Two-Step Classification using Recasted Data for Low Resource Settings

Shagun Uppal[1], Vivek Gupta[2], Avinash Swaminathan[1], Debanjan Mahata [3], Rakesh Gosangi [3], Haimin Zhang [3] Rajiv Ratn Shah[1], Amanda Stent[3]

[1] IIIT-Delhi, India, [2] University of Utah, [3] Bloomberg, New York
shagun16088@iiitd.ac.in, vgupta@cs.utah.edu, s.avinash.it.17@nsit.net.in,
{dmahata, rgosangi, hzhang449, astent}@bloomberg.net,
rajivratn@iiitd.ac.in

Abstract

An NLP model's ability to reason should be independent of language. Previous works utilize Natural Language Inference (NLI) to understand the reasoning ability of models, mostly focusing on high resource languages like English. To address scarcity of data in low-resource languages such as *Hindi*, we use data recasting to create four NLI datasets from existing four text classification datasets in *Hindi* language. Through experiments, we show that our recasted dataset[1] is devoid of statistical irregularities and spurious patterns. We study the consistency in predictions of the textual entailment models and propose a consistency regulariser to remove pairwise-inconsistencies in predictions. Furthermore, we propose a novel *two-step* classification method which uses textual-entailment predictions for classification task. We further improve the classification performance by jointly training the classification and textual entailment tasks together. We therefore highlight the benefits of data recasting and our approach [2] with supporting experimental results.

1 Introduction

Textual entailment (TE) is the task of determining if a *hypothesis* sentence can be inferred from a given *context* sentence. Figure 1 shows examples of *context-hypothesis* pairs for TE. Previous works (Wang and Zhang, 2009; Tatu and Moldovan, 2005; Sammons et al., 2010) investigated several semantic approaches for TE and demonstrated how they can be used to evaluate inference-related tasks such as Question Answering *(QA)*, reading comprehension *(RC)* and paraphrase acquisition *(PA)*.

Context-Hypothesis	Label
p : The kid exclaimed with joy. h : The kid is happy.	*entailed*
p : I am feeling happy. h : I am angry.	*not-entailed (contradictory)*

Table 1: Example illustrating context (c) - hypothesis (h) pairs for the task of textual entailment.

Researchers have curated many resources[3] and benchmark datasets for TE in English (Bowman et al., 2015; Williams et al., 2018; Khot et al., 2018). However, to our knowledge, there is only one TE dataset (XNLI) in Hindi, which was created by translating English data (Conneau et al., 2018) and another in Hindi-English code-switched setting (Khanuja et al., 2020). Hindi is the language with the fourth most native speakers in the world[4]. Despite its wide prevalence, Hindi is still considered a low-resource language by NLP practitioners because there are a rather limited number of publicly available annotated datasets. Developing models that can accurately process text from low-resource languages, such as Hindi, is critical for the proliferation and broader adoption of NLP technologies.

Creating a high-quality labeled corpus for TE in Hindi through crowd-sourcing could be challenging. In this paper, we employ a recasting technique from Poliak et al. (2018a,b) to convert four publicly available text classification datasets in Hindi and pose them as TE problems. In this recasting process, we build template hypotheses for each class in the label taxonomy. Then, we pair the original anno-

[1] https://github.com/midas-research/hindi-nli-data

[2] https://github.com/midas-research/hindi-nli-code

[3] https://aclweb.org/aclwiki/Textual_Entailment_Resource_Pool

[4] https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

tated sentence with each of the template hypotheses to create TE samples. Unlike XNLI, our dataset is based on the original Hindi text and is not translated. Furthermore, the multiple annotation artefacts (Tan et al., 2019) present in the original classification data are leveled out for the Textual entailment task on the recasted data due to label balance [5].

We evaluated state-of-the-art language models (Conneau et al., 2019) performance on the recasted TE data. We then combine the predictions of related pairs (same premise) from TE task to predict the classification labels of the original data (premise sentence), a *two-step classification*. We observed that a better TE performance on the recasted data leads to higher accuracy on the followed classification task. We also observed that TE models can make inconsistent predictions across samples derived from the same *context* sentence. Driven by these observations, we propose two improvements to TE and classification modeling. First, we introduce a regularisation constraint based on the work of (Li et al., 2019) that enforces consistency across pairs of training samples, thus correcting inconsistent predictions. Second, we propose a joint objective for training TE and classification simultaneously. Our results demonstrate that the regularization constraint and joint training helps improve the performance of both the TE models and the followed classification task. Though our work demonstrates the use of recasting and modeling improvements for TE in Hindi, we expect these techniques can be applied to other low-resource languages and other semantic phenomenon beyond textual classification.

Following are the main contributions of this work:

1. We develop new NLI datasets for a low-resource language *Hindi* using recasting (Section 3) and evaluated state-of-the-art language models on them (Section 4.1).

2. Based on our analysis of inconsistencies in the predictions of TE models, we propose a new regularisation constraint (Section 4.1.1).

3. We propose a two-step classification approach that uses TE predictions from *context-hypothesis* pairs to predict the labels of the original classification task (Section 4.2).

4. We propose a novel joint-training objective paired with consistency regularisation to obtain state-of-the-art performance for text classification on four Hindi datasets (Section 4.2.1).

[5] See Appendix Section A.4 for other benefits of recasting data.

2 Related Work

In this section, we list some of the related works in the field of NLI as well as challenges encountered in low-resource settings.

2.1 Natural Language Inference

Recent studies in the field of NLI have emphasized the role of TE for estimating language comprehensibility of the models. White et al. (2017) takes into consideration the need to leverage the existing pool of annotated collections as targeted textual inference examples (such as pronoun resolution and sentence paraphrasing). Poliak et al. (2018b) discussed existing biases in NLI datasets which helps the models to perform well on Hypothesis-only baselines. Poliak et al. (2018a) analysed NLI datasets based on various semantic phenomenon to verify the ability of a model to perform unique, varied levels of reasoning. It performs data recasting on existing classification datasets to obtain a conventional context/hypothesis/label for common NLI tasks. Several modifications have been tried over baseline models for enhanced NLI and NLU. Liu et al. (2019) focuses on NLU over cross-task data to achieve generalisability over new unseen tasks. Li et al. (2018) incorporates attention mechanism to capture semantic relations in between individual words of the sentence for robust encodings.

However, NLI has mostly revolved around English language. Our approach is motivated by such studies to analyse NLU using current embeddings for low-resource languages like *Hindi*. Bhattacharyya (2012) discusses some of the key challenges associated with *Hindi*, for example, grammatical constraints for most words to be masculine/feminine (similar to French and unlike English), which makes

semantic tasks like pronoun resolution, paraphrasing tough.

2.2 NLP for Low-Resource Languages

In a plethora of diverse languages, only a handful of them have plenty of labeled resources for data-driven analysis and advancements (Joshi et al., 2020). Data in low-resource languages is either unlabeled or resides in spoken dialect than texts. There have been recent efforts using curriculum learning for making pretrained language models for several multi-lingual tasks (Conneau et al., 2018, 2019). However, many such languages give rise to creoles, building new mixed languages at the interface of existing languages. One such example is Hinglish (Hindi + English) that has widely been taken over in the form of tweets and social media messages. Attempts have been made to study linguistic tasks like language identification, NER (Singh et al., 2018) and detection of hate speech from social media (Mathur et al., 2018). (Sitaram et al., 2019) looks at the challenges and opportunities of code-switching.

Joshi et al. (2019) compares the current deep learning methods for classification tasks in Hindi and concludes the need of more efficient models for the same. Apart from that, low-resource languages also challenge us to shift from data-driven modelling to intelligent neural modelling. This improves language understanding from limited available data and also diminishes the need of hand-engineered feature representations similar to generative modelling. Some such efforts have been put forth by Kumar et al. (2019) and Akhtar et al. (2016). Keeping these challenges in mind, this work is a step towards understanding of a low-resource language - *Hindi* using TE.

3 Recasting Classification Datasets

One of the main challenges for TE evaluation for low-resource languages is the lack of labeled data. In this work, we employ recasting to convert annotated classification datasets in *Hindi* to labeled TE samples. As in (Poliak et al., 2018a), we selected four different datasets for recasting thus introducing linguistic diversity in the resulting TE dataset.

Product Review - The first dataset (*PR*) contains 5,417 samples of online user reviews in Hindi for different products (Akhtar et al., 2016). These samples were annotated into one of the following four sentiment classes: *positive*, *negative*, *neutral*, and *conflict*. For recasting the samples in this dataset, we first built 8 hypothesis templates: 2 per class label. For each label, we create one positive and one negative hypothesis which roughly translate to: *'This product got <label> reviews'* and *'This product did not get <label> reviews'*.

Given a sample from the PR dataset, we treat it as the context sentence and combine with the 8 hypotheses sentences to create NLI samples. If the *<label>* of the premise matches that of the positive hypothesis, then the NLI sample is marked as '*entailed*'. Likewise, if the *<label>* of the premise does not match the negative hypothesis, then the NLI sample is also marked as '*entailed*'. For the remaining cases, the sample is marked as '*non-entailed*'. This process is summarized with an example in Figure 1. For more detailed recasting illustration, see Appendix Section A.1 Figure 5.

BHAAV - The second dataset BHAAV (*BH*) (Kumar et al., 2019) contains 20,304 sentences from Hindi short stories annotated for one of the following five emotion categories: *joy*, *anger*, *suspense*, *sad*, and *neutral*. We used a similar process as *PR* to recast *BH* using the following templates to create the hypothesis: *'It is a matter of great <label>'* and *'It is not a matter of great <label>'*.

Hindi Discourse Modes Dataset (HDA) - This dataset (Dhanwal et al., 2020) consists of 10,472 sentences from Hindi short stories annotated for five different discourse modes `argumentative`, `narrative`, `descriptive`, `dialogic` and `informative`.

Hindi BBC News Dataset (BBC) - This dataset[6] contains 4,335 Hindi news headlines tagged across 14 categories: *India*, *Pakistan*, *news*, *International*, *entertainment*, *sport*, *science*, *China*, *learning english*, *social*, *southasia*, *business*, *institutional*, *multimedia*. We processed this dataset to combine two sets of relevant but low prevalence classes. Namely, we merged the samples from *Pakistan*, *China*, *international*, and *southasia* as one class called

[6] https://tinyurl.com/y8hxtbn8

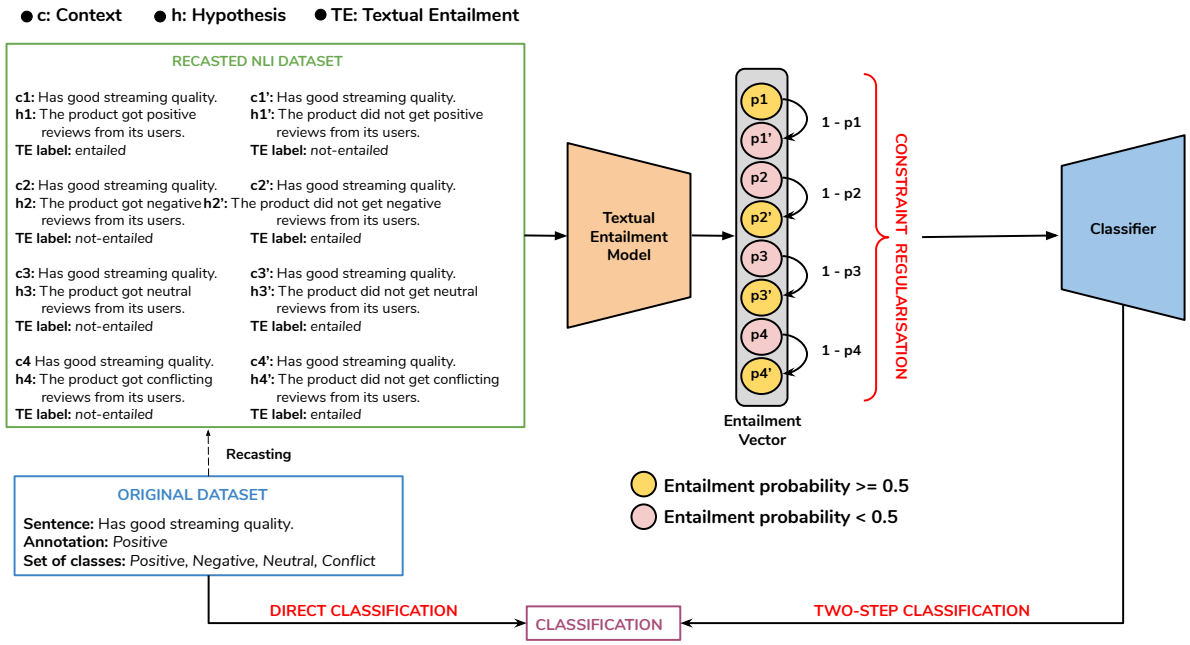

Figure 1: Illustration of the proposed approach

`international`. Likewise, we also merged samples from *news*, *business*, *social*, *learning english*, and *institutional* as `news`. Lastly, we also removed the class *multimedia* because there were very few samples.

Table 2 shows statistics about the datasets and Table 3 shows examples from each.

	Datasets			
	PR	**BH**	**HDA**	**BBC**
Original datasets				
# Classes	4	5	5	6
# Train	4334	16243	8377	3889
# Dev	541	2030	1047	216
# Test	542	2031	1048	217
Recasted TE data				
# Classes	2	2	2	2
# Train	17336	64972	33508	15556
# Dev	4328	20300	10470	2592
# Test	4336	20310	10480	2604

Table 2: Statistics of the original classification data and recasted NLI data.

4 Methodology

Our objective in this paper is not only to use recasting to create a NLI dataset in low-resource settings but also to understand how different models are effective in both TE and classification task. Furthermore, we also discuss our novel *two-step* classification technique with joint objective and regularization constraints.

4.1 Textual Entailment

One straightforward application of NLI comes with evaluating the task of Textual Entailment (TE). It analyses if the TE model can draw reasonable inferences from the context to hypothesise over other related/unrelated data, as shown in Table 1.

However, apart from being correct/incorrect, certain times, TE models are not always consistent with their own beliefs (Li et al., 2019) due to spurious patterns in the dataset (Poliak et al., 2018a). Consider two *context-hypothesis* pairs P and P' generated from the same context sentence and opposing hypotheses statements (as illustrated in Figure 1). Consequently, P and P' would have opposing TE labels. When a TE model makes predictions on these two pairs, there are three possibilities (Table 5). The model can get both predictions right, in which case the predictions are consistent. It can also get both predictions wrong but still they are consistent. Lastly, it can get one of the predictions wrong, in which case they are inconsistent[7]. To mitigate this inconsistency problem, we propose consistency regularisation loss.

[7]See Appendix Section A.3 Table 11 for additional inconsistency examples.

Dataset	Sentence (Hindi)	Sentence (English)	Sentiment
PR	फिलहाल , इसमें कोई वीडियो या वॉयस कॉल सपोर्ट नहीं है।	At the moment, there is no video or voice call support.	*negative*
BH	इतनी मिठाइयाँ लीं, मुझे किसी ने एक भी न दी।	Took so many sweets, nobody gave me one.	*anger*
HDA	सौर मंडल के सारे ग्रहे बृहस्पति में समां सकते हैं ।	All the planets in the solar system can be contained within the Jupiter.	*informative*
BBC	अखबार ने बताया कि फेसबुक पर मिलेगी असल जादू की झप्पी।	The newspaper said that real magic hug will be found on Facebook.	*entertainment*

Table 3: Sample sentences from the four datasets and the corresponding annotation labels.

4.1.1 Consistency Regularisation (CR)

To enforce this pairwise-consistency, we add a regularisation loss[8], inspired from (Li et al., 2019), for our settings, where the entailment probabilities p and p' of pairs P and P' respectively, is required to always sum up to one as illustrated in Figure 1. Mathematically, we define the regularisation term as depicted in Equation 1.

$$\mathcal{L}_{reg} = \left\| p + p' - 1 \right\|_2^2 \quad (1)$$

Our regularisation is different from (Li et al., 2019) in terms of different consistency problem being considered, which in-term diversifies a very different inductive bias from former.

4.2 Two-step classification

We further extend the knowledge accumulated by TE predictions for multi-class classification. Consider a TE model with binary output where 1 (*entailed*) represents *entailed* and 0 (*not-entailed*) represents *not-entailed*. One can co-relate model predictions for related TE pairs with same context but different hypothesis during prediction (inference) to retrieve the classification label. This is depicted by an example in Table 4. We call our approach a *two-step classification* method, where we obtain TE predictions in the first step and use them to obtain classification label in step two. For demarcation, we refer to the straightforward task (without the recasted data) as *direct classification*.

Therefore, a perfect TE model would lead to a 100% accuracy over the *two-step classification* task. However, having a completely accurate TE model is often a bottleneck due to inaccurate and inconsistent predictions. Here, inconsistency can even occur across pairs, for example, two different pairs can predict two different labels. So instead of binary outputs, we use soft TE probabilities (p_i) of each context-hypothesis pair (c_i-h_i) and concatenate them together to form an *entailment vector* ($\mathcal{E}$), see Figure 1. The classifier $\mathcal{C} : \mathcal{E} \rightarrow \mathcal{Y}$, then takes as input the *entailment vector* ($\mathcal{E}$) to retrieve the classification label ($\mathcal{Y}$). Here, the *entailment vector* works as an added weaker supervision at the group level (group of all recasted pairs for a given context) to the classifier. Thus the classifier identify the correct boundary for the final classification task.

Furthermore, *two-step classification* adds an

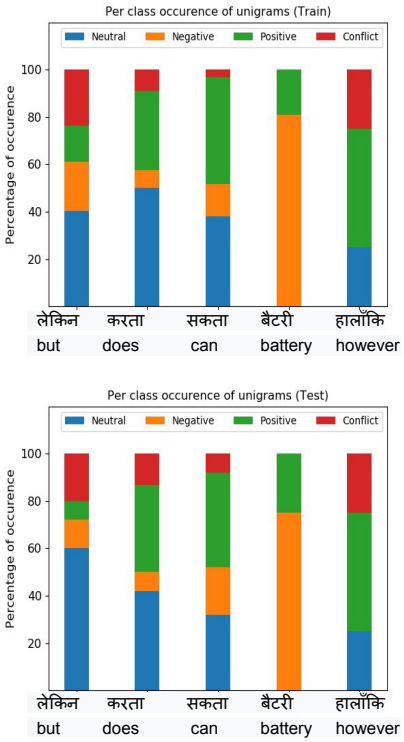

Figure 2: Plot showing statistics of unigram patterns in PR dataset for train (top) and test (bottom) across different classes for some sentiment as well as non-sentiment keywords. The x-axis represents the keyword with the percentage of occurrence on the y-axis.

[8]Other suitable loss function also works (Li et al., 2019).

interpretable advantage over the direct classification. This is because, direct-classifcation is driven by a lot of spurious unigram patterns present in the original dataset. These patterns are leveled in the *two-step classification* approach due to the balanced set of text tokens for both entailed and not-entailed pairs (both labels) with data recasting. Figure 2 shows some of the unigram statistics for PR dataset over some sentiment as well as non-sentiment words to depict the type of artefact patterns in the classification datasets, similar to (Tan et al., 2019). These annotation aretefacts are nullified in the recasted TE task due to balanced label balanced for every premise tokens.

4.2.1 Joint Objective (JO)

One simple method for *two-step classification* is to first train a TE model and then train the classifier on its predictions. However, using a fixed TE model prediction imposes a prior bottleneck on the classification accuracy. Since both the tasks i.e. the TE and the follow-up classification, can influence each other, thus we propose a joint training objective as shown in Equation 2

$$\mathcal{L}_{joint} = \mathcal{L}_{TE} + \lambda \mathcal{L}_{clf} \tag{2}$$

where λ is the weight of the follow-up classification loss, $\mathcal{L}_{TE}$ and $\mathcal{L}_{clf}$ are cross-entropy loss for the task of TE and classification respectively as defined in Equations 3 and 4.

$$\mathcal{L}_{TE} = \sum_{k} \sum_{j=1}^{m} -p_{k,j}^{true} \log p_{k,j} \tag{3}$$

$$\mathcal{L}_{clf} = \sum_{k} \sum_{j=1}^{m} -c_{k,j}^{true} \log c_{k,j} \tag{4}$$

Here, m represents the total classes, $p_{k,j}^{true}$ and $c_{k,j}^{true}$ represent the binary label of sample k to belong to class j, and $p_{k,j}$ and $c_{k,j}$ represent the probability of predicted label for sample k to be class j.

Benefit of Joint Objective. Satisfying the joint objective not only ensures that the model predictions are correct but also ensures that they are correct for the right reasons. The true classification label can be retrieved from the entailment vector only when the model draws necessary inferences correctly. Otherwise the multi-class classification would fail. Furthermore, combining the joint objective (Equation 2) with consistency regulariser (Equation 1) for the intermediate TE prediction further force pairwise-consistency between prediction of related TE pairs.

Context sentence: He cried over his lost pet.	
Hypotheses	**TE Prediction**
1. He is happy.	*not-entailed*
2. He is not happy.	*entailed*
3. He is angry.	*not-entailed*
4. He is not angry.	*entailed*
5. He is sad.	*entailed*
6. He is not sad.	*not-entailed*
Inferred label: *Sad*	

Table 4: An example demonstrating inference of the label for the original classification task based on predictions from TE model.

5 Experiments

Most of the sentence embedding models have been designed and evaluated to perform well on *English* language. The experiments in this work are motivated to answer the following questions for a low-resource language, *Hindi*:

- Are these representations effective to derive logical entailment in context-hypothesis pairs on recasted data?. Furthermore, how consistent/inconsistent are such models with their own decisions? Also, does consistency regulariser help to mitigate model inconsistency?
- Do sentence representation models work well for direct classification? Can models trained on recasted NLI data be used to retrieve ground truth classification annotations using *two-step classification*? Does our joint training objective with consistency regularization improve performance?

Baselines - For evaluating our approach, we use the following baselines: InferSent (Conneau et al., 2017), Sent2Vec (Pagliardini et al., 2018), Bag-of-words (BoW) and XLM-RoBERTa (Conneau et al., 2019) which is state-of-the-art for multilingual language modelling. Also, we evaluate a hypothesis-only analogue for each one of them as well. For experiments with recasted data, we use embeddings of *context-hypothesis* pair for baselines whereas for the hypothesis-only (Poliak

Context (Hindi): वह रोया जब उसने अपना पालतू खो दिया **(English):** He cried over his lost pet.			**Emotion class (Hindi):** दुख **(English):** Sad	
Hypothesis (Hindi)	**Hypothesis (English)**	**TE label**	**Consistency**	**Prediction**
$h1$: वह खुश है	*h1:* He is happy.	*not-entailed*	Consistent	Correct
$h1'$: वह खुश नहीं है	$h1'$: He is not happy.	*entailed*		Correct
$h1$: वह खुश है	$h1$: He is happy.	*not-entailed*	Inconsistent	Correct
$h1'$: वह खुश नहीं है	$h1'$: He is not happy.	*not-entailed*		Incorrect
$h1$: वह खुश है	$h1$: He is happy.	*entailed*	Inconsistent	Incorrect
$h1'$: वह खुश नहीं है	$h1'$: He is not happy.	*entailed*		Correct
$h1$: वह खुश है	$h1$: He is happy.	*entailed*	Consistent	Incorrect
$h1'$: वह खुश नहीं है	$h1'$: He is not happy.	*not-entailed*		Incorrect

Table 5: A simple example illustrating the concept of consistency in model prediction for TE task for the task of emotion analysis.

et al., 2018b) models, we only use embeddings of the *hypothesis* sentence, keeping it blind to the *context*.

Hypothesis only Baselines - Evaluating hypothesis-only models is motivated by irregularities and biases presented in entailment datasets. Such biases often lead to high performance over NLI tasks without completely comprehending the semantic reasonings in data and language. When the accuracy of a hypothesis-only model is much lower than the baseline and closer to random (50%), it exhibits that learning is not boosted due to statistical irregularities in data such as word count, unigram/bi-gram pattern or any other spurious pattern (artefacts). We achieve this using our approach since recasting ensures label balance for the augmentations of each class label for every sentence and its tokens.

Experimental Settings - For each of the models, we use the initial learning rate 1 x 10^{-3} and a decay rate of 0.9, using Adam optimizer with the embedding dimension kept as 1024 for all the models. For all the experiments associated with XLM-RoBERTa, We use XLM-RoBERTa large with 1024-hidden. For InferSent and Sent2Vec we use the default parameter for NLI model architecture as stated in the paper. For hypothesis only baseline we use the single sent model of XLM-RoBERTa, InferSent and Sent2Vec as reported in paper for binary classification.

After the embeddings are obtained, we use an MLP classifier for performing all the classification experiments. For a hypothesis-only baseline, only the hypothesis embedding is passed as an input to the MLP, whereas for a premise-hypothesis baseline, we concatenate the embeddings of premise, hypothesis, as well as their element-wise product and element-wise subtraction. For the joint objective training (see Eq. 2), we use λ=2.0. We train our model for 15 epochs on a machine with GeForce RTX 2080 GPU using the PyTorch framework.

5.1 Textual Entailment Results

For all four semantic phenomenon considered, we use recasted data to predict the performance on textual entailment task. While training, we use four *context-hypothesis* pairs - with hypothesis having true classification label, its negation (hypothesis 5 and 6 in Table 4), a random label from the remaining classes and its negation (hypothesis 1 and 2 in Table 4). This ensures that neither original classification label nor the negation (we choose only one random pair) correlate with entailment labels. For development and test sets, we use all possible $2n$ recasted pairs (where n is the number of classes in classification data) since ideally, while testing we have no prior knowledge of the ground-truth label.

Context-Hypothesis Baselines				
Sentence	**Dataset**			
Representation	**PR**	**BH**	**HDA**	**BBC**
BoW	47.32	51.00	54.20	57.00
Sent2Vec	61.21	62.67	64.00	65.42
InferSent	68.00	65.04	67.9	68.84
XLM-RoBERTa	**74.02**	**74.48**	**75.29**	**73.56**
Hypothesis-only Baselines				
BoW	44.89	47.01	44.82	43.00
Sent2Vec	51.91	50.84	50.88	48.80
InferSent	54.32	52.14	53.54	51.08
XLM-RoBERTa	**55.00**	**52.60**	**53.92**	**55.00**

Table 6: TE classification accuracies using different sentence embeddings for all four datasets.

With Table 6, we establish that XLM-

RoBERTa (Conneau et al., 2019) gives the best performance as compared to all the other baselines. Therefore, we use it for all the following experiments. Also, random performance on hypothesis-only baseline ensures that our recasted data does not contain hypothesis-bias.

Consistency - We analyse the effect of consistency regulariser (CR) by comparing the percentage of inconsistent model predictions for TE models with and without CR. Figure 3 clearly depicts that the constraint regularisation helps in reducing the percentage of inconsistent pairs and hence makes the model predictions congruent with its own internal representation in the model parameters.

5.2 Two-step Classification Results

We now use the TE model to perform *two-step classification* as explained in section 4.2. Table 10 shows the classification accuracies obtained via direct as well as *two-step* classification with consistency regularisation and joint-objective. As reported in Table 9 and 10, we observe a jump in both the TE as well as *two-step classification* accuracies with the addition of consistency regularisation. Such a constraint restricts the model predictions to be either correct or incorrect but not pairwise-inconsistent with its other beliefs.

Joint Objective - In Table 9 and 10, we observe that joint objective proves to be much more beneficial than independent TE and classifier training. The *two-step classification* accuracy with joint-objective (+JO+CR) surpasses the direct classification performance.

We observe an increment of 5% in TE and 2% in classification accuracy across all the datasets. Furthermore, from Figure 3, we observe that, JO also improve the prediction consistency across all the datasets. Table 7 shows the exact percentage of correct/incorrect and inconsistent pairs.

Improved Performance Analysis - The two-step classification is able to achieve overall improvement over direct classification approach mainly due to following two factors. Firstly, the joint objective (JO) helps in creating a feedback loop with the two tasks of textual entailment and classification, which enforce consistency in the model predictions for the two tasks. Secondly, the consistency regularisation (CR) for the TE helps in making the model decisions congruent across same context premise but different related hypothesis. Thus, both the JO and CR imposes indirect and direct inductive bias through constrained loss objective which improves model performance compared to the direct classification task.

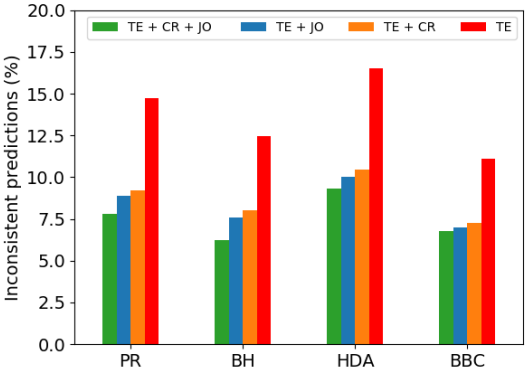

Figure 3: Plot depicting percentage (%) of inconsistent predictions for all the datasets using *XLM-RoBERTa* with and without consistency regularisation (CR) and Joint Objective (JO).

5.3 Direct vs Two-Step Classification

We analyse the classification predictions obtained by direct as well as two-step classification to compare the differences. Figure 4 shows the percentage (%) of correct and incorrect predictions obtained for the two approaches considered. More generally, we see a maximum consensus across the main diagonal between the two approaches. However, there are irregularities wherein one of the predictions contradicts the other.

As illustrated in Table 8, we depict qualitative examples corresponding to these irregularities. We analyse their entailment vectors to interpret intermediate predictions and realise that the high entailments corresponding to the gold label and certain incorrect label lead to incorrect predictions. For example, for the first sentence in Table 8, we observe that the context-hypothesis pairs with hypothesis corresponding to *The product received negative reviews from its users*, and *'The product received conflicting reviews from its users'* get the entailment probabilities 0.64 and 0.58, respectively. This shows that apart from the

Dataset	Correct				Incorrect				Inconsistent			
	TE	+CR	+JO	+CR +JO	TE	+CR	+JO	+CR +JO	TE	+CR	+JO	+CR +JO
PR	71.43	72.18	72.50	**74.00**	**13.82**	18.6	18.6	18.2	14.75	9.22	8.90	**7.80**
BH	73.20	74.50	74.76	**75.80**	**14.32**	17.50	17.66	17.99	12.48	8.00	7.58	**6.21**
HDA	72.00	74.88	75.22	**76.8**	**11.50**	14.66	14.78	13.9	16.50	10.46	10.00	**9.30**
BBC	71.17	74.56	74.84	**76.00**	**17.75**	18.2	18.16	17.2	11.08	7.24	7.00	**6.80**

Table 7: Percentage (%) of correct, incorrect and inconsistent prediction pairs for all the datasets using XLM-RoBERTa.

Sentence	True Label	Direct clf.	Two-step clf.
यहाँ खाना पीना उतना मँहगा नहीं पर रहना जेब को काफी भारी पड़ता है । **English:** Drinking here is not that expensive but living on the pocket is very heavy.	*negative*	*conflict*	*negative*
राजगुरु , महाराज कृष्णदेव राय को कहते है के तेनालीराम झूठ बोल रहे है । **English:** Rajguru tells Maharaja Krishnadeva Raya that Tenaliram is lying.	*anger*	*anger*	*sad*

Table 8: Qualitative examples where direct and two-step classification methods contradict predictions.

Dataset	Textual Entailment			
	w/o CR/JO	+CR	+JO	+CR+JO
PR	74.02	77.80	78.40	**81.40**
BH	74.48	76.57	77.01	**80.05**
HDA	75.29	78.00	78.22	**81.67**
BBC	73.56	76.24	77.69	**79.22**

Table 9: TE accuracies for all the four datasets using XLM-RoBERTa (Conneau et al., 2019).

Dataset	Direct clf.	Two-step clf.			
		TE	TE+ CR	TE+ JO	TE+ CR+JO
PR	71.65	66.24	69.38	70.58	**73.70**
BH	73.03	68.06	70.91	71.82	**74.80**
HDA	74.25	68.22	71.45	72.45	**75.96**
BBC	70.22	65.98	68.20	70.30	**72.18**

Table 10: Classification (direct and two-step) accuracies for all the four datasets using XLM-RoBERTa (Conneau et al., 2019).

gold label i.e. *negative* here, there is an inclination towards the class label *conflict*.

Moreover, we see certain statistical word patterns like the usage of the keyword *but* in most of the sentences corresponding to the class *conflict*, thereby ensuring a certain degree of artefact learning which governs the decisions in direct classification. One advantage of two-step classification is that it is more transparent about it's predictions. This ensures more interpretability in the model decisions. We also compare class-wise accuracies of both the approaches for each of the datasets and see improvements with the two-step method in all classes[9].

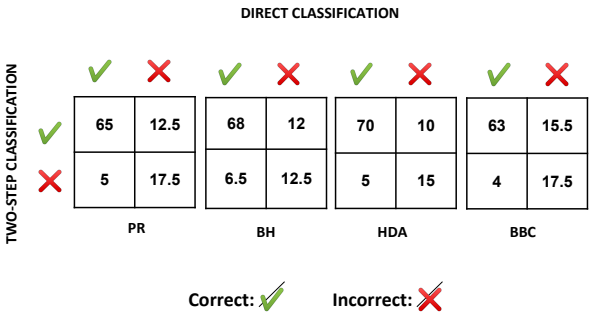

Figure 4: Correct vs Incorrect Predictions (%) for Direct and Two-Step classification.

6 Conclusion

In this work, we share the first recasted NLI dataset in a low-resource language Hindi, and show how a large-scale NLI data can be developed for low-resource languages without undergoing costly and time taking human annotations. We perform TE experiments and introduce a consistency regulariser to avoid pairwise-inconsistent TE predictions. Furthermore, we propose a *two-step* classification approach with a joint training objective. Our results with the joint objective shows significant improvement in performance.

As a future work, we aim to analyse the proposed methodology which is language independent on other low-resource languages. We

[9]See Appendix Section A.2 Figure 6 for class-wise results

also aim to use more generalisable templates for linguistic diversity in recating data. It would be interesting to analyse how extending textual entailment knowledge especially the consistency regularization constraint affect other downstream NLP tasks apart from textual classification, not only in terms of the performance, but also in enhancing the model interpretability.

References

Md Shad Akhtar, Ayush Kumar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. A hybrid deep learning architecture for sentiment analysis. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 482–493.

Pushpak Bhattacharyya. 2012. Natural language processing: A perspective from computation in presence of ambiguity, resource constraint and multilinguality. *CSI journal of computing*, 1(2):1–13.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, F. Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *ArXiv*, abs/1911.02116.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Swapnil Dhanwal, Hritwik Dutta, Hitesh Nankani, Nilay Shrivastava, Yaman Kumar, Junyi Jessy Li, Debanjan Mahata, Rakesh Gosangi, Haimin Zhang, Rajiv Ratn Shah, and Amanda Stent. 2020. An annotated dataset of discourse modes in Hindi stories. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1191–1196, Marseille, France. European Language Resources Association.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Ramchandra Joshi, Purvi Goel, and Raviraj Joshi. 2019. Deep learning for hindi text classification: A comparison. In *International Conference on Intelligent Human Computer Interaction*, pages 94–101. Springer.

Simran Khanuja, S. Dandapat, S. Sitaram, and M. Choudhury. 2020. A new dataset for natural language inference from code-mixed conversations. In *CodeSwitch@LREC*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Yaman Kumar, Debanjan Mahata, Sagar Aggarwal, Anmol Chugh, Rajat Maheshwari, and Rajiv Ratn Shah. 2019. Bhaav- a text corpus for emotion analysis from hindi stories. *ArXiv*, abs/1910.04073.

Guanyu Li, Pengfei Zhang, and Caiyan Jia. 2018. Attention boosted sequential inference model. *CoRR*, abs/1812.01840.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A logic-driven framework for consistency of neural models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in hindi-english code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. 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 528–540.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018a. Collecting diverse natural language inference problems for sentence representation evaluation. In *BlackboxNLP@EMNLP*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018b. Hypothesis only baselines in natural language inference. In **SEM@NAACL-HLT*.

Mark Sammons, V.G.Vinod Vydiswaran, and Dan Roth. 2010. "ask not what textual entailment can do for you...". In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1199–1208, Uppsala, Sweden. Association for Computational Linguistics.

Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru. 2018. Language identification and named entity recognition in hinglish code mixed tweets. In *Proceedings of ACL 2018, Student Research Workshop*, pages 52–58.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2019. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*.

Shawn Tan, Yikang Shen, Chin-Wei Huang, and Aaron C. Courville. 2019. Investigating biases in textual entailment datasets. *ArXiv*, abs/1906.09635.

Marta Tatu and Dan Moldovan. 2005. A semantic approach to recognizing textual entailment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 371–378, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Rui Wang and Yi Zhang. 2009. Recognizing textual relatedness with predicate-argument structures. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 784–792. Association for Computational Linguistics.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

A Appendix

A.1 Illustration of Recasting Approach

We illustrate the proposed recasting approach in more detail with example templates in Figure 5. We show how each classification sentence is used to create a context-hypothesis pair for NLI task for different datasets corresponding to the diverse semantic phenomenon considered.

A.2 Additional Results

Development Set Results - We report the results on development set for textual entailment as well as classification in Table 12 and 13 respectively. We observe similar trends in the development set as depicted in the test set performance for both the tasks of textual entailment as well as the *two-step* classification task.

Class-wise Performance - In Figure 6, we show class-wise accuracies obtained by the two classification approaches - direct vs two-step. Broadly, we obtain a considerable improvement in the performance of two-step classification over direct classification, over all classes across all the four datasets. This ensures that the obtained performance improvement is balanced across all classes.

Semi-supervised setting - We extend our analysis to a semi-supervised setting (with fewer labels) wherein we retain the true labels for only 40%, 60% and 80% of the data while training and analyse its effect on the performance of TE and classification tasks.

Table 14, 16 and 18 show the results obtained with different ablations with 80%, 60% and 40% of the labelled data respectively for the TE task. Similarly, Table 15, 17 and 19 report the results for direct and two-step classification in the semi-supervised approach highlighting the effect of joint objective and consistency regularisation in obtaining improvement.

Although, we utilize the consistency regularisation, since it does not depend on the true label, rather operated on pairwise context-hypothesis groupings. We observe that TE with consistency regularisation and joint objective surpasses the trivial TE task without any added constraints. This depicts that our regularisation and joint objective approach add robust improvements in TE model performance even with minimum supervision.

A.3 Another Inconsistency Example

In Table 11, we explain the concept of pairwise consistencies and inconsistencies in the context-hypothesis pairs in the recasted data with an example. It depicts how different entailment results for the same context but different hypothesis can lead to inconsistencies within the model predictions.

A.4 Benefits of Data Recasting

There are several benefits of data recasting (Conneau et al., 2019) especially for low-resource languages

- Recasting is an automated process and hence remove the need of expensive human annotation to labelled data.
- Uniform procedure of recasting data has equal number of *context-hypothesis* pairs for each label, hence making it neutral to statistical irregularities (see hypothesis bias experiments in Section 5).
- Diverse semantic phenomenon for various classification tasks can be unified as a single task using data recasting.

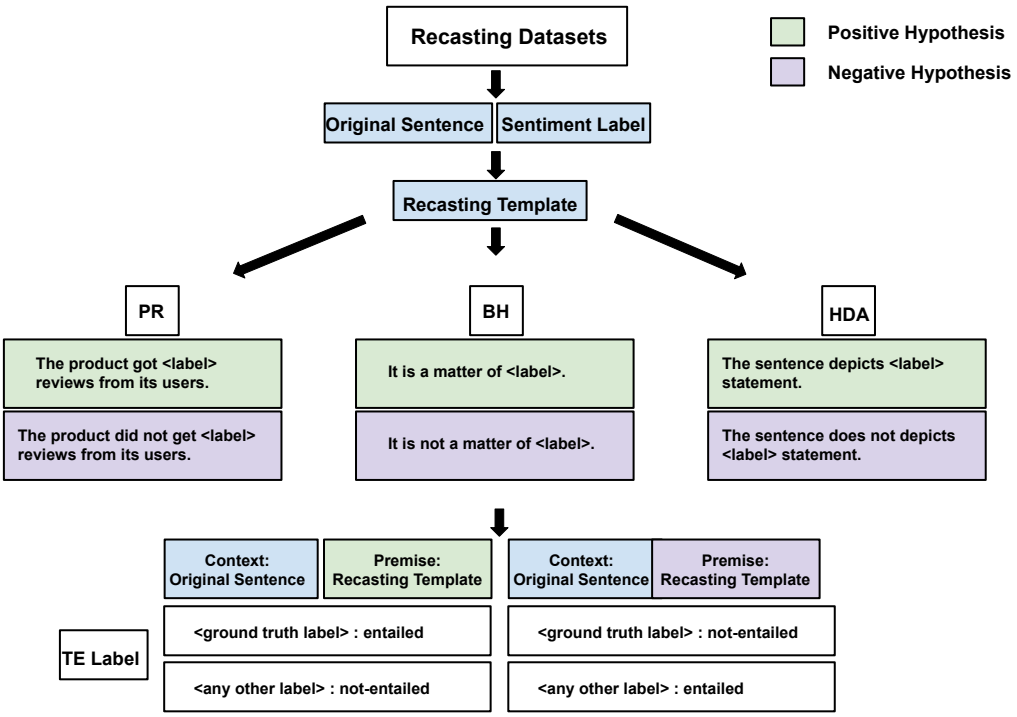

Figure 5: Illustration of the proposed recasting approach.

Original Sentence(Hindi)	**Original Sentence (English)**	**Sentiment**		
इन पवित्र भावों से उसकी आत्मा विह्वल हो गयी।	His soul was overwhelmed by these holy feelings.	Joy		
Model Consistency/Inconsistency				
Contradictory TE pairs (Hindi)	**Contradictory TE pairs (English)**	**Prediction**		**Label**
		p-*h*1	*p*-*h*2	
p: इन पवित्र भावों से उसकी आत्मा विह्वल हो गयी।	*p*: His soul was overwhelmed by these holy feelings.	*e*	*e*	Inconsistent
*h*1: क्या यह खुशी की बात है?	*h*1: Is this a matter of joy?	*e*	*ne*	Correct
p: इन पवित्र भावों से उसकी आत्मा विह्वल हो गयी।	*p*: His soul was overwhelmed by these holy feelings.	*ne*	*e*	Incorrect
*h*2: क्या यह खुशी की बात नहीं है?	*h*2: Is this not a matter of joy?	*ne*	*ne*	Inconsistent

Table 11: Example sentences for contradictory premise (p) - (h) pairs for measuring inconsistency in the recasted model predictions with e representing *entailed* and ne representing *not-entailed*.

Dataset	**Textual Entailment** ↑			
	w/o	+CR	+JO	+CR+JO
PR	74.26	78.44	78.02	**80.60**
BH	73.88	76.46	76.82	**80.95**
HDA	75.90	78.54	78.48	**81.86**
BBC	73.45	76.48	77.96	**79.02**

Table 12: TE accuracies for all the four datasets using XLM-RoBERTa on the development set.

Dataset	**Direct clf.**	**Two-step clf.** ↑			
		TE	TE+ CR	TE+ JO	TE+ CR+JO
PR	71.40	65.48	68.76	70.84	**72.98**
BH	73.50	69.24	70.88	71.46	**75.66**
HDA	74.85	68.46	72.34	73.50	**75.56**
BBC	71.36	66.40	68.38	70.47	**73.08**

Table 13: Classification (direct and two-step) accuracies for all the four datasets using XLM-RoBERTa on the development set.

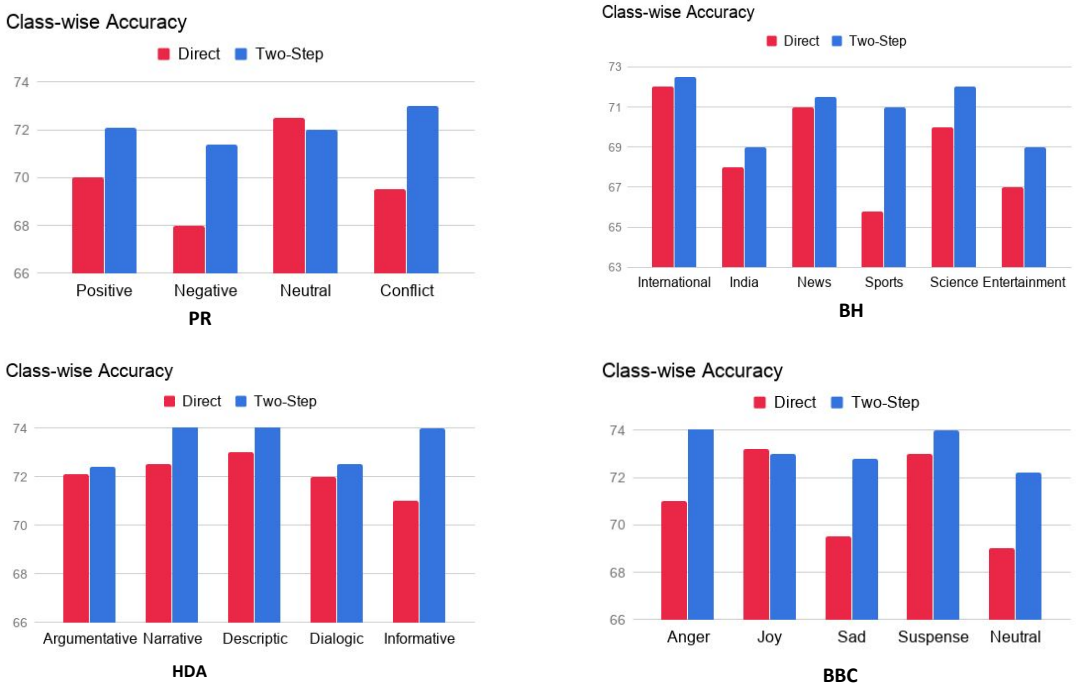

Figure 6: Class-wise comparison of Direct vs Two-Step Classification.

Dataset	Textual Entailment ↑			
	w/o	+CR	+JO	+CR+JO
PR	69.23	72.68	70.48	**74.04**
BH	70.65	71.09	70.99	**73.98**
HDA	70.29	72.23	71.32	**74.67**
BBC	70.36	73.84	71.65	**74.52**

Table 14: TE accuracies for all the four datasets using XLM-RoBERTa with fewer labels (80%).

Dataset	Direct clf.	Two-step clf. ↑			
		TE	TE+ CR	TE+ JO	TE+ CR+JO
PR	67.20	61.28	64.87	62.49	**68.98**
BH	68.51	64.22	66.71	71.46	**69.46**
HDA	68.82	62.62	65.13	63.75	**69.95**
BBC	66.93	60.94	63.14	61.47	**67.73**

Table 15: Classification (direct and two-step) accuracies for all the four datasets using XLM-RoBERTa with fewer labels (80%).

Dataset	Textual Entailment ↑			
	w/o	+CR	+JO	+CR+JO
PR	65.12	67.46	65.58	**70.06**
BH	66.12	68.57	67.22	**70.69**
HDA	65.29	67.25	66.34	**70.59**
BBC	66.87	68.22	67.19	**71.42**

Table 16: TE accuracies for all the four datasets using XLM-RoBERTa with fewer labels (60%).

Dataset	Direct clf.	Two-step clf. ↑			
		TE	TE+ CR	TE+ JO	TE+ CR+JO
PR	60.29	61.82	62.37	62.00	**63.98**
BH	61.52	62.14	64.18	62.45	**64.81**
HDA	61.82	63.47	63.94	63.33	**65.56**
BBC	60.23	61.24	62.16	62.09	**64.73**

Table 17: Classification (direct and two-step) accuracies for all the four datasets using XLM-RoBERTa with fewer labels (60%).

Dataset	Textual Entailment ↑			
	w/o	+CR	+JO	+CR+JO
PR	57.12	58.46	58.08	**59.56**
BH	59.12	59.57	59.22	**60.69**
HDA	59.29	59.25	60.19	**60.78**
BBC	58.42	58.70	58.10	**59.02**

Table 18: TE accuracies for all the four datasets using XLM-RoBERTa with fewer labels (40%).

Dataset	Direct clf.	Two-step clf. ↑			
		TE	TE+ CR	TE+ JO	TE+ CR+JO
PR	55.29	56.28	56.48	57.00	**59.89**
BH	58.52	59.17	59.18	59.59	**60.11**
HDA	58.82	58.43	58.94	59.23	**60.68**
BBC	55.23	57.24	56.46	58.01	**60.78**

Table 19: Classification (direct and two-step) accuracies for all the four datasets using XLM-RoBERTa with fewer labels (40%).