Automatic Economic Activity Classification using ML and DL Techniques

Roi Santos-Rios[1] and Jesús Vilares[1]

Universidade da Coruña, Centro de Investigación CITIC - Grupo LYS, Facultad de Informática, Campus de Elviña, 15071 - A Coruña (España)
roi.santos.rios@udc.es
https://dunque.github.io/

Abstract. Automatic classification of companies according to their economic activity is a key task in the production of official statistics. Within the CIDMEFEO project, developed in collaboration with the Spanish National Statistics Institute (INE), this doctoral research[1] investigates Natural Language Processing techniques for automatically assigning CNAE codes from short company descriptions.
The research addresses challenges such as the hierarchical structure of the CNAE taxonomy, severe class imbalance, and the limited variability of real-world company descriptions. To overcome these issues, the proposed work explores transformer-based models, hierarchical text classification, and the use of Large Language Models for synthetic data generation.
A central objective is to achieve high classification performance while maintaining computational efficiency following the principles of Green AI. Preliminary results show that compact transformer encoders outperform both traditional machine learning methods and the evaluated small language model, while hierarchical approaches improve the structural consistency of predictions. These findings motivate the continued development of efficient hierarchical classifiers and data augmentation strategies for sustainable industrial text classification.

Keywords: Natural Language Processing · Text Classification · Machine Learning · Data Generation · Hierarchical Classification

1 Introduction

Natural Language Processing (NLP) has become a key technology for automatically extracting structured information from textual data. This doctoral research is developed within the CIDMEFEO project, in collaboration with the Spanish National Statistics Institute (INE, from *Instituto Nacional de Estadística*).[2], with the objective of automatically classifying Spanish companies according to the *Clasificación Nacional de Actividades Económicas* (CNAE). Given a short

[1] The thesis started development in September of 2024
[2] https://www.ine.es

textual description of a company's activity, the system must assign one of 629 economic activity codes organized into a four-level hierarchy.

Despite being a text classification problem, automatic CNAE coding presents several challenges. Company descriptions are typically short, ambiguous, and highly imbalanced between classes, while the hierarchical organization of the CNAE taxonomy introduces semantic relationships that conventional flat classifiers often ignore.

Recent transformer-based language models have significantly advanced text classification performance. However, larger models also entail higher computational and deployment costs, which may limit their applicability in public institutions. Consequently, this research investigates whether compact transformer encoders can provide an effective balance between predictive performance and computational efficiency. Large Language Models are therefore explored primarily as supporting tools for tasks such as synthetic data generation rather than as the main classification architecture.

This paper presents the current progress of the doctoral research, including the preliminary work, research objectives, methodology, experimental results, and future research directions.

2 Preliminary Work

The work presented in this paper builds upon the extensive literature on text classification and hierarchical text classification while also incorporating the preliminary research already carried out within the CIDMEFEO project. This section summarizes both the state of the art that motivates the proposed research and the progress achieved during the first stage of the doctoral work.

2.1 Related work

Text classification is one of the fundamental tasks in Natural Language Processing (NLP) and has experienced substantial advances over the last decades. Early approaches relied on traditional machine learning algorithms such as Naive Bayes [8], Support Vector Machines (SVM) [7], and decision trees [10], typically operating over handcrafted features and sparse representations such as bag-of-words or TF-IDF vectors. While these methods proved effective in many practical scenarios, their ability to capture semantic relationships and contextual information remained limited. FastText [1] represented an important step forward by incorporating distributed word representations and subword information, enabling efficient and scalable text classification while improving generalization to unseen words. Nevertheless, its shallow architecture still constrained its capacity to model complex contextual dependencies.

The emergence of deep learning techniques further transformed the field. Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) demonstrated improved performance by capturing sequential and local patterns within text. More recently, transformer-based architectures have become the dominant paradigm for text classification. Models such as BERT [4]

and RoBERTa [9] learn contextualized representations through large-scale self-supervised pretraining, allowing them to effectively model long-range dependencies and adapt to downstream tasks with limited amounts of labeled data. These models have consistently achieved state-of-the-art results across a wide range of classification benchmarks and have become the foundation of most modern NLP systems.

Beyond conventional flat classification, many real-world applications require labels organized according to hierarchical taxonomies. In such settings, hierarchical text classification (HTC) seeks to exploit relationships between classes in order to improve predictive performance and maintain consistency across different levels of the hierarchy [5, 11]. Early HTC approaches adapted traditional machine learning methods to hierarchical label structures through strategies such as Local Classifier per Node (LCPN), Local Classifier per Parent Node (LCPPN), and Local Classifier per Level (LCPL), which decompose the classification task into a set of smaller decisions aligned with the taxonomy structure [2]. These methods demonstrated that explicitly incorporating hierarchical information can reduce classification errors and improve interpretability, particularly in large taxonomies.

More recently, neural approaches have extended these ideas by learning hierarchical representations directly from data. Architectures such as HiAGM [13] model label dependencies through graph-based representations of the hierarchy, while subsequent methods including HTCInfoMax [3], HGCLR [12], and HBGL [6] make use of contrastive learning, information maximization, and hierarchical graph structures to capture semantic relationships between categories. These approaches have consistently shown that exploiting hierarchical information can improve both classification accuracy and hierarchical consistency when compared to flat classification models.

Given that the CNAE standard defines 629 economic activities organized according to a multi-level taxonomy, the problem addressed in this work can naturally be formulated both as a multiclass text classification task and as a hierarchical classification task.

2.2 Current Progress

The first stage of the doctoral research focused on preparing the INE datasets and establishing a common experimental framework. Baseline classifiers, including FastText, BETO, and MrBERT, have been implemented and evaluated, while preliminary experiments have explored Large Language Models for both classification and synthetic data generation. These developments provide the foundation for the remaining stages of the thesis.

3 Research Goals

The objective of this doctoral research is to develop efficient Natural Language Processing techniques for the automatic classification of Spanish economic activities according to the CNAE taxonomy. The proposed research combines high

predictive performance with computational efficiency to produce practical classification systems suitable for deployment within official statistical institutions.

More specifically, the research pursues the following objectives:

- Develop transformer-based models for automatically assigning CNAE codes from short company descriptions.
- Investigate hierarchical classification strategies that exploit the CNAE taxonomy to improve prediction consistency.
- Study data augmentation techniques, including LLM-based synthetic data generation, to mitigate class imbalance.
- Evaluate the trade-off between predictive performance and computational efficiency following the principles of Green AI.
- Assess the proposed approaches using predictive, hierarchical, and efficiency-oriented evaluation metrics.

The thesis aims to advance efficient NLP methods for hierarchical industrial text classification, balancing predictive performance, computational efficiency, and practical applicability.

4 Methodology

The proposed research follows an iterative methodology in which NLP models are progressively developed and evaluated for automatic CNAE classification. The methodology combines dataset preparation, hierarchical classification, data augmentation, and systematic evaluation, while prioritizing computationally efficient solutions under the principles of Green AI.

4.1 Dataset

As mentioned in Section 1, CNAE is the Spanish standard classification system for economic activities. Most of its structure is inherited from the European Community standard, the so-called *Statistical Classification of Economic Activities* (NACE, from its French initials). This project started when the current work version was CNAE-2009,[3], although nowadays the standard is CNAE-2025. Thus, all experiments were performed on CNAE-2009, even though with minimal changes, the developed systems can be adapted for CNAE-2025.

CNAE-2009 is structured in four hierarchical levels, as shown in Figure 1:

1. **Section**: represented by a letter (21 categories).
2. **Division**: represented by 2 digits (88 categories).
3. **Group**: represented by 3 digits, and where the first 2 digits indicate the corresponding division (272 categories).
4. **Class**: represented by 4 digits, and where the first 3 digits indicate the corresponding group (629 categories).

[3] `https://www.ine.es/daco/daco42/clasificaciones/cnae09/notasex_cnae_09.pdf`

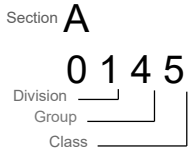


Fig. 1. Example of the hierarchy of a CNAE code.

The regulations state that every company must be associated with a single category on each level. This means that, if the company develops more than one economic activity, the most relevant one must be chosen. This, together with the fact that the descriptions are usually incomplete or ambiguous, and that low-level categories are sometimes complex and full of exceptions, makes it a hard classification problem, even for domain expert humans.

The hierarchical organization of the CNAE taxonomy naturally defines two complementary research scenarios: A multiclass classification problem where each company is assigned directly to one of the 629 activity classes, or the hierarchical dependencies between categories can be explicitly exploited through hierarchical classification models capable of producing coherent predictions across all taxonomy levels. Investigating the advantages and limitations of these two paradigms constitutes one of the main research directions of this doctoral work.

4.2 Preprocessing

The first step consisted of preprocessing the data to obtain a unified and usable format. We merged multiple datasets provided by the INE into a single corpus. We then applied a series of preprocessing and augmentation steps to homogenize the textual descriptions while preserving relevant information. Specifically, all text was lowercased, accents and non-UTF characters were normalized to UTF equivalents (except for the ñ character due to its high frequency in Spanish), and numbers, punctuation, line breaks, and extra whitespace were removed.

We further constructed synthetic descriptions from the titles and explanatory notes of CNAE-2009 categories. This subset was augmented using a domain-specific synonym dictionary employed by INE for manual annotation, replacing words with semantically equivalent alternatives to increase lexical diversity without altering meaning. This process yielded a training set of approximately 3.15 million instances, whose hierarchical distribution is illustrated in Figure 2.

Despite this augmentation, the resulting dataset remains highly imbalanced due to the large number of classes. The most frequent class accounts for 10.1% of the data, whereas the least frequent represents only 0.00013%. As shown in Table 1, a substantial long-tail distribution persists, with 45.63% of classes containing fewer than 1,000 samples. This imbalance reflects the underlying structure of the Spanish economy, where economic activities are inherently unevenly distributed.

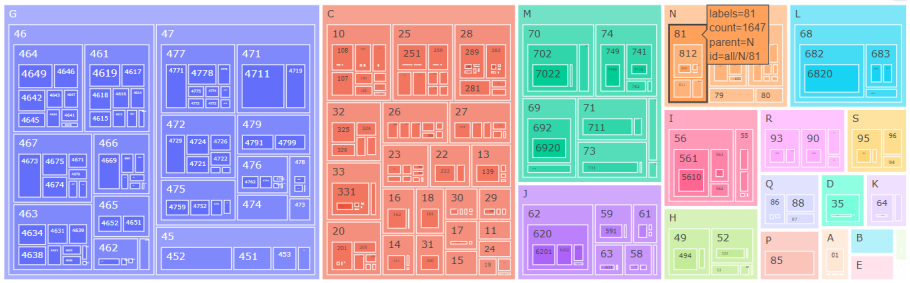


Fig. 2. Treemap showing the hierarchical structure of the dataset. The numbers represent each different class, and the rectangle size is proportional to the amount of data in each category.

# samples per class	# classes	% classes
> 1,000	342	54.37%
> 5,000	125	19.87%
> 10,000	62	9.85%

Table 1. Sample distribution in the resulting working dataset.

The resulting corpus provides a unified experimental framework for the remainder of the thesis. Besides improving the overall quality of the data, the adopted preprocessing pipeline facilitates reproducibility across experiments and enables fair comparisons between alternative machine learning approaches.

4.3 Data Augmentation

The CNAE datasets exhibit a severe class imbalance, limiting the performance of supervised classifiers. To mitigate this issue, the research investigates data augmentation techniques based on both lexical substitution and LLMs. Initially, expert-defined synonym dictionaries provided by INE were used to generate additional training samples. More recently, LLMs have been employed offline to generate realistic company descriptions for underrepresented classes. This approach enriches the training data while maintaining the computational efficiency of the final classifier.

4.4 Classification Models

The proposed research evaluates three families of text classification models: FastText as an efficient baseline, the transformer encoders BETO and MrBERT as the primary classification architectures, and instruction-tuned Large Language Models. While LLMs are evaluated as direct classifiers, they are mainly investigated as tools for synthetic data generation. Comparing these approaches enables

the analysis of the trade-off between predictive performance and computational efficiency.

4.5 Hierarchical Classification

A central hypothesis of this research is that explicitly modelling the hierarchical structure of the CNAE taxonomy improves prediction consistency. To investigate this, hierarchical variants of BETO and MrBERT are developed alongside their flat counterparts.

Hierarchical consistency is enforced during both training and inference. During training, invalid child classes are masked according to the ground-truth parent label, and an alignment loss encourages coherent predictions across hierarchy levels. During inference, predictions are generated in a top-down manner, restricting each decision to the valid children of the previously predicted class. This guarantees taxonomy-consistent predictions while preserving the hierarchical relationships between economic activities.

4.6 Evaluation methods

To gauge the performance of the models, we used a standarized test dataset that INE developed, so all comparisons between models are fair. This set is made of **1,809 samples** distributed among four categories:

- `confident_learning` **(742 samples)**: This subset has been developed using confidence learning algorithms on the `base_data` dataset to extract the most relevant and varied samples.
- `handwritten` **(36 samples)**: This subset contains very few entries that have been hand-written by experts at INE. They specifically target categories that have very similar entries, in order to determine if the models are able to classify samples with subtle differences.
- `queries` **(416 samples)**: This subset is made of real world examples of submitted company activity descriptions. This subset is the most important, as the models' performance over it reflect more closely their theoretical performance when deployed.
- `train_set` **(615 samples)**: This subset is made of examples from `base_data`, thus when the models are trained with either `base_data` or `base_filtered`, they should show solid results. The entries in this subset can be the result of data augmentation, so their quality isn't expected to be the best.

For the multiclass classification setting, we report standard evaluation metrics including accuracy, macro-averaged F1, and weighted F1. Accuracy measures the proportion of correctly predicted labels over all samples. Macro F1 computes the unweighted mean of F1 scores across all classes, treating each class equally regardless of frequency, and is therefore sensitive to performance on minority classes. Weighted F1 similarly averages class-wise F1 scores but weights each

class by its support, reflecting the overall performance under the dataset's natural class distribution.

We evaluate hierarchical models using path-based metrics that measure correctness across all levels of the hierarchy. For each sample, predictions and ground-truth labels are represented as full label paths spanning the four hierarchy levels. Hierarchical precision, recall, and F1 are computed by aggregating the proportion of correctly predicted nodes along these paths across all samples.

5 Current Results

The first stage of the doctoral research focused on establishing a solid experimental framework for automatic CNAE classification. The experiments evaluate different classification models, data augmentation strategies, and hierarchical learning approaches, providing the basis for the subsequent stages of the thesis.

5.1 Baseline Models

The initial experiments compared three representative classification paradigms: FastText as a lightweight baseline, the transformer encoders BETO and MrBERT, and the instruction-tuned LLM Salamandra-2B. This comparison assesses the trade-off between predictive performance and computational efficiency across conventional and generative language models.

Model	Conf Learn		Handwritten		Queries		Train_set		Average	
	Acc	wF1	Acc	wF1	Acc	wF1	Acc	wF1	Acc	wF1
FastText	65.41	65.65	51.61	52.53	45.89	45.63	72.44	73.37	63.13	63.69
BETO	**66.08**	**66.26**	**64.52**	**64.42**	47.10	46.49	**74.20**	**74.57**	**64.51**	**64.96**
MrBERT	65.68	65.91	58.06	61.40	**50.24**	**48.89**	71.96	73.11	64.23	64.94
Salamandra-2B	33.35	31.46	42.23	43.58	24.56	22.73	50.10	48.34	37.14	35.65

Table 2. Performance of FastText, BETO, MrBERT and Salamandra-2B trained with codauto_data against each subsection of the standarized test set. Metrics used are Accuracy and Weighted F1 Score

The results indicate that LLM-generated descriptions cannot fully replace real company descriptions, as models trained exclusively on synthetic data achieve substantially lower performance. However, combining synthetic and real samples consistently improves several experimental configurations. These findings suggest that synthetic data is best used to complement existing datasets, motivating future work on targeted generation strategies for underrepresented classes.

5.2 Data Augmentation

The severe imbalance of the CNAE datasets motivated a second research line focused on data augmentation. Since collecting additional manually labelled company descriptions is both expensive and time-consuming, the thesis investigates whether synthetic data generated through Large Language Models can improve classification performance while reducing the need for manual annotation.

Dataset	Conf Learn		Handwritten		Queries		Train_set		Average	
	Acc	wF1	Acc	wF1	Acc	wF1	Acc	wF1	Acc	wF1
gen-data-qwen3	28.65	30.97	25.81	26.34	19.08	18.54	28.69	31.81	26.42	28.31
gen-data-qwen25	19.59	19.40	16.13	17.10	31.64	30.73	25.16	25.32	24.21	24.12
gen-data	38.51	40.53	25.81	26.88	32.37	32.41	46.31	48.37	39.58	40.99

Table 3. Performance of BETO trained with the newly generated datasets, using qwen 2.5 and qwen 3, and a concatenation of both (gen-data), tested against the test set with no modification. Metrics used are Accuracy and Weighted F1 Score

In addition to generating new entries, we decided to reduce the number of samples in the dominating classes, and find the threshold were their performance starts to decrease. This would make training take less time, and may allow the models to be able to focus more on the minority classes. Thus, we created the `clean_real_data` set.

Then, we combined the generated data with `codauto_data` and `clean_real_data`, leaving us with five working datasets to train our models on:

- `codauto_data` **(3.15M samples)**: This is the dataset explained in detail in Section 4.1. It is severely imbalanced.
- `clean_real_data` **(571K samples)**: Dataset derived from `codauto_data`, but removing all samples that were produced originally by data augmentation techniques. This is to assess the improvements brought by adding them in the first place. The imbalance of the dataset follows a similar distribution to `codauto_data`, albeit with far less data.
- `gen_data` **(131k samples)**: Dataset generated with Qwen 2.5 and 3, containing 270 entries per CNAE class. Its entries consist exclusively of better-written sentences and are generally longer than the vast majority of real-life examples of CNAE descriptions provided by companies.
- `codauto_data_gen` **(3.28M samples)**: Combination of `codauto_data` and `gen_data`. Made to see if there is an improvement on adding even more data, regardless of origin.
- `clean_real_data_gen` **(702K samples)**: Combination of `clean_real_data` and `gen_data`. Made to assess the improvements brought by only augmenting the dataset with LLM-generated data.

The obtained results indicate that synthetic descriptions generated by LLMs contain useful semantic information but cannot fully replace naturally occurring company descriptions. Models trained exclusively on generated data exhibit substantially lower predictive performance, suggesting that synthetic datasets do not capture the linguistic variability present in real-world company activity descriptions.

Nevertheless, combining generated and real data consistently improves several experimental configurations. These findings indicate that synthetic data should be regarded as a complementary resource capable of enriching existing corpora rather than replacing them entirely. This conclusion motivates future research on more targeted and controlled generation strategies specifically designed for minority classes.

5.3 Hierarchical Classification

The final experimental stage completed so far investigates whether explicitly incorporating the hierarchical structure of the CNAE taxonomy improves classification quality. Both BETO and MrBERT were extended with hierarchical learning strategies and evaluated using the datasets described in the previous section.

Across most experimental settings, hierarchical variants achieve comparable flat classification performance while consistently improving hierarchical evaluation metrics, as shown in Table 4. Although the gains in conventional accuracy are generally modest, the improvements in hierarchical F1 demonstrate that explicitly modelling the taxonomy leads to predictions that are structurally more coherent.

These observations support the continuation of hierarchical learning as one of the principal research directions of the thesis. Beyond improving predictive performance, hierarchical classifiers produce outputs that better reflect the semantic organisation of economic activities, making them particularly suitable for practical deployment in official statistical systems.

Acknowledgments

CIDMEFEO funded by the Spanish National Statistics Institute (INE); as well as funding by Xunta de Galicia (ED431C 2024/02), and Centro de Investigación de Galicia "CITIC", funded by the *Xunta de Galicia* through the collaboration agreement between the *Consellería de Cultura, Educación, Formación Profesional e Universidades* and the Galician universities for the reinforcement of the research centres of the Galician University System (CIGUS).

CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref.ED431G 2023/01)

Dataset	Model	Acc	wF1	hF1
Clean Real	FastText	52.68	54.21	63.71
	BETO (MC)	56.99	57.42	67.88
	BETO (H)	57.10	57.76	69.72
	MrBERT (MC)	57.10	58.35	68.09
	MrBERT (H)	57.05	57.47	70.45
Clean Real + Gen	FastText	40.85	40.75	57.35
	BETO (MC)	60.81	61.62	71.19
	BETO (H)	59.37	59.58	72.53
	MrBERT (MC)	60.59	60.88	71.63
	MrBERT (H)	60.75	61.39	72.44
Codauto	FastText	63.13	63.69	72.87
	BETO (MC)	64.51	64.96	74.49
	BETO (H)	64.62	64.73	74.77
	MrBERT (MC)	64.23	64.94	74.03
	MrBERT (H)	62.85	63.45	72.89
Codauto + Gen	FastText	33.78	33.73	48.01
	BETO (MC)	66.72	67.15	75.98
	BETO (H)	66.17	66.63	75.21
	MrBERT (MC)	65.12	65.50	74.57
	MrBERT (H)	64.73	65.36	75.22
Gen Only	FastText	25.54	28.45	41.29
	BETO (MC)	39.80	41.32	56.74
	BETO (H)	41.79	43.34	60.13
	MrBERT (MC)	31.18	33.24	46.50
	MrBERT (H)	28.14	29.64	49.90

Table 4. Performance across datasets and model configurations. MC denotes multiclass classification and H denotes hierarchical classification.

Declaration on Generative AI

We have used ChatGPT for minor copy-editing and Grammarly for grammar and spelling check. After using these tools, we have reviewed and edited the content as needed and take full responsibility for the publication's content.

References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017). https://doi.org/10.1162/tacl_a_00051
2. Cesa-Bianchi, N., Gentile, C., Zaniboni, L.: Hierarchical classification: Combining bayes with svm. Proceedings of the 23rd International Conference on Machine Learning pp. 177–184 (2006)

3. Deng, Z., Peng, H., He, D., Li, J., Yu, P.: HTCInfoMax: A global model for hierarchical text classification via information maximization. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3259–3265. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.260, https://aclanthology.org/2021.naacl-main.260/
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding (2019), https://arxiv.org/abs/1810.04805
5. Dumais, S., Chen, H.: Hierarchical classification of web content. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 256–263 (2000)
6. Jiang, T., Wang, D., Sun, L., Chen, Z., Zhuang, F., Yang, Q.: Exploiting Global and Local Hierarchies for Hierarchical Text Classification. arXiv (May 2022). https://doi.org/10.48550/arXiv.2205.02613
7. Joachims, T.: Text categorization with Support Vector Machines: Learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) Machine Learning: ECML-98. pp. 137–142. Springer Berlin Heidelberg, Berlin, Heidelberg (1998)
8. Kibriya, A.M., Frank, E., Pfahringer, B., Holmes, G.: Multinomial naive Bayes for text categorization revisited. In: Australasian Joint Conference on Artificial Intelligence. pp. 488–499. Springer (2004)
9. Liu, Y., Ott, M., Goyal, N., Du, J.: RoBERTa: A robustly optimized BERT pretraining approach (2019), https://arxiv.org/abs/1907.11692
10. Quinlan, J.R.: Induction of decision trees. Machine learning **1**, 81–106 (1986)
11. Silla Jr, C.N., Freitas, A.A.: A survey of hierarchical classification across different application domains. Data mining and knowledge discovery **22**(1), 31–72 (2011)
12. Wang, Z., Wang, P., Huang, L., Sun, X., Wang, H.: Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7109–7119. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.491, https://aclanthology.org/2022.acl-long.491/
13. Zhou, J., Ma, C., Long, D., Xu, G., Ding, N., Zhang, H., Xie, P., Liu, G.: Hierarchy-aware global model for hierarchical text classification. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1106–1117. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.104, https://aclanthology.org/2020.acl-main.104/