Intermediate-Task Transfer Learning: Leveraging Sarcasm Detection for Stance Detection

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

Abstract—Stance Detection (SD) in the context of social media has emerged as a prominent area of interest with implications for social, business, and political applications, thereby garnering escalating research attention within the realm of Natural Language Processing (NLP). The inherent subtlety, nuance, and complexity of texts procured from online platforms via crowd-sourcing pose challenges for SD algorithms in accurately discerning the author's stance. Particularly, the inclusion of sarcastic and figurative language drastically impacts the performance of SD models. This paper addresses this challenge by employing sarcasm detection intermediate-task transfer learning tailored for SD. The proposed methodology involves the fine-tuning of BERT and RoBERTa and the sequential concatenation of convolutional, bidirectional LSTM, and dense layers. Rigorous experiments are conducted on publicly available benchmark datasets to evaluate our transfer-learning framework. The performance of the approach is assessed against various State-Of-The-Art (SOTA) baselines for SD, providing empirical evidence of its effectiveness. Notably, our model outperforms the best SOTA models, achieving average F1-score gaps of 0.038 and 0.053 on the SemEval 2016 Task 6A Dataset (SemEval) and Multi-Perspective Consumer Health Query Data (MPCHI), respectively, even prior to sarcasm-detection pre-training. The integration of sarcasm knowledge into the model proves instrumental in mitigating misclassifications of sarcastic textual elements in SD. Our model accurately predicts 85% of texts that were previously misclassified by the model without sarcasm-detection pre-training, thereby amplifying the average F1-score of the model. Furthermore, our experiments revealed that the success of the transfer-learning framework is contingent upon the correlation of lexical attributes between 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 in the context of SD and simultaneously exploits the concatenation of BERT or RoBERTa with other deep-learning techniques, establishing the proposed approach as a foundational baseline for future research endeavors in this domain.

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

I. INTRODUCTION

Social media platforms, increasingly popular, enable individuals to freely express opinions and connect globally for real-time updates on diverse topics [1]–[3]. Discourse on emerging subjects yields substantial data valuable for Natural Language Processing (NLP) tasks, notably Stance Detection (SD). SD is the automated identification of an individual's stance on a specific topic based solely on their utterance or authored material [2][4]–[6]. Stance labels categorize expressions into *InFavor*, *Against*, or *None*. This phenomenon, particularly on social media, is a burgeoning focus in social, business, and political applications [3][7].

Previous SD research has been evaluated using the publicly available datasets crowd-sourced from online platforms [2][3][5][8]. However, texts procured from online platforms are often characterized by subtlety, nuance, and complexity, featuring inherent sarcastic and figurative language. This complexity poses challenges for SD algorithms in accurately discerning the author's stance [2]. Additionally, targets are not consistently mentioned in text [4], and stances are not explicitly transparent. Consequently, inferring the author's stance becomes further complicated, often necessitating implicit inference through a combination of interaction, historical context, and social linguistic attributes, such as sarcasm or irony.

Prior work has explored intermediate-task transfer learning, involving fine-tuning a model on a secondary task before its application to the primary task to address the aforementioned challenge [1][9]–[13]. Specifically, [10] and [13] utilized sentiment classification to enhance their models for SD. In a similar vein, [1] incorporated emotion and sentiment classification prior to sarcasm detection. The study by [1] suggested that pre-training a model with sentiment analysis before sarcasm detection enhances overall performance, attributing this improvement to the correlation between sarcasm and an implied negative sentiment. This finding aligns with one of our experimental observations in Section IV, wherein most sarcastic sentences with an "Against" stance were initially misclassified as "InFavor" before the integration of sarcasm pre-training into our model. Nonetheless, sarcasm language in the target tasks has detrimentally affected performance, and previous research has not explored the sarcasm phenomenon for enhancing SD models. In this study, our focus is to experiment with and employ sarcasm detection as an intermediate task tailored to improve SD performance.

Sarcasm detection involves inferring intention or secondary meanings from an utterance, discounting literal meaning [14]. It employs positive words and emotions to convey negative or undesirable figurative attributes, serving as a mechanism to express opinions using seemingly conflicting language [15]–[17]. Sarcasm can alter the stance of a text from *Against* to *InFavor* and vice versa [16][18]. Thus, we propose infusing sarcasm knowledge into the model before SD fine-tuning to enhance performance.

This work employs a model framework consisting of BERT [19] or RoBERTa [20], two convolutional layers (Conv), a Bidirectional LSTM layer (BiLSTM), and a dense layer. Experimental results affirm the efficacy of our approach, demonstrated by improved F1-scores upon the inclusion of sarcasm detection in the model framework. Furthermore, the significance of this approach is emphasized by presenting a sample of sarcastic texts from datasets during a failure analysis of the original SD model results, prior to the incorporation of sarcasm intermediate-task pre-training. Exploring three publicly available sarcasm datasets, we find that different sarcasm detection tasks impact SD performance variably, depending on linguistic and quantitative attributes. 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.
- *Performance Superiority:* Demonstrating superior performance against State-Of-The-Art (SOTA) SD baselines, even without sarcasm detection pre-training, indicated by higher 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 conclusion and recommendations for further study are provided in Section V. The final section critically examines the limitations inherent in our study.

II. Related work

This section conducts a literature review on SD and intermediate-task transfer learning.

A. Stance Detection (SD)

The literature on SD has traditionally explored two primary perspectives: Target-Specific SD (TSSD), focusing on individual targets [2][3][21][22], and Multi-Target SD (MTSD), concurrently inferring stances towards multiple related subjects [22]–[25]. Early SD approaches utilized rule-based methods [21][26], followed by classical machine learning techniques [27][28]. Later, the emergence of deep learning models led to neural networks supplanting classical

approaches [4][13][29][30]. For instance, a neural ensemble model incorporating BiLSTM, attention mechanism, and multi-kernel convolution was presented in [29], evaluated on both TSSD and MTSD. While our work shares similarities in model framework, it distinctively employs BERT or RoBERTa and introduces an intermediate-task transfer learning technique, deviating from ensemble approaches and multi-kernel usage.

Recent efforts have explored the use of pre-trained language models for SD. While [2] conducted a comparative study, fine-tuning pre-trained BERT against classical SD approaches, [22] employed BERT as an embedding layer to encode textual features in a zero-shot deep learning setting, yielding promising results; however, both studies observed challenges in accurately classifying sarcastic examples. On the other hand, [21] experimented with ChatGPT, prompting the model directly with test cases to discern their stances.

B. Intermediate-Task Transfer Learning

Recent studies have also embraced intermediate-task transfer learning to transfer knowledge from a data-rich auxiliary task to a primary task [12]. This technique has shown significant success in various NLP tasks. For instance, [9] employed supervised pre-training with four-example intermediate training tasks to enhance performance on the primary task evaluated using the GLUE benchmark suite [31]. Furthermore, [13] introduced few-shot learning, utilizing sentiment-based annotation to improve cross-lingual SD performance. Additionally, [1] employed transfer learning by separately fine-tuning pre-trained BERT on emotion and sentiment classification before fine-tuning the model on the primary task of sarcasm detection, leveraging the correlation between sarcasm and negative sentiment polarity.

To our knowledge, prior work 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 TSSD within a model framework comprising BERT, Conv, BiLSTM, and a dense layer.

III. Methodology

This section outlines our approach, encompassing the intermediate-task transfer learning and the underlying model architecture.

A. Intermediate-Task Transfer Learning

Our model adopts a single intermediate-task training, which consists of two phases: pre-training on an intermediate task and fine-tuning on a target task.

1) Target Task: The focal task in this study is SD, where the objective is to predict the stance expressed in a given text, such as a tweet, towards a specified target, like '*Feminist Movement*'. A tweet, denoted as T, is represented as a word sequence $(w_1, w_2, w_3, ...w_L)$, with L denoting the sequence length. Stance labels are categorized as *InFavor* (supporting the target/topic/claim), *Against* (opposing the topic), or *None* (indicating neutrality towards the target).

2) Intermediate Task: The intermediate task in this study is sarcasm detection. As prior research has not employed sarcasm as an intermediate task, we investigate the following three sarcasm-detection tasks to gain insights into the crucial linguistic attributes for a model to learn from the intermediate task, aiming to enhance SD performance.

Sarcasm V2 Corpus (SaV2C). The SaV2C dataset, introduced by [32], presents a diverse corpus of sarcasm, utilizing syntactical cues and crowd-sourced from the Internet Argument Corpus (IAC 2.0). It comprises 4,692 lines extracted from Quote and Response sentences in political debate dialogues. Our exploration focuses on the General Sarcasm category within the dataset, containing 3,260 instances each of sarcastic or non-sarcastic comments.

The Self-Annotated Reddit Corpus (SARC). The SARC dataset [33] is derived from Reddit. In contrast to the other datasets, sarcasm annotations in SARC are directly provided by the authors, ensuring reliable and trustworthy data. Due to accessibility issues with the original website, we obtained the Main Balanced first version of the dataset directly from the author of [1]. This version comprises 1,010,826 training samples, evenly distributed between sarcastic and non-sarcastic instances.

SARCTwitter (ST). The ST dataset [34] is designed to predict readers' sarcasm understandability using features, including eye movement. In our study, we utilized the dataset variant employed by [35], excluding the eye movement feature. Crowd-sourced from Twitter (X), ST is manually annotated by seven readers and contains 350 sarcastic and 644 non-sarcastic tweets. Our intermediate-task transfer learning pipeline is depicted in Figure 1.
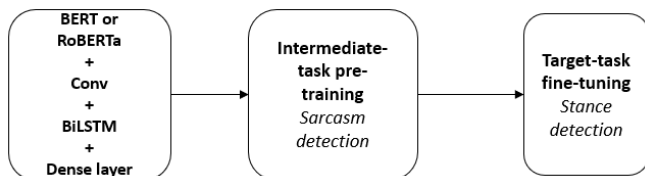

Figure 1. Intermediate-task transfer learning pipeline.

B. Underlying Model Architecture

The entire model framework primarily comprises an input layer, an embedding layer, and deep neural networks.

1) Input Layer: This layer takes a text S encoding the stance information and comprising n words. S is transformed into a vector of words and passed to the embedding layer.

2) The Embedding Layer: We employ BERT [19] and RoBERTa [20] for textual input encoding into hidden state H in our experimentation. Noteworthy achievements of these language models in the literature [1][2][9][12][30][36] motivate their exploration to identify the most suitable model for alignment with our research objectives.

3) Deep Neural Networks: This module utilizes Conv, a BiLSTM layer, and a dense layer, positioned atop the embedding layer. The purpose of incorporating convolution is to

discern specific sequential word patterns within a sentence, generating a composite feature map from H. This feature map facilitates the BiLSTM layer in acquiring nuanced higher-level stance representations, which are subsequently mapped into a more differentiable space by the dense layer. Figure 2 depicts the overall model framework.

IV. EXPERIMENTS

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

A. Datasets

For evaluation purposes, we employed two publicly available SD datasets: 1) the well-established SemEval 2016 Task 6A Dataset (SemEval); and 2) the Multi-Perspective Consumer Health Query Data (MPCHI).

1) SemEval: The SemEval [37] task encompasses tweets manually annotated for stance towards a specified target, a target of opinion, and sentiment. Our experiments exclusively utilize tweets and their corresponding stance annotations. The dataset comprises tweet data associated with five distinct targets: Atheism (AT), Climate Change (CC), Feminist Movement (FM), Hillary Clinton (HC), and Legalization of Abortion (LA).

2) MPCHI: MPCHI [38] serves as a dataset for stance classification to enhance Consumer Health Information (CHI) query search results. Comprising formal texts extracted from top-ranked articles corresponding to queries on a specific web search engine, the dataset includes sentences related to five distinct queries, which are also the targets for stance classification: MMR vaccination can cause autism (MMR), E-cigarettes are safer than normal Cigarettes (EC), women should take HRT post menopause (HRT), Vitamin C prevents common cold (VC), and Sun exposure leads to skin Cancer (SC).

Consistent with [2], the datasets are partitioned into training and test sets following similar proportions. Each text in the datasets is annotated with one of three classes: *InFavor*, *Against*, and *None*. Table I provides statistical details describing the datasets.

B. Data pre-processing

We conducted standard data pre-processing steps, including case-folding, stemming, stop-word removal, and deletion of null entries, across all datasets. Text normalization, following the approach by [39], and hashtag pre-processing, using Wordninja [40], were also performed. However, for neural network models relying on pre-trained embeddings, stemming and stopword removal were omitted, as stemmed versions of terms might not be present in the pre-trained embeddings. The default tokenizer for the corresponding pre-trained language model was employed to tokenize words in tweets before supplying them to the classifier.
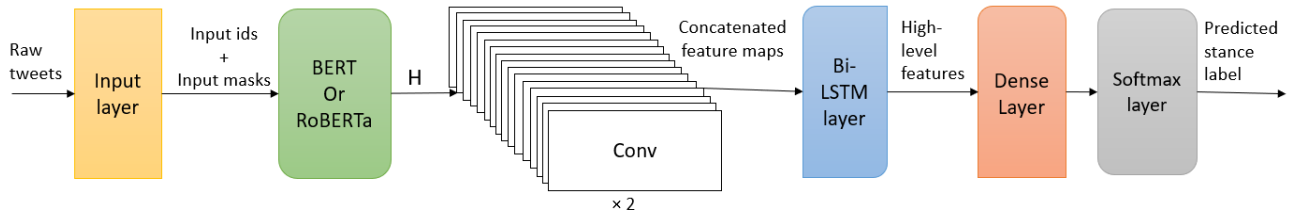
Figure 2. Proposed model framework.

TABLE I
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

C. Baseline models

Our model is evaluated against the top-performing results from the SemEval challenge [41], as reproduced in [2] with minor modifications. Additionally, we compare our model's performance with the most recent SOTA methods in SD.

1) SemEval models: We select the Target-Specific Attention Neural Network (TAN-) proposed by [42], and the 1-D sem-CNN introduced by [43]. Additionally, we adopt Com-BiLSTM and Com-BERT, implementations provided solely by [2].

2) ChatGPT and ZSSD: The work by [21] investigated ChatGPT for SD by directly probing the generative language model for the stance of a given piece of text, focusing on the SemEval task with specific targets: FM, LA, and HC. On the other hand, the Zero-Short SD (ZSSD) technique [22], employing contrastive learning, was similarly implemented on SemEval only.

D. Experimental settings

The inductive approach to transfer learning was applied to the entire model framework, initializing the target task model with parameters learned during sarcasm-detection pre-training. Given the primary focus on enhancing model efficacy for the target task, intermediate tasks were divided into training and validation sets solely for model pre-training on sarcasm detection. In contrast, the target task featured a separate test set for final evaluations and comparisons. As Sav2C and ST are the smallest intermediate-task datasets, five-fold cross-validation was employed on both, while SARC, with its larger size, undergoes an 80/20 train/validation split.

A kernel of size 3, 16 filters, and a ReLu activation function have been employed for the convolutional layer. The BiLSTM layer has been used with a hidden state of 768, matching the hidden state size of the pre-trained language models. The dense layer has employed an output size of 3 and a softmax activation function. All experiments have been conducted on an NVIDIA Quadro RTX 4000 GPU.

Hyperparameter tuning has been performed through multiple experiments, selecting the best intermediate-task training scheme based on holdout development set results. The optimal per-task model has been then evaluated on the test set. Iterating over datasets with a mini-batch of 16 samples, the Adam optimizer [44] has been used for parameter learning, employing cross-entropy loss as the cost function. Training runs span 10 to 50 epochs, with early stopping triggered if validation accuracy on holdout data stagnates for five consecutive epochs. The training schedule involves an initial learning rate of 3e-5, decayed to a final learning rate of 1e-9 for the intermediate-task and 1e-10 for the target task. A dropout of 0.25 is introduced between model layers to address overfitting. Due to imbalanced class distributions, class weights are incorporated during training to enhance model generalization on underrepresented classes. Experimental setups adhere to the original papers for baseline models unless otherwise specified, in which case our experimental configurations are adopted.

E. Evaluation metrics

For consistency with prior works [2][4][41], the evaluation of our model employs the macro-average F1-score for the *InFavor* and *Against* classes.

F. Results

We report averaged results from five experiment runs on each target task. Table II shows experimental outcomes before sarcasm pre-training in our model. Results for ChatGPT and ZSSD are directly transcribed from their original papers, while the results for other baseline models have been replicated in our experimentation. The table illustrates the commendable performance of our BERT-based model across various targets, with notable superiority in all aspects except for HC and CC, where ChatGPT and our RoBERTa-based model excel, respectively. Consequently, we opt to proceed with our BERT-based model in subsequent experimental results.

Table III presents experimental results involving sarcasm detection pre-training with our model only. Model performance improves by **0.050** and **0.003** on SemEval and MPCHI, respectively, when pre-trained with ST, surpassing all baseline models in Table II, but diminishes with Sav2C and SARC.

Table IV presents results of an ablation study using ST only. Different base model components were systematically removed to assess the contribution of each constituent module to the entire model framework. As shown in the table, the model with all components—BERT, Conv, BiLSTM, and sarcasm pre-training—performs the best with average F1-scores of **0.775** and **0.724** on SemEval and MPCHI, respectively.

G. Failure Analysis and Discussion

Subsequent to obtaining results in Table II, a failure analysis was conducted on misclassified test samples. Predominantly, misclassifications on SemEval were associated with texts containing sarcastic content, aligning with prior findings [2]. This observation substantiated the motivation for considering sarcasm-detection pre-training before fine-tuning on SD. On the contrary, misclassifications on MPCHI were associated with samples encompassing colossal and generic health-related facts neutral to the respective target under study. Additional observations stemming from the experiments and results across all tasks are outlined below.

1) Our model outperforms SOTA models even without sarcasm detection: Specifically, it outperforms ChatGPT and Com-BERT, the best models, on SemEval and MPCHI, by **0.038** and **0.053** on average F1-scores, respectively. While Com-BERT employs only BERT and a dense layer as a classifier, our model incorporates Conv and BiLSTM before the dense layer, contributing to the observed performance improvement. Additionally, it was noted that the inclusion of the BiLSTM module in our model yielded better performance than using pooling layers after the Conv module. This suggests the effectiveness of our model architecture and its ability to capture nuanced representations, leading to proper generalization on SD tasks.

2) Sarcasm detection is correlated with SD: Consider an illustrative misclassified example: "*I like girls. They just need to know their place. #SemST*", a sarcastic comment from the FM target in SemEval. The ground truth for this example is *Against*, but it was predicted as *InFavor* before sarcasm-detection pre-training. Notably, most sarcastic samples in the

Against class were misclassified as *InFavor* due to their explicit positive content. After incorporating sarcasm knowledge into the model through pre-training, 85% of misclassified sarcastic samples were predicted correctly. This observation underscores the relevance of sarcasm-detection pre-training in improving the performance of SD models in our experimentation.

3) Not every sarcasm detection model is a good candidate for intermediate-task transfer learning on SD: The inclusion of SARC and SaV2C knowledge in the model pipeline introduced noise and adversely affected model performance on SD compared to incorporating ST knowledge. An analysis of Sav2C and SARC revealed several discrepancies between the intermediate-task datasets and the target tasks. Firstly, the average sentence length in Sav2C and SARC is longer than in SemEval and MPCHI. Secondly, SARC is sourced from different domains than both SemEval and MPCHI, leading to disparities in topic coverage, vocabulary overlap, and the framing of ideas across datasets. Additionally, SARC, being the largest intermediate task, covers a wide range of topics through various subreddits. In contrast, ST, the best-performing intermediate task, shares a similar average sentence length with the target tasks. Moreover, both ST and SemEval are crowd-sourced from Twitter, which likely contributes to the strong performance observed when using ST as an intermediate task on the SemEval dataset. Consequently, the mismatched attributes render certain intermediate tasks less commensurated and less correlated with target tasks, resulting in a negative impact on model performance. Careful consideration and experimentation are essential when selecting a suitable sarcasm model for transfer learning in the context of SD.

4) Ablation study regarding sarcasm knowledge: The variations in the ablation study results in Table IV help to isolate the effects of each module and determine their individual contributions to the overall improvement in SD performance through sarcasm detection pre-training. Comparing our best average results in Table II and Table IV, the infusion of sarcasm knowledge significantly enhances model performance on the SemEval task compared to the MPCHI task. The SemEval task comprises extensive opinionated and sarcastic texts. Conversely, the majority of examples in the MPCHI dataset encompass extensive health-related facts, unrelated to specific targets, aside from occasional sarcasm-related expressions. Consequently, there is a modest increase in performance on MPCHI even when sarcasm detection is utilized. This observation prompts the consideration of exploring variants of BERT or RoBERTa embeddings pre-trained on health-related data specifically for SD on MPCHI as a potential avenue for future work.

V. CONCLUSION AND FUTURE WORK

In this study, we introduced a transfer-learning framework that leverages sarcasm detection for SD. RoBERTa and BERT were individually fine-tuned and sequentially concatenated with other deep neural networks, with BERT delivering

TABLE II
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 III
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 IV
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**

promising results. The model underwent separate pre-training on three sarcasm-detection tasks before fine-tuning on two target SD tasks. Evaluation against SOTA models demonstrated superior performance, even prior to incorporating sarcasm knowledge. We established the correlation between sarcasm detection and SD, with the infusion of sarcasm knowledge boosting model performance, accurately predicting 85% of misclassified samples in the SemEval task. Failure analysis revealed SemEval's abundance of opinionated sarcastic samples, underscoring the efficacy of sarcasm pre-training, compared to MPCHI, characterized by generic health-related facts unrelated to specific targets. Additionally, we showed that not every sarcasm-detection intermediate task improved SD due to incongruous linguistic attributes. Finally, an ablation study highlighted that optimal model performance is achieved when utilizing all model constituents.

To the best of our knowledge, this is the inaugural exploration of sarcasm-detection pre-training applied to the BERT(RoBERTa)+Conv+BiLSTM architecture before fine-tuning for SD. Serving as a foundational reference, our approach establishes a baseline for future researchers in this domain. Future investigations will assess variant BERT or RoBERTa embeddings tailored to health-related text data for the MPCHI task. The research will also concentrate on cross-target SD for both tasks and a more comprehensive examination of other intermediate tasks, including sentiment and emotion knowledge.

VI. LIMITATIONS

Despite the significant advancements this study brings to NLP applied to social media contexts, several limitations merit consideration. Firstly, the extent of model performance enhancement is contingent upon the attributes of both the intermediary sarcasm detection task and the ultimate SD task. The divergence in linguistic characteristics across datasets utilized for sarcasm detection and SD potentially constrains the broader applicability of the study's outcomes. Secondly, although the integration of BERT or RoBERTa with other deep-learning methodologies represents an innovative approach, the intricate nature of the model architecture may present computational resource challenges and interoperability issues in certain contexts. Lastly, the extensive reliance on fine-tuning techniques and specific datasets raises concerns regarding the model's capacity to generalize effectively across diverse text types or domains not encompassed within the training data corpus.

ACKNOWLEDGMENT

REFERENCES

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

[2] 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.

[3] 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.

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

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

[6] 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.

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

[8] 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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.

[31] A. Wang et al., "Glue: A multi-task benchmark and analysis platform for natural language understanding," *arXiv preprint arXiv:1804.07461*, ACL, 2018.

[32] 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.

[33] 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.

[34] 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.

[35] N. Majumder et al., "Sentiment and sarcasm classification with multitask learning," *IEEE Intelligent Systems*, vol. 34, no. 3, pp. 38–43, IEEE, 2019.

[36] G. Nkhata, "Movie reviews sentiment analysis using bert," Masters thesis, University of Arkansas, Fayetteville, AR, USA, December 2022.

[37] 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.

[38] 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.

[39] 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.

[40] Keredson, "Wordninja," https://github.com/keredson/wordninja, 2017, [Online; accessed 19-April-2024].

[41] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "Semeval-2016 task 6: Detecting stance in tweets," in *Proceedings of*

the 0th international workshop on semantic evaluation (SemEval-2016), ACL, 2016, pp. 31–41.

[42] 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.

[43] 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.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, ArXiv. /abs/1412.6980[cs.LG], 2014.