Exploration of Contrastive Learning Strategies toward more Robust Stance Detection

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

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

Abstract

Stance Detection is the task of identifying the position of an author of a text towards an issue or a target. Previous studies on Stance Detection indicate that the existing systems are non-robust to the variations and errors in input sentences. Our proposed methodology uses Contrastive Learning to learn sentence representations by bringing semantically similar sentences and sentences implying the same stance closer to each other in the embedding space. We compare our approach to a pretrained transformer model directly finetuned with the stance datasets. We use char-level and word-level adversarial perturbation attacks to measure the resilience of the models and we show that our approach achieves better performances and is more robust to the different adversarial perturbations introduced to the test data. The results indicate that our approach performs better on small-sized and class-imbalanced stance datasets.

1 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 (Schiller et al., 2021). 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 as a token-level objective to learn textual representations (see Figure 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 conduct a comprehensive empirical investigation using various settings and datasets for stance detection, analyzing the results and pro-

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

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 431–440
July 14, 2023 ©2023 Association for Computational Linguistics

viding valuable insights into effective strategies for different contexts.

2 Related Work

Many approaches (Darwish et al., 2017; Matero et al., 2021; Zhang et al., 2021; Landwehr et al., 2005; Sobhani et al., 2017; Aldayel and Magdy, 2019; Rashed et al., 2020; Lai et al., 2020; Liang et al., 2022) were proposed to tackle different problems in the Stance Detection task. However, the existing Stance Detection models are sensitive to adversarial errors, and changes in the vocabulary of the input sentences (Schiller et al., 2021).

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 (Dong et al., 2021; Zhang et al., 2022; Wang et al., 2020). Moradi and Samwald (2021) used various perturbations for Char-level such as Insertion, Deletion, Replacement, etc., and word-level perturbations such as Replacement with Synonyms, Negation, etc. Schiller et al. (2021) used the resilience score introduced by Thorne et al. (2019) to measure the robustness of the model.

CL is used to acquire better representations of text for many natural language tasks such as Question-Answering (Yue et al., 2021), multiple choice video questions, text-to-video retrieval (Xu et al., 2021), text summarization (Wu et al., 2020a; Du et al., 2021; Cao and Wang, 2021) etc. Wu et al. (2020b) used Contrastive Learning to learn noise invariant sentence representation with the help of different sentence-level augmentation strategies like span deletion, substitution, and reordering. Liang et al. (2022) introduced a hierarchical contrastive learning strategy to improve the Zero-shot Stance Detection (ZSSD) task by capturing the relationships not only between target-invariant and target-specific features but also among various stance labels.

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.

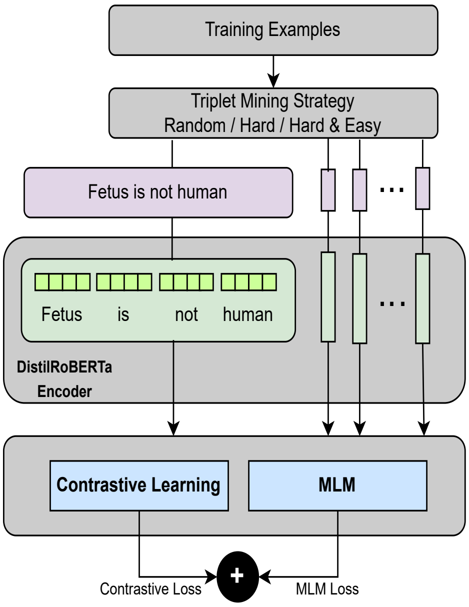

Figure 1: Architecture diagram for learning sentence representations using CL and MLM objectives to further use in the Stance Detection task.

3 Method

3.1 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 in the representation space from the anchor example (Ostendorff et al., 2022). The goal of the contrastive loss function (loss_{CL}) given by Eqn. 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_-). We use pairwise Euclidean distance measure for calculating the contrastive loss in the Eqn. 1. m in the 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. For instance, the examples *‘Fetus is not human’* and *‘Bunch of cells in a woman’s womb’* are having the same stance as *support* though the lexicons used in these examples are completely different. The example, *‘Really? Fetus is not human?’* is a rhetorical question, having an opposite stance compared to the example *‘Fetus is not human’*, however, both these examples are similar in

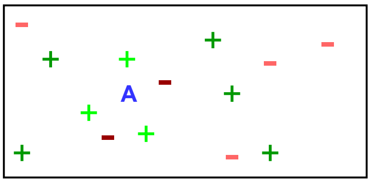

Figure 2: Illustration of Easy Positive + and Negative -, Hard Positive + and Negative - samples for an Anchor sample A in the representation space.

terms of lexicons. The contrastive learning method attempt to make the examples *'Fetus is not human'* and *'Bunch of cells in a woman's womb'* have similar representations by bringing the two examples closer to each other in the representation space. The final loss is the sum of CL loss and MLM loss.

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

3.2 Contrastive Learning Strategies

We use different strategies to select positives and negatives for an anchor for CL. 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.

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 Figure 2 for a graphical illustration of the hard and easy positives and negatives for an anchor in the representation space).

3.3 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. (2019) 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 three adversarial attacks *spelling errors*, *adding tautology*, and *synonym replacements* (see Table 1). The correctness ratio c_a of an adversarial attack a gives the total number of correctly transformed examples from the number of examples considered for perturbation.

Spelling error. We introduced spelling errors to perturb all the original sentences in the test set. We select two words randomly from a sentence to introduce misspellings by replacing a letter in one of the selected words and by swapping the position of two letters in the other word.

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

Synonyms replacement. We consider 15 words that are frequent in the test dataset for the Synonyms replacement adversarial attack. We use WordNet (Miller, 1994), a Lexical Database for English, to select the synonyms for the 15 frequent words in the test dataset. We select a maximum of 2 words from a sentence (selected words fall under the frequent words) to replace 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.

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

3.4 Learning and Leveraging Robust Representations

Let F be the transformer model (DistilRoBERTa), for each of the input sequences $x^{(\mathrm{i})}$ 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 $x^{(\mathrm{i})}$ (anchor) based on the strategies explained in Section 3.2. 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 $\mathrm{D_p}$. 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

Adv. 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

Table 1: Illustration of the different types of adversarial attacks for perturbing the test set to measure the robustness and reliability of the model.

adversarial attacks spelling errors, tautology, and synonyms respectively.

4 Experiments

We have chosen seven Stance Detection datasets, DebateForum (DF) (Hasan and Ng, 2013), SemEval2016 (SE) (Mohammad et al., 2016), ARC (Habernal et al., 2018), Perspectrum (Chen et al., 2019), FNC-1 (Pomerleau and Rao), KSD-Biden and KSD-Trump (Kawintiranon and Singh, 2021) for the experiments. We have retained only the examples that have support/against equivalent labels in the datasets 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 2 for more information on the statistics of these datasets. Table 4 describes the datasets, the domain of the corresponding datasets, and an example from the dataset to show the input and the stance output.

4.1 Setups

The setups below vary according to the level of information leveraged to train and evaluate the conceived models. To further validate our evaluation of resilience, we only perturbed the instances that were correctly classified (*Partial Perturbation*) by the models from the original test dataset and assessed its resilience in relation to those perturbations.

Mixed Topics. We consider the examples of all topics from a dataset as a whole for the experiments. The evaluation of models is carried out by perturbating all the examples in the test dataset while testing the model against an adversarial attack.

Mixed Topics + Partial Perturbation (PP). The models are constructed based on all topics similar to the *Mixed Topics* setup but 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.

Individual Topics. The models are constructed and evaluated based on individual topic-related sub-datasets. We consider topics from DF and SE datasets for this setup (see Table 3).

Individual Topics + Partial Perturbation (PP). The models are constructed based on individual topic-related sub-datasets similar to the *Individual Topics* setup but 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.

4.2 Models

We have used the DistilRoBERTa (Sanh et al., 2019) as the transformer model which is twice as fast as RoBERTa-base (Liu et al., 2019) for all our experiments. Inspired by the work of Giorgi et al. (Giorgi et al., 2020), 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$_{\textbf{Baseline}}$ is pretrained DistilRoBERTa model finetuned with stance datasets.

Model$_{\textbf{Random}}$. Randomly formed triplets from a batch are used in CL.

Model$_{\textbf{Random2}}$. Two random triplets from a batch are used in CL.

Model$_{\textbf{Hard}}$. One Hard triplet is used in CL.

Model$_{\textbf{H\&E}}$. One Hard and one Easy triplets are

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 2: Statistics about the different datasets used for the experiments

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

Table 3: 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 4: Illustrates the domain of the different datasets used for the experiments and an example from each of the datasets

used in CL.

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 5: Hyperparameters for the training with CL

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

Table 6: Hyperparameters for the Objectives Contrastive Learning and Masked Language Modeling

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

Table 7: Hyperparameters for finetuning the DistilRoBERTa model with stance dataset

4.3 Settings

The number of characters and words used in social media posts is usually restricted to cut out the fluff. For example, currently, Twitter (Twitter, 2022) 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. We then finetune the model with stance datasets for 4 epochs. See Tables 5, 6 and 7 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 (Kincaid et al., 1975) 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 resilience of models is measured by perturbing **all** the examples in the test dataset with the adversarial attacks individually for the experiment setups **Mixed Topics** and **Individual Topics**, see under Section 4.1. 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 example, the model$_{\text{Hard}}$ is evaluated on the original non-perturbed dataset initially, then a dataset is prepared by perturbing (with an adversarial attack, 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 *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 example, from the original non-perturbed test set, if Model 1 predicts 60% of the examples correctly

Dataset	Model$_{\text{Baseline}}$	Model$_{\text{Random}}$	Model$_{\text{Random2}}$	Model$_{\text{Hard}}$	Model$_{\text{H\&E}}$
DebateForum	93.24 (64.06)	98.33 (**68.68**)	98.12 (65.73)	98.42 (62.22)	**98.53** (62.97)
SemEval2016	98.31 (**74.04**)	99.24 (72.21)	**99.66** (73.31)	99.5 (71.18)	99.49 (71.27)
ARC	**99.71** (60.94)	98.19 (61.77)	99.02 (**62.97**)	95.92 (62.21)	99.35 (62.25)
Perspectrum	92.91 (65.5)	95.16 (**66.05**)	96.54 (65.81)	95.55 (64.75)	**98.82** (63.15)
FNC-1	93.77 (48.86)	97.61 (**52.87**)	96.58 (52.22)	**99.06** (52.63)	97.95 (52.2)
KSD-Biden	93.32 (82.08)	98.38 (**88.77**)	98.25 (87.87)	**98.47** (85.22)	98.16 (84.21)
KSD-Trump	98.97 (86.95)	**99.72** (**88.81**)	99.19 (82.86)	98.97 (85.97)	98.84 (83.58)
Average	95.74 (68.91)	98.09 (**71.30**)	98.19 (70.11)	97.98 (69.16)	**98.73** (68.51)

Table 8: 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$_{\text{Baseline}}$	Model$_{\text{Random}}$	Model$_{\text{Random2}}$	Model$_{\text{Hard}}$	Model$_{\text{H\&E}}$
Abortion$_{DF}$	97.22 (67.01)	98.24 (68.39)	**98.92** (65.66)	98.54 (**68.78**)	98.58 (66.29)
Marijuana$_{DF}$	**98.65** (40.14)	97.56 (45.31)	98.55 (42.29)	95.79 (50.94)	96.99 (**53.19**)
Gay Rights$_{DF}$	96.61 (67.14)	95.66 (60.75)	94.1 (60.06)	**98.74** (58.51)	96.85 (**67.75**)
Obama$_{DF}$	98.91 (64.07)	99.10 (**68.2**)	98.59 (68.17)	98.65 (61.48)	**99.64** (64.8)
Abortion$_{SE}$	97.33 (71.39)	**98.89** (74.3)	97.39 (74.59)	96.31 (**81.19**)	96.63 (78.68)
Atheism$_{SE}$	96.23 (77.14)	95.70 (78.18)	95.22 (79.54)	**97.56** (**80.43**)	96.07 (77.14)
Climate$_{SE}$	93.89 (61.81)	79.54 (68.57)	**94.12** (68.57)	91.89 (**82.37**)	90.60 (72.97)
Feminism$_{SE}$	**99.60** (64.32)	93.72 (**65.06**)	99.36 (60.82)	95.51 (62.97)	85.12 (63.97)
Hillary Clinton$_{SE}$	86.17 (**84.63**)	92.49 (82.37)	94.56 (80.3)	**98.19** (71.52)	96.01 (73.46)
Average	96.06 (66.40)	94.54 (67.90)	96.75 (66.67)	**96.79** (**68.69**)	95.16 (**68.69**)

Table 9: 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

Dataset	Model$_{\text{Baseline}}$	Model$_{\text{Random}}$	Model$_{\text{Hard}}$	Model$_{\text{H\&E}}$
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 ± 9.2	92.39 ± 6.26	93.12 ± 5.61	**93.77** ± 4.06

Table 10: 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$_{\text{Baseline}}$	Model$_{\text{Random}}$	Model$_{\text{Hard}}$	Model$_{\text{H\&E}}$
Abortion$_{\text{DF}}$	90.26	93.32	**95.34**	94.96
Marijuana$_{\text{DF}}$	**98.77**	95.06	93.57	96.32
GayRights$_{\text{DF}}$	88.25	**92.97**	90.19	80.15
Obama$_{\text{DF}}$	92.9	**95.64**	94.92	94.24
Abortion$_{\text{SE}}$	79.08	88.64	**90.03**	87.53
Atheism$_{\text{SE}}$	85.7	**93.59**	90.97	90.96
Climate$_{\text{SE}}$	86.3	96.17	**97.37**	92.08
Feminism$_{\text{SE}}$	**87.15**	79.64	84.84	80.47
Hillary Clinton$_{\text{SE}}$	74.24	80.10	**92.55**	90.23
Average	86.96 ± 7.18	90.57 ± 6.44	**92.2** ± 3.71	89.66 ± 5.92

Table 11: 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.

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.

4.4 Results

Mixed Topics Our proposed method outperforms the Model$_{\text{Baseline}}$, in terms of F1-score in 6 out of 7 original, non-perturbed datasets (see Table 8). 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, Model$_{\text{Random}}$ achieved the best overall classification performance by learning from multiple randomly selected examples, while Model$_{\text{Random2}}$, which selects only two random triplets that may belong to different topics, performed worse. However, 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.

In terms of resilience to perturbations, all of our models show a higher average resilience compared to the baseline (see Table 8). Model$_{\text{H\&E}}$ 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, or a combination of both "hard" and standard "easy" examples, leads to more robust models when training examples belong to different topics (see Tables 8 and 10). On the other hand, although the baseline has a slightly better resilience score for the ARC dataset, all of our contrastive models perform better for highly unbalanced datasets like FNC-1, as well as for slightly less unbalanced datasets such as DebateForum and SemEval2016.

Mixed Topics + Partial Perturbation To validate previous 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 10). There was a significant increase of more than 10% for unbalanced datasets FNC-1 and DebateForum. Training with Model$_{\text{H\&E}}$ and Model$_{\text{Hard}}$ produced more robust models in general.

Individual Topics 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 Model$_{\text{Baseline}}$ on average, and outperform it in eight out of nine non-perturbed test sets (refer to Table 9). Model$_{\text{H\&E}}$ and Model$_{\text{Hard}}$, achieved better performance compared to the Random models in 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 9). This is particularly evident when we only perturb correctly classified instances (see Table 11). 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.

Our proposed models exhibit better resilience scores than $\text{Model}_{\text{Baseline}}$ in 7 out of 9 datasets and also perform better in terms of resilience for the smaller and more imbalanced SE datasets, such as Abortion, Atheism, and Hillary Clinton. While the average resilience score of $\text{Model}_{\text{Random2}}$ and $\text{Model}_{\text{Hard}}$ is comparable, $\text{Model}_{\text{Hard}}$ achieves the best average F1-score among all the models on the original, non-perturbed test set.

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 11). Once again, $\text{Model}_{\text{Hard}}$ appears to be the most robust among the proposed models.

5 Conclusion

In this work, we have adopted the combination of CL + MLM method 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 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.

Limitations

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.

References

Abeer Aldayel and Walid Magdy. 2019. Your stance is exposed! analysing possible factors for stance detection on social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–20.

Shuyang Cao and Lu Wang. 2021. Cliff: Contrastive learning for improving faithfulness and factuality in abstractive summarization.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims.

Kareem Darwish, Walid Magdy, and Tahar Zanouda. 2017. Improved stance prediction in a user similarity feature space.

Xinshuai Dong, Anh Tuan Luu, Min Lin, Shuicheng Yan, and Hanwang Zhang. 2021. How should pre-trained language models be fine-tuned towards adversarial robustness? In *Advances in Neural Information Processing Systems*, volume 34, page 4356–4369. Curran Associates, Inc.

Yangkai Du, Tengfei Ma, Lingfei Wu, Fangli Xu, Xuhong Zhang, Bo Long, and Shouling Ji. 2021. Constructing contrastive samples via summarization for text classification with limited annotations.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2020. Declutr: Deep contrastive learning for unsupervised textual representations.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1930–1940.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the sixth international joint conference on natural language processing*, pages 1348–1356.

Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4725–4735, Online. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

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

Niels Landwehr, Mark Hall, and Eibe Frank. 2005. Logistic model trees. *Machine Learning*, 59(1):161–205.

Bin Liang, Zixiao Chen, Lin Gui, Yulan He, Min Yang, and Ruifeng Xu. 2022. Zero-shot stance detection via contrastive learning. In *Proceedings of the ACM Web Conference 2022*, pages 2738–2747.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Matthew Matero, Nikita Soni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2021. Melt: Message-level transformer with masked document representations as pre-training for stance detection.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

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

Milad Moradi and Matthias Samwald. 2021. Evaluating the robustness of neural language models to input perturbations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1570, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings.

Dean Pomerleau and Delip Rao. Exploring how artificial intelligence technologies could be leveraged to combat fake news.

Ammar Rashed, Mucahid Kutlu, Kareem Darwish, Tamer Elsayed, and Cansın Bayrak. 2020. Embeddings-based clustering for target specific stances: The case of a polarized turkey.

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

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *KI - Künstliche Intelligenz*, 35(3–4):329–341.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Evaluating adversarial attacks against multiple fact verification systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2944–2953, Hong Kong, China. Association for Computational Linguistics.

Twitter. 2022. Twitter. it's what's happening.

Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. 2020. Infobert: Improving robustness of language models from an information theoretic perspective.

Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. 2020a. Unsupervised reference-free summary quality evaluation via contrastive learning.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020b. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding.

Zhenrui Yue, Bernhard Kratzwald, and Stefan Feuerriegel. 2021. Contrastive domain adaptation for question answering using limited text corpora.

Cenyuan Zhang, Xiang Zhou, Yixin Wan, Xiaoqing Zheng, Kai-Wei Chang, and Cho-Jui Hsieh. 2022. Improving the adversarial robustness of NLP models by information bottleneck. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3588–3598, Dublin, Ireland. Association for Computational Linguistics.

Zhiwei Zhang, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. 2021. Abstract, rationale, stance: A joint model for scientific claim verification.