2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA) | 979-8-3503-4503-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/DSAA60987.2023.10302466

Learning Representations through Contrastive Strategies for a more Robust Stance Detection

Udhaya Kumar Rajendran
Department of Computer Science
Lakehead University
Canada
rajendranu@lakeheadu.ca

Amir Ben Khalifa[§]
École Nationale des Sciences de l'Informatique
Université de La Manouba
Tunisia
amir.benkhalifa@ensi-uma.tn

Amine Trabelsi
Department of Computer Science
Université de Sherbrooke
Canada
amine.trabelsi@usherbrooke.ca

***Abstract*—Stance Detection refers to the process of determining an author's position towards a particular issue or target in a text. Previous research suggests that existing systems for Stance Detection are not resilient enough to handle variations and errors in input sentences. In our proposed methodology, we utilize Contrastive Learning to learn sentence representations. We achieve this by bringing semantically similar sentences and those implying the same stance closer to each other in the embedding space. To compare our approach, we use a pretrained transformer model that is directly finetuned with the stance datasets. We evaluate the resilience of the models using char-level and word-level adversarial perturbation attacks and show that our approach performs better and is more robust to the different adversarial perturbations introduced to the test data. Our approach is also shown to perform better on small-sized and class-imbalanced stance datasets. We further experiment with unlabeled stance datasets to make the representation learning independent of domain-specific labels, and the models trained with our approach on unlabeled datasets are still robust and perform comparably to those trained with labeled data.**

I. Introduction

A controversial topic divides people into two groups with different views (support/against) on the topic of discussion. Some popular, controversial topics include the Legalization of Abortion, Concern about Climate Change, Gay Marriage, Obama, the Legalization of Marijuana, Feminism, and Atheism. The existing Stance Detection models are non-robust, and even simple perturbations in the input sentences affect the model's performance [1]. For example, the input sentence 'Fetus is not human' has the stance label of 'support' for the topic of 'Legalization of Abortion.' However, when there is a variation to the same input sentence, such as 'A bunch of cells is not human,' it will confuse the model in reproducing the same stance label of 'support.' Also, spelling errors, missing words, repetition of words, and other commonly occurring errors in the text are the adversarial errors that make the Stance Detection models fall short in detecting the stance compared to humans. We aim to make the Stance Detection system more robust to adversarial perturbations by accommodating the variations and errors in the text when detecting the stance. We primarily concentrate on binary stances (e.g., support/against) in social media for English texts, such as tweets, news comments, and discussion forums.

We use the Contrastive Learning (CL) approach to construct more robust sentence representations for the Stance Detection task. Given an example we call anchor, the CL technique brings the similar example closer to the anchor and drives the dissimilar example away from the anchor in the representation space. We build similar (positive) and dissimilar (negative) examples for CL by considering the stance label of the examples. We mainly explored different strategies for building positive and negative examples for an anchor example to learn the sentence representations in a contrastive fashion. Along with CL, we use Masked Language Modeling [42] as a token-level objective to learn textual representations (see Fig. 1). Our code is available in the GitHub repository [1]. We make the following contributions.

- We develop an approach using a CL framework with different positive and negative pairs selection strategies to learn more robust sentence representations to use in the Stance Detection task. To the best of our knowledge, this work is the first to employ a Contrastive Learning framework to learn robust sentence representations in the context of Stance Detection task.
- We develop an approach using CL to learn sentence representations from unlabeled examples to use in the Stance Detection task.
- We conduct a comprehensive empirical investigation using various settings and datasets for stance detection, analyzing the results and providing valuable insights into effective strategies for different contexts.

II. Background and Related Work

A. Stance Detection

Many approaches [2]–[8], [31], [32] were proposed to tackle different problems in the Stance Detection task.

However, the existing Stance Detection models have been shown to be sensitive to adversarial errors, and changes in the vocabulary of the input sentences [1]. The adversarial robustness of the model is measured by making the model predict against the test set with char-level, and sequence-level modifications to the input as well as with the word substitutions [9]–[11]. Moradi and Samwald [12] used various

[§]Work done as undergraduate intern

[1]https://github.com/rajendranu4/stance-detection

perturbations for Char-level such as Insertion, Deletion, Replacement, etc., and word-level perturbations such as Replacement with Synonyms, Negation, etc. Schiller et al. [1] used the resilience score introduced by Thorne et al. [19] to measure the robustness of the model. Jayaram et al. [40] tested the faithfulness of a model's prediction by introducing attention weights for the words in the text and shown that attribution prior as well as the attention weights improve the model's rationales. Yang et al. [41] tested the model's reliability and filtered out doubtful predictions by having a negative version of the original Perspective in the training dataset, i.e., for some of the original perspectives, a negative version of the same perspective is generated by some methods to include as part of the training.

B. Contrastive Learning

CL is used to acquire better representations of text for many natural language tasks such as Question-Answering [13], multiple choice video questions, text-to-video retrieval [14], text summarization [15]–[17] etc. Sun et al. [38] used the Contrastive Learning framework to alleviate the exposure bias problem (discrepancy between training and inference) in text summarization. Wu et al. [39] used Contrastive Learning to learn noise invariant sentence representation with the help of different sentence-level augmentation strategies like span deletion, substitution, and reordering.

In this study, our objective is to develop and explore a range of strategies encompassed within contrastive learning. Our aim is to enhance the quality of document representations specifically for the task of stance detection, consequently bolstering the robustness of stance detection classification models.

III. Method

As depicted in Fig. 1, our methodology primarily utilizes the DistilRoBERTa Transformer model [29], along with Contrastive Learning and Masked Language Modeling (MLM) techniques [42], for the pretraining phase on the stance dataset. Subsequently, the model undergoes fine-tuning specifically for the Stance Detection downstream task. The inclusion of the MLM objective aids in capturing word-level representations, whereas the contrastive learning objective focuses on learning more about sentence-level meaning.

In this section, we detail the devised contrastive learning strategies to enhance the robustness of our stance detection models. We elaborate on the methods used to measure the robustness of these models and describe an adapted approach for our proposed methodology when labeled data is not accessible.

A. Contrastive Learning

Contrastive Learning maps the representations of 'similar' patterns closer to each other while pushing the representations of 'different' patterns farther away in the embedding space. CL learns from the examples that are hard to distinguish

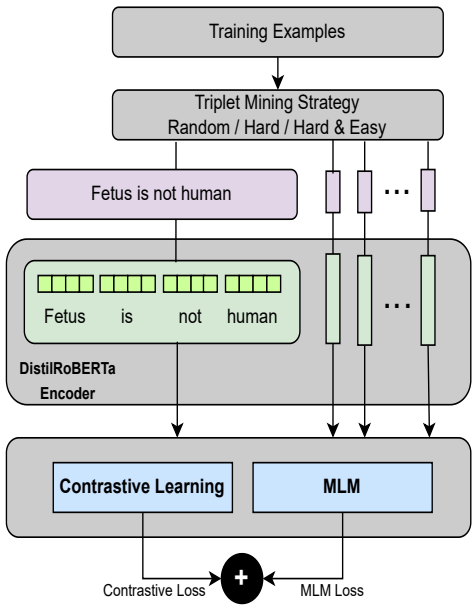

Fig. 1: Architecture diagram for learning sentence representations using CL and MLM objectives to further use in the Stance Detection task.

in the representation space from the anchor example [18]. The goal of the contrastive loss function (loss_{CL}) given by Equation (1) is to minimize the distance between the anchor-positive pair (d_{A}, d_{+}) and to maximize the distance between the anchor-negative pair (d_{A}, d_{-}). m in Equation (1) is the margin and is the desired difference between the anchor-positive and anchor-negative distances. CL makes similar examples have similar representations in the representation space, which makes the language model less sensitive (more robust) to adversarial errors, including changes in the text's vocabulary. Similar to BERT pretraining [42], we use MLM to learn word-level representations by masking a percentage of words in a sentence and allowing the model to predict the masked words given the context of the surrounding words. The final loss is the sum of CL loss and MLM loss.

$$loss_{\text{CL}} = max\{|d_{\text{A}} - d_{+}| - |d_{\text{A}} - d_{-}| + m, 0\} \qquad (1)$$

B. Contrastive Learning Strategies

In our experiments, different strategies, as described below (Random, Hard, and Hard & Easy strategies), are used to select positives and negatives for a particular anchor for Contrastive Learning. The combination of anchor, positive and negative, is called a triplet.

Random Strategy The triplets are formed randomly, satisfying the anchor-positive and anchor-negative selections.

Hard Strategy Hard positive (same ground truth label as the anchor but far away from it) and hard negative (different ground truth label from the anchor but close to it) are chosen for an anchor.

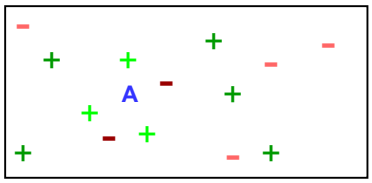

Fig. 2: Illustration of Easy Positive + and Negative -, Hard Positive + and Negative - samples for an Anchor sample A in the representation space.

Hard & Easy (H&E) Strategy One Hard triplet similar to the Hard strategy and one Easy triplet (easy positive and easy negative) are chosen for an anchor (see Fig. 2 for a graphical illustration of the hard and easy positives and negatives for an anchor in the representation space).

C. Robustness of Stance Detection Systems

We measure the robustness of the model with the resilience score Res in Equation (2) introduced by Thorne et al. [19] by identifying the deviation between the performances of the model with the original test set $p(s,t)$, also called as ***non-perturbed test set*** and the adversarial ***perturbed test set*** $p(s,a)$ with adversarial attack a for a natural language system s.

We use four adversarial attacks *spelling errors*, *adding tautology*, *synonym replacements* and *paraphrasing* (see Table II). The **correctness ratio** c_a of an adversarial attack a gives the total number of correctly transformed examples (preserving readability, semantics, and grammar) from the number of examples considered for perturbation.

Spelling error. We introduced spelling errors to perturb all the original sentences in the test set.

Adding tautology. All the input sentences in the test set are appended with '***False is not true and,***'.

Synonyms replacement. We select 15 words that are frequent in the dataset and replaced them with their synonymous words which do not change the meaning of the sentences. Since the frequent words are selected for the *synonyms replacement* attack, the words that are selected may or may not be in a given example. Hence not necessarily all the examples are perturbed for the *synonyms replacement* adversarial attack though all the examples are candidates for this attack.

Paraphrasing. We utilized GPT 3.5 [43] to paraphrase the sentences. See Table I for the GPT series, GPT model and the instruction used to paraphrase the sentences. Since we have restrictions in the usage of GPT 3.5 model, we have taken 20 examples from each of the datasets for our experiments.

$$Res = \left| \frac{\sum_{a \in A} c_a * (p(s,t) - p(s,a))}{\sum_{a \in A} c_a} \right| \quad (2)$$

D. Learning and Leveraging Robust Representations

Let F be the transformer model (DistilRoBERTa), for each of the input sequences $x^{(i)}$ from batch j, the MLM objective masks a percentage of tokens, and the model predicts the masked token with the help of the surrounding tokens. Again, for the same input sequences from batch j, the Contrastive Learning framework identifies the triplets for each $x^{(i)}$ (anchor) based on the strategies explained in Section III-B. Given an anchor sentence, positive example and a negative example, the model learns the sentence representations by training the classifier to distinguish positive example from the negative example. The combined loss (Contrastive Learning + MLM) is backpropagated to adjust the weights of the transformer model. Now the transformer model F trained with the Contrastive Learning and MLM objectives is added with a classification layer on top and finetuned with the stance datasets. Let $P^{(o)}$ be the model's performance after finetuning with the stance dataset D. The robustness of model F is identified by testing the finetuned model F against the perturbed test set D_p. Let $P_p^{(se)}$, $P_p^{(n)}$, and $P_p^{(sm)}$ be the performances of the model against the perturbed test sets generated with the adversarial attacks spelling errors, tautology, and synonyms respectively.

E. Proposed Methodology on Unlabeled dataset

It is standard that subject matter experts manually annotate the stance labels for all the examples in the dataset. The effort needed to manually annotate the stance labels is huge, indicating that data labeling is costly. Also, the model based on labeled data tends to be domain specific. To avoid this manual effort, we explore the dataset without ground truth labels. We remove the ground truth labels from DebateForum dataset. The pseudo-ground truth labels are assigned for the dataset with the help of the clustering method using the user's metadata proposed by Li et al. [22]. The method provides two clusters from the input dataset; all the examples in a cluster are assigned the same stance labels. Here the size of the dataset is reduced as the examples selected for clustering are required to have the threshold such as the number of posts made by any user (at least 2) and the number of tokens in the example (at least 4). The transformer model is trained with the input text sequence and the stance labels assigned from clustering and use the Contrastive Learning and Masked Language Modeling objectives to learn the representations of the input text sequence. The pretrained model is then finetuned with the original ground truth label of the stance dataset.

IV. Experiments

We have utilized seven, commonly experimented, Stance Detection datasets, DebateForum (DF) [23], SemEval2016 (SE) [24], ARC [25], Perspectrum [26], FNC-1 [27], KSD-Biden and KSD-Trump [28]. We have retained only the examples that have support/against equivalent labels in the datasets

GPT Series	GPT Model	Instruction
GPT 3.5	text-davinci-003	Rephrase the following text sentence by sentence without modifying the mentions and hashtags and without adding other mentions and hashtags, DO NOT summarize the text, ONLY rephrase it and Ensure that the meaning of the text remains the same, but TRY TO USE DIFFERENT words, sentence structures and DO NOT USE ANY PRIOR KNOWLEDGE.

TABLE I: Paraphrasing sentences with GPT model.

Adversarial Attack	Original Sample	Perturbed Sample
Spelling Error	Green is the way **forward**	Green is the way **ferward**
Adding Tautology	The Olympics create a sense of national pride	**False is not True and** the Olympics create a sense of national pride
Synonyms	Golf is one of **independent** sports	Golf is one of **stand-alone** sports
Paraphrasing	The NHL has no plans to expand, said NHL. commissioner Gary Bettman	Gary Bettman, the NHL commissioner, declared that the league has no intention to expand.

TABLE II: Illustration of the different types of adversarial attacks for perturbing the test set to measure the robustness and reliability of the model.

as we mainly focus on binary stances. Out of the seven chosen datasets, the Perspectrum dataset has more instances (11825), KSD-Biden has the least number of instances (766) and FNC-1 is the most imbalanced dataset (78/22). See Table III for more information on the statistics of these datasets. Table V describes the datasets, the domain of the corresponding datasets, and an example from the dataset to show the input and the stance output.

A. Setups

The setups below vary according to the level of information leveraged to train and evaluate the conceived models. **With Ground Truth (WGT)**. The dataset has ground truth stance labels for every example.

Without Ground Truth. The ground truth label is removed from the dataset during training with Contrastive Learning and Masked Language Modeling. Pseudo-labels are assigned for the dataset as explained in Section III-E

Partial Perturbation (PP). The evaluation of models is carried out by perturbing with an adversarial attack only the examples that are correctly classified by the models from the original test dataset run.

Mixed Topics. Since all the datasets described in Table III consist of more than one topic, in this setting, we do not construct or evaluate the models on separate topics, i.e., by topicality. Instead, we consider all the examples of all topics as a whole during our experiments.

Individual Topics. In this setup, in contrast to the previous setting of Mixed Topics, the models are constructed and evaluated based on individual topic-related sub-dataset, e.g., Abortion for DebateForum, see Table IV. Each topic is an individual dataset with its own train/dev/test splits. We mainly consider DebateForum and SemEval2016 topics.

B. Models

We have used the DistilRoBERTa [29] as the transformer model for all our experiments. Inspired by the work of Giorgi et al. [21], we have used the code architecture and modified the loss objectives and the pipeline according to our experiment setup. The transformer model in our proposed methodology is not pre-trained from scratch. We use DistilRoBERTa pre-trained weights as the initial weights for the DistilRoBERTa model. We compare our proposed models described below with a baseline model.

Model$_{\text{Baseline}}$. is pretrained DistilRoBERTa model finetuned with stance datasets. Let B be the baseline model and ($x^{(i)}$,$y^{(i)}$) be the i^{th} instance with $x^{(i)}$ be the input sequence and $y^{(i)}$ be the corresponding label. B($x^{(i)}$) $\longrightarrow$ $\hat{y}^{(i)}$ is the prediction for the input sequence $x^{(i)}$ by the baseline model B, and the performance of the baseline model is to be compared with that of our proposed methodology.

Model$_{\text{Random}}$. For the input examples from a batch, the Contrastive Learning framework identifies all possible random triplets for each $x^{(i)}$ (anchor) to use in Contrastive Learning.

Model$_{\text{Hard}}$. is trained with a hard triplet which is formed with a hard positive and a hard negative for an anchor example.

Model$_{\text{H\&E}}$. One hard triplet and one easy triplet are used in Contrastive Learning

Model$_{\text{Random2}}$. The Hard and H&E strategies use only one and two triplets respectively from a batch of examples. We experiment with the Random strategy but with only two randomly formed triplets from a batch to compare it with the Hard and H&E strategy.

C. Settings

The number of characters and words used in social media posts is usually restricted to cut out the fluff. For example,

Dataset	# Examples	Classes	Splits		
			Train	Dev	Test
DebateForum	4904	for(60%), against(40%)	3431	884	589
SemEval2016	3170	favor(35%), against(65%)	2149	205	816
ARC	3368	agree(47%), disagree(53%)	2660	283	425
Perspectrum	11825	support(52%), undermine(48%)	6979	2072	2774
FNC-1	7121	agree(78%), disagree(22%)	4519	1301	1301
KSD-Biden	766	favor(50%), against(50%)	546	110	110
KSD-Trump	843	favor(41%), against(59%)	591	126	126

TABLE III: Statistics about the different datasets used for the experiments

Topic	Class Ratio	# Examples	Splits		
			Train	Dev	Test
$\text{Abortion}_{\text{DF}}$	56 / 44	1918	1341	288	289
$\text{GayRights}_{\text{DF}}$	64 / 36	1378	963	207	208
$\text{Marijuana}_{\text{DF}}$	71 / 29	629	439	95	95
Obama_{DF}	53 / 47	988	690	149	149
$\text{Abortion}_{\text{SE}}$	24 / 76	714	498	108	108
$\text{Atheism}_{\text{SE}}$	21 / 78	591	412	89	90
$\text{Climate}_{\text{SE}}$	90 / 10	364	253	55	56
$\text{Feminism}_{\text{SE}}$	35 / 65	782	546	118	118
$\text{HillaryClintion}_{\text{SE}}$	23 / 77	730	510	110	110

TABLE IV: The topicwise distribution of the datasets DebateForum and SemEval2016

Dataset	Domain	Example	Topic	Stance Label
DebateForum	Debating Forum	Passive smoking is harmful and secondhand smoke from the use of marijuana increases the chances of others suffering the damage by inhaling the smoke.	Marijuana	against
Arc		This is a great move by Wal-Mart. I hope they take out all the high fructose corn syrup out of their products as well. I avoid anything with high fructose corn syrup and as a result I have lost 37 pounds.	Wal-Mart can make us healthier	agree
Perspectrum		A game is less enjoyable if there is video replay.	There should be video replays for refs in football	undermine
SemEval2016	Social Media	Today Europe is breaking heat records, while Asia is breaking the lowest temperature records!! Should we not be concerned	Climate Change is a Real Concern	favor
KSD-Biden		i miss having a president that speaks eloquently. that has empathy and hope for a better tomorrow. fortunately, we will soon have that again with #bidenharris2020.	Biden	favor
KSD-Trump		not everyone in oklahoma is welcoming the president's visit	Trump	against
FNC-1	News	Tesla is reportedly choosing Nevada for its new battery factory.	Tesla to choose Nevada for Battery Factory	agree

TABLE V: Illustrates the domain of the different datasets used for the experiments and an example from each of the datasets

currently, Twitter [30] has a character limit of 280 characters per post to express the user's thoughts. In all our experiments, we use a word limit of 100 to capture the valuable meaning of the user's post. To allow maximum participation of different examples in CL, the training batch size is reduced from 16 to 8 as the strategies Hard and H&E mine one and two triplets, respectively, from a batch of examples for CL. All the other hyperparameters for the models are as per the transformer model's predefined values. We train the DistilRoBERTa model using CL (0.5 as margin, m) and MLM objectives (15% tokens masked) for 20 epochs to learn the sentence representations.

For unlabeled dataset setup, we generate pseudo-labels with the help of a clustering method introduced by Li et al. [22] to use them during CL. We then finetune the model with stance datasets for 4 epochs. See Tables VI, VII and VIII for more details on hyperparameters for pretraining and finetuning.

The Correctness Ratio for the adversarial attack '*adding tautology*' is 1 as the data is perturbed by prefixing the example sentence with the words ***False is not True and*** which does not change the truth value of the sentence, hence the stance labels for the sentence remains the same.

The Correctness Ratio for the adversarial attack '*synonyms replacement*' is also 1 as the words in a sentence are replaced with their synonyms which does not change the sentence's truth value and hence the stance labels for the sentences remain the same.

We use Flesch–Kincaid grade level [20] to check if the transformed sentence with the adversarial attack '*spelling error*' is readable. We consider the example after perturbation which has the same readability grade level as the original example as a correctly perturbed example. The Correctness Ratio of adversarial attack '*spelling error*' is 1 as all the

examples used in the experiments are correctly perturbed for all the datasets.

The correctness ratio of the '*paraphrasing*' adversarial attack is calculated by selecting 5 examples randomly from all the test sets, and manually annotated for semantic equivalence between the original texts and their paraphrased counterparts. The manual annotation task is conducted by three annotators with background knowledge in natural language processing. The inter-annotator agreement is calculated with Fleiss Kappa [44] and the agreement between the annotators is substantial with $\kappa = 0.67$. The correctness ratio for the '*paraphrasing*' attack is 0.8857 and 0.9333 for the Mixed Topics setup and Individual Topics setup, respectively.

The resilience of models is measured by perturbing the examples in the test dataset with the adversarial attacks individually for the experiment setups **Mixed Topics** and **Individual Topics**, see under Section IV-A. For example, the model $\text{Model}_{\text{Random}}$ is evaluated on the original non-perturbed dataset and the four datasets that are perturbed with the four adversarial attacks (*spelling*, *adiding tautology*, *synonyms replacement* and *paraphrasing*). For the experiment setups, **Mixed Topics + PP** and **Individual Topics + PP** , the resilience of the model is measured by making the model predict on the test set in which the perturbations are introduced on the examples that are correctly classified in the original non-perturbed test. For e.g., the $\text{model}_{\text{Hard}}$ is evaluated on the original non-perturbed dataset initially, then a dataset is prepared by perturbing (with an adversarial attack, for e.g., spelling attack) only the correctly classified examples from the original non-perturbed test run and finally, the model is evaluated on the prepared dataset to measure the resilience of the model.

We consider only the spelling and negation adversarial attacks for the experiments **Mixed Topics + PP** and **Individual Topics + PP** since not all the examples in a given set of examples are perturbed in *paraphrasing* and *synonyms replacement* adversarial attack. The difference in the performance of the models between the original non-perturbed test set and the adversarial test sets is measured to identify the robustness of the model. The percentage of examples perturbed from a given set of examples needs to be consistent across the different adversarial attacks as well as the different models. For e.g., from the original non-perturbed test set, if Model 1 predicts 60% of the examples correctly and Model 2 predicts 70% of the examples correctly, then all the 60% of the examples for Model 1 and 70% of the examples for Model 2 need to be perturbed with an adversarial attack to maintain the consistency in measuring the difference in the performance of the models Model 1 and Model 2 against the corresponding adversarial attack. The models are pre-trained on NVIDIA 8GB GPUs.

D. Results

Mixed Topics + WGT Our proposed method outperforms the $\text{Model}_{\text{Baseline}}$, in terms of F1-score in 6 out of 7 original, non-perturbed datasets (see Table IX). All of our models achieve a higher or comparable average F1-score than the baseline. In addition, our models consistently outperform the baseline on the highly unbalanced FNC-1 dataset. When comparing our proposed models, $\text{Model}_{\text{Random}}$ achieved the best overall classification performance by learning from multiple randomly selected examples, while $\text{Model}_{\text{Random2}}$, which selects only two random triplets that may belong to different topics, performed worse. However, $\text{Model}_{\text{Random2}}$ still outperformed models **Hard** and **H&E**, which use only a few contrastive examples (one or two triplets) based on their label and similarity or dissimilarity to the anchor. This approach makes it less likely for them to cover a wider range of mixed-topic examples compared to both of the random strategies.

In terms of resilience to perturbations, all of our models show a higher average resilience compared to the baseline (see Table IX). $\text{Model}_{\text{Hard}}$ achieves a better average resilience score compared to all other models while maintaining a comparable average F1-score to the baseline. Indeed, the results suggest that using contrastive learning with only extreme or unorthodox "hard" examples (as positive and negative examples), or a combination of both "hard" and standard "easy" examples, leads to robust models when training examples belong to different topics (see Tables IX and XII). $\text{Model}_{\text{Random}}$ achieves a comparable resilience to these models but requires several triplets examples for each anchor during training, while "Hard" and "Hard & Easy" strategies require one and two triplets, respectively. The two triplets' version of the Random strategy, $\text{Model}_{\text{Random2}}$, seems to also lead to comparable resilience results in this mixed-topics setting.

On a more fine-grained analysis, as shown in Table IX, when comparing our contrastive models to the baseline on specific datasets, we notice that all of them showcase

Hyperparameter	Value
Batch Size	8
Epochs	20
Max. Seq. Length	100
Optimizer	Adam
Learning Rate	5e-5
Gradient Clipping	max norm: 1.0
Epsilon	1e-6
Weight Decay	0.1

TABLE VI: Hyperparameters for the training with CL

Objective	Hyperparameter	Value
MLM	% of tokens masked	15%
CL	Margin (m)	0.5

TABLE VII: Hyperparameters for the Objectives Contrastive Learning and Masked Language Modeling

Hyperparameter	Value
Batch Size	16
Epochs	4
Optimizer	Adam
Learning Rate	5e-5

TABLE VIII: Hyperparameters for finetuning the DistilRoBERTa model with stance dataset

Models	$Model_{Baseline}$	$Model_{Random}$	$Model_{Random2}$	$Model_{Hard}$	$Model_{H\&E}$
DebateForum	94.78 (64.06)	**97.56** (**68.68**)	96.25 (65.73)	94.32 (62.22)	96.41 (62.97)
SemEval2016	97.57 (**74.04**)	96.82 (72.21)	**98.58** (73.31)	98.49 (71.18)	97.3 (71.27)
ARC	97.65 (60.94)	97.35 (61.77)	**98.34** (**62.97**)	96.45 (62.21)	97.21 (62.25)
Perspectrum	91.15 (65.5)	92.88 (**66.05**)	92.55 (65.81)	95.52 (64.75)	**97.81** (63.15)
FNC-1	93.62 (48.86)	98.16 (**52.87**)	95.35 (52.22)	**99.27** (52.63)	98.42 (52.2)
KSD-Biden	93.71 (82.08)	98.75 (**88.77**)	98.63 (87.87)	**98.82** (85.22)	96.29 (84.21)
KSD-Trump	99.21 (86.95)	**99.78** (**88.81**)	99.38 (82.86)	99.21 (85.97)	97.97 (83.58)
Average	95.38 (68.91)	97.32 (**71.30**)	97.01 (70.11)	**97.44** (69.16)	97.34 (68.51)

TABLE IX: Resilience and F1-score (within parenthesis) of all the models for all the datasets in ***Mixed Topic*** setup. The F1-scores are reported in % on all the original, non-perturbated datasets. Bold numbers in **Purple** and **Blue** colors indicate the model with the best Resilience score and F1-score respectively

Dataset	$Model_{Baseline}$	$Model_{Random}$	$Model_{Random2}$	$Model_{Hard}$	$Model_{H\&E}$
$Abortion_{DF}$	94.25 (67.01)	97.34 (68.39)	**99.18** (65.66)	98.83 (**68.78**)	98.92 (66.29)
$Marijuana_{DF}$	96.73 (40.14)	97.55 (45.31)	**98.9** (42.29)	92.13 (50.94)	95.96 (**53.19**)
$Gay\ Rights_{DF}$	97.58 (67.14)	96.36 (60.75)	94.62 (60.06)	**99.67** (58.51)	95.39 (**67.75**)
$Obama_{DF}$	94.27 (64.07)	93.04 (**68.2**)	90.23 (68.17)	98.13 (61.48)	**98.64** (64.8)
$Abortion_{SE}$	96.32 (71.39)	**98.52** (74.3)	95.74 (74.59)	92.83 (**81.19**)	91.33 (78.68)
$Atheism_{SE}$	96.72 (77.14)	93.48 (78.18)	91.68 (79.54)	**98.09** (**80.43**)	95.66 (77.14)
$Climate_{SE}$	95.34 (61.81)	84.39 (68.57)	**95.52** (68.57)	93.82 (**82.37**)	89.2 (72.97)
$Feminism_{SE}$	98.12 (64.32)	95.21 (**65.06**)	**99.52** (60.82)	94.94 (62.97)	88.39 (63.97)
$Hillary\ Clinton_{SE}$	86.98 (**84.63**)	90.7 (82.37)	92.33 (80.3)	**98.62** (71.52)	96.96 (73.46)
Average	95.14 (66.40)	94.06 (67.90)	95.39 (66.67)	**96.34** (**68.69**)	94.49 (**68.69**)

TABLE X: Resilience and F1-score (within parenthesis) of all the models for all the datasets in ***Individiual Topic*** setup. The F1-scores are reported in % on all the original, non-perturbated datasets. Bold numbers in **Purple** and **Blue** colors indicate the model with the best Resilience score and F1-score respectively

	Original Non-Perturbed (F1-score)	Resilience
$Model_{Random}$	64.68	99.36
$Model_{Hard}$	62.34	99.3
$Model_{H\&E}$	61.08	99.21

TABLE XI: F1-score and Resilience of all the proposed models for the unlabeled DebateForum dataset. The F1-scores and Resilience are reported in %.

Dataset	$Model_{Baseline}$	$Model_{Random}$	$Model_{Hard}$	$Model_{H\&E}$
DebateForum	82.05	90.68	**95.15**	93.37
SemEval2016	88.98	**91.69**	91.16	91
ARC	**96.96**	95.98	95.84	95.86
Perspectrum	95.80	96.26	**96.47**	**96.47**
FNC-1	75.15	79.36	81.62	**86.08**
KSD-Biden	98.19	95.29	**98.62**	97.76
KSD-Trump	**98.96**	97.49	92.97	95.88
Average	90.87 $\pm$ 9.2	92.39 $\pm$ 6.26	93.12 $\pm$ 5.61	**93.77** $\pm$ 4.06

TABLE XII: Reslience of all the models for all the datasets in ***Mixed Topic + Partial Perturbation*** setup. Bold numbers in **Purple** color indicate the model with the best Resilience score. The last row shows the models' average resilience over all datasets including standard deviation.

Dataset	$Model_{Baseline}$	$Model_{Random}$	$Model_{Hard}$	$Model_{H\&E}$
$Abortion_{DF}$	90.26	93.32	**95.34**	94.96
$Marijuana_{DF}$	**98.77**	95.06	93.57	96.32
$GayRights_{DF}$	88.25	**92.97**	90.19	80.15
$Obama_{DF}$	92.9	**95.64**	94.92	94.24
$Abortion_{SE}$	79.08	88.64	**90.03**	87.53
$Atheism_{SE}$	85.7	**93.59**	90.97	90.96
$Climate_{SE}$	86.3	96.17	**97.37**	92.08
$Feminism_{SE}$	**87.15**	79.64	84.84	80.47
$Hillary\ Clinton_{SE}$	74.24	80.10	**92.55**	90.23
Average	86.96 $\pm$ 7.18	90.57 $\pm$ 6.44	**92.2** $\pm$ 3.71	89.66 $\pm$ 5.92

TABLE XIII: Reslience of all the models for all the datasets in ***Individual Topic + Partial Perturbation*** setup. Bold numbers in **Purple** color indicate the model with the best Resilience score. The last row shows the models' average resilience over all datasets including standard deviation.

enhanced resilience on the highly unbalanced dataset FNC-1. Moreover, the majority of our models (3 out of 4) exhibit improved resilience on slightly less unbalanced datasets such as DebateForum and SemEval2016.

Mixed Topics + Partial Perturbation To validate previous robustness results, we performed experiments where we only perturbed instances that were correctly classified by the models in the original test dataset. We observed similar results, with our proposed contrastive models exhibiting better resilience than the baseline overall (see Table XII). There was a significant increase of more than 10% for unbalanced datasets FNC-1 and DebateForum. Training with $\text{Model}_{\text{H\&E}}$ and $\text{Model}_{\text{Hard}}$ produced more robust models in general.

Mixed Topics + Without Ground Truth The resilience of all the models is comparable, and the $\text{Model}_{\text{Random}}$ has better resilience (99.36) than the other models. Though the dataset size is different for the unlabeled dataset setup as described in Section III-E, the $\text{Model}_{\text{Random}}$ (unlabeled setup) has better F1-score of 64.68% than $\text{Model}_{\text{Baseline}}$ for the same dataset with ground truth (WGT) labels setup (64.06%) (see Table XI for the performance of the individual models).

Individual Topics + WGT In this setting where the training data consists of examples from the same topic and dataset, our proposed models demonstrate comparable or superior F1-scores compared to the $\text{Model}_{\text{Baseline}}$ on average, and outperform it in eight out of nine non-perturbed test sets (refer to Table X). $\text{Model}_{\text{H\&E}}$ and $\text{Model}_{\text{Hard}}$, achieved better F1 performance compared to all the models including $\text{Model}_{\text{Random}}$, unlike the mixed topics settings. Specifically, the "Hard" contrastive training strategy, which selects a dissimilar example with the same stance and a similar example with an opposite stance from the "same topic" in this case, appears to give the model a better ability not only to generalize but also to exhibit better stability, as evidenced by the resilience score of $\text{Model}_{\text{Hard}}$ (see Table X). This is particularly evident when we only perturb previously correctly classified instances in the source (original) test dataset (see Table XIII). For the smallest and most unbalanced topic dataset, $\text{Climate}_{\text{SE}}$, all our models outperform the baseline, with $\text{Model}_{\text{Hard}}$ achieving more than 20% increase in classification performance. Similarly, a notable increase in F1-score is observed with our models, specifically $\text{Model}_{\text{Hard}}$, for Marijuana_{DF}, Abortion_{SE}, and Atheism_{SE}. These datasets are highly imbalanced and relatively small, containing less than 750 examples.

At least one of the proposed models exhibits better resilience scores than $\text{Model}_{\text{Baseline}}$ in all of the experimented 9 datasets. Among all the models, $\text{Model}_{\text{Hard}}$ achieves the highest average resilience score across the datasets. It is worth noting that $\text{Model}_{\text{Random2}}$ exhibits the strongest resilience on four out of the nine datasets. However, in contrast with $\text{Model}_{\text{Hard}}$, it has the lowest average F1-score among all the contrastive models.

Individual Topics + Partial Perturbation When perturbing only the correctly classified examples of a model, as in the previous setting, we observe a significant increase in the resilience score for our proposed models compared to the $\text{Model}_{\text{Baseline}}$ for the small and unbalanced topic datasets, namely Abortion, Atheism, Climate, and Hillary Clinton, as well as on average (see Table XIII). Once again, $\text{Model}_{\text{Hard}}$ appears to be the most robust among the proposed models.

V. Conclusion & Future Work

In this work, we have adopted the combination of Contrastive Learning + Masked Language Model methods and explored different triplet strategies to learn more robust sentence representations to use in the Stance Detection task. Experiment results show that our proposed methodology with labelled examples setting is more resilient to errors and variations. Also, the experiments with different setups show that our proposed methodology is effective for small-sized as well as class-imbalanced datasets. Furthermore, our proposed methodology on unlabelled datasets has comparable performances with the models trained with labelled data. This shows that the stance of examples can be effectively identified even when the stance annotations are not associated with them.

We considered the binary stances examples topics mainly i.e. for/against, support/refute, or agree/disagree. The proposed methodology leverages the Contrastive Learning framework which is conditioned to work with two stance labels examples to identify whether the author of the text is in favor of or against the topic of discussion. However, social media such as Twitter and online forums like Reddit will have threads discussing topics having more than two stances such as for/against/neither, or support/refute/comment. In future work, we propose to accommodate more than two stance labels in the proposed methodology for the Stance Detection task.

References

[1] Schiller, B., Daxenberger, J. & Gurevych, I. Stance Detection Benchmark: How Robust is Your Stance Detection?. *KI - Künstliche Intelligenz*. **35**, 329-341 (2021,3)

[2] Darwish, K., Magdy, W. & Zanouda, T. Improved Stance Prediction in a User Similarity Feature Space. *Proceedings Of The 2017 IEEE/ACM International Conference On Advances In Social Networks Analysis And Mining 2017*. (2017,7), http://dx.doi.org/10.1145/3110025.3110112

[3] Matero, M., Soni, N., Balasubramanian, N. & Schwartz, H. MeLT: Message-Level Transformer with Masked Document Representations as Pre-Training for Stance Detection. *Findings Of The Association For Computational Linguistics: EMNLP 2021*. (2021), http://dx.doi.org/10.18653/v1/2021.findings-emnlp.253

[4] Zhang, Z., Li, J., Fukumoto, F. & Ye, Y. Abstract, Rationale, Stance: A Joint Model for Scientific Claim Verification. *Proceedings Of The 2021 Conference On Empirical Methods In Natural Language Processing*. (2021), http://dx.doi.org/10.18653/v1/2021.emnlp-main.290

[5] Landwehr, N., Hall, M. & Frank, E. Logistic Model Trees. *Machine Learning*. **59**, 161-205 (2005,5)

[6] Sobhani, P., Inkpen, D. & Zhu, X. A Dataset for Multi-Target Stance Detection. *Proceedings Of The 15th Conference Of The European Chapter Of The Association For Computational Linguistics: Volume 2, Short Papers*. (2017), http://dx.doi.org/10.18653/v1/e17-2088

[7] Aldayel, A. & Magdy, W. Your Stance is Exposed! Analysing Possible Factors for Stance Detection on Social Media. *Proceedings Of The ACM On Human-Computer Interaction*. **3**, 1-20 (2019,11)

[8] Rashed, A., Kutlu, M., Darwish, K., Elsayed, T. & Bayrak, C. Embeddings-Based Clustering for Target Specific Stances: The Case of a Polarized Turkey. (2020,5)

[9] Dong, X., Luu, A., Lin, M., Yan, S. & Zhang, H. How Should Pre-Trained Language Models Be Fine-Tuned Towards Adversarial Robustness?. *Advances In Neural Information Processing Systems*. **34** pp. 4356-4369 (2021), https://proceedings.neurips.cc/paper/2021/file/22b1f2e0983160db6f7bb9f62f4dbb39-Paper.pdf

[10] Zhang, C., Zhou, X., Wan, Y., Zheng, X., Chang, K. & Hsieh, C. Improving the Adversarial Robustness of NLP Models by Information Bottleneck. *Findings Of The Association For Computational Linguistics: ACL 2022*. pp. 3588-3598 (2022,5), https://aclanthology.org/2022.findings-acl.284

[11] Wang, B., Wang, S., Cheng, Y., Gan, Z., Jia, R., Li, B. & Liu, J. InfoBERT: Improving Robustness of Language Models from An Information Theoretic Perspective. (arXiv,2020), https://arxiv.org/abs/2010.02329

[12] Moradi, M. & Samwald, M. Evaluating the Robustness of Neural Language Models to Input Perturbations. *Proceedings Of The 2021 Conference On Empirical Methods In Natural Language Processing*. pp. 1558-1570 (2021,11), https://aclanthology.org/2021.emnlp-main.117

[13] Yue, Z., Kratzwald, B. & Feuerriegel, S. Contrastive Domain Adaptation for Question Answering using Limited Text Corpora. *Proceedings Of The 2021 Conference On Empirical Methods In Natural Language Processing*. (2021), http://dx.doi.org/10.18653/v1/2021.emnlp-main.754

[14] Xu, H., Ghosh, G., Huang, P., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L. & Feichtenhofer, C. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. *Proceedings Of The 2021 Conference On Empirical Methods In Natural Language Processing*. (2021), http://dx.doi.org/10.18653/v1/2021.emnlp-main.544

[15] Wu, H., Ma, T., Wu, L., Manyumwa, T. & Ji, S. Unsupervised Reference-Free Summary Quality Evaluation via Contrastive Learning. *Proceedings Of The 2020 Conference On Empirical Methods In Natural Language Processing (EMNLP)*. (2020), http://dx.doi.org/10.18653/v1/2020.emnlp-main.294

[16] Du, Y., Ma, T., Wu, L., Xu, F., Zhang, X., Long, B. & Ji, S. Constructing contrastive samples via summarization for text classification with limited annotations. *Findings Of The Association For Computational Linguistics: EMNLP 2021*. (2021), http://dx.doi.org/10.18653/v1/2021.findings-emnlp.118

[17] Cao, S. & Wang, L. CLIFF: Contrastive Learning for Improving Faithfulness and Factuality in Abstractive Summarization. *Proceedings Of The 2021 Conference On Empirical Methods In Natural Language Processing*. (2021), http://dx.doi.org/10.18653/v1/2021.emnlp-main.532

[18] Ostendorff, M., Rethmeier, N., Augenstein, I., Gipp, B. & Rehm, G. Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings. (arXiv,2022), https://arxiv.org/abs/2202.06671

[19] Thorne, J., Vlachos, A., Christodoulopoulos, C. & Mittal, A. Evaluating adversarial attacks against multiple fact verification systems. *Proceedings Of The 2019 Conference On Empirical Methods In Natural Language Processing And The 9th International Joint Conference On Natural Language Processing (EMNLP-IJCNLP)*. pp. 2944-2953 (2019,11), https://aclanthology.org/D19-1292

[20] Kincaid, J., Fishburne Jr, R., Rogers, R. & Chissom, B. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. (Naval Technical Training Command Millington TN Research Branch,1975)

[21] Giorgi, J., Nitski, O., Wang, B. & Bader, G. DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations. (arXiv,2020), https://arxiv.org/abs/2006.03659

[22] Li, J., Shao, H., Sun, D., Wang, R., Yan, Y., Li, J., Liu, S., Tong, H. & Abdelzaher, T. Unsupervised Belief Representation Learning with Information-Theoretic Variational Graph Auto-Encoders. *Proceedings Of The 45th International ACM SIGIR Conference On Research And Development In Information Retrieval*. (2022,7), https://doi.org/10.1145

[23] Hasan, K. & Ng, V. Stance classification of ideological debates: Data, models, features, and constraints. *Proceedings Of The Sixth International Joint Conference On Natural Language Processing*. pp. 1348-1356 (2013)

[24] Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X. & Cherry, C. Semeval-2016 task 6: Detecting stance in tweets. *Proceedings Of The 10th International Workshop On Semantic Evaluation (SemEval-2016)*. pp. 31-41 (2016)

[25] Habernal, I., Wachsmuth, H., Gurevych, I. & Stein, B. The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants. *Proceedings Of The 2018 Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. pp. 1930-1940 (2018)

[26] Chen, S., Khashabi, D., Yin, W., Callison-Burch, C. & Roth, D. Seeing Things from a Different Angle: Discovering Diverse Perspectives about Claims. (arXiv,2019), https://arxiv.org/abs/1906.03538

[27] Pomerleau, D. & Rao, D. Exploring how artificial intelligence technologies could be leveraged to combat fake news.. *Fake News Challenge*., http://www.fakenewschallenge.org/

[28] Kawintiranon, K. & Singh, L. Knowledge Enhanced Masked Language Model for Stance Detection. *Proceedings Of The 2021 Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies*. pp. 4725-4735 (2021,6), https://aclanthology.org/2021.naacl-main.376

[29] Sanh, V., Debut, L., Chaumond, J. & Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*. **abs/1910.01108** (2019)

[30] Twitter Twitter. It's what's happening.. *Twitter.com*. (2022), https://twitter.com/?lang=en

[31] Lai, M., Cignarella, A., Hernández Farías, D., Bosco, C., Patti, V. & Rosso, P. Multilingual stance detection in social media political debates. *Computer Speech & Language*. **63** pp. 101075 (2020), https://www.sciencedirect.com/science/article/pii/S0885230820300085

[32] Liang, B., Chen, Z., Gui, L., He, Y., Yang, M. & Xu, R. Zero-shot stance detection via contrastive learning. *Proceedings Of The ACM Web Conference 2022*. pp. 2738-2747 (2022)

[33] Wei, P., Lin, J. & Mao, W. Multi-Target Stance Detection via a Dynamic Memory-Augmented Network. *The 41st International ACM SIGIR Conference On Research & Development In Information Retrieval*. (2018)

[34] Hardalov, M., Arora, A., Nakov, P. & Augenstein, I. Cross-Domain Label-Adaptive Stance Detection. *Proceedings Of The 2021 Conference On Empirical Methods In Natural Language Processing*. (2021), http://dx.doi.org/10.18653/v1/2021.emnlp-main.710

[35] Joseph, K., Shugars, S., Gallagher, R., Green, J., Quintana Mathé, A., An, Z. & Lazer, D. (Mis)alignment Between Stance Expressed in Social Media Data and Public Opinion Surveys. *Proceedings Of The 2021 Conference On Empirical Methods In Natural Language Processing*. pp. 312-324 (2021,11), https://aclanthology.org/2021.emnlp-main.27

[36] Wei, P., Lin, J. & Mao, W. Multi-Target Stance Detection via a Dynamic Memory-Augmented Network. *The 41st International ACM SIGIR Conference On Research & Development In Information Retrieval*. (2018,6), http://dx.doi.org/10.1145/3209978.3210145

[37] Li, Y., Zhao, C. & Caragea, C. Improving Stance Detection with Multi-Dataset Learning and Knowledge Distillation. *Proceedings Of The 2021 Conference On Empirical Methods In Natural Language Processing*. (2021), http://dx.doi.org/10.18653/v1/2021.emnlp-main.511

[38] Sun, S. & Li, W. Alleviating Exposure Bias via Contrastive Learning for Abstractive Text Summarization. (2021,8)

[39] Wu, Z., Wang, S., Gu, J., Khabsa, M., Sun, F. & Ma, H. CLEAR: Contrastive Learning for Sentence Representation. (2020,12)

[40] Jayaram, S. & Allaway, E. Human Rationales as Attribution Priors for Explainable Stance Detection. *Proceedings Of The 2021 Conference On Empirical Methods In Natural Language Processing*. (2021), http://dx.doi.org/10.18653/v1/2021.emnlp-main.450

[41] Yang, S. & Urbani, J. Tribrid: Stance Classification with Neural Inconsistency Detection. *Proceedings Of The 2021 Conference On Empirical Methods In Natural Language Processing*. (2021), http://dx.doi.org/10.18653/v1/2021.emnlp-main.547

[42] Delvin, J., Chang, M., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. (2019), https://aclanthology.org/N19-1423

[43] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. & Amodei, D. Language Models are Few-Shot Learners. (2020)

[44] Fleiss, J. Measuring nominal scale agreement among many raters.. *Psychological Bulletin*. **76**, 378 (1971)