORT WITHOUT REASON. The is because such tweets tend not to include a rationale (Example 5). However, similar words for #DefundPP, e.g. “defund" and “stop [funding Planned Parenthood]", do not have high SHAP values with respect to SUPPORT WITHOUT REASON, as they often appear with additional explanations (Example 4). Other than “please" (Example 6), there are not many indicators of SUPPORT WITHOUT REASON for #DefundPP. There are not many SUPPORT WITHOUT REASON tweets to begin with as shown in Table 2. Again, we suspect that non-NO EXPLICIT SUPPORT tweets for #DefundPP tend to contain a reason, as they have to be convincing enough to change the status quo.

Note that the informative features are not always helpful. Non-SUPPORT WITH REASON tweets that contain top feature words of SUPPORT WITH REASON can be incorrectly classified. For example, the tweet can be a news tweet reporting the state of affairs. Such tweet does not always reveal the stance of the user posting the tweet (Example 8). The tweet can also be part of a conversation where the reason for supporting the claim cannot be determined without knowing the tweet being replied to (Example 9).

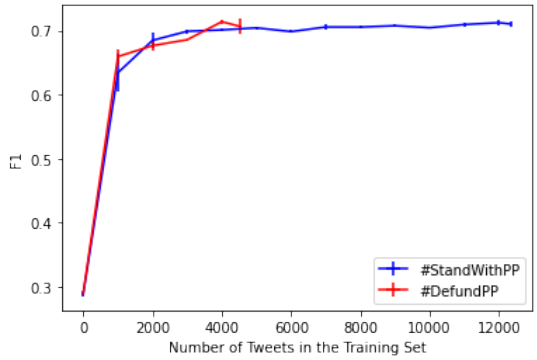

Figure 3: Impact of the training set size on the performance. Pre-trained BERT was fine-tuned on randomly subsampled training sets. Training set sizes in increments of 1,000 were tested and averaged over 3 runs.

6 Limitations and Implications

There are two main limitations of this work that need to be considered in future research. First, labeled training data is required to train a premise-tweet classifier for each claim of interest. The most informative features for our classifiers are specific to the claim-hashtag they are trained for; even though #StandWithPP and #DefundPP are on the same topic, the informative features are drastically different. This suggests that a classifier trained for a claim-hashtag likely is not effective for identifying premise-tweets for other claims-hashtags. In fact, this was confirmed through cross-domain testing, i.e., we fined-tuned BERT on the #StandWithPP training set and tested on the #DefundPP test set, and vice versa. There was a significant drop in the F1 score from 71% to 56% in both scenarios.

To determine how much training data is necessary, we fine-tuned BERT on randomly selected subsets of the training sets. We tested training set sizes in increments of 1k as shown in Figure 3. The same pattern can be observed for both claim-hashtags: There is a drastic improvement in performance after fine-tuning with even a small training set of size 1k; and the performance plateaus after increasing the size to about 3k to 4k. Based on the result, we suggest that a labeled dataset of at least 3k tweets is prepared to train a premise-tweet classifier for a claim-hashtag of your interest.

Second, our results are based on experiments with tweets containing two specific claim-hashtags. Future work should consider a more diverse set of claim-hashtags. It will not only test the generalizabilty of this approach, but may also reveal informative features that are claim-independent.

Manually compiling a list of diverse claim-hashtags can be laborious, however. To alleviate this issue, we have identified a class of claim-hashtags that can be automatically recognized. These hashtags represent so called *policy propositions*, meaning they suggest policies, or courses of action to be taken (Park et al., 2015). They typically take the form of an imperative—starting with a verb and ending with a noun, e.g. #StandWithPP, #DefundPP, #FightFor15, #LegalizeMarijuana, and #BanGuns. Hashtags do not contain spaces, but the *CamelCase* capitalization can be used for tokenization—a capitalized letter marks the beginning of a new word, unless several capitalized letters appear in succession to denote a proper noun. The repetitive use of hashtags in tweets is helpful in this regard, as it is very likely that at least one variation of a given hashtag is in CamelCase. Thus, to identify a diverse set of claim-hashtags, we suggest the method of identifying trending hashtags representing policy propositions.

7 Conclusion

Twitter is a popular platform to share opinions, which may be accompanied by the underlying rationale. However, the effort to automatically extract arguments from Twitter has been limited, mainly due to tweets typically not containing both a claim and a premise. The brevity renders it difficult to apply argument mining techniques designed for other domains, where claims and premises can be extracted together. In this paper, we proposed a novel problem formulation to mine arguments from Twitter: We formulated argument mining on Twitter as a text classification task to identify tweets serving as premises for hashtags that represent claims. We demonstrated the efficacy of this formulation by mining arguments for and against funding Planned Parenthood expressed on Twitter. We achieved the best performance of 71% F_1 with fine-tuned BERT. We also showed that domain specific words serve as the most important features, which in turn reveal prominent arguments in support of the given claim. In future work, we would like to continue the effort addressing the issues discussed in Section 6.

Acknowledgments

We thank the University of Richmond and the Thomas F. and Kate Miller Jeffress Memorial Trust, Bank of America, Trustee for their generous support for this project. We also thank Jamison Poland.

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.

Aseel Addawood, Jodi Schneider, and Masooda Bashir. 2017. Stance classification of twitter debates: The encryption debate as a use case. In *Proceedings of the 8th International Conference on Social Media & Society*, pages 1–10.

Pierpaolo Basile, Valerio Basile, Elena Cabrio, and Serena Villata. 2016. Argument Mining on Italian News Blogs. In *Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*, Naples, Italy.

Tom Bosc, Elena Cabrio, and Serena Villata. 2016a. 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).

Tom Bosc, Elena Cabrio, and Serena Villata. 2016b. Tweeties squabbling: Positive and negative results in applying argument mining on social media. *COMMA*, 2016:21–32.

Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014. Towards argument mining from dialogue. In *COMMA*, pages 185–196.

Katarzyna Budzynska and Serena Villata. 2016. Argument mining. *IEEE Intell. Informatics Bull.*, 17(1):1–6.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: Argument mining for PERSuAsive oNline discussions. 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 2933–2943, Hong Kong, China. Association for Computational Linguistics.

Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. Gan-bert: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119.

Sharmila Devi. 2015. Anti-abortion groups target funding of planned parenthood. *The Lancet*, 386(9997):941.

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.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument extraction from news, blogs, and social media. In *Artificial Intelligence: Methods and Applications*, pages 287–299. Springer.

Glen A Halva-Neubauer and Sara L Zeigler. 2010. Promoting fetal personhood: The rhetorical and legislative strategies of the pro-life movement after planned parenthood v. casey. *Feminist Formations*, pages 101–123.

Elizabeth A. House and Sadja Goldsmith. 1972. Planned parenthood services for the young teenager. *Family Planning Perspectives*, 4(2):27–31.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. 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 2131–2137, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Marco Lippi and Paolo Torroni. 2015. Argument mining: A machine learning perspective. In *International Workshop on Theory and Applications of Formal Argumentation*, pages 163–176. Springer.

Marco Lippi and Paolo Torroni. 2016. Argument mining from speech: Detecting claims in political debates. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Tobias Mayer, Elena Cabrio, Marco Lippi, Paolo Torroni, and Serena Villata. 2018. Argument mining on clinical trials. In *COMMA*, pages 137–148.

Raquel Mochales and Marie-Francine Moens. 2011a. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Raquel Mochales and Marie-Francine Moens. 2011b. Argumentation mining. *Artif. Intell. Law*, 19(1):1–22.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, pages 225–230, New York, NY, USA. ACM.

G. Morio and K. Fujita. 2018. Annotating online civic discussion threads for argument mining. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 546–553.

Joonsuk Park, Cheryl Blake, and Claire Cardie. 2015. Toward machine-assisted participation in erulemaking: An argumentation model of evaluability. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, ICAIL '15, pages 206–210, New York, NY, USA. ACM.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland. Association for Computational Linguistics.

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

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Sarah Primrose. 2012. The attack on planned parenthood: A historical analysis. *UCLA Women's LJ*, 19:165.

Zhang Qi. 2020. The text classification of theft crime based on tf-idf and xgboost model. In *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pages 1241–1246. IEEE.

Iyad Rahwan, Guillermo R Simari, and Johan van Benthem. 2009. *Argumentation in artificial intelligence*, volume 47. Springer.

Chris Reed, Raquel Mochales Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language resources for studying argument. In *LREC*. European Language Resources Association.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

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.

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

Diana Silver and Farzana Kapadia. 2017. Planned parenthood is health care, and health care must defend it: a call to action.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Roger Alan Stein, Patricia A Jaques, and Joao Francisco Valiati. 2019. An analysis of hierarchical text classification using word embeddings. *Information Sciences*, 471:216–232.

Amanda J Stevenson, Imelda M Flores-Vazquez, Richard L Allgeyer, Pete Schenkkan, and Joseph E Potter. 2016. Effect of removal of planned parenthood from the texas women's health program. *New England Journal of Medicine*, 374(9):853–860.

Lin Tian, Xiuzhen Zhang, Yan Wang, and Huan Liu. 2020. Early detection of rumours on twitter via stance transfer learning. In *European Conference on Information Retrieval*, pages 575–588. Springer.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691.

Linda J. Wharton, Susan Frietsche, and Kathryn Kolbert. 2006. Preserving the core of roe: Reflections on planned parenthood v. casey. *Yale Journal of Law and Feminism*, 18:2.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Amelie Wührl and Roman Klinger. 2021. Claim detection in biomedical Twitter posts. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 131–142, Online. Association for Computational Linguistics.

Adam Wyner, Raquel Mochales-Palau, Marie-Francine Moens, and David Milward. 2010. Semantic processing of legal texts. chapter Approaches to Text Mining Arguments from Legal Cases, pages 60–79. Springer-Verlag, Berlin, Heidelberg.

Adam Wyner, Jodi Schneider, Katie Atkinson, and Trevor JM Bench-Capon. 2012. Semi-automated argumentative analysis of online product reviews. *COMMA*, 245:43–50.

Mary Ziegler. 2012. Sexing harris: The law and politics of the movement to defund planned parenthood. *Buff. L. Rev.*, 60:701.