Benchmark BERT models for dialog act classification in French

Arthur Guédon
CentraleSupélec
arthur.guedon@student-cs.fr

Thomas Brilland
CentraleSupélec
thomas.brilland@student-cs.fr

Abstract

In this study, we explore the efficacy of fine-tuning pre-trained BERT models for intent classification in French, using the MIAM dataset. We fine-tuned three BERT models, including two that were specifically trained for French (FlauBERT and CamemBERT), and one multilingual BERT model. We evaluated the models' performance using accuracy, F1 score, and confusion matrices. Our findings indicate that all three models perform similarly on the classification task, with CamemBERT exhibiting the most satisfactory results. Although the multilingual BERT model obtained lower scores, it remains a viable option compared to models that are specifically trained on French. Our experience affirms the robustness of BERT models for a range of downstream tasks. Additionally, this paper underscores the suitability of the MIAM dataset for fine-tuning pre-trained models for dialog act classification in various languages, including French.

1 Problem Framing

Dialog act classification is a fundamental task in Natural Language Processing(NLP), which involves identifying the communicative intent behind a speaker's utterances in a conversation [36; 7]. Dialog act classification can provide valuable insights into the structure of conversations and help build more intelligent systems for automated dialogue management, sentiment analysis, and machine translation, among other applications [37; 22; 14; 42; 20; 38; 24; 34].

In recent years, deep learning models have shown promising results in NLP classification tasks [33; 17; 27]. Among these models, Bidirectional Encoder Representations from Transformers (BERT) [23] has emerged as one of the most effective and widely used models for NLP tasks, including dialog act classification. BERT is a pre-trained language model that learns to encode the meaning of words in the context of a sentence, allowing it to capture the semantic nuances of language more effectively than traditional machine learning models.

However, most studies on dialog act classification using BERT have focused on English-language data [21; 4; 9; 39; 16; 12; 35; 5; 8; 19; 18; 11; 6; 10; 25], with little research on the performance of BERT models for dialog act classification in other languages, such as French. The ability to classify dialog acts accurately in French could enable more sophisticated applications of NLP in French-language conversations.

Therefore, the goal of this paper is to evaluate and compare the performance of different BERT models that we fine tune for dialog act classification in French in order to identify which model is more appropriate for this task.

2 Experiments Protocol

Using the definition in [39], let's introduce the concepts. We have dialogues D defined as sequences of contexts (truncated conversations)

$$D = (C_1, C_2, C_3, ..., C_{|D|})$$

Each context is composed of utterances U and can be defined as follows:

$$C_i = (U_1, U_2, ..., U_{|C_i|})$$

For Dialog Act classification, each utterance U_i is associated with a unique DA label y_i.
We decided to benchmark three pre-trained BERT Models and apply them to the DialogAct Benchmark (MIAM) dataset. We then compare the performances of the three different models.

2.1 MIAM dataset

For the dialog act classification task we decided to work with the DialogAct Benchmark (MIAM) dataset [26].It is itself divided in five datasets in five different languages with annotated dialog acts. We work with the French version of the dataset which contains 10.5k rows and 31 different labels of dialog acts. Due to its multilingual specificity this dataset can be used for various applications in dialog act classification in languages other than English. [39] We used the official train, validation and test splits of this dataset which contain respectively 8465, 942 and 1047 labelled utterances.

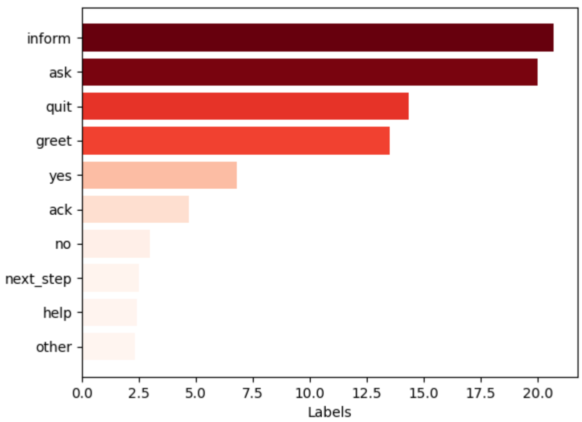

Figure 1: Most frequent labels in the MIAM dataset

Utterance	Dialog Act
Mon métier peut plaire autant aux filles qu'aux garçons.	inform
Voulez vous que je vous décrive un peu mon métier ?	ask
Non, ça va.	no
merci de votre aide, au revoir !	quit
D'accord	yes
Bravo! Vous avez été rapides!	greet
Ok, je vois.	ack

Table 1: Examples of utterances in the MIAM dataset

2.2 FlauBERT

FlauBERT [32] is a French language model trained on a large and heterogeneous French corpus which has the structure of the BERT model. BERT is a type of neural network architecture based on transformers, a self-attention mechanism that enables the model to capture dependencies between different words in a sentence. BERT is pre-trained on a large corpus of text data, such as Wikipedia, using unsupervised learning techniques. This model can be fine-tuned on specific downstream tasks such as dialog act classification.

2.3 CamemBERT

CamemBERT [29] is a French language model based on the RoBERTa (Robustly Optimized BERTPretraining Approach)[28] architecture. RoBERTa is a modified version of BERT, it is based on the transformer architecture and pre-trained on a large corpus of text data using a masked language modeling task similarly to BERT. However, RoBERTa incorporates several key improvements over BERT. One of the main differences between RoBERTa and BERT is the training data since it is pre-trained on a much larger and diverse dataset, including web pages, books, and articles, allowing it to capture a wider range of linguistic patterns and improve its ability to generalize to new tasks and domains.

RoBERTa also uses a different token masking strategy, it randomly masks tokens instead of doing it statically. This forces the model to use all available context to predict the masked tokens, improving its understanding of contextual information.

2.4 mBERT

Multilingual BERT Model (mBERT) reproduces the architecture of the initial BERT model[23] and is trained on a large corpus of over 100 languages. Overall, the Multilingual BERT model has been shown to be highly effective for a range of multilingual NLP tasks and has become a popular choice for researchers and practitioners working with multilingual data. We chose this multilingual model in order to compare it to models that have been trained specifically for French.

2.5 Cross-Entropy Loss

The cross-entropy loss, also known as log loss, is a commonly used loss function in machine learning for classification tasks. It measures the dissimilarity between the predicted class probabilities and the true class labels. Specifically, given a set of n training examples $x_i, y_i{}_{i=1}^{n}$, where x_i is an input feature vector and y_i is a one-hot encoded label vector, the cross-entropy loss is defined as:

$$CE = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{C} y_{i,j} log(p_{i,j})$$

where C is the number of classes, $y_{i,j}$ is the j-th element of the one-hot encoded label vector for the i-th example, and $p_{i,j}$ is the predicted probability of the j-th class for the i-th example.

2.6 Adam Optimizer

The Adam optimizer [15] is a popular stochastic gradient descent (SGD) optimization algorithm that uses adaptive learning rates to update the model parameters during training. It combines the advantages of two other SGD algorithms, namely AdaGrad and RMSProp, by adapting the learning rate based on the first and second moments of the gradients. Specifically, Adam computes an exponential moving average of the gradient and its squared values, and uses these to scale the learning rate for each weight. This results in faster convergence and better generalization performance compared to traditional SGD optimization algorithms. Adam has become a popular choice for optimizing deep neural networks in various machine learning tasks, including natural language processing and computer vision.

3 Results

We used the same process to fine-tune the 3 models. We loaded the pre-trained versions using the huggingface library, and reset the classifier head with the proper number of labels (31 labels for our classification task). We then trained the models for 10 epochs on Google Collab with a batch size of 16. We used an Adam Optimizer with a learning rate of 3e-5, and a CrossEntropy loss function.

We chose to study 2 metrics for the evaluation, the accuracy and the f1-score. While the accuracy measures the proportion of correctly classified samples out of the total number of samples in the dataset, the f1-score is the harmonic mean of precision and recall. Both are common metrics to evaluate classification models.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

Where TP and TN are the number of correctly classified elements by the model, while FP and FN are the utterances that have been incorrectly classified.

We used the same process to fine-tune the 3 models. First, we loaded the pre-trained versions using the huggingface library. We replaced the classifier head with one that has the proper number of labels (31 for our classification task). Then, we trained the models for 10 epochs using Google Collab with a batch size of 16.

We used the Adam Optimizer with a learning rate of 10^{-5}, and a CrossEntropy loss function. The training time was equivalent for the three models (around 2 hours).

We obtained the following results:

Model	Accuracy	F1 score
FlauBERT	0.8806	0.6291
CamemBERT	**0.8825**	**0.6645**
mBERT	0.8691	0.6062

Table 2: Test metrics for the BERT models fine-tuned on MIAM dataset

We can observe that the 3 models have very close accuracy scores. The "best" model in terms of accuracy is CamemBERT, and it is only 0.0134 ahead of the "worst", mBERT. However, the difference is more significant with the F1 score. Indeed, CamemBERT is better than the other two and is 0.06 ahead of mBERT, which is therefore the worse performer for both metrics.

Below is the confusion matrix for the best model, CamemBERT. We only show this one because of how close the three are. Since there are 31 labels, we chose to limit our matrix to the 6 most frequent labels, which are ack, ask, greet, inform, quit, and yes.

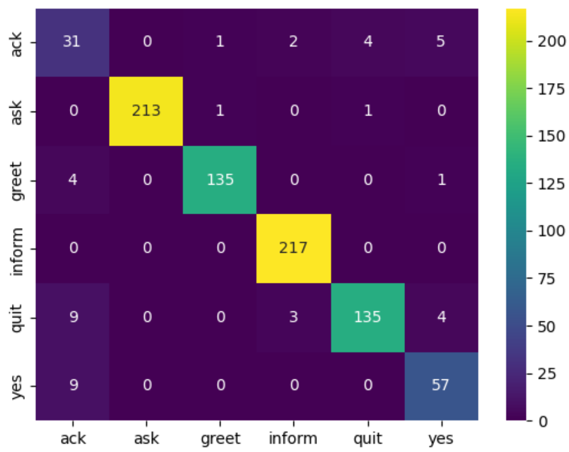

Figure 2: CamemBERT confusion matrix on the 6 most frequent labels

We can note that the confusion matrix is very satisfactory, which confirms the good results in terms of accuracy. While the model is performant for the ask label with precision and recall close to 1, we notice it is underperforming on the 'ack' utterances, which it confuses with the 'yes' and 'quit' labels. The same tendency is observed on FlauBERT and mBERT models. Finally, it is important to mention that the dataset is unbalanced, and there are labels that are almost absent, thus leading to a poor performing model on these labels.

4 Discussion Conclusion

In this study, we investigated the effectiveness of fine-tuning three pre-trained BERT models for intent classification in French using the MIAM dataset. The three models included two that were specifically trained for French (FlauBERT and CamemBERT) and one multilingual BERT model. We evaluated their performance using accuracy, F1 score, and confusion matrices, with results presented in Table 2 and Figure 2. Fine-tuning each model took the same amount of time.

Our results showed that all three models had comparable performance on the classification task, with CamemBERT demonstrating the most satisfactory results. Despite obtaining lower scores, the multilingual BERT model remains a promising option compared to models trained specifically on French. Our study reinforces the robustness of BERT models for a wide range of downstream tasks.

Moreover, this paper highlights the suitability of the MIAM dataset for fine-tuning pre-trained models for dialog act classification in various languages, including French. By demonstrating the efficacy of these models on the MIAM dataset, we contribute to the growing body of research exploring the use of pre-trained models for natural language processing tasks in different languages.

As future research directions, it would be interesting to explore the application of fairness [45; 44; 30; 41] and out-of-distribution (OOD) classifiers [46; 3; 31; 1; 43; 13; 47; 2] to the fine-tuned BERT models. Fairness concerns the avoidance of biases in models, and OOD classifiers can help prevent models from making incorrect predictions on data that is significantly different from the training data. Additionally, future work could investigate the fine-tuning of BERT models on multimodal [40] and diverse datasets to further test their effectiveness on different languages and domains.

.

References

[1] Marine Picot, Nathan Noiry, Pablo Piantanida, and Pierre Colombo. . Adversarial attack detection under realistic constraints.

[2] Marine Picot, Federica Granese, Guillaume Staerman, Marco Romanelli, Francisco Messina, Pablo Piantanida, and Pierre Colombo. . A halfspace-mass depth-based method for adversarial attack detection. *Transactions on Machine Learning Research*.

[3] Marine Picot, Guillaume Staerman, Federica Granese, Nathan Noiry, Francisco Messina, Pablo Piantanida, and Pierre Colombo. . A simple unsupervised data depth-based method to detect adversarial images.

[4] John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, page 517–520, USA. IEEE Computer Society.

[5] Henry Thompson, Anne Anderson, Ellen Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo. 1993. The hcrc map task corpus: natural dialogue for speech recognition.

[6] Daniel Salber and Joëlle Coutaz. 1993. A wizard of oz platform for the study of multimodal systems. In *INTERACT'93 and CHI'93 Conference Companion on Human Factors in Computing Systems*, pages 95–96.

[7] Herbert H Clark and Jean E Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

[8] Geoffrey Leech and Martin Weisser. 2003. Generic speech act annotation for task-oriented dialogues.

[9] Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

[10] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

[11] Gary Mckeown, Michel Valstar, Roddy Cowie, Maja Pantic, and M. Schroder. 2013. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3:5–17.

[12] R. Passonneau and E. Sachar. 2014. Loqui human-human dialogue corpus (transcriptions and annotations).

[13] Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

[14] Ondřej Dušek and Filip Jurčíček. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 45–51, Berlin, Germany. Association for Computational Linguistics.

[15] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

[16] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset.

[17] Wojciech Witon, Pierre Colombo, Ashutosh Modi, and Mubbasir Kapadia. 2018. Disney at IEST 2018: Predicting emotions using an ensemble. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 248–253, Brussels, Belgium. Association for Computational Linguistics.

[18] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

[19] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations.

[20] Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137, Melbourne, Australia. Association for Computational Linguistics.

[21] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Ultes Stefan, Ramadan Osman, and Milica Gašić. 2018. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[22] Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3734–3743, Minneapolis, Minnesota. Association for Computational Linguistics.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

[24] Wei Wei, Jiayi Liu, Xianling Mao, Guibing Guo, Feida Zhu, Pan Zhou, and Yuchong Hu. 2019. Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1401–1410.

[25] Alexandre Garcia, Pierre Colombo, Florence d'Alché Buc, Slim Essid, and Chloé Clavel. 2019. From the token to the review: A hierarchical multimodal approach to opinion mining. 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 5539–5548, Hong Kong, China. Association for Computational Linguistics.

[26] Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 2019. A multilingual and multidomain study on dialog act recognition using character-level tokenization. *Information*, 10:94.

[27] Yazhou Zhang, Qiuchi Li, Dawei Song, Peng Zhang, and Panpan Wang. 2019. Quantum-inspired interactive networks for conversational sentiment analysis.

[28] 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. *arXiv preprint arXiv:1907.11692*.

[29] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suá rez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

[30] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR.

[31] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475.

[32] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french.

[33] Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloé Clavel. 2020. Guiding attention in sequence-to-sequence models for dialogue act prediction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7594–7601. AAAI Press.

[34] Hamid Jalalzai, Pierre Colombo, Chloé Clavel, Eric Gaussier, Giovanna Varni, Emmanuel Vignon, and Anne Sabourin. 2020. Heavy-tailed representations, text polarity classification & data augmentation. In *Advances in Neural Information Processing Systems*, volume 33, pages 4295–4307. Curran Associates, Inc.

[35] Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloé Clavel. 2020. Hierarchical pre-training for sequence labelling in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2636–2648, Online. Association for Computational Linguistics.

[36] Tanvi Dinkar, Pierre Colombo, Matthieu Labeau, and Chloé Clavel. 2020. The importance of fillers for text representations of speech transcripts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7985–7993, Online. Association for Computational Linguistics.

[37] Kai Wang, Junfeng Tian, Rui Wang, Xiaojun Quan, and Jianxing Yu. 2020. Multi-domain dialogue acts and response co-generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7125–7134, Online. Association for Computational Linguistics.

[38] Pierre Colombo, Chloé Clavel, Chouchang Yack, and Giovanna Varni. 2021. Beam search with bidirectional strategies for neural response generation. In *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 139–146, Trento, Italy. Association for Computational Linguistics.

[39] Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloé Clavel. 2021. Code-switched inspired losses for spoken dialog representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8320–8337, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[40] Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloé Clavel. 2021. Improving multimodal fusion via mutual dependency maximisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 231–245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[41] Pierre Colombo. 2021. *Learning to represent and generate text using information measures*. Ph.D. thesis, Ph. D. thesis, Institut polytechnique de Paris.

[42] Pierre Colombo, Pablo Piantanida, and Chloé Clavel. 2021. A novel estimator of mutual information for learning to disentangle textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6539–6550, Online. Association for Computational Linguistics.

[43] Pierre Colombo, Eduardo Dadalto, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. 2022. Beyond mahalanobis distance for textual ood detection. In *Advances in Neural Information Processing Systems*, volume 35, pages 17744–17759. Curran Associates, Inc.

[44] Georg Pichler, Pierre Jean A. Colombo, Malik Boudiaf, Günther Koliander, and Pablo Piantanida. 2022. A differential entropy estimator for training neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17691–17715. PMLR.

[45] Pierre Colombo, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. 2022. Learning disentangled textual representations via statistical measures of similarity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2614–2630, Dublin, Ireland. Association for Computational Linguistics.

[46] Maxime Darrin, Pablo Piantanida, and Pierre Colombo. 2022. Rainproof: An umbrella to shield text generators from out-of-distribution data. *arXiv preprint arXiv:2212.09171*.

[47] Maxime Darrin, Guillaume Staerman, Eduardo Dadalto Câmara Gomes, Jackie CK Cheung, Pablo Piantanida, and Pierre Colombo. 2023. Unsupervised layer-wise score aggregation for textual ood detection. *arXiv preprint arXiv:2302.09852*.