# Learning from Disagreements

**Aditi Hande** **Shanay Ghag** **Tanmay Jain** **Dipali Telavane**

**Venkat Sumanth Reddy Vadde**

University of Southern California

## Abstract

In recent years, the assumption that natural language (NL) expressions have a single and identifiable interpretation in a given context is increasingly recognized as just a convenient idealization. The Learning with Disagreement shared task aims to provide a unified testing framework for learning from disagreements.

## 1 Introduction

Recently, people have realized that it's not always true that when we use words or sentences, they only have one specific meaning in a particular situation. It's a simplified idea we sometimes use for convenience. The *Learning with Disagreement* shared task aims to address the problem of learning from disagreements in natural language interpretation by providing a standardized testing framework. Through this project, we hope to build models that successfully predict the ground truth (gold) target label and models that can effectively capture the disagreement information in the dataset. For these tasks, we plan to explore contrastive learning, which has traditionally been used for computer vision tasks but has also gained much traction in NLP research.

Addressing the challenge of learning from disagreements in natural language interpretation is vital because it reflects the complexity of human communication. People often have different perspectives, opinions, and interpretations of language, even when communicating about the same topic. It is necessary to account for these variations in interpretation to develop accurate and effective natural language processing systems.

Furthermore, many real-world applications of natural language processing require a nuanced understanding of language beyond simple, one-dimensional interpretations. For instance - in healthcare, a system that can accurately interpret patient notes and medical reports could help healthcare providers make more informed decisions about patient care. In finance, a system that can accurately analyze financial news and reports could help investors make more informed investment decisions. Addressing the challenge of learning from disagreements in natural language interpretation can help us create more advanced and beneficial natural language processing systems to improve various aspects of our lives.

Meeting this objective could help researchers and developers in natural language processing to improve their understanding of how people interpret language and to develop more accurate models and algorithms for language understanding. This could lead to advancements in machine translation, sentiment analysis, and text classification. Additionally, it could facilitate the creation of more effective tools for natural language processing in various industries, such as healthcare, finance, and education.

We propose a two-fold training strategy to capture information from hard and soft target labels effectively. The first stage involves pre-training a transformer-based encoder model using only hard labels through supervised contrastive learning. This allows the model to learn highly discriminative representations that capture both the underlying structure of the data and class labels, resulting in better performance on downstream tasks.

In the second stage, we fine-tune the pre-trained encoder using soft-label probability distributions by treating it as a regression task. Specifically, we regress over the probability of one of the labels (0 or 1). This two-stage training strategy enables us to effectively utilize both target labels and perform better on our objective task. This training pipeline has been discussed further in section 4.

## 2 Related Work

The current methods for learning from crowd annotations can be broadly categorized into four categories.

Aggregation of coder judgments: The first category comprises techniques ((Dawid and Skene, 1979)) that automatically combine annotations from a crowd into a single label for each case. These techniques suppose a sole, objective "gold" truth exists for every instance and aim to evaluate this standard without depending on manual adjudication.

Filtering hard items: The second category involves techniques ((Reddy et al., 2015)) that also assume the presence of a gold label but relax the notion that it can always be retrieved. These methods use the disagreement information to filter or weigh hard items.

Learning directly from crowd annotations: The third category comprises techniques ((Plank et al., 2014), (Fornaciari et al., 2021)) that directly train a classifier from the crowd annotations, potentially using a probability distribution that assigns a score to each label obtained from the crowd annotations.

Augmenting hard labels with disagreements: The fourth category uses techniques ((Aroyo and Welty, 2013), (Sheng et al., 2008)) that train a classifier by combining hard and soft labels obtained from crowd annotations. These methods utilize either gold labels or approximated ground truths for training and supplementing them with data from the crowd annotations to weigh items by their difficulty or annotator ability.

## 3 Problem Description

This project aims to provide a unified testing framework for learning from disagreements in Natural Language tasks using datasets containing information about disagreements in interpreting language. The expectation being that unifying research on disagreement from different fields may lead to novel insights and impact ai widely. The input for the task is a large corpus of natural language text. Each of these corpora comes from different datasets aimed at solving different subjective binary classification tasks. In each dataset, we have two types of target labels - the hard labels (0 or 1) as well as the soft label probability distribution (for eg. 0.75 and 0.25 for labels 0 and 1, respectively), which gives us the disagreement information for each instance in the datasets. The aim is to use either soft labels, hard labels, or a combination to output a probability distribution that captures the disagreement information and models the soft-label probability distribution.

## 4 Methods

### 4.1 Data Preprocessing

A benchmark of four textual datasets have been compiled, each with unique characteristics, including different genres (social media and conversations), languages (English and Arabic), tasks (misogyny, hate speech, and offensiveness detection), and annotations meth- ods (experts, specific demographic groups, and AMT-crowd). The datasets encourage training with disaggregated labels and focus entirely on subjective tasks. Since these datasets are quite different in structure, an aggregation function was written to capture all different datasets into a single dataset with a composite structure.

We used a distilbert-base-multilingual-case pretrained tokenizer for the contrastive learning approach to convert the text into tokenized text and attention masks (to indicate which tokens are relevant to the model's attention). It also handles special tokens, such as adding [CLS] and [SEP] tokens and truncating/padding the text to a maximum length of 128. Apart from this, hard labels are normalized as part of data pre-processing.

### 4.2 Contrastive Learning to Warm-Up the Transformer Models

Transformer-based language models, such as BERT and its variants, have achieved remarkable success in a wide range of natural language processing tasks. However, fine-tuning these models on specific downstream tasks can still be computationally expensive and require much labeled data. To tackle this we have used contrastive learning as a warm-up strategy for pre-training transformer models, which can effectively leverage small amounts of labeled data and capture underlying data structure to improve downstream tasks' performance.

Our goal is to learn a text representation by maximizing agreement between inputs from positive pairs via a contrastive loss in the latent space and the learned representation can then be used for the downstream task. We used the Multilingual DistilBert model, a pre-trained transformer-based language model, as the base architecture for our study. The model was warmed-up on our task using a contrastive loss function to learn a mapping from text inputs to an embedding space where similar inputs are closer together and dissimilar inputs are further apart.

Specifically, we initialized the DistilBert model with pre-trained weights and then trained the model on our task-specific dataset using the triplet margin loss function. The loss function compares the distances between three examples: an anchor example, a positive example (with the same label as the anchor), and a negative example (with a different label than the anchor). The anchor, positive, and negative examples are selected from the same batch. The loss is calculated based on the difference between the distance between the anchor and positive examples and the distance between the anchor and negative examples. The loss is only incurred if the difference between these distances is less than a pre-specified margin value. The triplet margin loss encourages the model to learn embeddings that place examples with the same label closer together than examples with different labels, by at least the margin value. This helps the model to learn more discriminative embeddings that can be used for various downstream tasks, such as classification or retrieval.

#### 4.2.1 Triplet Margin Loss function:

The loss function for each sample in the mini batch is

$$\begin{aligned} L(a,p,n) = \max\{d(a_i,p_i)- \\ d(a_i,n_i) + margin, 0 \end{aligned} \tag{1}$$

where

$$d(x_i, y_i) = \|x_i - y_i\|_p \tag{2}$$

$a$ : anchor,
$p$: Positive sample from the same batch,
$n$: Negative sample from the same batch

To evaluate the effectiveness of our training approach, we also experimented with using a contrastive loss function in place of the triplet margin loss. This involved selecting pairs of similar and dissimilar examples from the same batch based on their labels and optimizing the model to minimize the distance between similar examples and maximize the distance between dissimilar examples in the embedding space.

$$yd^2 + (1-y)max(margin - d, 0)^2 \tag{3}$$

where,
L is the contrastive loss
py is a binary label that indicates whether the two inputs are similar (Y=0) or dissimilar (Y=1)
d is the distance between the two inputs in the feature space. We have used Lp distance.
Margin is a hyperparameter that specifies the minimum distance between similar examples and the maximum distance between dissimilar examples We found that the transformer model trained with triplet loss achieved better performance compared to the contrastive loss on our task-specific dataset. To further improve the performance of our model, we also used techniques such as dropout and weight decay during training to prevent overfitting.

### 4.3 Fine-Tuning the Transformer Models

After warming up the model, we added a linear layer and fine-tuned it for the regression task to regress the soft-labels indicating the disagreement. We evaluated the fine-tuned model using CrossEntropy Loss on a held-out test set.

Our experiments showed that fine-tuning the contrastive learning model significantly improved classification performance compared to training from scratch. The results suggest that contrastive learning can be an effective warmup strategy for downstream tasks.

## 5 Experimental Results

### 5.1 Experimental Setup

We will use an existing dataset released as a shared task in semeval 23. A benchmark of four textual datasets has been compiled, each with unique characteristics including different genres (social media and conversations), languages (English and Arabic), tasks (misogyny, hate speech, and offensiveness detection), and annotations methods (experts, specific demographics groups, and AMT-crowd). However, each dataset has multiple labels for each instance. Following are the four datasets we shall use:

HS-Brexit dataset: The HS-Brexit dataset contains tweets on Abusive Language on Brexit and is annotated for hate speech, aggressiveness, and offensiveness by six annotators belonging to two distinct groups: a target group of three Muslim immigrants in the UK and a control group of three other individuals.

ArMIS dataset: The ArMIS dataset is a new dataset of Arabic tweets annotated for misogyny detection by annotators with different demographic characteristics, including "Moderate Female," "Liberal Female," and "Conservative Male."

ConvAbuse Dataset: The ConvAbuse dataset consists of English dialogues between users and two conversational agents, annotated by experts in gender studies using a hierarchical labeling scheme.

MultiDomain Agreement dataset: The MultiDomain Agreement dataset contains English tweets from three domains (BLM, Election, Covid-19) and is annotated for offensiveness by five annotators via Amazon Mechanical Turk. Particular attention was given to pre-selecting annotated tweets that could lead to disagreement. Almost one-third of the dataset was annotated with a two vs. three annotator disagreement, and another third had an agreement of 1 vs. 4.

The datasets encourage training with disaggregated labels and focus entirely on subjective tasks. Dataset-specific information is also provided, which varies for each dataset, from demographics information of annotators (ArMIS and HS-Brexit datasets) to other annotations made by the same annotators within the same dataset (all datasets) and additional annotations given for the same item by the same annotator (HS-Brexit and ConvAbuse datasets). We are hoping to leverage this information to improve performance for specific datasets.

### 5.2 Baseline Methods

The project's goal is to utilize disagreement as a means to model the complexity and ambiguity of subjective tasks, thereby extracting meaningful information. We explored methods belonging to two subcategories - methods that directly learn from the crowd annotations by learning a probabilistic distribution and methods that use gold labels but supplement these labels with annotation labels. We implemented three existing methods which belong to these categories -

Firstly, we employ a Bayesian approach to model the underlying preference distribution, which allows for an accurate estimation of the ground truth preference ranking. We implemented the model using gpytorch (Gardner et al., 2018), which is a Gaussian Process library in Pytorch. Adam optimizer was used with a learning rate of 0.01

Next, we implemented a BERT-based architecture with a soft cross-entropy loss inspired by (Uma et al., 2021). This combination of probabilistic soft labels with a probability comparing loss function is the essence of a soft loss.

Our third approach entailed incorporating predictions generated from soft labels in conjunction with hard labels. The strategy combined the use of Soft Loss with Multi-Task Learning (MTL) methodologies inspired from (Fornaciari et al., 2021). Furthermore, this approach utilized a BERT-based architecture to facilitate the computation of the soft loss. In addition to BERT, linear layers were incorporated after the architecture. The calculation of the soft loss was based on cross-entropy loss. To further enhance the accuracy of the approach, mean values of hard labels were added as weights to the cross-entropy loss. Finally, we employed the Adam optimizer to optimize the model's performance.

### 5.3 Evaluation Protocols

A way of evaluating models as to their ability to capture disagreement was in need, especially for datasets with substantial extent of disagreement. The simplest 'soft' metric of this type is to evaluate ambiguity-aware models by treating the probability distribution of labels they produce as soft labels, and comparing that to the full distribution produced by annotators, using, for example, cross-entropy.

We evaluated the performance of different pipelines by calculating the cross-entropy loss between the predicted and the ground-truth soft-label distribution. This loss was calculated for the model's performance on the four tasks. The equation for this soft cross-entropy loss can be found in eq. 1.

$$-\sum_{i=1}^{n}\sum_{c} p_{hum}\,(y_i|x_i)\,logp_\theta(y_i = c|x_i) \qquad (4)$$

Where $p_\theta(x|y)$ is obtained by applying a probability function (softmax) over the logits produced by the classifier and $p_{hum}\,(y_i|x_i)$ is the human label distribution.

Since our approach essentially involves a regression-based model to predict the probability, we calculated the mean-squared loss (MSE Loss) between the ground truth probability and the predicted probability.

### 5.4 Results and Discussion

The evaluation findings for all used methodologies are shown in Table 1.

Our proposed model was compared against previously established baseline models in the final evaluation. Table 1 includes the Average Cross Entropy loss of different baseline methods across the four tasks. Our approach outperformed both the BERT + Soft Loss model and the Multi-Task Learning + Soft Loss model. This indicates that pretraining the distilBERT encoder proved to be

| Model | Cross Entropy |
|---|---|
| BERT + Soft Loss | 0.67 |
| Gaussian Processes | 0.54 |
| Muti-Task Learning + Soft Loss | 0.59 |
| **Contrastive Learning + Fine Tuning (our approach)** | **0.57** |

Table 1: Average results for different models

an effective strategy in improving the performance from the baseline models.

Fig. 1 shows our method's training triplet loss results for the four tasks. For ArMIS and HS-Brexit tasks, the figure shows that the triplet loss converges.

| Task/Dataset | Cross Entropy Loss |
|---|---|
| HS-Brexit | 0.48 |
| ArMIS | 0.67 |
| MD Agreement | 0.65 |
| ConvAbuse | 0.48 |

Table 2: Results on different datasets

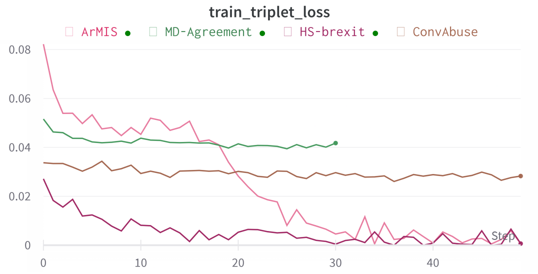

Figure 1: Train Triplet Loss Results

## 6 Conclusions and future work

### 6.1 Conclusion

In summary, addressing the challenge of learning from disagreements in natural language interpretation is vital. In this paper, we have adopted a contrastive learning approach in addition to the baseline models to capture the disagreements. The datasets encourage training with disaggregated labels and focus entirely on subjective tasks. Our proposed two-fold training strategy, involving pre-training with hard labels and fine-tuning with soft-label probability distributions, allows our transformer-based encoder model to effectively capture information from both types of labels. This approach enables us to leverage both hard and soft target labels, leading to better performance on our objective task.

### 6.2 Future Work

In the two-fold training strategy, we are currently using DistilBERT tokenizer to normalize the hard labels and pretrained DistilBERT transformer model is being used as the first part of training process. As the future scope we are proposing to experiment with other pre-trained models such as BERT, RoBERTa, GPT-2, ELECTRA and so on for better performance.

Apart from the models, we can leverage more context or information from the datasets. For example, the ArMIS dataset contains annotations for Misogyny and sexism detection in Arabic tweets and information about their gender (male or female) and political belief (moderate, liberal, and conservative). This data can be incorporated in the model training to improve the model's understanding further. Another potential direction for future research is to explore the effectiveness of the DualCL framework in semi-supervised and weakly-supervised learning scenarios. [Reference] Dual contrastive learning (DualCL) is a framework that simultaneously learns the features of input samples and the parameters of classifiers in the same space. In semi-supervised learning, where only a small fraction of labeled data is available, DualCL can leverage the augmented samples to learn more robust and discriminative features from both labeled and unlabeled data. In weakly-supervised learning, where only partial or noisy labels are given, DualCL can help to alleviate the impact of label noise by incorporating the augmented samples as a form of regularization.

## 7 Code Repository

All the contributions have been compiled and pushed onto this repository. **https://github.com/tanmayj000/544-nlp**

We have shared links for the contrastive learning approach in the README file. Please refer to that. The SRC folder consists of the code for baseline models

## 8 Individual Contributions

Everyone in the group had equal contributions to the project. Specifics of it are as follows:

**Tanmay Jain**: Implemented contrastive learning based model and using DistilBert as base model. Worked on the evaluation of the trained contrastive models. Dataset Preprocessing, Implemented MTL + Soft Loss method, Manuscript Writing.

**Shanay Ghag**: Implemented contrastive learning based model and Dual CONT model using DistilBert as base model and experimented with different contrastive loss functions. Implemented BERT + Soft Loss method for baseline, Manuscript writing.

**Venkat Sumanth Reddy Vadde**: Dataset Preprocessing, Implemented MTL + Soft Loss method, Manuscript Writing. Worked on the evaluation of the trained contrastive models.

**Dipali Telavane**: Researched different types of losses, Implemented Gaussian Processes method for baseline models and fine-tuning of the model, Manuscript writing. Worked on evaluation of trained contrastive models.

**Aditi Hande**: Dataset preprocessing, Implemented BERT + Soft Loss method for baseline, Identified Risks, Challenges and plan to mitigate, Manuscript Writing, fine-tuning and evaluation of contrastive models.

## References

Lora Aroyo and Christopher Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Eyal Beigman and Beata Beigman Klebanov. 2009. Learning with annotation noise. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 280–287, Suntec, Singapore. Association for Computational Linguistics.

A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

E Jayakiran Reddy, CNV Sridhar, and V Pandu Rangadu. 2015. Knowledge based engineering: notion, approaches and future trends. *American Journal of Intelligent Systems*, 5(1):1–17.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 614–622, New York, NY, USA. Association for Computing Machinery.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.