Sarcasm Detection as a Catalyst: Improving Stance Detection with Cross-Target Capabilities

Gibson Nkhata, Shi Yin Hong, Susan Gauch
Department of Electrical Engineering & Computer Science
University of Arkansas
Fayetteville, AR 72701, USA
Emails: gnkhata@uark.edu, syhong@uark.edu, sgauch@uark.edu

***Abstract*—Stance Detection (SD) in social media has become a critical area of interest due to its applications in social, business, and political contexts, leading to increased research within Natural Language Processing (NLP). However, the subtlety, nuance, and complexity of texts sourced from online platforms, often containing sarcasm and figurative language, pose significant challenges for SD algorithms in accurately determining the author's stance. This paper addresses these challenges by employing sarcasm detection as an intermediate-task transfer learning approach specifically designed for SD. Additionally, it tackles the issue of insufficient annotated data for training SD models on new targets by conducting many-to-one Cross-Target SD (CTSD). The proposed methodology involves fine-tuning BERT and RoBERTa models, followed by sequential concatenation with convolutional layers, Bidirectional Long Short Term Memory (BiLSTM), and dense layers. Rigorous experiments are conducted on publicly available benchmark datasets to evaluate the effectiveness of our transfer-learning framework. The approach is assessed against various State-Of-The-Art (SOTA) baselines for SD, demonstrating superior performance. Notably, our model outperforms the best SOTA models in both in-domain SD and CTSD tasks, even before the incorporation of sarcasm-detection pre-training. The integration of sarcasm knowledge into the model significantly reduces misclassifications of sarcastic text elements in SD, allowing our model to accurately predict 85% of texts that were previously misclassified without sarcasm-detection pre-training on in-domain SD. This enhancement contributes to an increase in the model's average macro F1-score. The CTSD task achieves performance comparable to that of the in-domain task, despite using a zero-shot fine-tuning approach, curtailing the lack of annotated samples for training unseen targets problem. Furthermore, our experiments reveal that the success of the transfer-learning framework depends on the correlation between the lexical attributes of the intermediate task (sarcasm detection) and the target task (SD). This study represents the first exploration of sarcasm detection as an intermediate transfer-learning task within the context of SD, while also leveraging the concatenation of BERT or RoBERTa with other deep-learning techniques. The proposed approach establishes a foundational baseline for future research in this domain.**

***Keywords*-Stance detection; sarcasm detection; transfer learning; BERT; RoBERTa.**

I. Introduction

This paper extends our previous research on intermediate-task transfer learning, specifically, leveraging sarcasm detection to enhance Stance Detection (SD) [1]. In our prior work, we focused on pretraining models on sarcasm detection before fine-tuning them on SD, utilizing in-domain training data of SD targets. This study further explores SD from two perspectives: in-domain SD, where a single target is used for training and evaluation, and Cross-Target SD (CTSD), which involves training a model on one or more targets and evaluating it on a different target. CTSD represents the latest research direction in this area.

The proliferation of the Internet and social media platforms such as Twitter (X), Facebook, microblogs, discussion forums, and online reviews has significantly altered how individuals communicate and share information [2][3]. These platforms allow users to express opinions and engage with global audiences on various topics, including current trends, products, and politics [4]–[6]. The vast amount of discourse generated on these platforms provides valuable data for Natural Language Processing (NLP) tasks, particularly SD.

SD is the automated identification of an individual's stance based solely on their utterance or written material [5][7]–[9]. Stance refers to the expression of a speaker's or author's position, attitude, or judgment toward a specific topic, target, or proposition [6][10]. Stance labels typically categorize expressions into *InFavor*, *Against*, or *None*. SD has become increasingly relevant in various domains such as opinion mining, fake news detection, rumor verification, election prediction, information retrieval, and text summarization [6][10].

SD research can be broadly classified into two perspectives [6]: detecting expressed views and predicting unexpressed views. The former involves categorizing an author's text to determine their current stance toward a given subject [6][60], while the latter aims to infer an author's position on an event or subject that they have not explicitly discussed [11][12]. Additionally, SD tasks can be categorized as either Target-Specific SD (TSSD) or Multi-Target SD (MTSD). TSSD focuses on individual subjects, whereas MTSD involves jointly inferring stances toward multiple related subjects [5][6][13][38]–[40]. This paper primarily addresses detecting expressed views within the TSSD framework, incorporating unexpressed views through the infusion of sarcasm knowledge into the model framework. Examples of the SD task are provided in Table I.

Previous SD research has primarily utilized publicly available datasets sourced from online platforms [5][6][8][14]. However, texts from these platforms often exhibit subtlety, nuance, and complexity, including sarcastic and figurative language. These characteristics present challenges for SD algorithms in accurately discerning the author's stance [5].

TABLE I
EXAMPLES OF THE STANCE DETECTION TASK

Target	Text	Stance
Feminist Movement	Women don't make 75% less than men for the same job. Women, on average, make less than men. Look it up feminazis. #EqualPayDay #SemST	Against
Feminist Movement	Congratulations to America for overcoming 1 battle for #equality. Now let's have women & all races treated equally #AllLivesMatter #SemST	Favor
Feminist Movement	Honoured to be followed by the truly inspirational Kon__K founder of @ASRC1 #realaustralianssaywelcome #thethingsthatmatter #SemST	None

Moreover, targets are not always explicitly mentioned in the text [7], and stances may not be overtly expressed, further complicating the task of inferring the author's stance. Due to this problem, some examples discussed do not necessarily reflect the authors' beliefs. This often requires implicit inference through a combination of interactions, historical context, and sociolinguistic attributes such as sarcasm or irony.

To address these challenges, prior work has explored intermediate-task transfer learning, involving the fine-tuning of a model on a secondary task before its application to the primary task [2][15]–[19]. For instance, [16] and [19] utilized sentiment classification to enhance their models for SD. Similarly, [2] incorporated emotion and sentiment classification prior to sarcasm detection, suggesting that pre-training with sentiment analysis before sarcasm detection improves overall performance due to the correlation between sarcasm and negative sentiment. This finding aligns with one of our experimental observations in Section IV, where most sarcastic sentences with an "Against" stance are initially misclassified as "InFavor" before incorporating sarcasm pre-training into our model. However, despite its potential, sarcasm has been relatively unexplored as a means of improving SD models. In this study, we experiment with sarcasm detection as an intermediate task tailored to enhance SD performance.

Sarcasm detection involves discounting literal meaning to infer intention or secondary meaning from an utterance [20]. Sarcasm often involves using positive words or emotions to convey negative, ironic, or figurative meanings [21][22]. For example, in the sarcastic sentence "*I like girls. They just need to know their place*," the word "like" is used figuratively to mock the subject, making it difficult for SD algorithms to detect the true stance without accounting for sarcasm. Thus, sarcasm can alter the stance of a text from *Against* to *InFavor* and vice versa if not properly addressed [22][24]. Based on these observations, we developed an SD approach that incorporates sarcasm detection.

This study employs a model framework consisting of BERT [27] or RoBERTa [28], convolutional layers (Conv), a Bidirectional Long Short Term Memory (BiLSTM) layer, and a dense layer. Our experimental results demonstrate the efficacy of this approach, evidenced by improved macro F1-scores when sarcasm detection is included in the model framework. Additionally, we explore the impact of different sarcasm detection approaches on SD performance, considering the linguistic and quantitative attributes inherent in sarcasm datasets. Furthermore, the significance of this approach is underscored through a failure analysis of sarcastic texts from datasets, revealing the limitations of the original SD model before sarcasm pre-training.

We extend the work from [1] by applying CTSD to our tasks using a leave-one-out training approach. This method explores zero-shot fine-tuning on the target of interest, where four targets are used for model training and the remaining one for evaluation. The goal is to transfer knowledge from other targets to the target with limited training examples, thereby circumventing the scarcity of training data and the challenges of annotating sufficient data for new targets [29].

CTSD can be approached in two traditional ways: one-to-one, where one source target is used for training and one destination target for evaluation, and many-to-one, where multiple source targets are used for training and one destination target for evaluation [29]–[32]. The former approach often underutilizes available targets and struggles with generalization to unrelated targets, while the latter addresses these issues but often relies on sophisticated meta-learning approaches and limited datasets have been explored. In this work, we explore the many-to-one CTSD approach on two competitive SD tasks, proposing a solution that integrates sarcasm detection while mitigating the challenges associated with limited annotated SD data through many-to-one CTSD on diverse datasets.

Our experimental results show that the cross-target approach achieves performance comparable to models trained on target-specific data. Further analysis, including correlation measures between training and evaluation targets using cosine similarity on pre-trained language model embeddings, suggests that the overlapping vocabulary between the targets contributes to this performance.

Our work makes the following key contributions:

- *Transfer-Learning Framework:* Introducing a novel transfer-learning framework incorporating sarcasm detection as an intermediate task before fine-tuning on SD, utilizing an integrated deep learning model.
- *Cross-Target Stance Detection:* Introducing the leave-one-out fine-tuning on the SD targets, using four targets in training and the remaining one during evaluation, giving performance on par with training on the latter's desig-

nated data and curtailing the lack of annotated samples for training unseen targets problem.
- *Performance Superiority:* Demonstrating superior performance against State-Of-The-Art (SOTA) SD baselines, even without sarcasm detection pre-training, as indicated by higher macro F1-scores.
- *Correlation Analysis:* Establishing and illustrating the correlation between sarcasm detection and SD, exemplified through a failure analysis, thereby emphasizing the improvement of SD through sarcasm detection.
- *Impact Assessment:* Measuring the impact of various sarcasm detection models on target tasks based on the correlation between linguistic and quantitative attributes in the datasets of the two tasks.
- *Ablation Study:* Conducting an ablation study to assess the contribution of each module to the overall model framework. The study also reveals a significant drop in performance without sarcasm knowledge, underscoring the importance of our proposed approach.

The remainder of this paper unfolds as follows: Section II reviews related work, Section III outlines our proposed approach, and Section IV delves into comprehensive experiments, covering datasets, results, and subsequent discussions. The limitations inherent in our study are critically examined in Section V. The final section provides the conclusion and recommendations for further research.

II. Related Work

This section comprehensively reviews the literature on SD and intermediate-task transfer learning.

A. Stance Detection (SD)

The research on SD has traditionally been explored from two primary perspectives: Target-Specific SD (TSSD), which focuses on individual targets [5][6][33][34], and Multi-Target SD (MTSD), which aims to infer stances toward multiple related subjects concurrently [34][38]–[40]. Early approaches to SD were based on rule-based methods [33][41], followed by classical machine learning techniques [42]–[45]. For instance, [44] applied Naive Bayes (NB) to SD using datasets and features derived from inter-post constraints in online debates. Similarly, [42] utilized features such as unigrams, bigrams, hashtags, external links, emoticons, and named entities in various Support Vector Machine (SVM) models, while [45] employed an SVM model with linguistic (n-grams) and sentiment features to predict stance. In contrast, [43] explored and compared linear SVM, Logistic Regression (LR), Multinomial NB, k-Nearest Neighbors (kNN), Decision Trees (DT), and Random Forests (RF) using the simple Bag-of-Words approach with term frequency-inverse document frequency (tf-idf) vectors of tweets as features for multi-modal SD.

While classical approaches relied on manually crafted features, the advent of deep learning models has seen neural networks gradually replace traditional methods [7][19][46][47]. For instance, the work by [7] investigated SD using Bidirectional Conditional Encoding (BCE) [48], incorporating an LSTM architecture to build a tweet representation dependent on the target. Similarly, [49] employed a CNN for SD, incorporating a voting scheme mechanism, while [16] utilized a bidirectional Gated Recurrent Unit (biGRU) within a multi-task framework that included a target-specific attention mechanism, leveraging sentiment classification to enhance SD performance. Moreover, [46] presented a neural ensemble model combining BiLSTM, an attention mechanism, and multi-kernel convolution, evaluated on both TSSD and MTSD. Although our work shares some similarities in model framework, it uniquely employs BERT or RoBERTa and introduces an intermediate-task transfer learning technique, diverging from ensemble approaches and multi-kernel usage.

Deep learning models necessitate large datasets for effective SD model training and generalization [27]. Consequently, recent research has explored the use of pre-trained language models for SD. For example, [47] proposed using BERT [27] in a cross-validation approach, developing a multi-dataset model from the aggregation of several datasets. Similarly, [5] conducted a comparative study, fine-tuning pre-trained BERT against classical SD approaches, while [34] employed BERT as an embedding layer to encode textual features in a zero-shot deep learning setting, yielding promising results. On the other hand, [33] experimented with ChatGPT, directly prompting the model with test cases to discern stances; however, all these studies reported difficulties in accurately classifying sarcastic examples.

B. Cross-Target Stance Detection (CTSD)

Research on CTSD can be divided into two main approaches. The first is the one-to-one approach, where a single source target and a single destination target share common words, which helps bridge the knowledge gap [29]–[31]. For example, [30] introduced the CrossNet model, utilizing an aspect attention layer to learn domain-specific aspects from a source target for generalization on a destination target. Similarly, [29] used external knowledge, such as semantic and emotion lexicons, to enable knowledge transfer between targets. Meanwhile, [31] explored few-shot learning by leveraging social network features alongside textual content, introducing 300+ training examples from the destination target. This line of research primarily explores related targets within a common domain.

The second approach is the many-to-one method, which involves using multiple source targets for a single destination target. For instance, [32] used many unrelated source targets to the destination target without leveraging external knowledge but instead employed a sophisticated meta-learning approach and did not utilize diverse datasets.

C. Intermediate-Task Transfer Learning

Recent studies have increasingly adopted intermediate-task transfer learning, which transfers knowledge from a data-rich auxiliary task to a primary task [18]. This technique has

proven highly effective across various NLP tasks. For example, [15] employed supervised pre-training on four-example intermediate tasks to enhance performance on primary tasks evaluated using the GLUE benchmark suite [50]. Additionally, [19] introduced few-shot learning, leveraging sentiment-based annotation to improve cross-lingual SD performance. Furthermore, [2] employed transfer learning by sequentially fine-tuning pre-trained BERT on emotion and sentiment classification before applying it to sarcasm detection, capitalizing on the correlation between sarcasm and negative sentiment polarity.

To the best of our knowledge, prior research has not explored sarcasm detection pre-training for SD, nor has it investigated the concatenation of BERT or RoBERTa with other deep learning techniques for SD. In this paper, we propose leveraging sarcasm detection for both in-domain SD and CTSD within a model framework comprising BERT, convolutional layers, BiLSTM, and a dense layer.

III. Methodology

This section delineates our approach, covering problem formulation, intermediate-task transfer learning, and the model architecture.

A. Problem Formulation

We denote the collection of labeled data in the source targets as $X^s = \{x_i^s, y_i^s, t_i^{sj}\}_{i=1}^N, j = \{1, 2, 3, ..., k\}$, where x represents the input text, y denotes the stance label, and t indicates the j^{th} target. Here, s represents a source target, and there are k source targets in X^s, comprising N data samples in total. Similarly, we denote the collection of data in the destination target as $X^d = \{x_i^d, y_i^d, t_i^d\}_{i=1}^M, d = \{1\}$, where d represents the destination target, with M data samples. Given an input text x from a destination target t^d, the objective is to predict the stance label of x towards t^d using the model trained on the labeled data X^s. For the in-domain task, $t^s = t^d$; for the CTSD task, $t^s \neq t^d$.

B. Intermediate-Task Transfer Learning

Our approach incorporates intermediate-task transfer learning, which involves two phases: pre-training on an intermediate task and fine-tuning on a target task.

1) Target Task: The primary task in this study is SD, aiming to predict the stance expressed in a given text, such as a tweet, towards a specific target (e.g., '*feminist movement*'). A tweet, denoted as x, is represented as a sequence of words $(w_1, w_2, w_3, \ldots, w_L)$, with L representing the sequence length. Stance labels are categorized as *InFavor* (supporting the target), *Against* (opposing the target), or *None* (neutral towards the target).

2) Intermediate Task: The intermediate task in this study is sarcasm detection, where the goal is to determine whether a given text S is sarcastic. Sarcasm detection labels are categorized as *Sarcastic* (the text is sarcastic) or *Non-Sarcastic* (the text is not sarcastic). As sarcasm has not previously been employed as an intermediate task, we explore three sarcasm-detection datasets to identify key linguistic features that can enhance SD performance:

Sarcasm V2 Corpus (SaV2C). The SaV2C dataset, introduced by [51], is a diverse corpus developed using syntactical cues and crowd-sourced from the Internet Argument Corpus (IAC 2.0). It comprises 4,692 lines containing quote and response sentences from political debates in IAC online forums. SaV2C is categorized into: 1) General Sarcasm (Gen, 3,260 sarcastic and 3,260 non-sarcastic comments); 2) Rhetorical Questions (RQ, 851 rhetorical and 851 non-rhetorical questions); and 3) Hyperbole (Hyp, 582 hyperboles and 582 non-hyperboles). Our focus is on the General Sarcasm category, which includes 3,260 sarcastic and 3,260 non-sarcastic comments.

The Self-Annotated Reddit Corpus (SARC). Created by [52], the SARC dataset contains over a million sarcastic and non-sarcastic statements from Reddit. This dataset features a balanced ratio of sarcastic and non-sarcastic comments, with 1,010,826 training and 251,608 evaluation statements. We utilized the Main Balanced variant, obtained directly from the author of [2].

SARCTwitter (ST). Released by [53], the ST dataset includes 350 sarcastic and 644 non-sarcastic tweets, annotated by seven readers. We used the variant of the dataset employed by [54], which consists of 994 tweets (350 sarcastic and 644 non-sarcastic), excluding eye movement data.

In this work, we implement two levels of transfer learning: first, from sarcasm detection to SD through intermediate-task pre-training; and second, from target-to-target through cross-target fine-tuning. The intermediate-task transfer learning pipeline is illustrated in Figure 1.

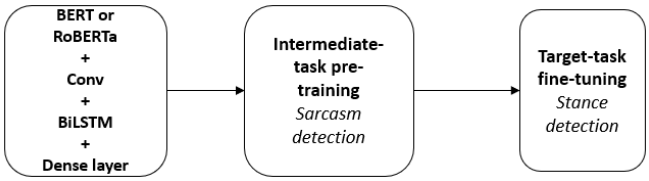

Figure 1. The intermediate-task transfer learning pipeline.

C. Underlying Model Architecture

The model framework consists of an input layer, an embedding layer, and deep neural networks.

1) Input Layer: The input layer processes text x encoding stance information, comprising L words. The text x is converted into a vector of words and passed to the embedding layer.

2) Embedding Layer: We utilize BERT [27] and RoBERTa [28] for encoding textual input into hidden states H. The choice of these language models is supported by their notable performance in the literature [2][5][15][18][47][55].

3) Deep Neural Networks: The deep neural network module includes two convolutional layers (Conv), a BiLSTM layer, and a dense layer, which are applied on top of the

embedding layer. The Conv layer identifies specific sequential word patterns within the text, creating a composite feature map from H. This feature map aids the BiLSTM layer in capturing higher-level stance representations, which are further refined by the dense layer. The overall model framework is depicted in Figure 2.

IV. Experiments

This section delineates the datasets employed, details the data preprocessing procedures, outlines the baseline models, presents experimental results, and engages in a subsequent discussion.

A. Datasets

For evaluation purposes, we utilize two publicly available SD datasets: 1) the SemEval 2016 Task 6A Dataset (SemEval) [56], and 2) the Multi-Perspective Consumer Health Query Data (MPCHI) [57].

1) SemEval: The SemEval dataset includes tweets manually annotated for stance towards specific targets, encompassing opinions and sentiments. For our experiments, we utilize tweets and their associated stance annotations. The dataset features tweets related to five distinct targets: Atheism (AT), Climate Change (CC), Feminist Movement (FM), Hillary Clinton (HC), and Legalization of Abortion (LA).

2) MPCHI: MPCHI is designed for stance classification to enhance Consumer Health Information (CHI) query search results. It comprises formal texts extracted from top-ranked articles corresponding to specific web search engine queries. The dataset includes sentences related to five distinct queries, which also serve as targets for stance classification: MMR vaccination and autism (MMR), E-cigarettes versus normal cigarettes (EC), Hormone Replacement Therapy post-menopause (HRT), Vitamin C and the common cold (VC), and sun exposure and skin cancer (SC).

Each text in the datasets is annotated with one of three classes: *InFavor*, *Against*, and *None*. Table II presents the original statistical details of the datasets.

B. Data Preprocessing

We employ standard data preprocessing steps, including case folding, stemming, stop-word removal, and deletion of null entries across all datasets. Text normalization is performed following the method described by [58], and hashtag processing utilized Wordninja [59]. For neural network models relying on pre-trained embeddings, stemming and stop-word removal are omitted, as stemmed forms of terms may not be present in the pre-trained embeddings. The default tokenizer of the respective pre-trained language model is used to tokenize words in tweets prior to inputting them into the classifier.

C. Baseline Models

For the in-domain SD task, we evaluate our model against the top-performing results from the SemEval challenge [60], as reproduced with minor modifications in [5]. Additionally, we compare our model's performance with recent SOTA methods in SD. The following first three baseline models are used for evaluating our model on the in-domain SD task, while the remaining models are used for evaluating the CTSD task.

1) SemEval Models: We select the Target-Specific Attention Neural Network (TAN-) proposed by [61] and the 1-D sem-CNN introduced by [62] from the SemEval competition. Additionally, we include Com-BiLSTM and Com-BERT, implementations provided solely by [5].

2) ChatGPT: The work by [33] explored the use of ChatGPT for SD by directly probing the generative language model to determine the stance of a given text, with a focus on specific targets from the SemEval task: FM, LA, and HC.

3) Zero-Shot Stance Detection (ZSSD): The ZSSD technique [34], which employs contrastive learning, was implemented for the SemEval dataset similarly to ChatGPT.

4) BiCond: An LSTM model that uses bidirectional conditional encoding to learn both input text and target representations for SD [35].

5) TextCNN-E: A variant of TextCNN [36] adapted for the CTSD task by incorporating semantic and emotional knowledge into each word and expanding the dimensionality of each word vector [32].

6) Semantic-Emotion Knowledge Transferring (SEKT): This model leverages external semantic and emotion lexicons to facilitate knowledge transfer across different targets [29].

7) Target-Adaptive Pragmatics Dependency Graphs (TPDG): This model constructs two graphs: an in-target graph to capture inherent pragmatic dependencies of words for a specific target, and a cross-target graph to enhance the versatility of words across all targets [37].

8) Refined Meta-Learning (REFL): A SOTA CTSD model that utilizes meta-learning by refining the model with a balanced, easy-to-hard learning pattern and adapting it according to target similarities [32].

D. Experimental Settings

The experimental setup adopts an inductive approach to transfer learning, where the target task model is initialized using parameters obtained from pre-training on sarcasm detection. This strategy is designed to enhance model performance for the target task. For the intermediate tasks, datasets are divided into training and validation sets solely for sarcasm detection pre-training. Given that Sav2C and ST are the smallest intermediate-task datasets, five-fold cross-validation is utilized for these, while SARC, being larger, employs an 80/20 train/validation split. In contrast, the target task featured a separate test set for final evaluations and comparisons.

Consistent with the methodologies of [5], datasets are divided into training and test sets using similar proportions for in-domain SD, while CTSD employs a leave-one-out strategy. In this approach, data from all source targets are used for model training, and the test data for the destination target is reserved for model evaluation. Each SD dataset consists of five targets; thus, during CTSD experimentation, four targets are used for training, and the remaining target is used for

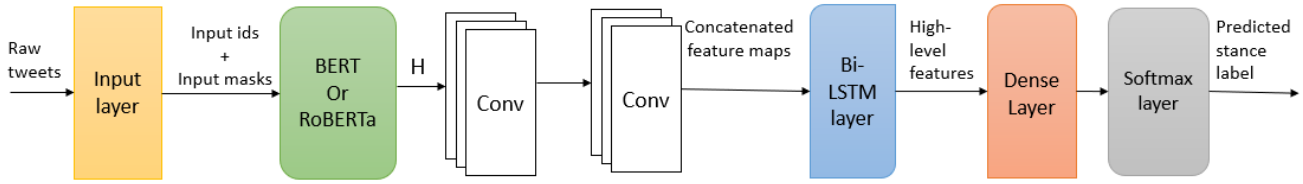

Figure 2. The proposed model framework.

TABLE II
ORIGINAL STATISTICS OF THE DATASETS DIVIDED INTO TRAINING AND TEST SETS

Dataset	Target	Training Samples			Test Samples		
		INFAVOR	AGAINST	NONE	INFAVOR	AGAINST	NONE
SemEval	AT	92	304	117	32	160	28
	CC	212	15	168	123	11	35
	FM	210	328	126	58	183	44
	HC	112	361	166	45	172	78
	LA	105	334	164	46	189	45
MPCHI	MMR	48	61	72	24	33	21
	SC	68	51	117	35	26	42
	EC	60	118	111	33	47	44
	VC	74	52	68	37	16	31
	HRT	33	95	44	9	41	24

TABLE III
STATISTICS OF THE DATASETS AFTER INCORPORATING CROSS-TARGET STANCE DETECTION

Dataset	Target	Training samples			Test samples		
		INFAVOR	AGAINST	NONE	INFAVOR	AGAINST	NONE
SemEval	AT	910	1593	826	32	160	28
	CC	699	2031	767	123	11	35
	FM	766	1546	800	58	183	44
	HC	878	1524	726	45	172	78
	LA	883	1534	761	46	189	45
MPCHI	MMR	314	402	425	24	33	21
	SC	279	417	365	35	26	42
	EC	301	343	376	33	47	44
	VC	276	424	421	37	16	31
	HRT	342	358	453	9	41	24

evaluation. Table III details the statistics of the datasets after incorporating the experimental settings of CTSD.

The Conv layer uses a kernel size of 3 with 16 filters and a ReLU activation function. A BiLSTM layer with a hidden state of 768, corresponding to the hidden state size of the pre-trained language models, is employed. The dense layer has an output size of 3 and utilizes a softmax activation function. All experiments are conducted on an NVIDIA Quadro RTX 4000 GPU.

Hyperparameter tuning involves multiple experiments to select the optimal intermediate-task training scheme based on results from a holdout development set. The best-performing per-task model is then evaluated on the test set. The training process uses a mini-batch size of 16 samples and the Adam optimizer [63], with cross-entropy loss as the cost function. Training epochs ranges from 10 to 50, with early stopping applied if validation accuracy on holdout data plateaus for five consecutive epochs. The learning rate is initially set to 3e-5, decaying to 1e-9 for the intermediate task and 1e-10 for the target task. A dropout rate of 0.25 is introduced between model layers to mitigate overfitting. To address class imbalance, class weights are incorporated during training to improve generalization for underrepresented classes. Experimental setups adhere to the configurations outlined in the original papers for baseline models unless otherwise specified, in which case our experimental configurations are applied.

E. Evaluation Metrics

In alignment with previous studies [5][7][60], the evaluation of our model is based on the average macro F1-score for the *InFavor* and *Against* classes.

F. Results

We first present the results for in-domain SD, followed by the CTSD results. Baseline results for CTSD are referenced from [32]. All results are averaged over five experimental runs per target task.

Table IV displays the experimental outcomes for in-domain SD before the incorporation of sarcasm detection pre-training. Results for ChatGPT and ZSSD are directly transcribed from their original publications, while other baseline results are replicated in our experiments. The table demonstrates the notable performance of our BERT-based model across various targets, achieving superior results in most metrics except HC and CC, where ChatGPT and our RoBERTa-based model excel. Consequently, we select our BERT-based model for subsequent experiments.

Table V reports the results of incorporating sarcasm detection pre-training with our model for in-domain SD. Performance improves by **0.550** on SemEval and **0.003** on MPCHI when pre-training with ST, surpassing all baseline models listed in Table IV. However, performance decreases with Sav2C and SARC. Therefore, subsequent results utilize the ST model.

Table VI presents the results of CTSD. Notably, no baseline models have been evaluated on the MPCHI dataset, focusing instead on SemEval with one target not addressed by the SEKT baseline. Our model outperforms all CTSD models listed in the table on the average macro F1 measure.

Table VII summarizes the results of an ablation study on the in-domain task. Various base model components are systematically excluded to evaluate their contributions to the overall model framework. The model integrating all components—BERT, Conv, BiLSTM, and sarcasm pre-training—achieves the highest average F1-scores of **0.775** and **0.724** on SemEval and MPCHI, respectively.

G. Failure Analysis and Discussion

Following the results presented in Table IV, a detailed failure analysis is conducted to investigate the misclassified test samples. The analysis reveals that misclassifications in the SemEval dataset are predominantly associated with texts containing sarcastic content, consistent with prior findings [5]. This observation supports the rationale for incorporating sarcasm-detection pre-training prior to fine-tuning for SD. Conversely, misclassifications in the MPCHI dataset are primarily linked to samples that contained large, generic health-related facts that are neutral with respect to the target under study. Additional insights derived from the experiments and results across all tasks are discussed below.

1) Performance of Our Model Relative to SOTA Models Without Sarcasm Detection: Our model demonstrates superior performance compared to SOTA models even in the absence of sarcasm detection. Specifically, it outperforms ChatGPT and Com-BERT, which are among the top-performing models, on both SemEval and MPCHI by **0.038** and **0.053** in average F1-scores, respectively, for the in-domain SD task. While Com-BERT utilizes only BERT and a dense layer for classification, our model benefits from additional Conv and BiLSTM layers preceding the dense layer, which contributes to the observed performance improvement. Furthermore, the inclusion of the BiLSTM module in our model results in better performance compared to using pooling layers after the Conv module. This finding highlights the effectiveness of our model architecture in capturing nuanced representations, leading to improved generalization for SD tasks.

2) Correlation Between Sarcasm Detection and SD: An illustrative example of misclassification involves the statement: "*I like girls. They just need to know their place. #SemST*", a sarcastic comment from the FM target in SemEval. The true label for this example is *Against*, but it is misclassified as *InFavor* before the incorporation of sarcasm-detection pre-training. Notably, sarcastic samples in the *Against* class are often misclassified as *InFavor* due to their overtly positive content. After integrating sarcasm detection through pre-training, 85% of these misclassified sarcastic samples are correctly predicted. This result underscores the importance of sarcasm-detection pre-training in enhancing the performance of SD models.

3) Challenges in Using Sarcasm Detection Models for Intermediate-Task Transfer Learning on SD: The integration of SARC and SaV2C knowledge into the model pipeline introduces noise and adversely affects model performance on SD compared to using ST knowledge. Analysis of Sav2C and SARC reveals several discrepancies with the target task. For instance, the average sentence length in Sav2C and SARC is longer compared to SemEval and MPCHI samples. Additionally, SARC is sourced from different domains than SemEval and MPCHI, leading to variations in topic coverage, vocabulary overlap, and the framing of ideas. SARC, being the largest intermediate task, spans a wide range of topics across various subreddits, while ST, which performs best, shares a similar average sentence length with the target tasks and is also crowd-sourced from Twitter (X). This alignment likely contributes to the superior performance observed when using ST as an intermediate task for SemEval. Consequently, the mismatched attributes of certain intermediate tasks can negatively impact model performance. This underscores the need for careful selection and experimentation when choosing a sarcasm model for transfer learning in SD.

4) Performance of Cross-Target Stance Detection: The CTSD task exhibits comparable performance to the in-domain task, despite using out-of-domain data during model fine-tuning. This suggests that our model effectively learns common features from various targets, thereby leveraging this data to perform well on new targets in CTSD. To further understand

TABLE IV
EXPERIMENTAL RESULTS WITHOUT SARCASM DETECTION PRE-TRAINING

Model	SemEval						MPCHI					
	AT	**CC**	**FM**	**HC**	**LA**	**Avg**	**MMR**	**SC**	**EC**	**VC**	**HRT**	**Avg**
Sem-TAN-	0.596	0.420	0.495	0.543	0.603	0.531	0.487	0.505	0.564	0.487	0.467	0.502
Sem-CNN	0.641	0.445	0.552	0.625	0.604	0.573	0.524	0.252	0.539	0.524	0.539	0.476
Com-BiLSTM	0.567	0.423	0.508	0.533	0.546	0.515	0.527	0.522	0.471	0.474	0.469	0.493
ZSSD	0.565	0.389	0.546	0.545	0.509	0.511	-	-	-	-	-	-
Com-BERT	0.704	0.466	0.627	0.620	0.673	0.618	0.701	0.691	0.710	0.617	0.621	0.668
ChatGPT	-	-	0.690	**0.780**	0.593	0.687	-	-	-	-	-	-
Ours-RoBERTa	0.740	**0.775**	0.689	0.683	0.696	0.712	0.692	0.687	0.700	0.701	0.698	0.695
Ours-BERT	**0.767**	0.755	**0.697**	0.704	**0.702**	**0.725**	**0.747**	**0.722**	**0.704**	**0.702**	**0.732**	**0.721**

TABLE V
EXPERIMENTAL RESULTS WITH SARCASM-DETECTION PRE-TRAINING

Task	SemEval						MPCHI					
	AT	**CC**	**FM**	**HC**	**LA**	**Avg**	**MMR**	**SC**	**EC**	**VC**	**HRT**	**Avg**
SaV2C	0.595	0.718	0.596	0.645	0.578	0.626	0.605	0.545	0.545	0.352	0.495	0.508
SARC	0.697	0.612	0.683	0.557	0.641	0.638	0.605	0.545	0.545	0.352	0.495	0.508
ST	**0.769**	**0.800**	**0.774**	**0.795**	**0.741**	**0.775**	**0.749**	**0.727**	**0.704**	**0.703**	**0.739**	**0.724**

TABLE VI
EXPERIMENTAL RESULTS OF CROSS-TARGET STANCE DETETCION WITH SARCASM-DETECTION PRE-TRAINING

Task	SemEval						MPCHI					
	AT	**CC**	**FM**	**HC**	**LA**	**Avg**	**MMR**	**SC**	**EC**	**VC**	**HRT**	**Avg**
BiCond	0.526	0.512	0.527	0.536	0.493	0.519	-	-	-	-	-	-
TextCNN-E	0.534	0.633	0.582	0.591	0.550	0.578	-	-	-	-	-	-
SEKT	0.623	0.600	0.648	-	0.649	0.630	-	-	-	-	-	-
TPDG	0.654	0.667	0.669	0.630	0.600	0.644	-	-	-	-	-	-
REFL	0.650	0.671	**0.734**	0.652	0.623	0.666	-	-	-	-	-	-
Ours	**0.689**	**0.697**	0.730	**0.682**	**0.656**	**0.691**	**0.699**	**0.687**	**0.695**	**0.701**	**0.700**	**0.696**

this observation, cosine similarity scores on the pre-trained BERT embeddings are analyzed. Figure 3 illustrates the cosine similarities between each target and the other targets in their respective datasets. In the figure, LAMMRSC should read as LA, MMR, and SC on the X axis. The figure demonstrates that all targets share common vocabulary with others, leading to shared features. Additionally, MPCHI targets have higher cosine similarity scores than SemEval targets, which aligns with the superior CTSD performance observed on the MPCHI task.

5) Ablation Study on Sarcasm Knowledge: The results of the ablation study presented in Table VII provide insights into the contribution of each module and the overall impact of sarcasm detection pre-training on SD performance. Comparing the results in Table IV and Table VII, the incorporation of sarcasm knowledge significantly enhances model performance on the SemEval task compared to the MPCHI task. SemEval includes a large volume of opinionated and sarcastic texts, whereas the MPCHI dataset primarily consists of health-related facts, with occasional sarcastic expressions. As a result,

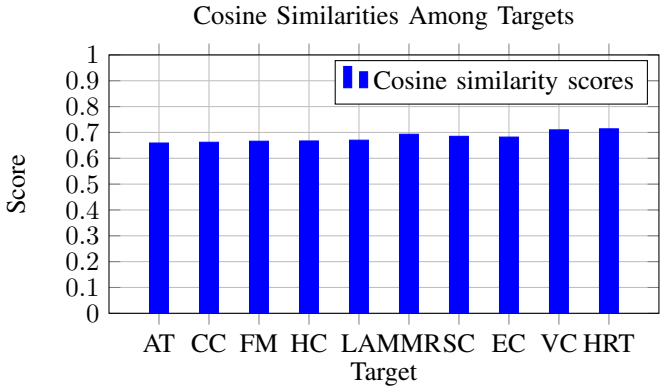

Figure 3. Cosine similarity scores for each target in comparison with other targets within their respective datasets.

there is a modest improvement in performance on MPCHI when sarcasm detection is used. This suggests the potential for exploring BERT or RoBERTa embeddings pre-trained

TABLE VII
EXPERIMENTAL RESULTS OF AN ABLATION STUDY

Model	SemEval						MPCHI					
	AT	**CC**	**FM**	**HC**	**LA**	**Avg**	**MMR**	**SC**	**EC**	**VC**	**HRT**	**Avg**
BERT	0.674	0.677	0.678	0.609	0.685	0.665	0.568	0.519	0.441	0.482	0.595	0.521
BERT+Conv+BiLSTM	0.767	0.755	0.697	0.704	0.702	0.725	0.747	0.722	**0.704**	0.702	0.732	0.721
ST+BERT	0.712	0.735	0.698	0.687	0.696	0.706	0.687	0.601	0.540	0.466	0.546	0.568
ST+BERT+Conv	**0.770**	0.759	0.689	0.683	0.694	0.719	0.458	0.535	0.479	0.350	0.524	0.469
ST+BERT+BiLSTM	0.747	0.765	0.675	0.657	0.678	0.704	0.640	0.618	0.573	0.528	0.633	0.598
ST+BERT+Conv+BiLSTM	0.769	**0.800**	**0.774**	**0.795**	**0.741**	**0.775**	**0.749**	**0.727**	**0.704**	**0.703**	**0.739**	**0.724**

on health-related data specifically for SD on MPCHI as a promising avenue for future research.

V. LIMITATIONS

Despite the significant contributions of this study to NLP in social media contexts, several limitations warrant consideration. Firstly, the extent of model performance improvement is dependent on the characteristics of both the intermediate sarcasm detection task and the ultimate SD task. Variations in linguistic features across datasets used for sarcasm detection and SD may limit the generalizability of the study's findings. Secondly, while the integration of BERT or RoBERTa with other deep-learning methodologies represents an innovative approach, the complexity of the model architecture may pose challenges in terms of computational resources and interoperability in certain contexts. Thirdly, the CTSD task presents additional challenges, as the language models employed may not be compatible across different targets. Lastly, the heavy reliance on fine-tuning techniques and specific datasets raises concerns about the model's ability to generalize effectively across diverse text types or domains not covered within the training data.

VI. CONCLUSION AND FUTURE WORK

In this study, we have proposed a transfer-learning framework that integrates sarcasm detection for SD. We have utilized pre-trained language models, RoBERTa and BERT, which have been individually fine-tuned and subsequently concatenated with other deep neural networks, with BERT demonstrating particularly promising results. The model has been pre-trained on three sarcasm-detection tasks before being fine-tuned on two target SD tasks. Our evaluations, including in-domain SD and CTSD, have shown that our approach outperformed SOTA models, even before incorporating sarcasm knowledge. The correlation between sarcasm detection and SD has been established, with the integration of sarcasm knowledge significantly enhancing model performance; notably, 85% of misclassified samples in the SemEval task have been accurately predicted after incorporating sarcasm knowledge. Failure analysis has indicated that the SemEval dataset, rich in opinionated sarcastic samples, has benefited significantly from sarcasm pre-training, in contrast to the MPCHI dataset, which primarily consists of generic health-related facts. Furthermore, our study has revealed that not all intermediate sarcasm-detection tasks have improved SD performance due to mismatched linguistic attributes. Additionally, the CTSD task has demonstrated performance on par with the in-domain task despite using a zero-shot fine-tuning approach, effectively addressing the issue of limited annotated samples from new targets. Finally, the ablation study has highlighted that the optimal performance of the model is achieved when all components are utilized.

To the best of our knowledge, this work represents the inaugural application of sarcasm-detection pre-training within a BERT (RoBERTa)+Conv+BiLSTM architecture before fine-tuning for SD. Our approach serves as a foundational reference, setting a baseline for future research in this domain. Future work will explore variant BERT or RoBERTa embeddings tailored to health-related text data for the MPCHI task and will focus on a more comprehensive evaluation of other intermediate tasks, including sentiment and emotion knowledge.

ACKNOWLEDGMENT

This work is supported by the National Science Foundation (NSF) under Award number OIA-1946391, Data Analytics that are Robust and Trusted (DART). We sincerely thank our anonymous reviewers for their valuable insights and constructive feedback. Additionally, we extend our gratitude to all individuals who contributed to this study in various capacities.

REFERENCES

[1] G. Nkhata and S. Gauch, "Intermediate-Task Transfer Learning: Leveraging Sarcasm Detection for Stance Detection," *The Sixteenth International Conference on Information, Process, and Knowledge Management*, eKNOW 2024, Barcelona, Spain, pp. 7-14.

[2] E. Savini and C. Caragea, "Intermediate-task transfer learning with bert for sarcasm detection," *Mathematics*, vol. 10, no. 5, p. 844, MDPI, 2022.

[3] M. Grčar, D. Cherepnalkoski, I. Mozetič, and P. Kralj Novak, "Stance and influence of Twitter users regarding the Brexit referendum," *Computational social networks*, Springer 2017, vol. 4, pp. 1-25.

[4] N. Newman, "Mainstream media and the distribution of news in the age of social media," *Reuters Institute for the Study of Journalism, Department of Politics*, 2011.

[5] S. Ghosh, P. Singhania, S. Singh, K. Rudra, and S. Ghosh, "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*. Springer, 2019, pp. 75–87.

[6] A. ALDayel and W. Magdy, "Stance detection on social media: State of the art and trends," *Information Processing & Management*, vol. 58, no. 4, p. 102597, Elsevier, 2021.

[7] I. Augenstein, T. Rocktäschel, A. Vlachos, and K. Bontcheva, "Stance detection with bidirectional conditional encoding," arXiv preprint arXiv:1606.05464, 2016.

[8] D. Küçük and F. Can, "Stance detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 1, pp. 1–37, ACM, NY, USA, 2020.

[9] D. Küçük and F. Can, "A tutorial on stance detection," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, ACM, 2022, pp. 1626–1628.

[10] D. Biber and E. Finegan, "Adverbial stance types in english," *Discourse processes*, vol. 11, no. 1, pp. 1–34, Taylor & Francis, 1988.

[11] W. Magdy, K. Darwish, N. Abokhodair, A. Rahimi, and T. Baldwin, "# isisisnotislam or# deportallmuslims? Predicting unspoken views," in *Proceedings of the 8th ACM Conference on Web Science*, 2016, pp. 95-106.

[12] K. Darwish et al., "Predicting online islamophobic behavior after# ParisAttacks," *The Journal of Web Science*, Now Publishers, Inc., 2018, vol. 4, pp. 34-52.

[13] P. Wei, J. Lin, and W. Mao, "Multi-target stance detection via a dynamic memory-augmented network," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 1229-1232.

[14] D. Küçük and F. Can, "Stance detection: Concepts, approaches, resources, and outstanding issues," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 2021, pp. 2673–2676

[15] J. Phang, T. Févry, and S. R. Bowman, "Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks," *arXiv preprint arXiv:1811.01088*, Cornell University, 2018.

[16] Y. Li and C. Caragea, "Multi-task stance detection with sentiment and stance lexicons," 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)*, ACL, 2019, pp. 6299–6305.

[17] M. Sap, H. Rashkin, D. Chen, R. LeBras, and Y. Choi, "Socialiqa: Commonsense reasoning about social interactions," *arXiv preprint arXiv:1904.09728*, Machine Learning, ICML, 2019.

[18] Y. Pruksachatkun et al., "Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work?" *arXiv preprint rXiv:2005.00628,* ACL, 2020.

[19] M. Hardalov, A. Arora, P. Nakov, and I. Augenstein, "Few-shot cross-lingual stance detection with sentiment-based pre-training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, AAAI, 2022, pp. 10 729–10 737.

[20] A. Ghosh and T. Veale, "Fracking sarcasm using neural network," in *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, ACL, 2016, pp. 161–169.

[21] S. M. Sarsam, H. Al-Samarraie, A. I. Alzahrani, and B. Wright, "Sarcasm detection using machine learning algorithms in twitter: A systematic review," *International Journal of Market Research*, vol. 62, no. 5, pp. 578–598, Sage Journals, 2020.

[22] R. Jamil et al., "Detecting sarcasm in multi-domain datasets using convolutional neural networks and long short term memory network model," *PeerJ Computer Science*, vol. 7, p. e645, National Library of Medicine, 2021.

[23] A. Kumar, V. T. Narapareddy, V. Aditya Srikanth, A. Malapati, and L. B. M. Neti, "Sarcasm detection using multi-head attention based bidirectional lstm," *IEEE Access*, vol. 8, pp. 6388–6397, IEEE, 2020.

[24] C. Liebrecht, F. Kunneman, and A. van Den Bosch, "The perfect solution for detecting sarcasm in tweets# not," in *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, ACL, 2013, pp. 29–37.

[25] B. Jang, M. Kim, G. Harerimana, S. Kang, and W. Jong, "Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism," *Applied Sciences*, MDPI, vol. 10, no. 17, pp. 5841, 2020.

[26] N. J. Prottasha et al., "Transfer learning for sentiment analysis using BERT based supervised fine-tuning," *Sensors*, MDPI, vol. 22, no. 11, pp. 5147, 2022.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, ACL, 2018.

[28] Y. Liu et al., "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, ACL, 2019.

[29] B. Zhang, M. Yang, X. Li, Y. Ye, X. Xu, and K. Dai, "Enhancing Cross-target Stance Detection with Transferable Semantic-Emotion Knowledge," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3188–3197, Online. Association for Computational Linguistics.

[30] C. Xu, C. Paris, S. Nepal, and R. Sparks, "Cross-Target Stance Classification with Self-Attention Networks," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, vol. 2, pp. 778–783, Melbourne, Australia. Association for Computational Linguistics.

[31] J. K. Parisa and Z. Arkaitz, "Few-shot Learning for Cross-Target Stance Detection by Aggregating Multimodal Embeddings," *arXiv*, 2023, URL. https://arxiv.org/abs/2301.04535.

[32] H. Ji, Z. Lin, P. Fu and W. Wang, "Cross-Target Stance Detection Via Refined Meta-Learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, ICASSP 2022 - 2022, Singapore, Singapore, 2022, pp. 7822-7826, doi: 10.1109/ICASSP43922.2022.9746302.

[33] B. Zhang, D. Ding, and L. Jing, "How would stance detection techniques evolve after the launch of chatgpt?" *arXiv preprint arXiv:2212.14548*, ArXiv. /abs/2212.14548, 2022.

[34] B. Liang et al., "Zero-shot stance detection via contrastive learning," in *Proceedings of the ACM Web Conference 2022,* ACM, 2022, pp. 2738–2747.

[35] I. Augenstein, T. Rocktäschel, A. Vlachos, and K. Bontcheva, 2016, "Stance Detection with Bidirectional Conditional Encoding," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 876–885, Austin, Texas, Association for Computational Linguistics.

[36] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1746–1751, Doha, Qatar, Association for Computational Linguistics.

[37] L. Bin et al., "Target-adaptive Graph for Cross-target Stance Detection," in *Proceedings of the Web Conference 2021*, 2021, pp. 3453–3464, Association for Computing Machinery, New York.

[38] P. Sobhani, D. Inkpen, and X. Zhu, "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*, ACL, 2017, pp. 551–557.

[39] H. Liu, S. Li, and G. Zhou, "Two-target stance detection with target-related zone modeling," in *Information Retrieval: 24th China Conference, CCIR 2018, Guilin, China, September 27–29, 2018, Proceedings 24,* Springer, 2018, pp. 170–182.

[40] P. Sobhani, D. Inkpen, and X. Zhu, "Exploring deep neural networks for multitarget stance detection," *Computational Intelligence*, vol. 35, no. 1, pp. 82–97, Computational Intelligence, 2019.

[41] M. Walker, P. Anand, R. Abbott, and R. Grant, "Stance classification using dialogic properties of persuasion," in *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, ACL, 2012, pp. 592–596.

[42] D. Küçük and F. Can, "Stance detection on tweets: An svm-based approach," *arXiv preprint arXiv:1803.08910*, ArXiv. /abs/1803.08910, cs. cL, 2018.

[43] I. Segura-Bedmar, "Labda's early steps toward multimodal stance detection." in *IberEval@ SEPLN*, ACL, 2018, pp. 180–186.

[44] K. S. Hasan and V. Ng, "Stance classification of ideological debates: Data, models, features, and constraints," in *Proceedings of the sixth international joint conference on natural language processing*, pp. 1348-1356, 2013.

[45] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, "Stance and sentiment in tweets," in *ACM Transactions on Internet Technology (TOIT)*, ACM New York, NY, USA, vol. 13, no. 3, pp. 1-23, 2017.

[46] U. A. Siddiqua, A. N. Chy, and M. Aono, "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)*, ACL, 2019, pp. 1868–1873.

[47] L. H. X. Ng and K. M. Carley, "Is my stance the same as your stance? a cross validation study of stance detection datasets," *Information Processing & Management*, vol. 59, no. 6, p. 103070, ACM, 2022.
[48] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural networks*, Elsevier, vol. 18, no. 5-6, pp. 602-610, 2005.
[49] W. Wei, X. Zhang, X. Liu, W. Chen, and T. Wang, "pkudblab at semeval-2016 task 6: A specific convolutional neural network system for effective stance detection," in *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pp. 384-388, 2016.
[50] A. Wang et al., "Glue: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint arXiv:1804.07461*, ACL, 2018.
[51] S. Oraby et al., "Creating and characterizing a diverse corpus of sarcasm in dialogue," *arXiv preprint arXiv:1709.05404*, ArXiv. /abs/1709.05404 [cs.CL], 2017.
[52] M. Khodak, N. Saunshi, and K. Vodrahalli, "A large self-annotated corpus for sarcasm. *arxiv," arXiv preprint arXiv:1704.05579,* arXiv:2312.04642v1 [cs.CL], 2018.
[53] A. Mishra, D. Kanojia, and P. Bhattacharyya, "Predicting readers' sarcasm understandability by modeling gaze behavior," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, AAAI, 2016, pp. 3747–3753.
[54] N. Majumder et al., "Sentiment and sarcasm classification with multitask learning," *IEEE Intelligent Systems*, vol. 34, no. 3, pp. 38–43, IEEE, 2019.
[55] G. Nkhata, "Movie reviews sentiment analysis using bert," Masters thesis, University of Arkansas, Fayetteville, AR, USA, December 2022.
[56] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, "Stance and sentiment in tweets," *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media*, vol. 17, no. 3, pp. 1–23, ACM, 2017.
[57] A. Sen, M. Sinha, S. Mannarswamy, and S. Roy, "Stance classification of multi-perspective consumer health information," in *Proceedings of the ACM India joint international conference on data science and management of data*, ACM, 2018, pp. 273–281.
[58] B. Han and T. Baldwin, "Lexical normalisation of short text messages: Makn sens a# twitter," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, ACL, 2011, pp. 368–378.
[59] Keredson, "Wordninja," https://github.com/keredson/wordninja, 2017, [Online; accessed 19-April-2024].
[60] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "Semeval-2016 task 6: Detecting stance in tweets," in *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, ACL, 2016, pp. 31–41.
[61] J. Du, R. Xu, Y. He, and L. Gui, "Stance classification with target-specific neural attention networks." *International Joint Conferences on Artificial Intelligence*, ICAI, 2017, pp. 3988–3994.
[62] Y. Kim, "Convolutional neural networks for sentence classification," *in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, ACL, 2014, pp. 1746–1751.
[63] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, ArXiv. /abs/1412.6980[cs.LG], 2014.