GunStance: Stance Detection for Gun Control and Gun Regulation

Nikesh Gyawali[*♥] **Iustin Sirbu**[*♦] **Tiberiu Sosea**[*♣] **Sarthak Khanal** [♥]

Doina Caragea[♥] **Traian Rebedea**[♦] **Cornelia Caragea**[♣]

[♥]Kansas State University [♦]University Politehnica of Bucharest [♣]University of Illinois Chicago
{gnikesh,sarthakk,dcaragea}@ksu.edu
{iustin.sirbu,traian.rebedea}@upb.ro
{tsosea2,cornelia}@uic.edu

Abstract

The debate surrounding gun control and gun regulation in the United States has intensified in the wake of numerous mass shooting events. As perspectives on this matter vary, it becomes increasingly important to comprehend individuals' positions. Stance detection, the task of determining an author's position towards a proposition or target, has gained attention for its potential use in understanding public perceptions towards controversial topics and identifying the best strategies to address public concerns. In this paper, we present GunStance, a dataset of tweets pertaining to shooting events, focusing specifically on the controversial topics of "banning guns" and "regulating guns." The tweets in the dataset are sourced from discussions on Twitter following various shooting incidents in the United States. Amazon Mechanical Turk was used to manually annotate a subset of the tweets relevant to the targets of interest ("banning guns" and "regulating guns") into three classes: *In-Favor*, *Against*, and *Neutral*. The remaining unlabeled tweets are included in the dataset to facilitate studies on semi-supervised learning (SSL) approaches that can help address the scarcity of the labeled data in stance detection tasks. Furthermore, we propose a hybrid approach that combines curriculum-based SSL and Large Language Models (LLM), and show that the proposed approach outperforms supervised, semi-supervised, and LLM-based zero-shot models in most experiments on our assembled dataset. The dataset and code are available on GitHub.[1]

1 Introduction

In the aftermath of the multitude of mass shooting events in the United States, topics regarding tighter gun regulations and control, expanding mental health services, and upgrading security in public places have been strongly debated over political campaigns on various social media platforms. People are generally on two opposing sides with respect to gun regulations and control: those who favor such regulations and those who strongly oppose them. Such differences in opinions on gun regulations have prevailed in society for over four decades, resulting in very little legislative success (Lin and Chung, 2020). Social media, especially Twitter (now X), has been a melting pot for such debates. It would be useful to automatically analyze the intricate dynamics, challenges, and complexities surrounding gun control topics.

Stance detection on social media is an emerging research area in opinion mining for various applications where sentiment analysis may be suboptimal (ALDayel and Magdy, 2021). Stance detection is the task of automatically determining if an author of a text is *In-Favor*, *Against*, or *Neutral* towards a proposition or target (Mohammad et al., 2017). While research on stance detection has seen significant progress in various domains, such as politics, healthcare, and finance (Mohammad et al., 2017; Ghosh et al., 2019; Conforti et al., 2020a; Glandt et al., 2021), there remains a notable gap in analyzing stance dynamics in the context of gun control and regulations on social media, despite the importance of this topic in the contemporary society. This is mainly due to a lack of datasets labeled with respect to stance towards the controversial topics of gun control and regulation.

In this work, we aim to address this need by creating GunStance, a dataset of tweets collected during mass shooting events in the United States. Specifically, we focus on seven shooting events and identify two controversial, interconnected gun-related stance targets, a stronger stance target - "banning guns" and a softer stance target - "regulating guns." We collect tweets posted during those shooting events and filter tweets potentially relevant to each of the targets considered. Note that some tweets can express (either the same or

[*] Equal contribution.

[1] https://github.com/gnikesh/gunstance

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12027–12044
August 11-16, 2024 ©2024 Association for Computational Linguistics

different) stances towards both targets, therefore the sets of tweets filtered for the two targets are not disjoint. Using Amazon Mechanical Turk, we annotate 2,700 tweets according to the author's stance toward the "banning guns" target, and 2,800 tweets according to the author's stance toward the "regulating guns" target. In addition, the GUNSTANCE dataset includes 7,334 unlabeled tweets for the "banning guns" target and 8,824 tweets for the "regulating guns" target. The unlabeled tweets can be utilized by semi-supervised learning approaches. More details about data collection and statistics are provided in Section 3.

In addition to assembling the GUNSTANCE dataset consisting of labeled and unlabeled tweets for the two targets, we first establish strong baselines for this dataset by employing supervised, semi-supervised and zero-shot LLM-based approaches. To push the performance on our task and increase the real-world applicability of the stance detection models, we propose a hybrid SSL/LLM approach that utilizes unlabeled examples from a stream of data (i.e., posted as events unfold), while leveraging the zero-shot capabilities of LLMs to improve the model. Critically, compared to traditional SSL, our method incurs *minimal additional training costs*, due to our novel approach of identifying informative unlabeled examples, and *no inference costs*. Detailed information on our methods is provided in Section 4, while the results of the models are presented and discussed in Section 5.

To summarize, the main contributions of this work consist in providing the research community with a dataset, a novel low-cost hybrid SSL/LLM approach, and comparisons with meaningful baseline models for analyzing social media data with respect to stance towards controversial topics regarding gun regulations and control. Our dataset and proposed hybrid model are designed to help uncover prevalent arguments and opinions related to these topics, and to foster a deeper understanding of the intricate dynamics surrounding gun control and regulations. By enabling the examination of diverse perspectives, our work has the potential to lead to advancements in addressing the challenges posed by gun control and regulation.

2 Related Work

In the last decade, stance detection has been an area of active research and numerous studies focused on introducing novel datasets or proposing stance detection models have been carried out.

2.1 Stance Detection Datasets

SemEval2016 dataset introduced by Mohammad et al. (2016, 2017) consists of tweets about US politics collected during the lead-up to the 2016 US Presidential election. Conforti et al. (2020a) presented a large dataset of approximately 50,000 tweets for stance detection. Villa-Cox et al. (2020) assembled a dataset for identifying stance in Twitter replies and quotes. Multi-target and multilingual datasets are contributed by several works, including (Sobhani et al., 2017a; Zotova et al., 2020; Vamvas and Sennrich, 2020; Lai et al., 2020; Glandt et al., 2021). Several non-English datasets have also been published (Alturayeif et al., 2022; Lozhnikov et al., 2020; Küçük and Can, 2018; Lai et al., 2018). Furthermore, Allaway and McKeown (2020a); Xu et al. (2022); Allaway and McKeown (2020b) curated zero-shot and few-shot stance detection datasets. A summary of existing datasets is provided in Table A1 in the Appendix.

While some previous works have incorporated the generic topic of "guns" as a target in their stance detection datasets (Sakketou et al., 2022; Cinelli et al., 2021), there has been limited differentiation between the intricate relationships surrounding the topics of "banning guns" and "regulating guns." Furthermore, prior studies have not targeted such topics, specifically in the aftermath of mass shooting events in the United States. As another important difference, our dataset includes unlabeled tweets, which makes it possible to study semi-supervised, and zero-shot/few-shot approaches, while most of the existing datasets only include labeled data. Therefore, our dataset provides a novel contribution in terms of stance targets included and approaches enabled.

2.2 Stance Detection Approaches

Support vector machines (SVM) with manually engineered features were established as strong baselines for the SemEval2016 Task 6 datasets (Mohammad et al., 2016, 2017). Subsequently, deep learning approaches, including recurrent neural networks (RNNs) and convolutional neural networks (CNNs), that incorporate both input and target-specific information using attention mechanisms (Augenstein et al., 2016; Du et al., 2017; Zhou et al., 2017; Sun et al., 2018; Siddiqua et al., 2019) outperformed the prior SVM strong baselines. With the advent of transformers (Vaswani et al., 2017) and bidirectional encoder representation from transformers (BERT) (Devlin et al., 2018) in NLP tasks, recent works (Slovikovskaya, 2019; Li and Caragea, 2021) have investigated the use of BERT for stance detection. Ghosh et al. (2019) found BERT to be the best model overall for stance detection on the SemEval2016 Task 6.

In addition to utilizing target-specific information, several studies have demonstrated that incorporating auxiliary information, such as sentiment, emotion, Wikipedia knowledge, etc. can enhance the performance of stance detection models as compared to using only tweet/target information (Mohammad et al., 2017; Sun et al., 2019; Li and Caragea, 2021; Hosseinia et al., 2020a; Xu et al., 2020; Luo et al., 2022; Fu et al., 2022; He et al., 2022). Moreover, in scenarios where the labeled data for the target task is limited, various approaches have been employed to enhance performance. These include techniques such as weak supervision (Wei and Zou, 2019), knowledge distillation (Miao et al., 2020), knowledge graphs (Liu et al., 2021), transfer learning using pre-trained models (Ebner et al., 2019; Hosseinia et al., 2020a), distant supervision (Zarrella and Marsh, 2016), and domain adaptation from a source to a target task (Xue and Li, 2018; Xu et al., 2020).

Following prior works, to address the scarcity of labeled data, we propose a hybrid SSL/LLM approach that makes use of unlabeled data by leveraging the capabilities of LLMs. We compare the proposed hybrid approach with supervised, SSL and LLM-based zero-shot strong baselines on the GunStance dataset.

3 GunStance Dataset

Dataset Collection and Filtering: Our data collection process involved gathering tweets related to mass shooting incidents across different locations in the United States. Using Wikipedia pages that list mass shooting events in the United States in 2021[2] and 2022[3], respectively, we identified and focused on 7 mass shooting events (3 in 2021 and 4 in the first half of 2022), each with at least 5 dead people. The events selected (shown in Table 2) cover a variety of locations/settings (including groceries, a parade, a school, massage parlors, etc).

Past tweets were retrieved in chronological order by using the Twitter's v2 full-archive search endpoint[4]. For each event, we collected tweets posted the day of the event and up to eight weeks after the event. We started the search using a basic set of seed rules specific to the location of each event. As an example, the seed rule for the Robb Elementary School shooting in Uvalde, Texas was: "((uvalde OR texas) shooting) OR uvaldeshooting OR uvaldetexas". The seed rules for the other events are shown in Table A2 in Appendix A.2. The rule for an event was updated with the most frequent hashtags daily during the first week of the event and weekly during the following weeks. On each update of the rule, we identified the 15 most frequent hashtags during the update period (a day or a week), and added those hashtags as additional *OR* terms to the current rule. In addition, we also maintained a list of terms to ignore, specifically terms that were observed among the frequent hashtags but were too general and added noise to the collection (e.g., *mass, dead, breaking, update, usa, america*). The terms added with each update are shown in Table A3 in Appendix A.2. The tweets collected were capped at a maximum of 4,500 tweets per day and 25,000 tweets per week for each event. The final dataset collected included a total of 395,485 original tweets, replies, and quoted tweets (retweets were ignored and are thus not included in this count).

A preliminary analysis of the data showed that 77.48% of the tweets collected (i.e., 306,422 tweets) contained URLs (e.g., links to news articles). While the tweets containing URLs may express a stance towards the targets of interest in our study, given that they link to external resources (many of them related to media outlets), the stance expressed may not be the general public's stance but rather the stance of media outlets, political organizations, etc. As we aim to focus on detecting the general people's stance towards gun regulations, we removed such tweets. We also removed tweets from a list of 2,834 Twitter handles corresponding to media outlets (e.g., PulpNews, NYDailyNews, ABC, etc.). The 2,834 handles were identified based on post frequency (as media outlets have a large number of posts as shown in Figure A.1 in Appendix A.2) and/or the use of the "news" keyword in their username. Together, we filtered out 309,395 tweets out of the total 395,485 collected tweets, leaving us with a set of 86,090 tweets.

To identify tweets most relevant to the targets

[2] https://en.wikipedia.org/wiki/List_of_mass_shootings_in_the_United_States_in_2021

[3] https://en.wikipedia.org/wiki/List_of_mass_shootings_in_the_United_States_in_2022

[4] https://developer.twitter.com/en/docs/twitter-api/tweets/search/introduction

Tweet	Target	Stance (Labelled)
I am following reports of a shooting at #SixFlag in the #Chicago suburb of #Gurnee #Illinois. #GunControl NOW!	Regulating guns	*In-Favor*
Illinois has some of the strictest gun laws in the country, yet policymakers are scrambling to explain the series of missed warning signs and communication failures that led to the Fourth of July mass shooting in Highland Park.	Regulating guns	*Against*
One would wonder how this Highland Park shooting suspect got his guns with all the gun laws in IL to flag with his mental sickness?	Regulating guns	*Neutral*
The shooter in Tulsa bought an AR-15, just hours before this shooting... shooters in both Buffalo and Uvalde... both just turned 18 yrs old... also bought their rifles just before going out to kill... we need a moratorium on sales of AR/AK rifles for now... that w/ save lives...	Banning guns	*In-Favor*
With your logic please explain why Buffalo NY shooting happened if gun laws will stop bad guys shooting?	Banning guns	*Against*
Which of these laws would have stopped the buffalo shooting or the uvalde shooting? In the ulvade shooting the door was open but was supposed to be closed. Can't legislate doors to work!	Banning guns	*Neutral*

Table 1: Examples of Annotated Tweets.

considered, we used *Sentence-BERT* (Reimers and Gurevych, 2019) to embed all tweets in the dataset, as well as the targets (expressed as "guns should be banned" versus "guns should be regulated") and ranked the tweets with respect to the targets based on cosine similarity. The similarity between these two targets themselves was 0.6980. We filtered out the least relevant tweets to these targets by applying a similarity threshold of 0.40. Consequently, we obtained 11,621 tweets for "banning guns" and 13,476 tweets for "regulating guns."

Dataset Preprocessing: As part of the data preprocessing phase, we applied several steps to clean and enhance the dataset. These steps involved removing duplicate tweets, emoticons, and non-ASCII characters from the tweet texts. We also filtered out all user mentions (e.g., @username) from the tweets. We only kept English-language tweets, filtering out all non-English-language ones. For this, we selectively retained tweets with the language code "en" as specified in the tweet metadata during the crawling process. Figure A.2 in Appendix A.3 shows the distribution of the number of words and length of tweets in our preprocessed dataset. A significant portion of the dataset consists of tweets containing 45-50 words, with the peak at 47 words. Similarly, the majority of the tweets have 270-280 characters, with the peak at 278 characters.

Data Annotation: We annotated 2,700 randomly selected tweets for the "banning guns" target and 2,800 randomly selected tweets for the "regulating guns" target using the Amazon Mechanical Turk crowd-sourcing platform. We ran the annotation task in several iterations in order to develop our quality control steps. Initially, we annotated 100 examples internally (50 for "banning guns" and 50 for "regulating guns") to use as a qualification task. Then, we asked all annotators to annotate this set of 100 examples to measure how well they perform.

		"banning guns"				"regulating guns"			
Shooting Events	**Date**	**A**	**F**	**N**	**U**	**A**	**F**	**N**	**U**
Atlanta, Georgia	03/16/2021	19	54	58	340	25	88	25	376
Boulder, Colorado	03/22/2021	73	117	126	572	86	188	78	704
San Jose, California	05/26/2021	20	24	21	221	13	35	9	224
Buffalo, New York	05/14/2022	330	354	477	3,215	373	518	342	4,033
Uvalde, Texas	05/24/2022	190	241	315	2,178	179	383	205	2,610
Tulsa, Oklahoma	06/01/2022	22	61	48	371	11	92	23	392
Highland Park, Illinois	07/04/2022	32	58	60	437	31	51	45	485
Total		686	909	1,105	7,334	718	1,355	727	8.824

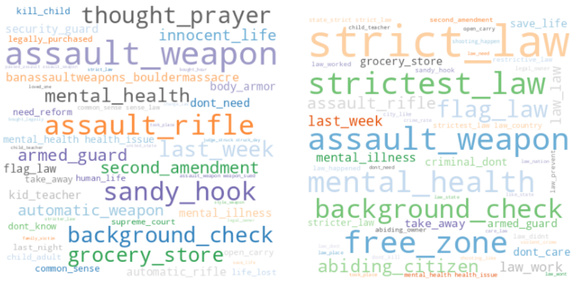
Table 2: Statistics on the labeled/unlabeled tweets for each event and each target. Here, A=*Against*, F=*In-Favor*, N=*Neutral*, and U=*Unlabelled*.

Finally, we annotated all the data using three high-performing annotators and calculated the final label for a tweet using majority voting. We measured the inter-annotator agreement using Krippendorff Alpha, and obtained an average value of α=0.63, indicating moderate agreement.

Table 1 shows examples of annotated tweets. We observe that some tweets have an implicit stance (e.g., "*In-Favor*" and "*Against*" examples for "regulating guns"), whereas others have an explicit stance (e.g., "*In-Favor*" and "*Against*" examples for "banning guns").

Dataset Statistics: Table 2 provides statistics about the labeled and unlabeled tweets for each target and each shooting event. For each event, the table shows the number of tweets annotated as *Against* (A), *In-Favor* (F), *Neutral* (N), as well as the number of *Unlabelled* (U) tweets. The total number of labeled tweets for "banning guns" is 2,700, while the number of unlabeled tweets is 7,334. The number of labeled tweets for "regulating guns" is 2,800, while the number of unlabeled tweets is 8,824.

Figure 1 shows bi-gram word clouds for "*In-Favor*" and "*Against*" stances with respect to the "banning guns" target. While many bi-grams are shared between the two word clouds, we find that people *In-Favor* of banning guns employ more terms such as "thought prayer," "don't need," and "common sense," indicative of their inclination to-



(a) "In-Favor"

(b) "Against"

Figure 1: Bi-gram word clouds for (a) *In-Favor* and (b) *Against* stances for the "banning guns" target.

wards advocating gun prohibition. Conversely, individuals *Against* banning guns employ more often terminology such as "strict law," "free zone," and "mental health." In particular, the term "strict law" is used to argue that despite strict gun laws in states such as Illinois and New York, instances of crime persist, thereby challenging the efficacy of banning guns as a viable solution to prevent gun violence, as shown in the examples in Table 1.

We also performed a lexical analysis of tweets pertaining to the "*Against*" and "*In-Favor*" stances of the "banning guns" target in terms of the five basic moral foundations (harm/care, cheating/fairness, betrayal/loyalty, subversion/authority and degradation/purity) of the Moral Foundation Theory (Araque et al., 2020, 2022; Graham et al., 2013), as shown in the Figure 2 (Left). Each dimension is rated on a continuous scale ranging from 1 to 9 (1 for "harm" and 9 for "care"; 1 for "cheating" and 9 for "fairness"; 1 for "betrayal" and 9 for "loyalty"; 1 for "subversion" and 9 for "authority"; and 1 for "degradation" and 9 for "purity"). We find that individuals *In-Favor* of "banning guns" tend to express sentiments with higher score than those *Against* "banning guns" in the dimensions of care, loyalty, authority, and purity. In contrast, those *Against* "banning guns" tend to articulate their viewpoints more in alignment with the dimension of fairness.

Employing VADER (Valence Aware Dictionary and sEntiment Reasoner) (Hutto and Gilbert, 2014), a lexicon-based sentiment analysis tool designed to capture sentiments in social media contexts, we performed a sentiment analysis on "*In-Favor*" and "*Against*" tweets pertaining to "banning guns", as shown in Figure 2 (Right). We did not find significant differences in sentiment between the "*In-Favor*" and "*Against*" stance categories. This result indicates substantial intricacies and nuances within this topic, thereby posing a significant challenge for machine learning models.

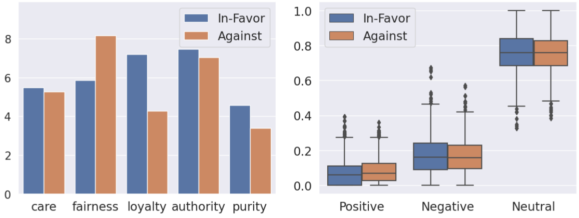


Figure 2: Scores for five basic moral foundation dimensions (Left); and sentiment scores (Right), corresponding to tweets expressing "Against" and "In-Favor" stances towards the "banning guns" target.

		All		Boulder		Buffalo		Uvalde	
		Ban	Reg.	Ban	Reg.	Ban	Reg.	Ban	Reg.
Train	Against	228	243	25	32	110	128	64	56
	In-favor	304	449	39	62	118	168	81	132
	Neutral	368	244	42	24	159	115	105	69
Test	Against	228	247	24	29	110	124	63	64
	In-favor	303	433	39	59	118	171	80	118
	Neutral	369	252	42	29	159	116	105	73
Dev	Against	230	228	24	25	110	121	63	59
	In-favor	302	473	39	67	118	179	80	133
	Neutral	368	231	42	25	159	111	105	63
	Total	2,700	2,800	316	352	1,161	1,233	746	767

Table 3: Statistics of the labeled dataset splits in GUNSTANCE. "All" refers to the tweet count from all events, "Ban" refers to "banning guns" and "Reg." refers to "regulating guns."

Benchmark subsets: To enable comparisons between our baselines and new models developed for the GUNSTANCE dataset, we randomly split each event in the dataset (using stratified sampling) into 33% training (`train`), 33% development (`dev`) and 33% test (`test`) subsets. We assemble benchmark subsets for the whole dataset by combining the `train`, `dev` and `test` subsets, respectively, of the constituent events. Class distributions over each subset are shown in Table 3 under the "All" columns. The table also shows the class distributions for the three largest events in our dataset (specifically, "Boulder", "Buffalo" and "Uvalde"), as we also experiment with the datasets of these specific events in our study.

4 Methodology

4.1 Proposed Approach: Hybrid SSL/LLM

We leverage the Area Under the Margin Self Training (AUM-ST), an SSL model introduced by Sosea and Caragea (2022), and propose a novel approach by aiding AUM-ST (Sosea and Caragea, 2022) with ChatGPT during the training process. We call this model AUM-ST$_{+ChatGPT}$. As shooting events unfold, the web is constantly flooded with new streams of unlabeled data which can be an important source of information to increase model performance and ensure up-to-date information is

Algorithm 1 AUM-ST$_{+LLM}$

```
Require: Labeled data $L = \{(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)\}$, unlabeled data $U = \{\hat{x}_1, \hat{x}_2, \dots \hat{x}_m\}$; $\tau$ confidence threshold; and $\gamma$ AvgMargin threshold.
1: Learn teacher model $\theta^t$ on labeled data minimizing the cross entropy loss
       $L_{\theta^t} = \frac{1}{n} \sum_{i=1}^{n} H(y_i, p(y|\pi(x_i); \theta^t))$
2: Use the teacher model to generate hard pseudo labels for unlabeled examples
       $\hat{y}_i = \text{argmax}(p(y|\pi(\hat{x}_i); \theta^t)), \forall i = 1, \cdots, m$
3: Train model $\theta^{AUM}$ on labeled and unlabeled examples, and monitor the training dynamics of unlabeled examples over $\mathcal{T}$ epochs using the Margin of each example averaged across the epochs: $AvgMargin(\hat{x}_i, \hat{y}_i) = \frac{1}{\mathcal{T}} \sum_{1}^{\mathcal{T}} [z_{\hat{y}_i} - max_{\hat{y}_i != j}(z_j)]$,
   $z_{\hat{y}_i}$ and $z_j$ are the logits corresponding to the pseudo-label $\hat{y}_i$ and the largest other logit.
4: Rank $\hat{x}_i$ based on $\phi(\hat{x}_i, \hat{y}_i) = Abs(AvgMargin(\hat{x}_i, \hat{y}_i))$ and select $k$ unlabeled examples $U^{LLM} = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_k\}$ with low $\phi$ values.
5: Use an LLM model $\theta^{LLM}$ to generate pseudo-labels for examples in $U^{LLM}$ (i.e. those examples that are perceived as ambiguous or hard-to-learn by the model) and overwrite the pseudo-labels generated by the teacher network with those generated by the LLM.
6: Train a student model $\theta^s$ on the labeled and high-confidence, high-average-Margin, and LLM-provided (i.e., the data pseudo-labeled using the LLM) pseudo-labeled data by minimizing the cross entropy loss:
       $L_{\theta^s} = \frac{1}{n} \sum_{i=1}^{n} H(y_i, p(y|\pi(x_i); \theta^s)) + \frac{1}{m} 1(max(p(y|\pi(\hat{x}_i) \geq \tau)) 1(AvgMargin(\pi(\hat{x}_i), \hat{y}_i) > \gamma) 1(\hat{x}_i \notin U^{LLM}) \sum_{i=1}^{m} H(\hat{y}_i, p(y|\Pi(\hat{x}_i); \theta^s))$
       $+ \sum_{\hat{x}_i \in U^{LLM}} H(argmax(p|\pi(\hat{x}_i); \theta^{LLM})), p(y|\Pi(\hat{x}_i); \theta^s))$
7: Use the student as a teacher and go back to Step 2
```

encoded into the models. Typically, SSL methods such as AUM-ST can use this stream of unlabeled data to improve model performance by adding the new data to the unlabeled set of examples. However, due to distribution shifts between the labeled and the recently collected unlabeled data, the model may struggle to generalize to the ongoing events. One alternative is to leverage the knowledge of LLM models, such as ChatGPT, by using them in a zero-shot manner. Unfortunately, using LLMs at large scale in these online, streaming setups is problematic as well, due to their high computational costs. We propose to leverage the high quality predictions of LLMs in combination with AUM-ST to obtain an approach with low computational cost and better generalization performance. We show our approach in Algorithm 1 and highlight our changes to the vanilla AUM-ST in blue.

During SSL training, AUM-ST leverages AUM (Pleiss et al., 2020) to measure the fluctuation of predictions of the model on unlabeled data and to characterize the unlabeled examples based on the correctness of their pseudo-labels (Step 3). Examples with high AUM scores are likely to be pseudo-labeled correctly while low-AUM examples are likely incorrectly pseudo-labeled. On the other hand, examples with neither high nor low AUMs are examples whose pseudo-label correctness is uncertain. We argue that such examples are vital for model learning and providing the model with correct pseudo-labels for these types of examples can significantly improve the performance. Since an AUM of zero indicates high uncertainty of pseudo-label correctness, we first propose to rank all unlabeled examples based on how close their AUM values are to an AUM of zero (Step 4). Then, we select uncertain examples and utilize ChatGPT as an external knowledge source to produce high-quality pseudo-labels for this particular category of unlabeled examples (Step 5). Concretely, at an arbitrary self-training iteration t, given the AUM of each unlabeled example (computed during AUM-ST training), we select a small percentage of unlabeled data with AUMs close to zero and use ChatGPT to obtain pseudo-labels for these examples. We then add these examples to the pseudo-labeled set for student model training (Step 6). We emphasize that the examples selected for LLM pseudo-labeling change from one iteration to another depending on the learning status of the model.

Note that our algorithm involves performing inference with ChatGPT on a very small fraction of the unlabeled set during the training process (only those that the teacher identifies as hard examples and has a high uncertainty on the label). Moreover, our method incurs no additional costs in practice

	Zero-shot Models			Supervised Models							Semi-supervised Models			
	BART-NLI	FLAN	ChatGPT	BERTweet	ATGRU	BiLSTM	GCAE	Kim-CNN	PGCNN	TAN	BERT-NS	FixMatch	AUM-ST	AUM-ST$_{+ChatGPT}$
Target: Ban														
Accuracy	46.11	52.33	64.11	58.07	48.96	49.89	49.96	52.41	51.30	51.07	58.28	58.44	60.78	66.38
Precision	56.42	59.82	64.22	58.45	49.08	49.99	50.15	52.62	41.48	51.19	58.27	58.77	60.86	66.32
Recall	46.11	52.33	64.11	58.07	48.96	49.89	49.96	52.41	51.30	51.07	58.28	58.44	60.78	66.72
F1-Score	38.74	40.93	63.62	57.89	48.95	49.82	49.90	52.43	51.23	51.07	58.25	58.38	60.58	66.50
ECE	N/A	N/A	N/A	3.30	3.88	3.70	3.85	3.18	3.77	3.47	4.67	3.08*	3.99	4.37
Target: Regulate														
Accuracy	42.81	56.65	65.02	63.71	53.51	58.33	55.22	59.33	57.44	57.15	64.10	64.52	65.77	66.37
Precision	57.27	52.92	65.34	63.40	53.11	58.11	54.66	58.57	56.61	56.54	64.50	64.17	65.40	66.22
Recall	42.81	56.65	65.02	63.71	53.51	58.33	55.22	59.33	57.44	57.15	65.09	64.52	65.77	66.05
F1-Score	38.61	47.28	64.67	63.09	53.08	58.13	54.57	57.99	56.73	56.64	64.27	63.90	65.45	66.10
ECE	N/A	N/A	N/A	4.08	3.67	3.94	4.29	3.17	2.49*	3.18	3.73	3.80	3.42	3.75

Table 4: Performance comparison for baseline models trained and evaluated independently for the "ban" and "regulate" targets. The models are trained using the combined training data of all events, and evaluated on the combined test data of all events. The best performance on each row is HIGHLIGHTED. For ECE, the lower the better.

since AUM-ST uses BERTweet as the base model.

4.2 Baseline Models

We employ supervised, semi-supervised, and zero-shot models to establish baseline results for the proposed AUM-ST$_{+ChatGPT}$ model on our dataset. In the supervised learning setting, we only use labeled data to train the models. Similar to prior works in stance detection, e.g., (Glandt et al., 2021; Ghosh et al., 2019; Li and Caragea, 2021), we used **BiLSTM** (Schuster and Paliwal, 1997), **Kim-CNN** (Kim, 2014), **TAN** (Du et al., 2017), **PGCNN** (Huang and Carley, 2019), **AT-GRU** (Zhou et al., 2017), **GCAE** (Xue and Li, 2018), and **BERTweet** (Loureiro et al., 2022) as our supervised baseline models.

Semi-supervised models are capable of leveraging unlabeled data to improve the performance of a teacher model. We utilized **BERT-NS** (Glandt et al., 2021; Xie et al., 2020; Hinton et al., 2015), **FixMatch** (Sohn et al., 2020), and **AUM-ST** (Sosea and Caragea, 2022) as SSL baselines.

For zero-shot (or few-shot) setups, a recent, powerful technique for addressing NLP tasks is to use large pre-trained language models (Brown et al., 2020). We employ **BART-NLI** (Lewis et al., 2019; Li et al., 2023; Williams et al., 2017), **FLAN** (Wei et al., 2021), and **ChatGPT** as zero-shot baselines.

A brief discussion of all the baseline models used is provided in Appendix A.4.

4.3 Experimental Setup

The details of our experimental setup, including hyperparameters used for each of the baseline models are presented in Appendix A.5. Each model was run three times using three different random seeds. The results reported represent averages over the three runs.

We report the accuracy, precision, recall and F1-score as defined in Appendix A.5. In addition, we also report the expected calibration error (ECE). ECE (Naeini et al., 2015) measures how reliably a model's predicted probabilities reflect the true probabilities. The ECE values for a well-calibrated model should be low. Well-calibrated models increase the trustworthiness in the model's predictions which is important to foster a deeper understanding of the intricate dynamics, challenges, and complexities surrounding gun control topics.

5 Results and Discussion

We divide our experiments into three sets based on the data used to train and evaluate the models.

Experiment 1: In this experiment, we train and evaluate models independently for our two targets, "ban" and "regulate". Zero-shot models do not use any training data. Supervised models are trained using the combined labeled training data of all events. Semi-supervised models are trained using the combined labeled training data and unlabeled data of all events. Subsequently, the trained models are evaluated on the combined test data of all events. Table 4 presents the Accuracy, Precision, Recall, F1-score, and Expected Calibration Error (ECE) metrics for the zero-shot, supervised, and semi-supervised baseline models considered, including our proposed SSL/LLM model, when trained/evaluated on "ban"/"regulate" train/test data, respectively.

AUM-ST$_{+ChatGPT}$, our semi-supervised model guided by ChatGPT, achieves the best performance across all metrics except for the ECE metric. Notably, the semi-supervised model FixMatch obtains the lowest ECE score for the "ban" target, while the supervised model PGCNN achieves the lowest ECE score for the "regulate" target. In terms of zero-shot models, ChatGPT outperforms both BART-NLI and FLAN significantly in all metrics.

Event: Boulder														
							Target: Ban							
	Zero-shot Models			Supervised Models							Semi-supervised Models			
	BART-NLI	FLAN	ChatGPT	BERTweet	ATGRU	BiLSTM	GCAE	Kim-CNN	PGCNN	TAN	BERT-NS	FixMatch	AUM-ST	AUM-ST$_{+ChatGPT}$
Accuracy	49.52	54.29	66.67	55.09	56.51	46.67	47.94	50.79	49.21	49.52	57.61	58.41	58.10	67.40
Precision	64.76	72.10	66.64	55.88	56.91	46.00	48.10	50.43	48.95	48.99	57.75	59.20	58.03	67.20
Recall	49.52	54.29	66.67	55.23	56.51	46.67	47.94	50.79	49.21	49.52	57.61	58.41	58.10	67.03
F1-Score	44.40	43.40	65.77	54.93	56.44	45.61	48.89	50.23	48.46	49.18	57.64	57.64	57.98	67.10
ECE	N/A	N/A	N/A	4.45*	5.79	7.30	5.54	9.56	4.49	7.50	6.12	8.47	7.75	8.11
							Target: Regulate							
Accuracy	53.85	58.12	64.96	67.52	54.70	60.68	58.11	62.39	57.26	61.53	67.24	68.09	70.09	68.54
Precision	76.59	46.55	64.68	66.87	55.38	58.89	56.03	60.59	55.41	60.24	65.78	67.58	69.23	68.99
Recall	53.85	58.12	64.96	67.52	54.70	60.68	58.11	62.39	57.26	61.53	67.24	68.09	70.09	68.65
F1-Score	50.38	48.52	63.87	66.95	54.56	59.44	55.75	59.74	55.46	60.65	65.38	66.05	68.91	68.78
ECE	N/A	N/A	N/A	7.38	6.92	6.68	7.36	5.48*	11.32	8.69	7.36	9.27	6.10	7.44
Event: Buffalo														
							Target: Ban							
	BART-NLI	FLAN	ChatGPT	BERTweet	ATGRU	BiLSTM	GCAE	Kim-CNN	PGCNN	TAN	BERT-NS	FixMatch	AUM-ST	AUM-ST$_{+ChatGPT}$
Accuracy	45.74	50.13	62.27	51.93	46.25	48.32	47.80	49.09	49.10	45.22	57.61	52.28	55.04	63.10
Precision	54.92	44.01	62.07	51.97	46.55	48.69	48.36	49.44	49.49	45.53	57.74	53.29	59.51	64.74
Recall	45.74	50.13	62.27	51.94	46.25	48.32	47.80	49.10	49.10	45.22	57.71	52.28	55.03	63.12
F1-Score	38.05	38.41	61.66	51773	45.93	47.70	46.23	48.68	48.18	44.90	57.62	51.53	53.65	63.40
ECE	N/A	N/A	N/A	5.50	6.89	6.81	5.83	4.80*	5.49	6.01	6.63	6.65	7.71	8.99
							Target: Regulate							
Accuracy	39.42	53.53	61.80	60.58	48.66	52.80	51.09	55.47	50.36	50.85	61.71	60.34	61.31	63.54
Precision	49.73	48.24	64.49	61.03	48.34	52.54	51.52	55.74	52.00	50.58	62.56	60.59	61.75	61.22
Recall	39.42	53.53	61.80	60.58	48.27	52.79	51.09	55.47	50.36	50.85	61.71	60.34	61.31	67.50
F1-Score	33.24	44.51	62.23	60.03	48.34	52.25	49.53	54.32	47.61	50.07	60.36	57.82	58.99	64.50
ECE	N/A	N/A	N/A	6.67	6.35	4.21	5.46	5.18	7.62	5.68	6.60	4.99	3.57*	9.32
Event: Uvalde														
							Target: Ban							
	BART-NLI	FLAN	ChatGPT	BERTweet	ATGRU	BiLSTM	GCAE	Kim-CNN	PGCNN	TAN	BERT-NS	FixMatch	AUM-ST	AUM-ST$_{+ChatGPT}$
Accuracy	46.37	53.63	66.94	54.83	43.15	43.95	42.74	49.20	47.98	47.18	56.98	53.76	58.06	67.93
Precision	51.53	72.76	67.63	54.82	44.26	46.04	41.23	50.79	48.31	48.62	57.05	54.20	57.80	68.20
Recall	46.37	53.63	66.94	54.83	43.15	43.95	42.74	49.19	47.98	47.18	56.98	53.76	58.06	68.20
F1-Score	38.30	42.70	66.87	53.79	42.04	43.85	39.98	48.49	47.24	45.65	56.99	52.50	57.20	68.20
ECE	N/A	N/A	N/A	5.19	6.10	11.32	8.46	6.41	9.30	5.03	6.57	5.67	4.34*	6.31
							Target: Regulate							
Accuracy	42.75	57.25	64.71	55.68	47.45	51.37	53.72	50.98	49.02	50.98	55.51	52.03	57.64	65.32
Precision	63.24	71.15	64.04	55.67	47.30	50.58	54.40	49.18	47.28	48.82	56.42	50.66	57.58	66.30
Recall	42.75	57.25	64.71	55.69	47.45	51.37	53.72	50.98	49.02	50.98	55.51	52.03	57.64	65.18
F1-Score	38.65	47.80	64.20	51.45	46.76	49.80	51.20	48.37	46.52	48.08	52.28	48.84	54.80	65.80
ECE	N/A	N/A	N/A	9.60	8.41	5.72*	6.47	8.18	9.39	5.66	11.34	10.60	7.84	11.32

Table 5: Performance comparison for the leave-one-event-out models trained for the "ban" and "regulate" targets and evaluated on "Boulder", "Buffalo" and "Uvalde", respectively. The best performance in each row is HIGHLIGHTED.

For supervised models, we find a significant performance difference in terms of F1-Score between the non-transformer architecture-based models and the BERTweet model. This difference can be linked to the complexity inherent within our dataset, which presents a greater challenge for the supervised models to effectively model such complexities. In contrast, BERTweet was able to learn complexities present in our dataset better than the non-transformer supervised models. However, the zero-shot ChatGPT outperforms the best supervised BERTweet model, underscoring the power of LLMs, such as ChatGPT, in a zero-shot setting on a very challenging task.

Experiment 2: This experiment follows a leave-one-event-out setting. Models are trained for the "ban"/"regulate" targets independently using the training (labeled/unlabeled) data of all events except the held-out event, then tested on this event. The held-out events are "Boulder", "Buffalo", and "Uvalde". We did not consider other events as held-out because of their relatively small size.

Table 5 shows the results of the leave-one-event-out models for the "ban" and "regulate" targets evaluated on "Boulder", "Buffalo", and "Uvalde" events, respectively. Similar to Experiment 1, we find that the proposed AUM-ST$_{+ChatGPT}$ obtains impressive results across multiple events. Notably, AUM-ST$_{+ChatGPT}$ outperforms the vanilla AUM-ST by 9% F1 score on the "ban" target in the Boulder shooting and by almost 10% on the same target in the Buffalo shooting. Additionally, we emphasize that AUM-ST$_{+ChatGPT}$ consistently outperforms ChatGPT in all setups by approximately 2%, an impressive achievement, especially as the supervised baselines struggle on this challenging task. This result shows the value of the unlabeled data relevant to the task at hand, and suggests that our approach effectively leverages such unlabeled data (and the external knowledge captured through ChatGPT) to improve the generalization performance.

Experiment 3: In this experiment, we explore a cross-target setup, where we train the models on one stance target and test them on the other target. Specifically, we train the models on the training data of the "ban" target and evaluate their perfor-

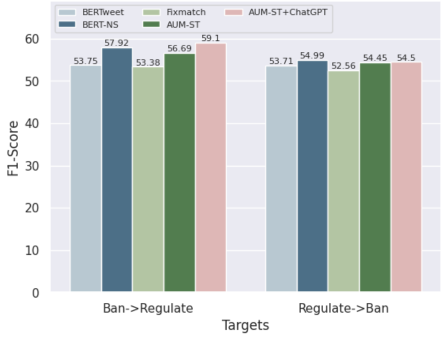


Figure 3: F1-scores of BERTweet, BERT-NS, FixMatch, AUM-ST, and AUM-ST$_{+ChatGPT}$ in the cross-target setup.

mance on the test data of the “regulate” target, and vice versa. This is a particularly difficult task, as differences between stances towards the “ban” and “regulate” targets can be very nuanced. To ensure good coverage with the unlabeled data, the semi-supervised models utilize the unlabeled data from both “ban” and “regulate” targets.

The results of this experiment are shown in Figure 3. As we can see from the figure, AUM-ST$_{+ChatGPT}$ outperforms both the supervised as well as other SSL methods for the “Ban→Regulate” setup and performs similarly with the other approaches for the “Regulate→Ban” setup.

6 Conclusions and Future Work

We presented GunStance, a stance detection dataset containing tweets pertaining to the controversial topics of “banning guns” and “regulating guns” in the aftermath of various shooting events in the United States. The dataset includes manually annotated tweets together with unlabeled tweets to facilitate supervised and semi-supervised learning. We provide baseline results using established and newer supervised and SSL models for stance detection. We also utilize LLMs in a zero-shot setting and find that LLMs, such as ChatGPT, are able to understand the complexities in our dataset and outperform supervised baselines. Furthermore, we propose a curriculum-based semi-supervised approach that utilizes ChatGPT as an external knowledge source on a small subset of hard-to-learn unlabeled examples when the SSL model is less confident, pushing both performance and computational efficiency of SSL methods, such as AUM-ST.

The stance detection task addressed is not only timely (given the large number of shooting events), but also challenging. Our work has the potential to enable researchers to use machine learning models to analyze mass shooting data, gain insights into prevalent arguments and opinions, and develop a deeper understanding of the complexities surrounding gun control and regulations.

While there are other targets of interest in the context of mass shooting and gun control topics (e.g., buyback programs, background checks, mental health, etc.), we believe our work represents an important step towards understanding the general public’s stance towards two extremely important and interconnected gun control targets. Other related targets are left as part of future work.

An evaluation of the proposed AUM-ST$_{+ChatGPT}$ model on other datasets is beyond the scope of this current study, but makes another excellent topic for future work. We also plan to focus on domain adaptation and transfer learning techniques that allow stance detection models to be adapted to different domains or targets. Finally, understanding similarities and differences between explicit and implicit stances in terms of LLM/SSL stance detection abilities is also of interest.

We make our dataset and code available on GitHub to spur further research in this area.

Acknowledgements

We thank the National Science Foundation for support from grants IIS-2107487, IIS-1741345 and ITE-2137846 which supported the research and the computation in this study. We also thank our reviewers for their insightful feedback and comments.

7 Limitations

Our work has a few limitations that should be acknowledged. First, the availability of data for certain shooting events is limited (e.g., San Jose, California and Tulsa, Oklahoma), while other events benefit from more extensive coverage (e.g., Buffalo, New York, and Uvalde, Texas), causing an imbalance in the amount of data that we were able to crawl for some events. Second, it is important to note that our dataset is comprised of publicly available tweets, which are often generated by a subset of users who are more active and vocal on social media. Therefore, the opinions expressed in these tweets may not necessarily represent the views of the majority. These limitations highlight other potential avenues for future research that could address the scarcity of data, improve the identification of implicit and explicit stances, and explore methods to capture a more diverse range of perspectives on the topic of gun control and regulations.

In our study, we removed tweets that contain external URLs as the goal was to focus on the general public's stance as opposed to the stance of media outlets, political organizations, etc. However, it would be interesting to detect the stance of such global entities and understand how their stance influences the stance of their followers.

As another limitation, we should note that the dataset that was used to pre-train ChatGPT is unknown. It is not possible to assess the extent to which tweets in our dataset may have been seen by ChatGPT and the effect on the performance. Information about the dataset could provide light on the impressive abilities of ChatGPT on the challenging task of detecting stance towards two important gun control and regulation targets.

Ethics Statement

Our work on the GUNSTANCE dataset is conducted with a strong commitment to ethical considerations. We prioritize privacy and collect only publicly available tweets and adhere to relevant guidelines for annotations and sharing the datasets. We do not share any private user information (or other metadata for that matter). Considering the nature of the dataset which consists of tweets posted during mass shooting events, we want to emphasize that we did remove tweets with external URLs from the dataset, which ensures that there are no graphic descriptions in the tweets in our dataset.

Additionally, we emphasize that our annotators were paid fairly. We prioritize transparency and open communication in our work and make our dataset and methodologies available to the research community, enabling peer review, validation, and further advancements in the field. While our models can be misused, as many other models in the literature, they are meant to be used to predict the general public's stance towards gun control targets, and can be very helpful in that respect.

References

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Emily Allaway and Kathleen McKeown. 2020a. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Emily Allaway and Kathleen McKeown. 2020b. Zero-shot stance detection: a dataset and model using generalized topic representations. *arXiv preprint arXiv:2010.03640*.

Nora Saleh Alturayeif, Hamzah Abdullah Luqman, and Moataz Aly Kamaleldin Ahmed. 2022. Mawqif: A multi-label arabic dataset for target-specific stance detection. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 174–184.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2020. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-based systems*, 191:105184.

Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2022. Libertymfd: A lexicon to assess the moral foundation of liberty. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, pages 154–160.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. *arXiv preprint arXiv:1606.05464*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020a. Will-they-won't-they: A very large dataset for stance detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724, Online. Association for Computational Linguistics.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020b. Will-they-won't-they: A very large dataset for stance detection on twitter. *arXiv preprint arXiv:2005.00388*.

RV Cox, Sumeet Kumar, Matthew Babcock, and Kathleen M Carley. 2020. Stance in replies and quotes (srq): A new dataset for learning stance in twitter conversations.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention networks. International Joint Conferences on Artificial Intelligence.

Seth Ebner, Felicity Wang, and Benjamin Van Durme. 2019. Bag-of-words transfer: Non-contextual techniques for multi-task learning. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 40–46.

Yujie Fu, Xiaoli Li, Yang Li, Suge Wang, Deyu Li, Jian Liao, and Jianxing Zheng. 2022. Incorporate opinion-towards for stance detection. *Knowledge-Based Systems*, 246:108657.

Shalmoli Ghosh, Prajwal Singhania, Siddharth Singh, Koustav Rudra, and Saptarshi Ghosh. 2019. Stance detection in web and social media: a comparative study. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*, pages 75–87. Springer.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in covid-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Long Papers)*, volume 1.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Omama Hamad, Ali Hamdi, Sayed Hamdi, and Khaled Shaban. 2022. Steducov: an explored and benchmarked dataset on stance detection in tweets towards online education during covid-19 pandemic. *Big Data and Cognitive Computing*, 6(3):88.

Zihao He, Negar Mokhberian, and Kristina Lerman. 2022. Infusing knowledge from wikipedia to enhance stance detection. *arXiv preprint arXiv:2204.03839*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Marjan Hosseinia, Eduard Dragut, and Arjun Mukherjee. 2020a. Stance prediction for contemporary issues: Data and experiments. *arXiv preprint arXiv:2006.00052*.

Marjan Hosseinia, Eduard Dragut, and Arjun Mukherjee. 2020b. Stance prediction for contemporary issues: Data and experiments. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 32–40, Online. Association for Computational Linguistics.

Binxuan Huang and Kathleen M Carley. 2019. Parameterized convolutional neural networks for aspect level sentiment classification. *arXiv preprint arXiv:1909.06276*.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

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.

Jonathan Kobbe, Ioana Hulpuș, and Heiner Stuckenschmidt. 2020. Unsupervised stance detection for arguments from consequences. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 50–60.

Dilek Küçük and Fazli Can. 2018. Stance detection on tweets: An svm-based approach. *arXiv preprint arXiv:1803.08910*.

Mirko Lai, Alessandra Teresa Cignarella, Delia Irazú Hernández Farías, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. Multilingual stance detection in social media political debates. *Computer Speech & Language*, 63:101075.

Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2018. Stance evolution and twitter interactions in an italian political debate. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*, pages 15–27. Springer.

Samuli Laine and Timo Aila. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yingjie Li and Cornelia Caragea. 2021. Target-aware data augmentation for stance detection.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021a. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365, Online. Association for Computational Linguistics.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021b. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365.

Yingjie Li, Chenye Zhao, and Cornelia Caragea. 2023. Tts: A target-based teacher-student framework for zero-shot stance detection. In *Proceedings of the ACM Web Conference 2023*, pages 1500–1509.

Yu-Ru Lin and Wen-Ting Chung. 2020. The dynamics of twitter users' gun narratives across major mass shooting events. *Humanities and social sciences communications*, 7(1):1–16.

Rui Liu, Zheng Lin, Yutong Tan, and Weiping Wang. 2021. Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3152–3157.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, page 100017.

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. *arXiv preprint arXiv:1907.11692*.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. *arXiv preprint arXiv:2202.03829*.

Nikita Lozhnikov, Leon Derczynski, and Manuel Mazzara. 2020. Stance prediction for russian: data and analysis. In *Proceedings of 6th International Conference in Software Engineering for Defence Applications: SEDA 2018 6*, pages 176–186. Springer.

Yun Luo, Zihan Liu, Yuefeng Shi, Stan Z Li, and Yue Zhang. 2022. Exploiting sentiment and common sense for zero-shot stance detection. *arXiv preprint arXiv:2208.08797*.

Geoffrey J McLachlan. 1975. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association*, 70(350):365–369.

Lin Miao, Mark Last, and Marina Litvak. 2020. Twitter data augmentation for monitoring public opinion on covid-19 intervention measures. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Fuqiang Niu, Min Yang, Ang Li, Baoquan Zhang, Xiaojiang Peng, and Bowen Zhang. 2024. A challenge dataset and effective models for conversational stance detection. *arXiv preprint arXiv:2403.11145*.

Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. 2020. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, 33:17044–17056.

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*. Association for Computational Linguistics.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29:1163–1171.

Flora Sakketou, Allison Lahnala, Liane Vogel, and Lucie Flek. 2022. Investigating user radicalization: A novel dataset for identifying fine-grained temporal shifts in opinion. *arXiv preprint arXiv:2204.10190*.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2019. Tweet stance detection using an attention based neural ensemble model. 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 1868–1873.

Iustin Sirbu, Tiberiu Sosea, Cornelia Caragea, Doina Caragea, and Traian Rebedea. 2022. Multimodal semi-supervised learning for disaster tweet classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2711–2723.

Valeriya Slovikovskaya. 2019. Transfer learning from transformers to fake news challenge stance detection (fnc-1) task. *arXiv preprint arXiv:1910.14353*.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017a. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017b. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557, Valencia, Spain. Association for Computational Linguistics.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608.

Tiberiu Sosea and Cornelia Caragea. 2022. Leveraging training dynamics and self-training for text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4750–4762, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Christian Stab, Tristan Miller, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources using attention-based neural networks. *arXiv preprint arXiv:1802.05758*.

Qingying Sun, Zhongqing Wang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Stance detection via sentiment information and neural network model. *Frontiers of Computer Science*, 13:127–138.

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In *Proceedings of the 27th international conference on computational linguistics*, pages 2399–2409.

Jannis Vamvas and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection. *arXiv preprint arXiv:2003.08385*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ramon Villa-Cox, Sumeet Kumar, Matthew Babcock, and Kathleen M Carley. 2020. Stance in replies and quotes (srq): A new dataset for learning stance in twitter conversations. *arXiv preprint arXiv:2006.00691*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698.

Chang Xu, Cécile Paris, Surya Nepal, Ross Sparks, Chong Long, and Yafang Wang. 2020. Dan: Dual-view representation learning for adapting stance classifiers to new domains. *arXiv preprint arXiv:2003.06514*.

Hanzi Xu, Slobodan Vucetic, and Wenpeng Yin. 2022. Openstance: Real-world zero-shot stance detection. *arXiv preprint arXiv:2210.14299*.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. *arXiv preprint arXiv:1805.07043*.

Guido Zarrella and Amy Marsh. 2016. Mitre at semeval-2016 task 6: Transfer learning for stance detection. *arXiv preprint arXiv:1606.03784*.

Bowen Zhang, Daijun Ding, and Liwen Jing. 2022. How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548*.

Yiwei Zhou, Alexandra I Cristea, and Lei Shi. 2017. Connecting targets to tweets: Semantic attention-based model for target-specific stance detection. In *Web Information Systems Engineering–WISE 2017: 18th International Conference, Puschino, Russia, October 7-11, 2017, Proceedings, Part I 18*, pages 18–32. Springer.

Elena Zotova, Rodrigo Agerri, Manuel Nuñez, and German Rigau. 2020. Multilingual stance detection in tweets: The catalonia independence corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1368–1375.

A Appendix

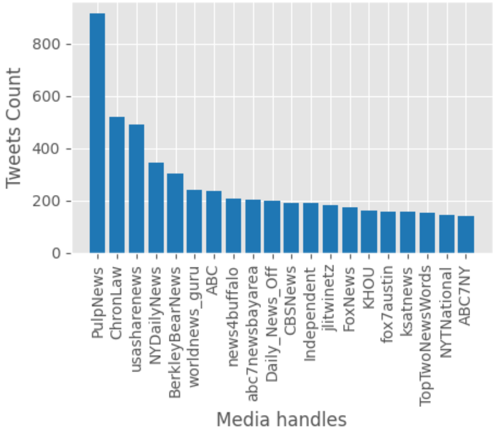
A.1 Existing Stance Detection Datasets

Table A1 shows the comparison of various Target-specific Stance detection datasets.

A.2 Crawling Rules/Terms

We use a set of seed rules[5] as shown in Table A2 to initiate the crawling of tweets for each event. Table A3 contains the additional terms we added to the rule based on frequent hashtags. In collecting these additional terms, we filtered out the terms containing *news* as a substring and terms like *mass, dead, breaking, update, usa, america* to reduce noise in the data.

Figure A.1: Plot of the top 20 news media outlets ordered by the number of tweets posted.



A.3 Tweets Statistics

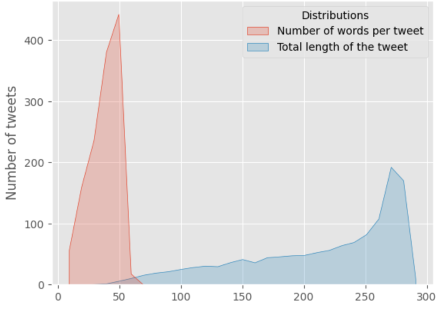


Figure A.2: Distributions for the number of words per tweet (red) and total character length of the tweet (blue). The mode of the word distribution is 47 words, while the mode of the length distribution is 278 characters.

A.4 Baseline Models

A.4.1 Supervised Baseline Models

BiLSTM: We utilize Bi-Directional Long Short-Term Memory Networks (BiLSTMs) (Schuster and Paliwal, 1997) as a supervised model. The model takes tweets as input and is trained to predict the stance toward a target without explicitly incorporating the target information.

Kim-CNN: We employ Convolutional Neural Networks (CNNs) for text, proposed by (Kim, 2014), as another supervised model. Similar to BiLSTM, the CNN model takes tweets as input and is trained to predict the stance towards a target without directly incorporating target information.

PGCNN: We utilize parameterized Convolutional Neural Networks proposed by (Huang and Carley, 2019). PGCNN utilizes parameterized filters and parameterized gates to capture the aspect specific features for sentiment classification.

TAN: We utilize Target-specific Attention Networks (Du et al., 2017), an attention-based BiLSTM model. As opposed to the BiLSTM and Kim-CNN models, TAN identifies features specific to the target of interest by explicitly incorporating the target information.

ATGRU: The Bi-Directional Gated Recurrent Unit (GRU) Network with Token-Level Attention Mechanism (Zhou et al., 2017) is an attention-based Bi-GRU model that also explicitly uses target information. It identifies specific target features using the attention mechanism.

GCAE: The Gated Convolutional Network with Aspect Embedding (Xue and Li, 2018) is a CNN-based model which, in addition to tweets, incorporates target information and also employs a gating mechanism to filter out information that is not related to the target.

BERTweet: We employ a pre-trained Transformer based model from (Loureiro et al., 2022). This model is based on RoBERTa (Liu et al., 2019) and is pre-trained on a large corpus of tweets. We fine-tuned the model using the labeled data.

Note: We have chosen to use well-established supervised methods for stance detection in our comparisons, as our focus was on zero-shot and semi-supervised methods, including a novel SSL/LLM hybrid model. While more recent, less established supervised models may perform better than those

[5] https://developer.twitter.com/en/docs/twitter-api/tweets/search/integrate/build-a-query

Dataset	Source	#Targets	Targets	Size
SemEval16 (Mohammad et al., 2016)	Twitter	6	Atheism, Climate change is a real concern, Feminist Movement, Hillary Clinton, Legalization of Abortion, Donald Trump	48,700
MultiTarget (Sobhani et al., 2017b)	Twitter	4	Clinton-Sanders, Clinton-Trump, Cruz-Trump	4,455
Cross Topic Argument Mining (AM) (Stab et al., 2018)	News reports, editorials, blogs, debate forums, and encyclopedia articles	8	Abortion, Cloning, Death penalty, Gun control, Marijuana legalization, Minimum wage, Nuclear energy, School uniforms	25,492
WTWT (Conforti et al., 2020b)	Twitter	5	Merger of companies: Cigna-Express Scripts, Aetna-Humana, CVS-Aetna, Anthem-Cigna, Disney-Fox	51,284
Stance in Replies and Quotes (SRQ) (Cox et al., 2020)	Twitter	4	General Terms, Iran Deal, Santa Fe, Shooting Student Marches	5,221
VAST (Allaway and McKeown, 2020b)	Comments	5,634	Various targets from comments collected from The New York Times 'Room for Debate' section, part of the Argument Reasoning Comprehension (ARC) Corpus	18,545
Procon20 (Hosseinia et al., 2020b)	Posts from procon.org	419	Issues (questions) and their related responses for Controversial issues	6,094
StArCon (Kobbe et al., 2020)	Debatepedia	190	Various topics	5,398
COVID-19 (Glandt et al., 2021)	Twitter	4	Anthony S. Fauci, M.D., Keeping Schools Closed, Stay at Home Orders, Wearing a Face Mask	6,133
P-Stance (Li et al., 2021a)	Twitter	3	Trump, Biden, Sanders	21,574
StEduCov (Hamad et al., 2022)	Twitter	1	Online Education during COVID-19 Pandemic	16,572
MT-CSD (Niu et al., 2024)	Reddit	5	Bitcoin, Tesla, SpaceX, Biden, Trump	15,876
GunStance (Ours)	Twitter	2	Banning guns, Regulating guns	5,500

Table A1: Summary of Stance Detection Datasets

Location	Seed Rules
Atlanta, Georgia	*((atlanta OR ackworth OR georgia OR atlantaga OR ackworthga) (shooting OR massacre))*
Boulder, Colorado	*((boulder OR colorado) (shooting OR massacre OR strong)) OR #bouldershooting OR #BoulderMassacre OR #BoulderStrong*
San Jose, California	*(((san jose) OR sanjose OR (santa clara) OR (santaclara) OR (vta)) (shooting OR #californiashooting))*
Buffalo, New York	*((buffalo OR ny OR (new york) OR newyork) shooting) OR buffaloshooting*
Uvalde, Texas	*((uvalde OR texas) shooting) OR uvaldeshooting OR uvaldetexas*
Tulsa, Oklahoma	*((tulsa OR (tulsaok) OR oklahoma OR (tulsaoklahoma)) shooting) OR tulsahooting*
Highland Park, Illinois	*((highlandpark OR (highland park) OR illinois OR (highlandparkillinois)) shooting) OR highlandparkshooting*

Table A2: Seed rules used to crawl tweets for each event.

we have used in our comparisons, the current trend in the literature is to focus on models that work with a small amount of data (if any), while leveraging pre-trained models and/or unlabeled data (e.g., zero-shot, few-shot, semi-supervised). We have followed this trend since the zero-shot and semi-supervised scenarios fit very well with the dataset that we assembled.

A.4.2 Semi-supervised Baselines

Given the availability of a significant amount of unlabeled data for each target in the dataset, we also investigate semi-supervised models capable of leveraging such unlabeled data. The following semi-supervised baseline models were employed:

BERT-NS: Similar to Glandt et al. (2021), we employ the self-training with Noisy Student (NS) method (Xie et al., 2020), a semi-supervised learning approach that leverages self-training and knowledge distillation (Hinton et al., 2015) to enhance the performance of a teacher model using unlabeled data. Initially, a teacher model is trained using the available labeled data and then utilized to generate pseudo-labels for the unlabeled data. Subsequently, a noisy student model is trained using both the labeled and pseudo-labeled data. This process can be iterated multiple times by replacing the teacher with the student. In our study, both the teacher and the student models represent fine-tuned BERTweet models.

FixMatch: Introduced by Sohn et al. (2020), FixMatch combines two common SSL techniques: consistency regularization (Sajjadi et al., 2016; Laine and Aila, 2016) and pseudo-labeling (McLachlan, 1975). As part of the FixMatch procedure, two augmented versions are created for the unlabeled examples: a weakly augmented version of an unlabeled example is passed through the model and the prediction is converted into a pseudo-label; then, the model is trained to predict that pseudo-label when fed with a strongly augmented version of the same unlabeled example. While originally designed for image classification, FixMatch has also shown good performance when used on

Location	Week 1	Weeks 2-7
Atlanta, Georgia	asian, massageparlor, activeshooter, woodstock, stopasianhate, stopaapihate, asianlivesmatter, atlantaspa, stopasianhatecrimes, aapi, hatecrime, racism, asianamericans, atlantamassacre	boulder, thalapathy65, master, thalapathyvijay
Boulder, Colorado	bouldercolorado, kingsoopers, guncontrolnow, activeshooter, coloradoshooting, police, zfgvideography, gunreform, gunsense, guncontrol, gunreformnow, gunviolence, atlantamassacre, nra	banassaultweapons, copolitics
San Jose, California	california, gunreformnow, guncontrolnow, massshooting, sanjosestrong, gunviolence, sanjoseshooting, guncontrol, gunsense, jose, transit, transportation, sf, sanfrancisco	sfmta, bart, muni
Buffalo, New York	buffalony, buffalostrong, massshooting, guncontrolnow, racism, hatecrime, paytongendron, whitesupremacist, whitesupremacy, buffalonewyork, supermarket, buffalomassacre, blacklivesmatter	uvalde, ushistory, texas, nra, endthefilibuster, gunviolence
Uvalde, Texas	guncontrolnow, guncontrol, robbelementaryschool, gunviolence, uvaldetx, uvaldemassacre, nra, robbelementary, texasschoolmassacre, uvaldeschoolmassacre	specialsessionnow, uvaldepolice, gunreformnow, endgunviolence, banassaultweaponsnow, enough, dosomething, uvaldecoverup, uvaldepolicecowards
Tulsa, Oklahoma	guncontrolnow, gunreformnow, guncontrol, uvalde, massshooting, enoughisenough, tulsamassacre, hospital,tulsahospital,care2,gunviolence	
Highland Park, Illinois	chicago, highlandparkparade, 4thofjuly, highland, massshooting, guncontrolnow, robertcrimo, guncontrol, highlandparkmassacre, prolifemyass, ushistory	park

Table A3: Additional terms added to the rule based on frequent hashtags.

text data in a multimodal setup (Sirbu et al., 2022). Hence, we also employ FixMatch as a strong semi-supervised baseline.

AUM-ST: AUM-ST is a novel SSL method that takes advantage of the training dynamics of unlabeled data to enhance a model's performance. This framework extends the concept of Area Under the Margin (Pleiss et al., 2020) to unlabeled data to identify potentially incorrect pseudo-labels. The authors show that removing examples with low AUM scores and using data augmentations such as synonym replacement, switchout, and back-translations can significantly boost the generalization performance.

A.4.3 Zero-shot Baseline Models

BART-NLI: Natural Language Inference (NLI) is the task of determining whether a given hypothesis entails, contradicts or is neutral to a given premise. As shown by (Li et al., 2023), the BART (Lewis et al., 2019) model pre-trained on the MultiNLI dataset (Williams et al., 2017) can be used for zero-shot stance detection by creating a semantic mapping between the stance labels and the NLI labels. Thus, we consider BART-NLI as a strong zero-shot baseline for our dataset.

FLAN: The main idea is to formulate the task in such a way that the language model generates the answer by completing the text. FLAN (Wei et al., 2021) improves the zero-shot learning capabilities of language models by introducing *instruction tuning*, and we consider it as another strong zero-shot baseline in our experiments.

ChatGPT: One of the most notable large language models (LLMs) that has been receiving increasing interest from the research community since its release in December 2022 is ChatGPT, as it demonstrated impressive performances in various domains (e.g. healthcare, education, scientific research, etc.) and on diverse downstream NLP tasks (e.g. question answering, text classification, text generation, code generation, etc.)(Liu et al., 2023). In particular, (Zhang et al., 2022) shows that ChatGPT obtains state-of-the-art results on commonly used stance detection datasets such as SemEval-2016 (Mohammad et al., 2016) and P-Stance (Li et al., 2021b), so we consider it as another strong zero-shot baseline for our dataset.

A.5 Experimental Setup

A.5.1 Hyperparameters

To determine suitable hyperparameters for the models, the development set was utilized. We utilized the Adam optimizer for all the supervised and semi-supervised models. However, for the supervised model, excluding BERTweet, we used a learning rate of 1e-5, weight decay of 4e-5, gradient clipping with the maximum norm of 4.0, and a dropout rate of 0.5 to the linear classification layer. The training process of those models (except for BERTweet) involved 120 epochs, with a mini-batch size of 32 for each iteration. For all models, we performed 3 runs (starting with different random seeds) and report average scores. Further details regarding the specific hyperparameters for each model are presented below:

BiLSTM, ATGRU, TAN: Each of these Recurrent Neural Network models used a hidden unit with 512 dimensions and a dropout of 0.2.

GCAE, Kim-CNN: These CNN-based models utilized filters of width 2, 3, 4, and 5. Each filter width was associated with 25 feature maps. After the convolutional layers, a linear classifier with a hidden dimension of 128 was employed.

BERTweet: This model was trained with a learning rate of 2e-5 using a linear scheduler with warmup steps equivalent to 10% of the total training steps. A batch size of 32 was employed, and the model was trained for 100 epochs. Early stopping criteria based on the F1-score were utilized with patience of 10 epochs. The model was trained using the Adam optimizer and cross-entropy loss.

BERT-NS: For BERT-NS, both teacher and student models were trained with hyperparameters similar to BERTweet. The confidence threshold for selecting pseudo-labels was set to 0.95.

AUM-ST: We utilized a supervised batch size of 32 and an unsupervised batch size of 128. To filter unlabeled data, we applied a threshold of 0.7 and an augmentation percentile of 0.5. The self-training process consisted of 50 steps. As augmentation strengths, we employed a mix weak augmentation strength of 0 and a max weak augmentation strength of 2. For strong augmentation, we utilized a min strength of 3 and a max strength of 40.

FixMatch: We utilized the BERTweet model in FixMatch. For the FixMatch-specific parameters, we used an unlabeled loss weight $\lambda_u = 1$, a ratio between unlabeled and labeled examples $\mu = 7$ and a threshold $\tau = 0.7$ for using only the pseudo-labels with high confidence. For the text augmentation, we used the same techniques and strengths employed for AUM-ST.

A.5.2 Prompt engineering

In the case of the zero-shot models employed, it is important to pay close attention to the design of the prompts used to obtain predictions.

BART-NLI: Similar to (Li et al., 2023), we consider a mapping between predicting stance labels (*Against*, *In-Favor*, *Neutral*) and the task of predicting entailment labels (*Contradiction*, *Entailment*, *Neutral*), by using the tweet as a premise and designing a prompt that contains the target for the hypothesis. We experimented with several prompt templates and chose “I think [target]!” as the best one, where [target] is either “guns should be banned” or “guns should be regulated.” We use the BART-large[6] version of the model.

FLAN: We experimented with several prompt templates and chose “[tweet] What is the stance of the writer? (A) [neg_target] (B) [target] (C) neutral” as the best performing one, where [target] is either “guns should be banned” or “guns should be regulated” and [neg_target] is the corresponding “guns should not be banned” or “guns should not be regulated.” We use the FLAN-T5 XXL[7] model.

ChatGPT: We use the Chat Completions API [8] provided by OpenAI. For the system message that sets the behaviour of the assistant we use the instruction ’You will be provided with a tweet, and your task is to classify its stance as "[target]", "neutral", or "[neg_target]".’ Then, for the user message we provide the request ’[tweet]’. The ChatGPT version used is ’gpt-3.5-turbo-0613’. Because the answer of the model may come in different forms (e.g. "guns should be banned", "Stance: guns should be banned"), the final label is considered the one stance ("[target]", "neutral", or "[neg_target]") that is a substring of the answer given by the model. When this is true for neither of them (e.g. "I’m sorry, but I can’t assist with that."), the final label for evaluation purposes is considered to be "neutral". In the streaming pipeline defined by AUM-ST$_{+LLM}$, the case where zero or more than one stances are substrings of the answer, We consider ChatGPT has provided an **inconclusive** label, so that tweet won’t be used. This happens for only about 0.5% of the cases.

A.5.3 Proposed Hybrid SSL/LLM Approach

For AUM-ST$_{+ChatGPT}$, we use the same setups as AUM-ST and we generate the ChatGPT pseudo-labels in the same manner as the zero-shot baseline. Additionally, at each self-training iteration we select around 5% of unlabeled examples (k in Step

[6] https://huggingface.co/facebook/bart-large-mnli

[7] https://huggingface.co/google/flan-t5-xxl

[8] https://platform.openai.com/docs/guides/text-generation/chat-completions-api

4 of the algorithm) with AUMs close to zero to be passed through ChatGPT.

A.5.4 Evaluation Metrics

In order to evaluate the supervised, semi-supervised and zero-shot models, we employ commonly used metrics for classification tasks: Accuracy, Precision, Recall, and F1 score. We denote true positives, true negatives, false positives, and false negatives by TP, TN, FP, and FN, respectively. **Accuracy** measures the overall correctness of the model's predictions by calculating the ratio of correctly predicted instances to the total number of instances. It is calculated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

Precision measures the proportion of correctly predicted positive instances out of all the instances that were predicted as positive. It is calculated as:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

Recall measures the proportion of correctly predicted positive instances out of all the actual positive instances and is calculated as:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

F1-Score is a commonly used metric that combines precision and recall into a single metric, providing a trade-off between these two metrics. The F1-score is calculated as:

$$\text{F1-score} = \frac{2 * TP}{2 * TP + FP + FN} \quad (4)$$

We also utilized the **Expected Calibration Error (ECE)** (Naeini et al., 2015) metric to measure how reliably a model's predicted probabilities reflect the true probabilities. ECE quantifies the difference between the predicted confidence of a model and its accuracy. A well-calibrated model should have a low ECE, implying that the model's predicted probabilities more accurately reflect the true probabilities. The ECE is calculated as:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)| \quad (5)$$

where, n is the number of samples, M is the equally-spaced bins, B_m is the set of indices of samples whose prediction confidence falls into the respective bins, and acc(B_m) and conf(B_m) represent the accuracy and confidence within the bin B_m, respectively (Guo et al., 2017).