# UMUTeam at SemEval-2023 Task 11: Ensemble Learning applied to Binary Supervised Classifiers with disagreements

**José Antonio García-Díaz**[1], **Ronghao Pan**[1], **Gema Alcaráz-Mármol**[2], **María José Marín-Pérez**[3], **Rafael Valencia-García**[1]
[1] Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain
{joseantonio.garcia8, ronghao.pan, valencia}@um.es
[2] Departamento de Filología Moderna, Universidad de Castilla La Mancha
Facultad de Educación, Avenida Carlos III, 45071
Gema.Alcaraz@uclm.es
[3]Facultad de Letras, Universidad de Murcia, Campus de la Merced, 30001, Murcia, Spain
{angelalm, mariajose.marin1}@um.es

## Abstract

This paper describes the participation of the UMUTeam in the Learning With Disagreements (Le-Wi-Di) shared task proposed at SemEval 2023, which objective is the development of supervised automatic classifiers that consider, during training, the agreements and disagreements among the annotators of the datasets. Specifically, this edition includes a multilingual dataset. Our proposal is grounded on the development of ensemble learning classifiers that combine the outputs of several Large Language Models. Our proposal ranked position 18 of a total of 30 participants. However, our proposal did not incorporate the information about the disagreements. In contrast, we compare the performance of building several classifiers for each dataset separately with a merged dataset.

## 1 Introduction

This manuscript summarizes the participation of the UMUTeam in the second edition of the Learning With Disagreements (Le-Wi-Di) shared task (Leonardellli et al., 2023), proposed in SemEval 2023. Le-Wi-Di focuses on the development of supervised automatic classifiers in which the training process considers the agreement and disagreement of the annotators. For this purpose, in this edition of Le-Wi-Di the organizers have released a multilingual dataset composed of four normalized datasets focused on entirely subjective tasks. According to the organizers, one of the underlying objectives of this task is to develop a unified testing framework for learning from disagreement.

Our proposal to solve the Le-Wi-Di shared task is based on the training of several deep learning classifiers and then combine their outputs using ensemble learning. To this end, we evaluate several Large Language Models (LLMs), and we also consider the usage of Linguistic Features (LFs) from UMUTextStats (García-Díaz et al., 2022c). Besides, we evaluate the training of a unique classifier for all datasets, or the strategy of training supervised classifiers specialized in each of the datasets. However, our major limitation is that we could not incorporate to the training of our models the information about disagreement.

Our participation outperformed the baseline and other participants, even though it does not incorporate disagreement. Accordingly, our proposal could be a more real baseline than the one proposed by the organizers of the task.

In addition, we have created a GitHub[1] repository with some resources concerning our participation.

## 2 Background

One of the key challenges of this edition of Le-Wi-Di is that it is not focused on a specific topic. Instead of this, the organizers of the task proposed different datasets, all based on binary classifications. The four datasets involved in this task are: i) the HS-brexit dataset (Akhtar et al., 2021), focused on hate-speech and abusive language on Brexit. This dataset was annotated by six annotators that were a target group of three Muslim immigrants in the UK, and a control group of three other persons; ii) the ArMIS dataset (Almanea and Poesio, 2022), composed of tweets written in Arabic, and annotated for misogyny detection. In this case, the annotators have different demographic traits; iii)

[1] https://github.com/NLP-UMUTeam/semeval-2023-lewidi



*Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1061–1066

the ConvAbuse dataset (Curry et al., 2021), composed by 4,185 dialogues written in English. These dialogues took place between users and two conversational agents. This dataset was annotated by three or more experts in gender studies; and iv) the MultiDomain Agreement dataset, composed by 10k English tweets concerning BLM[2], elections, and Covid-19. In this case, the dataset contains annotations concerning offensiveness performed by 5 annotators. It is worth noting that this dataset was annotated with disagreement in mind. For this, almost 1/3 of the dataset has then been annotated with a 2 vs 3 annotators disagreements, and a third one of the dataset has an agreement of 1 vs 4.

For each dataset, we reserve from the training split a 20% of the documents for performing a custom validation before the final submission. The dataset statistics are depicted in Table 1. All datasets are arranged as binary classification problems. As we deal with datasets for different topics, we renamed all labels to `true` and `false`. It can be observed that there are strong differences among the datasets. For instance, ConvAbuse and MD-Agreement are much larger than ArMIS and Brexit. Another relevant difference is the balance, as Brexit dataset contains very few positive labels.

## 3 System Overview

The system architecture proposed by our research group for solving the Le-Wi-Di task is depicted in Figure 1. In a nutshell, it can be described as follows. First, we clean all datasets by stripping hyperlinks, mentions and other similar stuff. Second, we use a dataset splitter that returns all the proposed datasets merged or separated. This is because we evaluate two different strategies. The first one is the training of a model from all the datasets combined. This strategy is like a multi-task training. The second one is the training of separate classifiers focused on each one of the specific datasets proposed. The third stage of our pipeline is the fine-tuning of three LLMs based on BERT and RoBERTa architectures. Once fine-tuned, we extract the sentence embeddings from each document and LLM for the next step. It is worth mentioning that we evaluate another feature set based on LFs, but we eventually decided to remove it from our pipeline because it downplayed the overall results. Fourth, we combine all the sentence embeddings from the fine-tuned models of each LLM into a unique classifier. For this, we evaluate, on the one hand, three ensemble learning based on different averaging strategies and, on the other, a Knowledge Integration (KI) strategy that consists in training a new classifier from the fine-tuned sentence embeddings from each model. Finally, according to the results of a custom validation split, we decided our final models for the official submission.

For the data cleaning stage, as some of the documents of all datasets are written in different languages, we apply a simple cleaning pipeline that removed hyperlinks, hashtags and mentions from the datasets.

The evaluated feature sets are described below:

- LFs from UMUTextStats (García-Díaz et al., 2022c). UMUTextStats is a similar tool as LIWC (Tausczik and Pennebaker, 2010) but designed from scratch for the Spanish Language. UMUTextStats considers several linguistic categories and more than 300 features. Some of these features are language-independent and they provided promising results separately or combined with embeddings from different LLMs (García-Díaz

Table 1: Dataset statistics

| ArMIS | | | | |
|---|---|---|---|---|
| | train | val | test | total |
| false | 387 | 84 | 83 | 554 |
| true | 270 | 57 | 61 | 388 |
| total | 657 | 141 | 144 | 942 |
| **Brexit** | | | | |
| false | 712 | 149 | 150 | 1011 |
| true | 72 | 19 | 18 | 109 |
| total | 784 | 168 | 168 | 1120 |
| **MD-Agreement** | | | | |
| | train | val | test | total |
| false | 4630 | 716 | 2039 | 7385 |
| true | 1962 | 388 | 1018 | 3368 |
| total | 6592 | 1104 | 3057 | 10753 |
| **ConvAbuse** | | | | |
| | train | val | test | total |
| false | 2009 | 679 | ? | 2688 |
| true | 389 | 133 | ? | 522 |
| total | 2398 | 812 | 840 | 4050 |

[2]Black Lives Matters

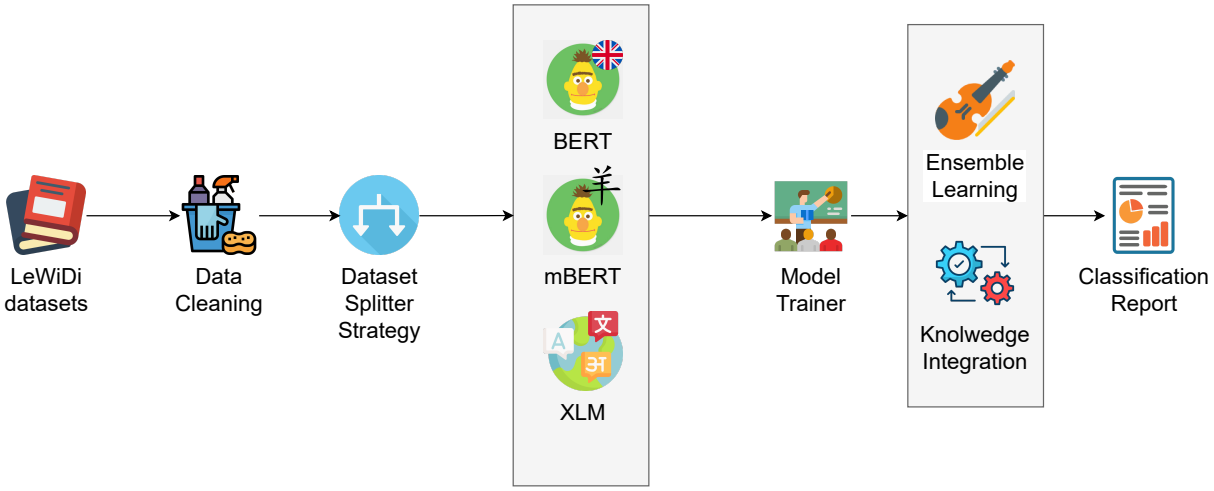


Figure 1: System architecture.

et al., 2022a; García-Díaz and Valencia-García, 2022). Besides, these features have been used in different tasks, related to the datasets from Le-Wi-Di, such as hate-speech identification (García-Díaz et al., 2022b).

- BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018). It is one of the first LLMs, developed by Google. BERT is a pretrained model that learns contextual word-embeddings; that is, embeddings that have different representation for a word, according to its surrounding words.

- multilingual BERT (Devlin et al., 2018). This is the multi-lingual version of BERT. We included this version of BERT as not all the documents in the dataset are written in English.

- XLM (Lample and Conneau, 2019). It is a multilingual LLM that has learned the embeddings using two tasks. On the one hand, it uses an unsupervised task concerning monolingual data and, on the other hand, it uses a supervised task that uses parallel data.

Next, we train a supervised classifier based on the LFs and we fine-tuned all the LLMs that have been evaluated. Once fine-tuned, we extract the sentence embeddings from the `[CLS]` token, in a similar fashion as Sentence BERT does(Reimers and Gurevych, 2019). The fine-tuned process is conducted by a hyperparameter optimization stage, which consisted in the training of 10 models in which the weight decay, the batch size, the warm-up speed, the number of epochs, and the learning rate vary using the Tree of Parzen Estimators (TPE) (Bergstra et al., 2013).

Next, we combine the outputs of each model using two strategies. The first strategy is based on KI, which consists in training a new model from sentence embeddings for each LLM. This new model is trained using Keras. The second strategy is the usage of ensemble learning, which consists in combining the predictions of probabilities output of each model separately to generate a more robust option. The ensemble learning strategies evaluated are:

- Highest probability (HIGHEST): Consists into output the highest probability of each evaluated model.

- Average of probabilities (MEAN): Consists into averaging the probabilities of each evaluated model.

- Hard voting (MODE): Consists into calculate the mode of the labels of each evaluated model.

## 4 Experimental Setup

First, we report our results with the custom validation split. As stated before, we evaluate two main strategies. Our first strategy consisted in training a unique model from all the datasets combined into a larger one. The results of these models with our custom validation set are reported in Table 2. It can be observed that the Knowledge Integration strategy achieves the overall better result, with an f1-score of 82.066%. However, all LLM and their

combination using Knowledge Integration and Ensemble Learning achieve similar results. It is only the LF which offers limited results, which suggests that stylometric and morphosyntax features are not enough to classify subjective information in which datasets concerning different topics are mixed.

Table 2: Classification report of all datasets merged with the custom validation split. The classification reports use macro-weighted measures.

| model | precision | recall | f1-score |
|---|---|---|---|
| LF | 36.584 | 50.000 | 42.253 |
| BERT | 80.414 | 78.863 | 79.574 |
| mBERT | 80.509 | 78.598 | 79.456 |
| XLM | 80.672 | 81.247 | 80.950 |
| EL (HIGHEST) | 81.794 | 81.072 | 81.419 |
| EL (MEAN) | 82.037 | 80.625 | 81.281 |
| EL (MODE) | 82.255 | 80.907 | 81.536 |
| **KI** | **82.429** | **81.728** | **82.066** |

Next, we report the results with our custom validation split concerning the training of supervised classifiers for each model separately. These results are depicted in Table 3. Due to paper-length restrictions, we only report the best model by dataset. It is worth noting that different models achieved similar results. In these cases, we have selected the simplest model. For the ArMIS dataset, the best result is achieved by a LLM separately: XLM. For the HS-Brexit, the best result is achieved by LF. However, this high result is due to the small number of positive instances in this dataset, as we have 149 negative instances and only 19 positive instances. For the MD-Agreement and the ConvAbuse datasets, the best results are achieved by combination of the features, using Knowledge Integration of MD-Agreement and Ensemble learning based on the mode for ConvAbuse.

## 5 Results

The evaluation of the Le-Wi-Di shared task is based on two metrics. The first metric is a soft metric, and it is the main evaluating metric. This metric considers how well the probability of each binary model reflects the level of agreement among annotators. The second metric is a hard metric, and it is the F1-score.

The participants were allowed to send a total of 5 runs. However, me commit a mistake generating the first runs and run out of time. Because of this, we could only send to the official leader board 3 runs, based on the ensemble learning models. We decided to send the results of the best supervised classifier trained from each dataset separately because we considered a better strategy based on the previous results.

From the results posted in the Codalab platform, we achieved a F1-score of 0.8176 using the ensemble learning strategy of the mode of the predictions, a F1-score of 0.8176 with the highest probability strategy, and a F1-score of 0.8303 averaging all the probabilities. We leave the latest submission as our official submission.

Table 4 depicts the official results on the leader board. Our team ranked position 18 from a total of 30 participants (including the organizers baseline), with an average cross entropy of 18.75. Our results are tied with the *Sana* team. The best score is achieved by the team *Duxy*, with a cross-entropy of 1.75

As expected, our results considering the F1-score metric are better. In fact, we achieve the position number 13 in the leader board. Our results considering the F1-score metric for each dataset are reported in Table 5. Our best result (position 11) is achieved for the HS-Brexit dataset and our most limited results (position 16) is with the ArMIS dataset.

Next, our results by dataset are reported in Table 6. It can be seen that our best result is achieved with the ConvAbuse dataset (position 16) and our most limited result with the MD-Agreement dataset (position 23).

## 6 Conclusion

In these working notes, we have described the participation of the UMUTeam in the Le-Wi-Di shared task, in which we have ranked position 18 from a total of 29 participants plus a baseline. Our participation is grounded on the development of several ensemble learning classification models based on three LLMs and LFs. We are happy with our participation, but we are aware that we achieved modest results. Besides, we could not have included information concerning the agreement among the annotators.

The differences in our ranking among the soft and hard metrics give value to the importance of considering the disagreements of the annotators in machine-learning experiments, as we achieve

Table 3: Classification report of all datasets evaluated separately with the custom validation split. Only the best feature set is reported for each dataset.

| dataset | model | precision | recall | f1-score |
|---|---|---|---|---|
| ArMIS | XLM | 74.266 | 73.810 | 74.000 |
| HS-Brexit | LF | 79.704 | 76.934 | 78.222 |
| MD-Agreement | KI | 80.202 | 79.835 | 80.010 |
| ConvAbuse | EL (MODE) | 92.567 | 89.947 | 91.189 |

Table 4: Official leaderboard, with the average across the Cross-Entropy ranking position in all datasets

| Rank | Team | Cross Entropy |
|---|---|---|
| 01 | Duxy | 1.75 |
| 02 | chencheng498 | 2.50 |
| 03 | king001 | 2.75 |
| 04 | PALI | 3.00 |
| 05 | Colain | 4.25 |
| 06 | stce | 6.75 |
| 07 | ymf924 | 7.25 |
| 08 | mjs227 | 9.00 |
| 09 | sadat1971 | 9.75 |
| 10 | sananc | 10.50 |
| 11 | EwelineCogSci | 11.00 |
| 12 | PeymanHosseini | 12.75 |
| 13 | CICL_DMS | 13.00 |
| 14 | ccasula | 14.75 |
| 15 | nasheedyasin | 17.00 |
| 16 | Nitrogen_pump | 18.00 |
| 17 | robvanderg | 18.50 |
| **18** | **UMUTEAM** | **18.75** |
| 18 | Sana | 18.75 |
| 20 | NiVi | 19.00 |
| 21 | xiacui | 19.50 |
| 22 | corner | 20.25 |
| 23 | babysong | 22.00 |
| 24 | Ankita | 22.25 |
| 25 | AlessandroAstorino | 22.50 |
| 25 | omaimah | 22.50 |
| 25 | rana1998al_essa | 22.50 |
| 28 | ruyuanwan | 24.00 |
| 29 | guneetsk99 | 24.25 |
| 30 | **BASELINE** | 24.50 |
| 30 | morlikowski | 24.50 |

much more better ranking considering hard metrics. As further work, we will try to incorporate this knowledge to our pipeline.

Table 5: Official leaderboard, with the average across the F1-score ranking position in all datasets

| Rank | Dataset | F1-score |
|---|---|---|
| 11 | HS-Brexit | 90.219 |
| 16 | ArMIS | 68.741 |
| 12 | ConvAbuse | 92.458 |
| 13 | MD-Agreement | 81.467 |

Table 6: Official results of the UMUTeam for each dataset separately

| Rank | Dataset | Cross Entropy |
|---|---|---|
| 17 | HS-Brexit | 0.474 |
| 19 | ArMIS | 0.713 |
| 16 | ConvAbuse | 0.302 |
| 23 | MD-Agreement | 0.729 |

## 7 Acknowledgments

This work is part of the research projects AIInFunds (PDC2021-121112-I00) and LT-SWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PSP (PID2019-107652RB-I00/AEI/ 10.13039/501100011033) funded by MCIN/AEI/10.13039/501100011033. In addition, José Antonio García-Díaz is supported by Banco Santander and the University of Murcia through the Doctorado Industrial programme.

## References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv:2106.15896*.

Dina Almanea and Massimo Poesio. 2022. Armis-the arabic misogyny and sexism corpus with annotator

subjective disagreements. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2282–2291.

James Bergstra, Daniel Yamins, and David Cox. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, pages 115–123. PMLR.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. Convabuse: Data, analysis, and benchmarks for nuanced detection in conversational ai. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

José Antonio García-Díaz, Ricardo Colomo-Palacios, and Rafael Valencia-García. 2022a. Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020. *Future Generation Computer Systems*, 130:59–74.

José Antonio García-Díaz, Salud María Jiménez-Zafra, Miguel Angel García-Cumbreras, and Rafael Valencia-García. 2022b. Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers. *Complex & Intelligent Systems*, pages 1–22.

José Antonio García-Díaz and Rafael Valencia-García. 2022. Compilation and evaluation of the spanish saticorpus 2021 for satire identification using linguistic features and transformers. *Complex & Intelligent Systems*, 8(2):1723–1736.

José Antonio García-Díaz, Pedro José Vivancos-Vicente, Angela Almela, and Rafael Valencia-García. 2022c. Umutextstats: A linguistic feature extraction tool for spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6035–6044.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Elisa Leonardellli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Massimo Poesio, Verena Rieser, and Alexandra Uma. 2023. SemEval-2023 Task 11: Learning With Disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.