Uncertain (Mis)Takes at LeWiDi-2025: Modeling Human Label Variation With Semantic Entropy

Ieva Raminta Staliūnaitė
University of Cambridge
irs38@cam.ac.uk

Andreas Vlachos
University of Cambridge
av308@cam.ac.uk

Abstract

The VariErrNLI task requires detecting the degree to which each Natural Language Inference (NLI) label is acceptable to a group of annotators. This paper presents an approach to VariErrNLI which incorporates measures of uncertainty, namely Semantic Entropy (SE), to model human label variation. Our method is based on the assumption that if two labels are plausible alternatives, then their explanations must be non-contradictory. We measure SE over Large Language Model (LLM)-generated explanations for a given NLI label, which represents the model uncertainty over the semantic space of possible explanations for that label. The system employs SE scores combined with an encoding of the inputs and generated explanations, and reaches a 0.31 Manhattan distance score on the test set, ranking joint first in the soft evaluation of VariErrNLI.[1]

1 Introduction

Annotator disagreement has recently received more attention in NLP research (Fornaciari et al., 2021; Leonardelli et al., 2021; Sandri et al., 2023). Human label variation has consequences for the data, modeling, and evaluation in ML tasks (Plank, 2022). The question of data quality is related to distinguishing legitimate human label variation, which stems from different interpretations or opinions, from errors. In the context where a single label is correct, the problem of determining annotation reliability has been addressed by Hovy et al. (2013), who propose to evaluate the trustworthiness of each annotator in predicting the correct label. Allowing human label variation adds a layer of difficulty to determining whether annotations are valid, since every combination of labels may be correct. Some work has used the difference between annotator entropy and model entropy to predict which instances may require more annotations in an active learning setup (Baumler et al., 2023).

In this work we propose to solve the VariErrNLI task with Uncertainty Quantification (UQ), specifically Semantic Entropy (SE) (Farquhar et al., 2024). This approach expands on the work by Baumler et al. (2023) by including the semantics of the input as well as sampled Large Language Model (LLM)-generated explanations, and applies it to predicting the soft labels themselves rather than quantifying additional annotation needs. SE has mostly been employed to detect hallucinations (Farquhar et al., 2024), where a prediction with a high SE is interpreted as likely to have been hallucinated, given that the model is uncertain over the semantic space of the output. This is in line with prior work on UQ, which focuses on model calibration (Gupta et al., 2006) and detecting noisy training data (Northcutt et al., 2021). Staliūnaitė et al. (2025) propose to use uncertainty metrics such as similarity-sensitive entropy (Cheng and Vlachos, 2024) for detecting bias in machine translation, by leveraging the fact that uncertainty can also arise from ambiguity (Baan et al., 2024). Models should be uncertain not only when they are not apt, but also when the input is ambiguous, where uncertainty is caused by more than one output being acceptable.

2 Task Summary

Weber-Genzel et al. (2024) introduced the Natural Language Inference (NLI)-inspired task VariErrNLI, which contains both 1) valid annotator disagreement and 2) annotation errors. The dataset builds on the ChaosNLI (Nie et al., 2020) dataset, which is composed of NLI items with soft labels. A subset of ChaosNLI instances is annotated from scratch in two rounds by Weber-Genzel et al. (2024), with four annotators providing initial NLI labels, and then returning to evaluate their own as well as their peers' judgments in a second round.

[1]The code is available at https://github.com/ieva-raminta/uncertain-mis-takes

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 256–262

Annotations that are self-corrected are interpreted as errors and are not included in the gold label sets.

VariErrNLI is one of the tasks in the LeWiDi shared task (Leonardelli et al., 2025). In this paper we discuss a system that solves VariErrNLI with soft label prediction. That is, for an instance of VariErrNLI, we predict the acceptance rate of each label after the second round of annotation. This creates a multilabel binary classification setup with soft targets, where the score for each label reflects the proportion of annotators who accepted it. The example below illustrates an instance where after the second round of annotations, half the annotators believe that the entailment label is appropriate, three quarters of the annotators accept Neutral as a valid label, and none support the Contradiction label:

Context: "The next year, he built himself a palace, Iolani, which can still be toured in Honolulu."
Statement: "Lolani was built in only 1 year."
Labels: Entailment: 0.5, Neutral: 0.75, Contradiction: 0.0

In the shared task, systems are evaluated with soft labels, measuring how well the predicted label distribution matches the acceptability ratings of the different possible interpretations for each instance, as introduced by Uma et al. (2022). Specifically, Manhattan distance is used to measure the difference between the predicted and target distributions, which has been shown to be particularly reliable for binary classification (Rizzi et al., 2024).

3 System Overview

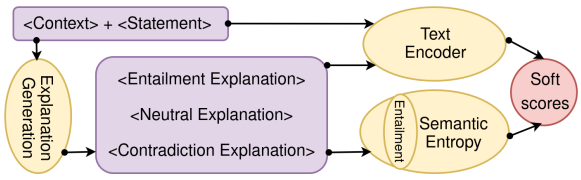

Figure 1: System pipeline: 1. An Explanation Generation (llama3-8B) model generates explanations for each combination of <Context>, <Statement>, and one of Entailment, Neutral, Contradiction labels; 2. Semantic Entropy is calculated for each set of explanations for a given instance using an Entailment model (finetuned bart-large-nli); 3. A Text Encoder (bart-large-nli) is used to embed the combination of <Context>, <Statement> and explanations for each label; 4. Soft scores are predicted from the SE and Text Encoder outputs.

The goal of our system is to be able to quantify the ambiguity in the NLI instances. We postulate that if an instance is ambiguous, the explanations for different labels are likely to not entail one another. For instance, in the example from Section 2, an explanation for the Entailment label could read *"The context states that he built himself a palace next year, which means that he finished it within the year"*, whereas the Neutral label could be explained with *"He may have started to build the palace the next year, but we do not know when he finished it"*. These explanations do not entail each other, which is indicative of ambiguity in the instance. In contrast, explanations for an instance which has only one valid interpretation should only have explanations which entail one another.

Thus, we build a pipeline that uses SE over the explanations for different labels, with the goal of representing the ambiguity of an instance. Predictive Entropy (PE) for an input x is calculated by taking the Shannon entropy of the model's predicted probability distribution over labels. SE is an extension of PE, which is calculated by sampling multiple model outputs, clustering them into sets of sequences of tokens that express the same meaning, and measuring the entropy between the clusters c (Farquhar et al., 2024):

$$\mathrm{SE}(x) = -\sum_{c} P(c \mid x) \log P(c \mid x) \quad (1)$$

Clustering is performed in such a way that any two samples are attributed to the same cluster if and only if they entail one another. The SE of model predictions is higher for instances where more than one interpretation is valid, as more contradictory generated explanations are likely to appear.

We combine the outputs of SE with an embedding of the input and generated explanations for each label. Figure 1 illustrates the full pipeline.

4 Experimental Setup

4.1 Models

This section describes the models used in each component of the pipeline.

Explanation Generation. First, to generate the explanations for each NLI label, we use llama3-8B (AI@Meta, 2024), chosen for its balance of efficiency and reasoning capabilities. The instructions for the model are as follows: *"You are an NLI*

assistant. Given a statement, context, and a judgment label (Entailment, Neutral, or Contradiction), explain why the label is appropriate.\n\n <*Examples*>\n\n *Now consider the following example:*\n *Statement:* <*Statement*>\n *Context:* <*Context*>\n *Judgment: Contradiction*\n *Explanation:".* <Examples> contains a 6-shot list of instances with explanations, two for each label.[2]

Text Encoder. Second, for embedding the inputs along with the generated explanations, we use bart-large-nli (Lewis et al., 2019). bart-large-nli is fine-tuned on the NLI task, which is highly relevant to the task we are solving, namely predicting soft scores for each NLI label.

Entailment. Third, the calculation of Semantic Entropy over the LLM-generated explanations requires an entailment model for the clustering step. We use bart-large-nli for this step as well. However, we further finetune bart-large-nli on the gold explanations in the VariErrNLI dataset. The data preprocessing for this step is described in Section 4.2. The LLM-generated explanations in our system pipeline are clustered by the finetuned bart-large-nli model to calculate SE.

Semantic Entropy. We follow the implementation of Semantic Entropy by Farquhar et al. (2024). We sample 128 generated explanations for all three NLI labels, cluster them together if and only if they entail each other, and calculate SE over the clusters obtained. We calculate seven SE scores, corresponding to each member of the powerset of NLI labels: ((Entailment), (Neutral), (Contradiction), (Entailment, Neutral), (Entailment, Contradiction), (Neutral, Contradiction), and (Entailment, Neutral, Contradiction)). This formulation allows us to isolate the contribution of each label to the total semantic uncertainty by comparing entropy values across subsets.

4.2 Data

For training the Text Encoder we use the full ChaosNLI dataset (Nie et al., 2020) and generate explanations for each label with an LLM.

For training the Entailment model for clustering explanations in SE calculation, we use the gold explanations from VariErrNLI dataset. Each data point is a set of two explanations from a single instance. We assume that two explanations have the Entailment relation if they explain the same label, and that two explanations have a Neutral relationship if they explain different labels but the annotators accept each others' judgments, and finally that two explanations are Contradictory if the annotators reject each others' judgments.

4.3 Configurations

Table 1 presents the different configuration values that we experiment with. We model the task as either a classification or regression task. In the regression setup we directly predict the probability of a given label being accepted, whereas in the classification task we either predict one of seven real values for each label: (0.0, 0.25, 0.33, 0.5, 0.66, 0.75, 1.0) or predict one of 20 combinations of real values which sum to one: ((0.0, 0.0, 1.0), (0.0, 0.25, 0.75), (0.0, 1.0, 0.0), (0.25, 0.0, 0.75), etc).[3] The classes cover the observed soft label distributions.

For the real-valued prediction setup we use either KL divergence or MSE loss, while for classification we use Cross Entropy loss, and we also experiment with a cross label loss function that incorporates dependencies between labels in multi-label classification (Ferreira and Vlachos, 2019).

Hyperparameter	**Values**
Learning Objective	classification, multilabel classification, regression
Dropout	0.1, 0.3, 0.5
Loss Function	Cross Entropy, Cross Label, KL Divergence, MSE
Learning Rate	1e-1 to 1e-5
Weight Decay	1e-2 to 1e-6
Unfrozen Layers	0, 1, 2, 3
Scheduler	step LR, cosine, linear, reduce on plateau
SE embedding size	8, 16
fusion layer size	128, 256
feature combination method	concatenation, fusion, fusion MLP
Entropy Penalty (β)	0, 0.05, 0.1
Temperature Annealing	1.0, 1.5, 2.0
Regularise Against Mean (λ)	0, 0.05, 0.1
Sum < 1 Penalty (γ)	0, 0.05, 0.1

Table 1: Search space for hyperparameter values, regularisation terms, and other model specifications.

We explore several regularisation methods in order to ensure that the predicted scores do not diverge from the targets. To begin with, in initial runs we observed rather similar predictions for instances where they should differ, and thus

[2]Please see Appendix A for the full list of examples.

[3]Please see Appendix B for the full list of the most common combinations of the binary soft labels.

experiment with (1) entropy penalty, which encourages the model to generate more diverse outputs by penalising low entropy (Grandvalet and Bengio, 2004) and (2) temperature annealing (Kirkpatrick et al., 1983; Hinton et al., 2015). Similarly, with many scores appearing close to the mean distribution of the target values, we add a (3) regularisation against the mean (Szegedy et al., 2016; Pereyra et al., 2017). Finally, in order to ensure that the sum of the predicted scores is no lower than one, we add a (4) penalty to the loss whenever the sum of the three scores is below one. All the penalties are applied to the loss, except for the temperature annealing, which is directly applied to the logits.

$$\mathcal{L}_{\text{entropy}} = \beta \cdot \sum_{i=1}^{N} \sum_{j=1}^{C} p_{ij} \log p_{ij} \tag{2}$$

$$p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \tag{3}$$

$$\mathcal{L}_{\text{mean}} = \lambda \cdot \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{p}_i - \bar{\mathbf{p}}\|_2 \tag{4}$$

$$\mathcal{L}_{\text{sum}} = \gamma \cdot \sum_{i=1}^{N} \max(0, 1 - \sum_{j=1}^{C} p_{ij}) \tag{5}$$

We use three different methods to combine the text embeddings with SE information. The first one is straightforward concatenation. The second one is a fusion model, where both representations are projected onto the same shape and summed with weights that are learned during training. The third one is a fusion Multilayer Perceptron (MLP), where the representations are first concatenated, followed by an MLP layer that learns non-linear interactions between the text and entropy modalities.

5 Results

Our best score on the test set is 0.31 Manhattan distance (lower is better), which is ranked number one in the LeWiDi VariErrNLI task (soft evaluation). It is substantially better than the most frequent baseline score of 0.59, and is only surpassed by a system that reaches 0.23, however the difference is not statistically significant. The configuration that led to the best score of our system is described in Appendix C.

We assess the contribution of each component of our pipeline by running an ablation study and excluding one of: the Semantic Entropy features

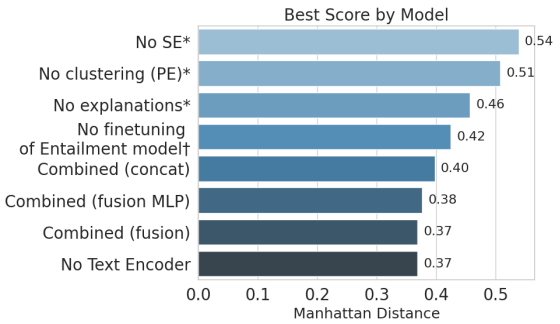

Figure 2: Ablation study results on the development set. Statistical significance between each model and the next best score is marked with * ($p < 0.05$ for all three labels), and † ($p < 0.05$ for at least one of the labels).

altogether, the clustering step for SE, the finetuning of the Entailment model, the complete Text Encoder output or the generated explanations. The results on the development set provide us with additional insights into the workings of the system (see Figure 2). We find that the SE component contributes the most to the performance of the model, as the performance drops from 0.37 to 0.54 in Manhattan distance without it. Furthermore, we find that the generated explanations as well as finetuning of the Entailment model are also instrumental in our pipeline. In addition, we find that the methods for incorporating different types of input do not significantly impact the outcomes. We further discover that the best result on the development set is achieved by completely excluding Text Encoding features. However, this SE-only model does not yield the best score on the test set, which we interpret as an indication that the model overfits.

6 Conclusion

This work presents an approach to soft label NLI, which proves to yield competitive results. The ablation study shows that SE is the most important module of the system, highlighting its versatility beyond hallucination detection and signal for further annotation needs. In future work this approach could be more specifically applied to detecting annotation errors by learning the different Semantic Entropy patterns associated with annotations that are incompatible with valid interpretations. The proposed method can further be applied to other tasks that include generation and ambiguity.

Limitations

The main limitation of this study is the requirement of an LLM for the explanation generation step. First, generating multiple explanations and calculating SE involves sampling and clustering steps that are computationally expensive, which may limit scalability or real-time applicability in practical settings. Second, our method relies on the quality of the explanations generated by the LLMs or alternatively on human generated explanations, which is labour-intensive.

Acknowledgments

Ieva Raminta Staliūnaitė is supported by Huawei. Andreas Vlachos is funded by the European Union's Horizon 2020 Research and Innovation programme grant AVeriTeC (Grant agreement No. 865958) and a grant from Translated.

References

AI@Meta. 2024. Llama 3 model card.

Joris Baan, Raquel Fernández, Barbara Plank, and Wilker Aziz. 2024. Interpreting predictive probabilities: Model confidence or human label variation? In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 268–277, St. Julian's, Malta. Association for Computational Linguistics.

Connor Baumler, Anna Sotnikova, and Hal Daumé III. 2023. Which examples should be multiply annotated? active learning when annotators may disagree. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10352–10371, Toronto, Canada. Association for Computational Linguistics.

Julius Cheng and Andreas Vlachos. 2024. Measuring uncertainty in neural machine translation with similarity-sensitive entropy. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2115–2128, St. Julian's, Malta. Association for Computational Linguistics.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

William Ferreira and Andreas Vlachos. 2019. Incorporating label dependencies in multilabel stance detection. 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 6350–6354, Hong Kong, China. Association for Computational Linguistics.

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.

Yves Grandvalet and Yoshua Bengio. 2004. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press.

Hoshin V Gupta, Keith J Beven, and Thorsten Wagener. 2006. Model calibration and uncertainty estimation. *Encyclopedia of hydrological sciences*.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science*, 220(4598):671–680.

Elisa Leonardelli, Silvia Casola, Siyao Peng, Giulia Rizzi, Valerio Basile, Elisabetta Fersini, Diego Frassinelli, Hyewon Jang, Maja Pavlovic, Barbara Plank, and Massimo Poesio. 2025. Lewidi-2025 at nlperspectives: third edition of the learning with disagreements shared task. In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP (NLPerspectives)*. Association for Computational Linguistics.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural*

Language Processing (EMNLP), pages 9131–9143, Online. Association for Computational Linguistics.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024. Soft metrics for evaluation with disagreements: an assessment. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 84–94, Torino, Italia. ELRA and ICCL.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Ieva Raminta Staliūnaitė, Julius Cheng, and Andreas Vlachos. 2025. Uncertainty quantification for evaluating machine translation bias. *Preprint*, arXiv:2507.18338.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. *J. Artif. Int. Res.*, 72:1385–1470.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

A Examples for 6-shot Setup

The following is a list of six examples, two for each label, and their corresponding explanations:

"Statement: Everything can be found inside a shopping mall." "Context: Enter the realm of shopping malls, where everything you're looking for is available without moving your car." "Judgment: Entailment" "Explanation: The context implies that the shopping mall has everything one might look for, as it can be found without moving your car."

"Statement: The matter of whether or not the Mass is a sacrifice for the remission of sins is controversial." "Context: As for the divisive issue of whether the Mass is a sacrifice for the remission of sins, the statement affirms that Christ's death upon the cross ..." "Judgment: Entailment" "Explanation: The context states that the Mass being a sacrifice for the remission of sins is divisive, which can be interpreted as a synonym for controversial."

"Statement: Most rock concerts take place in the Sultan's Pool amphitheatre." "Context: In the summer, the Sultan's Pool, a vast outdoor amphitheatre, stages rock concerts or other big-name events." "Judgment: Neutral" "Explanation: The context does not specify whether it is most or only some rock concerts that are staged in the Sultan's Pool."

"Statement: This information was developed thanks to extra federal funding." "Context: Additional information is provided to help managers incorporate the standards into their daily operations." "Judgment: Neutral" "Explanation: The context does not indicate where the information came from, which may or may not be federal funding."

"Statement: He had recently seen pictures depicting those things." "Context: He hadn't seen even pictures of such things since the few silent movies run in some of the little art theaters." "Judgment: Contradiction" "Explanation: If the pronoun 'he' and the object 'those things' refer to the same things in the statement and the context, then the statement negates the context. "

"Statement: Octavius Decatur Gass refers to four people. " "Context: One opportunist who stayed was Octavius Decatur Gass. " "Judgment: Contradiction" "Explanation: The context names one person as Octavius Decatur Gass, and does not mention additional referrents. "

B Score Combinations

Table 2 presents the most common combinations of binary soft labels.

Combination	Count
((1.0,0.0),(1.0,0.0),(0.0,1.0))	80
((1.0,0.0),(0.75,0.25),(0.25,0.75))	40
((1.0,0.0),(0.0,1.0),(1.0,0.0))	36
((0.75,0.25),(1.0,0.0),(0.25,0.75))	33
((1.0,0.0),(0.5,0.5),(0.5,0.5))	32
((0.0,1.0),(1.0,0.0),(1.0,0.0))	26
((0.5,0.5),(1.0,0.0),(0.5,0.5))	24
((1.0,0.0),(0.25,0.75),(0.75,0.25))	17
((0.25,0.75),(1.0,0.0),(0.75,0.25))	15
((1.0,0.0),(0.33,0.67),(0.67,0.33))	12
((0.75,0.25),(0.75,0.25),(0.5,0.5))	10
((0.67,0.33),(1.0,0.0),(0.33,0.67))	7
((0.75,0.25),(0.25,0.75),(1.0,0.0))	7
((1.0,0.0),(0.67,0.33),(0.33,0.67))	6
((0.75,0.25),(0.5,0.5),(0.75,0.25))	6
((0.5,0.5),(0.75,0.25),(0.75,0.25))	6
((0.25,0.75),(0.75,0.25),(1.0,0.0))	5
((0.5,0.5),(0.5,0.5),(1.0,0.0))	5
((0.33,0.67),(1.0,0.0),(0.67,0.33))	3
((0.33,0.67),(0.67,0.33),(1.0,0.0))	2

Table 2: Frequency of label distribution combinations

C Best Configuration

The best performing variant of our system had the following configuration: a multilabel classification task with seven classes for each label, cross-label loss, embedding dimension of 16 for the entropy module, using fusion MLP to combine the text and entropy features in a layer of size 256, dropout of 0.1, learning rate of 1e-5, weight decay of 1e-6, step LR scheduler, all embedding layers frozen, no regularisation against mean penalty, entropy penalty (β = 0.05), no temperature annealing, no sum < 1 penalty.