Tuebingen at SemEval-2025 Task 10: Class Weighting, External Knowledge and Data Augmentation in BERT Models

Özlem Karabulut, Ali Gharaee, Soudabeh Eslami, Matthew Kirk Andrews
University of Tübingen Tübingen, Germany
name.surname@student.uni-tuebingen.de

Abstract

The spread of disinformation and propaganda in online news presents a significant challenge to information integrity. As part of SemEval 2025 Task 10 on Multilingual Characterization and Extraction of Narratives from Online News, this study focuses on Subtask 1: Entity Framing, which involves assigning roles to named entities within news articles across multiple languages.

We investigate techniques such as data augmentation, external knowledge integration, and class weighting to improve classification performance. Our findings indicate that data augmentation was more effective than other approaches.

1 Introduction

The internet has opened new ways for communication but has also made consumers more vulnerable to misleading content and manipulation (SemEval2025-Task-10 (2025)). Recognizing propaganda strategies is crucial for combating disinformation, particularly in media analysis, politics, and online discussions. The SemEval 2025 Multilingual Characterization and Extraction of Narratives from Online News Task aims to automate the identification and classification of narratives, assisting analysts in addressing disinformation. This task comprises three subtasks: entity characterization, narrative classification, and narrative extraction. It is available in five languages: Bulgarian, English, Hindi, Portuguese, and Russian. More information can be found in the Task Description Document (Jakub Piskorski (2025)).

This paper discusses our experimental approach in Subtask 1. We evaluated several transformer-based models with minimal hyperparameter tuning. We experimented with data augmentation, external knowledge integration, and class weighting to improve performance.

Our model performed better than the baseline, ranking 21st out of 32 participants in the final submission. However, our results from the development set were better, suggesting that our models were likely overfitting and leading to overly optimistic results. Our code and setup are available at: https://github.com/cicl-iscl/SemEval25-Task10.

2 Background

The task covers news articles from two domains: the Ukraine- Russia War and Climate Change (SemEval2025-Task-10 (2025)). In this paper, we focus on Subtask 1, Entity-Framing. The goal is to assign one main role and one or more sub-roles to a pre-identified Named Entity (NE) in a news article, using a fine-grained entity role taxonomy (Stefanovitch et al. (2025)). The task is formulated as a multi-label, multi-class text-span classification problem and does not require Named Entity Recognition (NER) (Marrero et al. (2013)). An example of a system response to a news article is provided in Appendix A.

2.1 Related Work

Detecting frames in news articles has been a challenging task. Foundational studies by Card et al. (2015) and Boydstun et al. (2018) developed annotations for framed articles. A supervised approach by Naderi and Hirst (2017) applied deep neural networks to classify sentence-level frames using the Media Frames Corpus.

Our task builds on previous SemEval media analysis tasks. Da San Martino et al. (2020) in SemEval 2020 Task 11 focused on detecting propaganda techniques, where transformer-based models and ensembles performed well, particularly with contextual information. Similarly, Piskorski et al. (2023) in SemEval 2023 Task 3 addressed news categorization, framing, and persuasion techniques across nine languages. The multilingual aspect of their task aligns closely with our study. Our research extends these previous works by exploring various approaches to similar challenges.

2.2 Dataset

The input data comprises news and web articles. Details on the gold label and submission format are in Appendix B.
We used a training set, a development set without annotations to train and evaluate our models, and a test set for final submission. Initially, we trained our models on an English dataset with 328 training set articles. Thus, the results in Sections 4.1 and 4.2, as well as the class distribution in Table 2, are derived from this dataset. We later expanded the dataset through data augmentation on the multilingual dataset. The final data set, shown in Table 1, is used to augment the training data detailed in Section 3.5.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1248–1254

Language	Training Set	Development Set
English	399	27
Bulgarian	401	15
Hindi	366	35
Portuguese	400	31
Russian	133	28
Total	1,699	136

Table 1: Final dataset distribution (used in data augmentation)

The official evaluation metric is the exact match ratio, a metric that ignores partially correct results by considering them incorrect:

$$MR = \frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i)$$

where I is the indicator function (Sorower (2010)).

3 System Overview and Experimental Setup

This study evaluated transformer-based models with minimal hyperparameter tuning. We adopted a non-hierarchical approach, first classifying sub-labels and then assigning the main label within the algorithm. We employed BERT-family models, specifically RoBERTa (Liu et al. (2019)) and DistilBERT (Sanh et al. (2019)).

3.1 Model Training

Data Preprocessing . The raw data were restructured to align with the gold label and submission format. The English training data were split into 80% training and 20% validation. In data augmentation, this ratio was maintained while the dataset was modified. To preserve label distributions, we applied iterative stratification during the split.

Hyperparameter Tuning . Key hyperparameters tuned for optimal model performance included a batch size of 8 per device, an epoch count of 25 (based on epoch-based performance evaluation), and an overridden loss function parameter for class weighting.

Model Training . We fine-tuned models `distilbert-base-uncased` and `roberta-base` for multi-label classification using Hugging Face's API (HuggingFace, 2025a) on the tokenized data. Threshold was reduced from 30% to 20% to improve the recall of underrepresented labels, ensuring that entities with multiple roles were correctly classified. Also, binary cross-entropy loss was used to optimize multi-label predictions to handle overlapping labels.

3.2 Multi-label Classification

A sigmoid activation is applied to logits and threshold probabilities to generate binary predictions.

Evaluation Metrics . The system's training performance was evaluated using Exact Match Ratio, Hamming Loss, and F1-Score (Murat Arat (2020)).

3.3 Challenges and Experiments

The key challenge was underrepresentation (Chakraborty and Dey (2024)) of specific classes, namely the class imbalance. The class frequencies in the English dataset are presented as in the "Before" column of Table 2.

Most machine learning methods struggle with imbalanced datasets as they tend to favor majority-class samples, leading to lower accuracy for the minority class. There are two main approaches to address this problem (Chakraborty and Dey (2024)):

- **Algorithm Level Approach:** Class Weighting Algorithms
- **Data Level Approach:** Data Augmentation and External Knowledge

3.4 Class Weighting

We explored two weighting strategies: scikit-learn[1] and logarithmic weighting.

3.4.1 Pre-defined Class Weighting With Scikit Library

Many algorithms in scikit-learn support class weight adjustments (Chakraborty and Dey (2024)). We applied class weighting using `scikit-learn's compute_class_weight`[2] function, assigning weights inversely proportional to class frequencies. The computed weights are integrated into PyTorch's `BCEWithLogitsLoss` by extending HuggingFace's `Trainer`(HuggingFace (2025b)) class via modifying the `compute_loss` function to ensure that misclassifications in minority classes receive higher penalties.

3.4.2 Logarithmic Weighting

An alternative approach computes class weights using a logarithmic transformation relative to the most frequent class:

$$\text{class\_weights} = \log\left(1 + \frac{\max\_count}{\text{label\_counts} + \epsilon}\right)$$

where ϵ ensures numerical stability. This method smooths extreme weight differences, preventing biasing rare classes. The weights are applied via the `pos_weight` parameter in `BCEWithLogitsLoss`, offering a more gradual adjustment than traditional frequency-based weighting.

3.5 Data Augmentation

We leveraged data augmentation to enhance data diversity and model performance. Data augmentation is a general term used to increase robustness and accuracy

[1] https://scikit-learn.org/stable/whats_new/v1.2.html

[2] https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html

by allowing them to perform well on small, poorly representative data, according to (Mumuni and Mumuni (2022)). Specifically, we applied back translation and pivot translation to generate additional training samples for minority and moderate classes. The augmented dataset was then used for the models we mentioned earlier.

3.5.1 Dataset

Augmented data for minority classes (fewer than 200 instances) was stored separately, containing only back-translated examples and not initially merged with the original dataset. For moderate classes (200–500 instances), augmentation involved direct concatenation with the original data. This augmented dataset comprised 7,991 stratified instances (6,390 for training, 1,601 for validation). Validation data was incorporated into training to mitigate overfitting. The latest model we submitted was trained on 12,450 multilingual samples with a 20% validation split (10,731 training, 2,694 validation). The final sublabel distributions are detailed in Tables 2

Role	Before	After	
	Train	Train	Validation
Instigator	49	792	198
Guardian	40	822	206
Conspirator	38	565	141
Incompetent	35	658	165
Foreign Adversary	35	874	219
Victim	33	886	221
Tyrant	29	690	173
Deceiver	26	570	143
Saboteur	20	347	87
Virtuous	19	644	161
Corrupt	17	662	165
Peacemaker	15	474	118
Terrorist	14	558	139
Underdog	12	485	121
Rebel	11	566	142
Martyr	11	407	102
Bigot	9	375	94
Traitor	8	397	99
Scapegoat	8	428	107
Exploited	6	314	78
Spy	3	472	118
Forgotten	1	464	116

Table 2: Comparison of sub-label distributions before augmentation (train) and after augmentation (train and validation)

3.5.2 Dataset Preprocessing

The dataset augmentation pipeline follows these steps:

Back Translation and Pivot Translation: Sentences were translated to an intermediary language and back to the original language to introduce variability. The backtranslated texts are stored in a directory as separate files, one column including the modified text.

Entity Preservation: Placeholder tokens (`_[ENTITY_]`) were replaced with entity mentions in their respective languages. If the entity mention was altered during translation, contextual modifications were applied.

Duplicate Removal and Data Cleaning: Excessive augmentation and duplicate entries were removed.

Dataset Splitting and Stratifying

3.5.3 Model Training

The dataset was used in training models, notably the BERT family. Previous experimentation with the T5 (Raffel et al. (2019)) model on data showed minimal performance improvements.

3.6 External Knowledge

Incorporating external knowledge enhances model performance by providing additional context beyond the training data, improving generalization, robustness, and accuracy.((Jegierski and Saganowski (2019))). The key sources of external knowledge include:

- **Wikipedia / Wikidata**: Offers entity and factual knowledge, enriching the model's understanding of entities and relationships.
- **ConceptNet**: Provides commonsense knowledge and relationships between concepts, improving contextual relevance(Speer et al. (2017)).
- **Domain-Specific Knowledge Graphs**: Specialized databases such as medical, legal, or scientific knowledge graphs contribute domain-specific insights.

3.6.1 Implementation

In our system, the process of integrating external knowledge followed these key steps:

Data Preprocessing: We extracted relevant knowledge on entities from Wikidata and merged it with unstructured data.

Feature Engineering: We created knowledge-aware embeddings and augmented input representations with external data to enhance model features.

Model Training: We used our input data with DistilBERT.

4 Results

4.1 Initial Results

Table 3 summarizes the training evaluation results for DistilBERT and RoBERTa. Both models showed promising results, and RoBERTa significantly outperformed DistilBERT, also demonstrating a more consistent decline with better stability and lower validation loss.

Model	Exact Match Ratio	Hamming Loss	F1-score
DistilBERT	0.186	0.046	0.173
RoBERTa	0.300	0.041	0.256

Table 3: Training performance evaluation for different models

Threshold Adjustment Effect: Adjusting the classification threshold from 30% to 20% improved the Exact Match Ratio by 6% for DistilBERT, demonstrating the importance of dynamic threshold optimization.

4.2 Results for Class Weighting

4.2.1 RoBERTa

The following results belong to RoBERTa with different class-weighting methods on the initial English dataset consisting of 328 samples. The model is trained for 25 epochs. Table 4 summarizes the training evaluation results, and Table 5 illustrates the submission results.

Method	Exact Match Ratio	Hamming Loss	F1-score
Built-in Weighting	0.2093	0.0761	0.2693
Logaritmic Weighting	0.3139	0.0618	0.3113

Table 4: Training performance evaluation for different class-weighting methods

Method	EMR[1]	Micro P	Micro R	Micro F1	Acc[2]
Built-in weighting	0.10990	0.17480	0.18000	0.17730	0.67030
Logarithmic Weighting[3]	0.15384	0.96842	0.19000	0.98385	0.70329

Table 5: Submission results on development set for different class-weighting methods
[1] Exact Match Ratio, the official metric in evaluation, [2] Accuracy in main roles
[3] Not the official submission results, these are calculated by our code after the official leaderboard closed

The results of our experiments after the closing of the official submission leaderboard are presented in Table 16 in Appendix C. Surprisingly, this method performed better than our official submission results, which may be due to the reasons mentioned in Section 5.

4.2.2 DistilBERT

All the experiments on DistilBERT has been made after the closing of the official submission leaderboard. Tables 13 and 14 present the results in Appendix C.

4.3 Results for Data Augmentation

Table 6 summarizes the training evaluation results for DistilBERT and XLM-RoBERTa (Conneau et al. (2020)). Although our officially submitted model DistilBERT performed better in model training evaluation, XLM-RoBERTa outperformed DistilBERT in post-submission test set results. (See Table 7 and Table 15 in Appendix C).

Model	Exact Match Ratio	Hamming Loss	F1-score
DistilBERT	0.7112	0.0240	0.8157
XLM-RoBERTa	0.6911	0.0247	0.8189

Table 6: Training performance evaluation with augmented multilingual data

Prediction Set	EMR	Micro P	Micro R	Micro F1	Acc
Test set	0.13190	0.20580	0.21510	0.21030	0.74040
Development set	0.21980	0.29290	0.29000	0.29150	0.75820

Table 7: Submission result on different datasets with DistilBERT trained on augmented data

4.4 Results for External Knowledge

The training logs show a significant drop in training loss from 0.8024 to 0.0008 by Epochs 7-8, while validation loss decreases initially but then peaks at around 1.0679, indicating overfitting (Table 8). Despite this, the model achieves a validation accuracy of 0.8036 and a weighted score of 0.7885, and an exact match ratio of 0.8036 on the validation set, demonstrating some capability in identifying key features of the classification task (Table 9).
However, these metrics are based on our training performance evaluation, so they cannot be directly compared with the performance metrics of other methods, as those come from different code. Also, it cannot be compared to official submission results because this experiment has been submitted for official evaluation only once, which showed relatively poor results, though it did notably outperform other models in main role accuracy (Table 10).

Epoch	Training Loss	Validation Loss
1	0.802400	0.970969
2	0.534900	0.541840
3	0.311900	0.605355
4	0.074800	0.765731
5	0.073900	0.947722
6	0.001700	1.031157
7	0.000800	1.054733
8	0.000800	1.067893

Table 8: Training and Validation Loss per Epoch

Model	Exact Match Ratio	Hamming Loss	F1-score
DistilBERT	0.8036	0.1964	0.7457

Table 9: Evaluation performance metrics

Prediction Set	EMR	Micro P	Micro R	Micro F1	Acc
Development set[4]	0.05490	0.06590	0.0600	0.06280	0.80220

Table 10: Submission result on development dataset with DistilBERT

However, the increasing validation loss suggests poor generalization due to the small training dataset. To mitigate overfitting, strategies like regularization and early stopping are recommended, and incorporating external knowledge sources could improve the model's generalization ability by enhancing data representation.

5 Conclusion

Our results demonstrate the effectiveness of various approaches for entity role classification in multilingual

data. The Exact Match Ratio shows that while our models performed competitively, there is room for improvement. Before the closing of the official submission leaderboard, augmented data combined with RoBERTa achieved the highest EMR (0.13190). Data augmentation significantly enhanced performance for underrepresented classes, as seen in F1 scores. However, its failure to boost the EMR could be caused by back-translation-label noise, loss of contextual integrity, or inconsistencies in entity role assignments. Especially thinking our best results came from an intermediate data augmentation, in which we did not "over-augment" the data. However, a key oversight in our process was that we did not record the submission results at various stages. In short, the complexity of this task may require a more refined augmentation technique to reduce noisy data.

Another observation is that the validation results were likely inflated due to the similarity of training and validation data, leading to overly optimistic performance estimates. Test results revealed that our models were likely overfitted.

Also, the results of external knowledge also indicate potential overfitting, shown by high main role accuracy but low metrics. Other factors could be differences in dataset distribution, improved feature representation, or random initialization effects. Further examination with additional models is needed.

Beyond these, we explored prompting techniques at the entry level, which showed some promise in improving accuracy. However, due to time constraints, we could not further investigate prompting.

In closing, classifying with imbalanced datasets remains crucial. Future research should explore alternative models, such as GPT-based architectures, prompting or ensemble learning techniques (Jia et al. (2024)), as they combine multiple models to leverage their strengths.

Acknowledgments

We would like to express our heartfelt gratitude to Dr. Çağrı Çöltekin for his continuous guidance, insightful feedback, and support throughout this task.

References

Amber E. Boydstun, Dallas Card, Justin Gross, Paul Resnick, and Noah A. Smith. 2018. Tracking the Development of Media Frames within and across Policy Issues.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Sanjay Chakraborty and Lopamudra Dey. 2024. *Class Imbalance and Data Irregularities in Classification*, pages 23–49. Springer Nature Singapore, Singapore.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

HuggingFace. 2025a. HuggingFace Transformers Documentation.

HuggingFace. 2025b. HuggingFace Transformers Documentation.

Nikolaos Nikolaidis Ricardo Campos Alípio Jorge Dimitar Dimitrov Purificação Silvano Roman Yangarber Shivam Sharma Tanmoy Chakraborty Nuno Ricardo Guimarães Elisa Sartori Nicolas Stefanovitch Zhuohan Xie Preslav Nakov Giovanni Da San Martino Jakub Piskorski, Tarek Mahmoud. 2025. SemEval-2025 task 10: Multilingual characterization and extraction of narratives from online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation*, SemEval 2025, Vienna, Austria.

Hubert Jegierski and Stanislaw Saganowski. 2019. An "outside the box" solution for imbalanced data classification. *CoRR*, abs/1911.06965.

Jianguo Jia, Wen Liang, and Youzhi Liang. 2024. A review of hybrid and ensemble in deep learning for natural language processing. *Preprint*, arXiv:2312.05589.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís. 2013. Named Entity Recognition: Fallacies, challenges and opportunities. *Computer Standards Interfaces*, 35(5):482–489.

Alhassan Mumuni and Fuseini Mumuni. 2022. Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258.

Mehmet Murat Arat. 2020. Multi-label classification metrics. Accessed: 2025-02-27.

Nona Naderi and Graeme Hirst. 2017. Classifying frames at the sentence level in news articles. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 536–542, Varna, Bulgaria. INCOMA Ltd.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

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

SemEval2025-Task-10. 2025. Multilingual characterization and extraction of narratives from online news.

Mohammad S. Sorower. 2010. A literature survey on algorithms for multi-label learning. *A Literature Survey on Algorithms for Multi-label Learning*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Nicolas Stefanovitch, Tarek Mahmoud, Nikolaos Nikolaidis, Jorge Alípio, Ricardo Campos, Dimitar Dimitrov, Purificação Silvano, Shivam Sharma, Roman Yangarber, Nuno Guimarães, Elisa Sartori, Ana Filipa Pacheco, Cecília Ortiz, Cláudia Couto, Glória Reis de Oliveira, Ari Gonçalves, Ivan Koychev, Ivo Moravski, Nicolo Faggiani, Sopho Kharazi, Bonka Kotseva, Ion Androutsopoulos, John Pavlopoulos, Gayatri Oke, Kanupriya Pathak, Dhairya Suman, Sohini Mazumdar, Tanmoy Chakraborty, Zhuohan Xie, Denis Kvachev, Irina Gatsuk, Ksenia Semenova, Matilda Villanen, Aamos Waher, Daria Lyakhnovich, Giovanni Da San Martino, Preslav Nakov, and Jakub Piskorski. 2025. Multilingual Characterization and Extraction of Narratives from Online News: Annotation Guidelines. Technical Report JRC141322, European Commission Joint Research Centre, Ispra (Italy).

A Appendix: Example of Entity Roles for a given Article

Below is an example of how a system processes a news article for Subtask 1:

Met Office Should Put 2.5°C 'Uncertainties' Warning on All Future Temperature Claims

"It is "abundantly clear" that the Met Office cannot scientifically claim to know the current average temperature of the U.K. to a hundredth of a degree centigrade, given that it is using data that has a margin of error of up to 2.5°C, notes the climate journalist Paul Homewood. His comments follow recent disclosures in the Daily Sceptic that nearly eight out of ten of the Met's 380 measuring stations come with official 'uncertainties' of between 2-5°C. In addition, given the poor siting of the stations now and possibly in the past, the Met Office has no means of knowing whether it is comparing like with like when it publishes temperature trends going back to 1884.

There are five classes of measuring stations identified by the World Meteorological Office (WMO). Classes 4 and 5 come with uncertainties of 2°C and 5°C respectively and account for an astonishing 77% of the Met Office station total. Class 3 has an uncertainty rating of 1°C and accounts for another 8.4% of the total. The Class ratings identify potential corruptions in recordings caused by both human and natural involvement. Homewood calculates that the average uncertainty across the entire database is 2.5°C. In the graph below, he then calculates the range of annual U.K. temperatures going back to 2010 incorporating the margins of error.

The blue blocks show the annual temperature announced by the Met Office, while the red bars take account of the WMO uncertainties. It is highly unlikely that the red bars show the more accurate temperature, and there is much evidence to suggest temperatures are nearer the blue trend. But the point of the exercise is to note that the Met Office, in the interests of scientific exactitude, should disclose what could be large measurement inaccuracies. This is particularly important when it is making highly politicised statements using rising temperatures to promote the Net Zero fantasy. As Homewood observes, the Met Office "cannot say with any degree of scientific certainty that the last two years were the warmest on record, nor quantify how much, if any, the climate has warmed since 1884".

The U.K. figures are of course an important component of the Met Office's global temperature dataset known as HadCRUT. As we noted recently, there is ongoing concern about the accuracy of HadCRUT with large retrospective adjustments of warming in recent times and cooling further back in the record. In fact, this concern has been ongoing for some time. The late Christopher Booker was a great champion of climate scepticism and in February 2015 he suggested that the "fiddling" with temperature data "is the biggest science scandal ever". Writing in the Telegraph, he noted: "When future generations look back on the global warming scare of the past 30 years, nothing will shock them more than the extent to which official temperatures records – on which the entire panic rested – were systematically 'adjusted' to show the Earth as having warmed more than the actual data justified."

A.1 Entity Role Classification

This example illustrates how our system classifies entities and assigns their roles:

Entity	Role(s)
Met Office	Antagonist-[Deceiver]
Paul Homewood	Protagonist-[Guardian]
Daily Sceptic	Protagonist-[Guardian]
Christopher Booker	Protagonist-[Guardian, Virtuous]

Table 11: Entity roles assigned by the system for the given example.

B Gold Labels and Submission Format

B.1 Subtask 1 - Entity Framing

The format of a tab-separated line of the gold label and the submission files for Subtask 1 is:

article_id	entity_mention	start_offset	end_offset	main_role	fine_grained_roles
EN_10001.txt	Martin Luther King Jr.	10	32	Protagonist	Martyr
EN_10002.txt	Mahatma Gandhi	12	27	Protagonist	Martyr, Rebel
EN_10003.txt	ISIS	4	8	Antagonist	Terrorist, Deceiver

Table 12: Partial view of a gold label file for Subtask 1

The columns are defined as follows:

- **article_id**: The file name of the input article.
- **entity_mention**: The string representing the entity mention.
- **start_offset** and **end_offset**: Start and end position of the mention.
- **main_role**: A string representing the main entity role.
- **fine_grained_roles**: A tab-separated list of strings representing the fine-grained role(s).

Important Notes:

- For creating the submission file, a list of all entity mentions and their corresponding offsets for all the articles will be provided.
- The leaderboard evaluates predictions for both *main_role* and *fine_grained_roles*, but the official evaluation metric is based on the *fine_grained_roles*.
- *main_role* should take only one of three values from the 1st level of the taxonomy.
- *fine_grained_roles* should take one or more values from the 2nd level of the taxonomy.
- If you do not train a model to predict *main_role*, you must still provide a valid value under *main_role* to pass the format checker in the scorer.

C Post-Submission Results

DistilBERT All the experiments on DistilBERT has been made after the closing of the official submission. Tables 13 and 14 present the results.

Method	Exact Match Ratio	Hamming Loss	F1-score
Built-in Weighting	0.2906	0.0692	0.2410
Logaritmic Weighting	0.3372	0.0634	0.2931

Table 13: Training performance evaluation for different class-weighting methods

Method	EMR	Micro P	Micro R	Micro F1	Acc
Built-in weighting	0.12770	0.18110	0.16600	0.17320	0.77870
Logarithmic Weighting	0.10640	0.13930	0.12830	0.13360	0.66380

Table 14: Submission results on test set for different class-weighting

Prediction Set	EMR	Micro P	Micro R	Micro F1	Acc
Test set	0.15320	0.24420	0.27920	0.26060	0.77870
Development set[4]	-	-	-	-	-

Table 15: Submission result on different datasets with XLM-RoBERTa trained on augmented data

[4] The results of this table were submitted after the closing of the official leaderboard. Therefore, we do not have access to the development set results.

Method	EMR	Micro P	Micro R	Micro F1	Acc
Built-in weighting	0.15740	0.21340	0.19250	0.20240	0.79570
Logarithmic Weighting	0.17020	0.26340	0.26040	0.26190	0.75740

Table 16: Submission results on test set for different class-weighting for XLM-RoBERTa