Online News Classification Using Large Language Models with Semantic Enrichment

Joana Santos[1,2,*], Nuno Silva[2], Carlos Ferreira[2] and João Gama[3]

[1]Fac. Engineering, University of Porto, s/n, R. Dr. Roberto Frias, 4200-465 Porto, Portugal
[2]ISEP Instituto Superior de Engenharia do Porto, Rua Dr. António Bernardino de Almeida, 431 4249-015 Porto, Portugal
[3]Fac. Economics, University of Porto, R. Dr. Roberto Frias, 4200-464 Porto

Abstract

This paper addresses a critical gap in applying semantic enrichment for online news text classification using large language models (LLMs) in fast-paced newsroom environments. While LLMs excel in static text classification tasks, they struggle in real-time scenarios where news topics and narratives evolve rapidly. The dynamic nature of news, with frequent introductions of new concepts and events, challenges pre-trained models, which often fail to adapt quickly to changes. Additionally, the potential of ontology-based semantic enrichment to enhance model adaptability in these contexts has been underexplored.

To address these challenges, we propose a novel supervised news classification system that incorporates semantic enrichment to enhance real-time adaptability. This approach bridges the gap between static language models and the dynamic nature of modern newsrooms. The system operates on an adaptive prequential learning framework, continuously assessing model performance on incoming data streams to simulate real-time newsroom decision-making. It supports diverse content formats—text, images, audio, and video—and multiple languages, aligning with the demands of digital journalism.

We explore three strategies for deploying LLMs in this dynamic environment: using pre-trained models directly, fine-tuning classifier layers while freezing the initial layers to accommodate new data, and continuously fine-tuning the entire model using real-time feedback combined with data selected based on specified criteria to enhance adaptability and learning over time. These approaches are evaluated incrementally as new data is introduced, reflecting real-time news cycles. Our findings demonstrate that ontology-based semantic enrichment consistently improves classification performance, enabling models to adapt effectively to emerging topics and evolving contexts. This study highlights the critical role of semantic enrichment, prequential evaluation, and continuous learning in building robust and adaptive news classification systems capable of thriving in the rapidly evolving digital news landscape. By augmenting news content with third-party ontology-based knowledge, our system provides deeper contextual understanding, enabling LLMs to navigate emerging topics and shifting narratives more effectively.

Keywords

supervised online learning, semantic enrichment, NLP, LLM, news classification

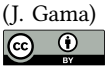
1. Introduction

News plays a vital role in society, keeping citizens informed about events in their city, country, and across the world. Effective news categorization is essential for delivering information to the right audience in an organized manner. Leveraging advanced tools, such as large language models (LLMs), can enhance this categorization by delivering high-quality results. However, as the concepts and topics in the news evolve, classification models must adapt to remain relevant. This article proposes a novel, online multi-class supervised news classification system that leverages semantic enrichment to enhance model adaptability and comprehension in real-time environments.

The primary objectives of this article are to compare the classification performance of raw news texts with semantically enriched texts. Additionally, the study examines the impact of incorporating

EKAW 2024: EKAW 2024 Workshops, Tutorials, Posters and Demos, 24th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2024), November 26-28, 2024, Amsterdam, The Netherlands

*Corresponding author.
†These authors contributed equally.
ffs@isep.ipp.pt (J. Santos); nps@isep.ipp.pt (N. Silva); cgf@isep.ipp.pt (C. Ferreira); jgama@fep.up.pt (J. Gama)
0009-0006-2656-4375 (J. Santos); 0000-0002-0556-0707 (N. Silva); 0000-0001-9933-8287 (C. Ferreira); 0000-0003-3357-1195 (J. Gama)



different types of information by applying retraining approaches with and without predefined criteria for selecting the data.

This paper is structured into five additional sections. Section 2 reviews key concepts and related works. Section 3 details the proposed methodology. Section 4 explains the system's setup and implementation. Section 5 presents the evaluation experiments and provides an analysis of the results. Finally, Section 6 concludes the paper and suggests directions for future research.

The developed code can be found at: https://pypi.org/project/online-news-classification/

2. Related work

The dynamic nature of news necessitates classification models that can adapt to ongoing changes. Models trained on large datasets and capable of generalization without requiring extensive modifications for each new analysis provide significant advantages. Large Language Models (LLMs) represent a key innovation in this regard. As described by Min et al.[1], LLMs are advanced neural network-based statistical models that excel in tasks such as Natural Language Processing (NLP) and text generation.

In NLP tasks like text classification, LLMs have demonstrated superior performance compared to traditional models, such as Naïve Bayes and Decision Trees. Unlike these conventional approaches, which often analyze words in isolation, LLMs process text holistically, capturing contextual relationships across sentences and documents. Their robust generalization capabilities enable them to handle a variety of tasks without requiring extensive task-specific fine-tuning. Techniques like few-shot classification—where models learn to differentiate classes with minimal examples—and zero-shot classification—where models classify text without prior exposure to specific classes—further enhance their versatility[2].

Prominent LLMs include GPT[3] and BERT[4], both of which have distinct strengths. GPT, widely recognized for text generation, also performs effectively in classification tasks, particularly with zero-shot techniques. In contrast, BERT excels at understanding word context within sentences, leveraging surrounding words to derive meaning. Variants of BERT, such as RoBERTa[5] and DistilBERT[6], have been developed to optimize specific NLP tasks. RoBERTa enhances BERT's performance on certain benchmarks, while DistilBERT offers a lighter, more efficient alternative with high performance.

Hybrid models like BART[7] combine the contextual understanding of BERT with GPT's capabilities in text generation and reconstruction. These models deliver a more comprehensive approach to NLP tasks, bridging the gap between classification and generative functionalities, making them particularly valuable for applications requiring both precision and adaptability.

Several studies have explored text classification [8, 9, 10, 11, 12] using large language models (LLMs). Many of these works integrate LLMs with natural language processing (NLP) tasks, including news topic classification. Within these tasks, some research targets specific areas, such as financial news classification [8, 9], while others adopt a more general approach [10, 11, 12]. These studies consistently demonstrate that fine-tuned LLMs outperform traditional models like BERT-base or DeBERTa in achieving superior results for this task.

Traditional metrics like Precision, Recall, F-measure, and Accuracy are commonly used in batch classification but are unsuitable for online environments where data and conditions change over time. In such cases, prequential error evaluation is more appropriate as it dynamically tracks error or accuracy for each instance, offering insight into the model's evolution. Prequential evaluation offers two approaches: prequential alpha, which incorporates a forgetting factor (α) to give more weight to recent errors, and prequential window, which uses a sliding window of fixed size (w) to calculate error rates based on the most recent subset of data while discarding older information [13].

A review of the state of the art highlights the advancement introduced in this work: the application of online classification techniques to semantically enriched news texts. Unlike traditional approaches that depend on static datasets and pre-trained models, this approach harnesses real-time adaptability, empowering the classification system to process dynamic and ever-evolving news content effectively.

3. Methodology

The system adopts an online processing approach, where each document is handled sequentially as it arrives. Each document undergoes enrichment and classification, and the model subsequently learns from this data, adhering to a prequential (predict-then-learn) framework [13].

The system comprises four key tasks (Figure 1):

- Pre-processing - this stage involves cleaning and preparing the text by removing stop words and tokenizing the title and abstract for further analysis
- Semantic Enrichment - this is the core task of the system, where the document is enriched semantically using an external knowledge base. It involves identifying semantic entities within the text and mapping these entities to their corresponding representations in the knowledge base
- Prediction - the system classifies the document by determining its most likely category based on the current model.
- Learning - the model learns from the actual class label assigned to the document, improving its performance over time.

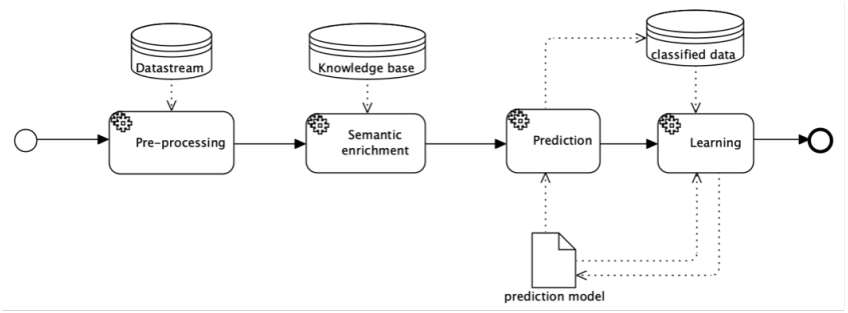


Figure 1: System proposed architecture

Semantic enrichment was performed using the ReFinED [14] library. This library is an entity linker that allows the identification of the entities present in the text and their relationship with the wikidata/wikipedia entities. Furthermore, this library has the ability to disambiguate the entities it identifies (e.g. 'The jaguar is in the zoo' returns Q35694 (referring to the animal 'jaguar') rather than Q30055 (which refers to 'Jaguar,' the automobile manufacturer)). The entities resulting from this process will be used in the classification of documents. In order to facilitate the interpretation of these entities by the LLM, the titles of the Wikipedia pages were used.

4. Setup

Three experiments were conducted to demonstrate the impact of semantic enrichment and the use of a prequential approach for classifying text documents from news datasets:

1. Utilizing the LLM in its original form, following a zero-shot classification approach
2. Fine-tuning the last layers of the model for each specific set of documents
3. Fine-tuning the final layers of the model using already known data from a specific category, selected based on predefined criteria that must be met for retraining.

All these experiments were carried out on a machine with the following characteristics: Intel Core I7-14700K, GeForce RTX 4090, Windows 11 Pro.

The following sections provide details on the datasets, the semantic enrichment process, and the LLMs employed in these experiments.

Various LLMs were employed to conduct the experiments, selected from the HuggingFace platform based on the following criteria: (i) Task - Zero-shot classification and (ii) Language - english.

Table 1
10 most downloaded LLM's

Models	Training datasets
MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli [15]	MultiNLI, Fever-NLI and Adversarial-NLI (ANLI)
cross-encoder/nli-deberta-v3-base [16]	SNLI and MultiNLI
cross-encoder/nli-roberta-base [17]	SNLI and MultiNLI
MoritzLaurer/mDeBERTa-v3-base-mnli-xnli [15]	XNLI and MNLI
DEPRECATED: sileod/deberta-v3-base-tasksource-nli has been replaced by tasksource/deberta-small-long-nli [18]	NLI
MoritzLaurer/DeBERTa-v3-xsmall-mnli-fever-anli-ling-binary [15]	MultiNLI, Fever-NLI, LingNLI and ANLI
MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli [15]	MultiNLI, Fever-NLI, Adversarial-NLI (ANLI), LingNLI and WANLI
joeddav/xlm-roberta-large-xnli [19]	NLI
MoritzLaurer/deberta-v3-large-zeroshot-v2.0 [20]	NLI
cross-encoder/nli-MiniLM2-L6-H768 [21]	SNLI and MultiNLI

Table 1 lists the ten most-downloaded models as of October 15, 2024, the date of the last consultation.

In these experiments, seven datasets about news were used. Table 2 presents them, as well as their characteristics, with regard to the number of documents and number of categories.

Table 2
Datasets

Datasets	N° of documents	N° of categories	Natural Language	Period
AG News [22]	120,000	4	english	since July 2004
BBC News [23]	15,488	84		2010
CNN [24]	4,076	6		2012 - 2022
MiND [25]	96,106	18		NA
News Category [26]	189,814	42		2012-2022
The Guardian [27]	1,422,200	297		2010-2024
The NY Times [28]	1,053,991	81		2010-2024

5. Results and Discussion

In the next sections, the results and respective interpretation for each experiment are presented.

5.1. Experiment 1

In Experiment 1, the LLM was utilized in its original form, without any fine-tuning. Table 3 presents the results for each dataset type (non-enriched (¬E) and enriched (E)), broken down by dataset and model. The highest accuracy values in the comparison between non-enriched and enriched datasets are highlighted in bold.

From Table 3, we observe that approximately half (35 out of 70) of the dataset and model combinations yield higher overall accuracy with the non-enriched dataset. Notably, the AG News and News Category datasets perform better with the non-enriched dataset in the majority of models, with 9 out of 10 and 8 out of 10 combinations, respectively. In contrast, The Guardian dataset shows improved results with the enriched dataset in 7 out of 10 model combinations.

Table 3
Results of experiment 1

		AG News	BBC News	CNN	MiND	News Category	The Guardian	The NY Times
MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli	¬E	**22.98**	1.58	15.78	0.98	**0.72**	0.11	**0.78**
	E	20.86	**1.85**	**25.64**	**1.07**	**0.72**	**0.12**	0.66
cross-encoder/nli-deberta-v3-base	¬E	**23.51**	**4.73**	20.17	1.31	2.62	0.07	0.14
	E	20.63	3.53	**20.53**	**2.37**	**4.17**	**0.16**	**0.47**
cross-encoder/nli-roberta-base	¬E	**26.67**	0.81	**26.74**	**20.77**	**1.39**	0.01	**6.64**
	E	18.46	**0.88**	12.19	15.58	1.35	**0.07**	3.52
MoritzLaurer/mDeBERTa-v3-base-mnli-xnli	¬E	**18.98**	0.87	2.75	4.06	**6.19**	**0.21**	**0.31**
	E	7.36	**1.23**	**5.25**	**4.57**	5.57	0.15	0.28
tasksource/deberta-small-long-nli	¬E	25.54	1.43	**1.42**	1.35	0.48	0.00	0.35
	E	**25.76**	**1.48**	**1.42**	**1.44**	**0.52**	0.00	**0.41**
MoritzLaurer/DeBERTa-v3-xsmall-mnli-fever-anli-ling-binary	¬E	24.93	0.89	3.83	**6.09**	2.24	0.00	**7.82**
	E	**24.94**	**0.90**	**4.12**	5.95	**2.28**	**0.01**	6.70
MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli	¬E	**14.82**	**1.63**	**5.27**	2.59	**1.89**	0.42	**1.78**
	E	11.76	**1.63**	4.59	**3.03**	1.65	**0.60**	1.45
joeddav/xlm-roberta-large-xnli	¬E	**1.56**	0.80	4.44	25.13	**2.42**	**0.03**	3.93
	E	0.85	**0.92**	**5.45**	**28.55**	2.38	0.02	**4.18**
MoritzLaurer/deberta-v3-large-zeroshot-v2.0	¬E	**22.72**	0.90	3.04	**16.21**	**1.24**	0.03	0.01
	E	21.31	**0.91**	**3.93**	10.45	1.21	**0.06**	**0.02**
cross-encoder/nli-MiniLM2-L6-H768	¬E	**24.07**	**1.09**	**23.28**	**4.45**	**2.38**	0.31	1.14
	E	16.58	0.91	21.22	3.60	2.08	**0.49**	**1.38**

5.2. Experiment 2

In Experiment 2, the final layers of the LLM, which correspond to the classifier, were updated using a prequential approach. This approach enables the LLM classifier to learn incrementally from previously unseen documents. The documents used for retraining the model were selected based on a strategy that considers five criteria, detailed in Table 4.

Table 4
Experiment 2 configurations values

Configuration	Definition	Value
ACCURACY_LIMIT	The minimum accuracy threshold a category must meet to avoid retraining	30
PERCENTAGE_OF_CLASSES	The percentage of classes meeting the accuracy threshold that will undergo retraining	50
NUMBER_OF_DOCUMENTS_TO_RETRAIN	The number of documents from a given class to include in the retraining process	500
WINDOW_FOR_RETRAINING	A sliding window specifying how many documents from each class must be considered for accuracy evaluation	50
DOCUMENTS_TO_RETRAIN	Specifies whether the retraining uses the most recent or oldest documents	newest

Table 5 presents the results for each dataset type (non-enriched (¬E) and enriched (E)), dataset and model. The highest accuracy values in the comparison between non-enriched and enriched datasets are highlighted in bold.

Table 5 reveals that only about 44% of cases (31 out of 70) achieve better results with the enriched dataset. However, compared to accuracy values obtained in previous experiments, there is significant improvement in certain instances, with gains exceeding 30% (e.g., the CNN dataset using the MoritzLaurer/mDeBERTa-v3-base-mnli-xnli model). Conversely, some cases show a decline in overall accuracy in this experiment (e.g., the CNN dataset with the cross-encoder/nli-deberta-v3-base model). Interestingly, both outcomes occur with the same dataset, suggesting that the initial training of the model prior to the described process may play a critical role in these discrepancies.

Table 5
Results of experiment 2

		AG News	BBC News	CNN	MiND	News Category	The Guardian	The NY Times
MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli	¬E	**22.98**	2.06	3.16	**3.82**	**2.91**	1.14	4.69
	E	20.86	**2.22**	**4.86**	1.64	2.27	**1.32**	**4.75**
cross-encoder/nli-deberta-v3-base	¬E	**23.51**	2.08	**12.02**	27.64	2.42	**1.47**	4.59
	E	20.63	**2.31**	5.27	**28.04**	**3.66**	0.96	**4.96**
cross-encoder/nli-roberta-base	¬E	**26.67**	1.84	4.91	**1.62**	**1.79**	**1.13**	**5.14**
	E	18.46	**1.96**	**52.26**	1.59	1.72	0.73	5.00
MoritzLaurer/mDeBERTa-v3-base-mnli-xnli	¬E	**18.98**	**1.87**	35.67	1.22	2.09	**1.34**	4.80
	E	7.36	1.84	**35.97**	**29.21**	**5.23**	1.31	**4.82**
tasksource/deberta-small-long-nli	¬E	25.54	2.60	**41.46**	29.74	**2.06**	**8.20**	5.68
	E	**25.76**	**2.81**	40.73	**29.75**	1.89	7.44	**5.84**
MoritzLaurer/DeBERTa-v3-xsmall-mnli-fever-anli-ling-binary	¬E	24.93	**2.12**	32.19	**3.66**	**4.07**	0.75	2.01
	E	**24.94**	2.02	**32.29**	3.46	2.21	**1.37**	**2.08**
MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli	¬E	**14.82**	**2.20**	**51.77**	**29.35**	**4.36**	**0.95**	**5.44**
	E	11.76	2.15	33.32	29.28	4.32	0.73	5.18
joeddav/xlm-roberta-large-xnli	¬E	**1.56**	1.68	51.77	**3.95**	**1.76**	**1.25**	**5.74**
	E	0.85	**1.96**	**51.91**	3.92	**1.76**	1.10	5.71
MoritzLaurer/deberta-v3-large-zeroshot-v2.0	¬E	**22.72**	**1.89**	35.35	1.57	**1.75**	**0.75**	4.92
	E	21.31	1.78	**48.41**	**1.97**	1.74	**0.75**	**5.06**
cross-encoder/nli-MiniLM2-L6-H768	¬E	**24.07**	2.07	5.18	**28.41**	4.47	0.92	4.81
	E	16.58	**2.34**	**54.17**	25.77	**4.50**	**1.11**	**5.77**

5.3. Experiment 3

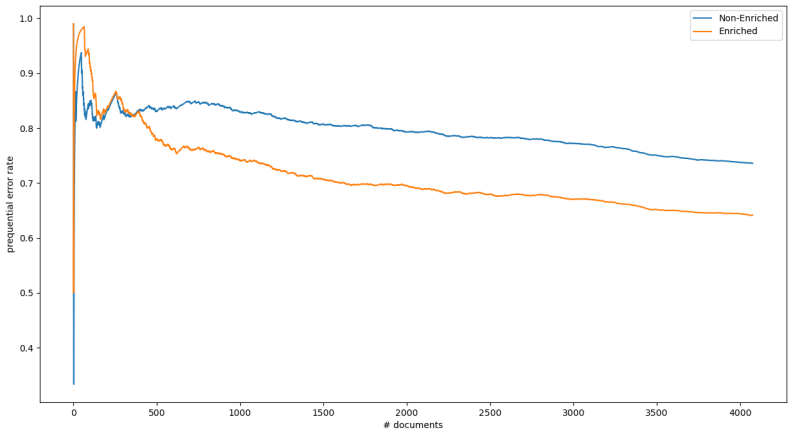
Similar to Experiment 2, a prequential approach is employed here as well. However, unlike Experiment 2, no selection or constraints were applied; instead, all documents were used for learning and retraining. This comprehensive approach ensures that the model leverages the full dataset for incremental updates, enhancing its exposure to diverse patterns. Table 6 presents the results of this experiment for both non-enriched (¬E) and enriched (E) datasets, categorized by dataset and model.

Table 6
Results of experiment 3

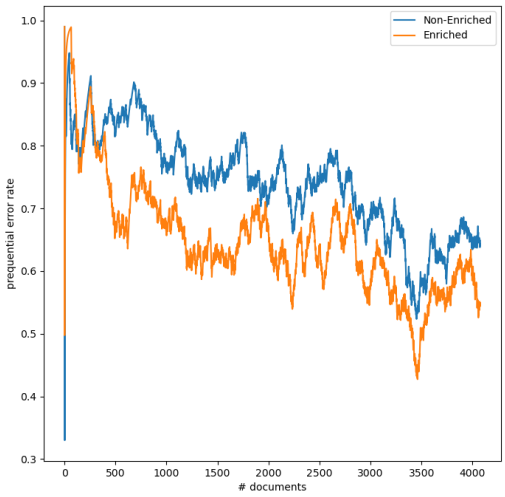
		AG News	BBC News	CNN	MiND	News Category	The Guardian	The NY Times
MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli	¬E	**11.55**	4.18	26.40	45.16	29.82	26.80	32.59
	E	11.31	**4.24**	**35.87**	**46.54**	**30.70**	**28.36**	**33.32**
cross-encoder/nli-deberta-v3-base	¬E	**14.74**	**5.02**	**35.67**	38.16	27.84	19.01	24.02
	E	14.50	4.33	28.53	**40.34**	**28.45**	**20.16**	**24.85**
cross-encoder/nli-roberta-base	¬E	**15.34**	2.69	34.37	39.14	27.45	21.42	25.48
	E	13.40	**2.81**	**37.39**	**42.87**	**28.47**	**26.25**	**29.26**
MoritzLaurer/mDeBERTa-v3-base-mnli-xnli	¬E	**13.95**	2.98	26.50	39.90	27.80	20.61	25.16
	E	13.71	**3.56**	**33.19**	**42.16**	**28.49**	**22.13**	**26.18**
tasksource/deberta-small-long-nli	¬E	**4.71**	2.96	**1.77**	62.02	43.38	43.08	**50.51**
	E	4.66	**3.27**	1.67	**62.36**	**43.70**	**43.62**	50.42
MoritzLaurer/DeBERTa-v3-xsmall-mnli-fever-anli-ling-binary	¬E	11.95	**1.15**	5.32	44.09	28.83	24.94	29.64
	E	**12.10**	1.04	**8.12**	**46.13**	**29.47**	**27.63**	**31.12**
MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli	¬E	9.76	4.58	32.21	51.10	34.79	37.06	41.60
	E	**9.99**	**4.76**	**33.78**	**52.23**	**35.80**	**37.47**	**41.78**
joeddav/xlm-roberta-large-xnli	¬E	**10.13**	3.98	33.71	47.83	31.76	34.77	39.09
	E	10.06	**4.07**	**37.29**	**50.79**	**32.97**	**38.13**	**41.01**
MoritzLaurer/deberta-v3-large-zeroshot-v2.0	¬E	**10.80**	**4.16**	32.65	48.21	32.42	32.82	36.95
	E	10.46	3.98	**33.24**	**49.38**	**33.27**	**34.49**	**37.69**
cross-encoder/nli-MiniLM2-L6-H768	¬E	16.37	2.18	**38.03**	37.81	26.96	17.98	22.28
	E	**16.39**	**2.69**	34.54	**38.59**	**27.36**	**20.90**	**22.72**

Analyzing Table 6, we find that only about 20% (14 out of 70) of the experiments conducted—encompassing dataset type, dataset, and model—achieve higher overall accuracy with the non-enriched dataset. When compared to the results from Experiment 1 and Experiment 2, there is a significant improvement of approximately 30%/35% in the number of cases where the enriched dataset delivers better accuracy. This highlights the effectiveness of combining a time-adaptive (online) learning approach with semantic enrichment. For instance, in the case of the MiND dataset and the tasksource/deberta-small-long-nli model, the improvement between the baseline model and the enriched learning model reaches 60%.

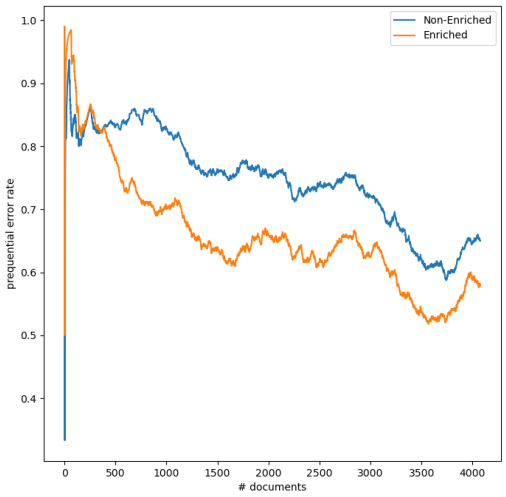
Figure 2 complements the findings in Table 6 by presenting the prequential analysis results for the most significant difference observed in this experiment. It highlights the contrast between the non-enriched and enriched datasets, focusing on the CNN dataset processed with the MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli model.



(a) Prequential



(b) Prequential alpha



(c) Prequential window

Figure 2: Prequential error results for MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli on the CNN dataset

For the CNN dataset using the MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli model, the enriched dataset achieves higher overall accuracy compared to the non-enriched dataset. In Figure 2a, the blue line for non-enriched data (0.8/0.9) is above the orange line for the enriched dataset (0.7/0.8), reflecting a higher error rate for the non-enriched dataset and thus lower accuracy. Figures 2b and 2c provide

prequential analyses under different conditions: Figure 2b applies a forgetting factor of 0.99 to emphasize recent errors, while Figure 2c uses a sliding window of 500 documents to examine error patterns. In both cases, the non-enriched dataset shows a consistently higher error than the enriched dataset.

6. Conclusion

Given the constant emergence of new concepts in the news, classification models must continually adapt. To address this challenge, the paper introduces a novel framework that combines LLMs with online retraining for continuous updates, enabling the model to classify documents with finner granularity and precision over time.

Based on the results obtained, it can be concluded that semantic enrichment significantly improves classification performance. Additionally, the prequential approach enabled the models to adapt to the specific data of each dataset, resulting in a substantial improvement compared to the original model. By integrating semantic enrichment, the system enhances contextual understanding, addressing challenges such as the introduction of new entities, topics, and shifting narratives typical of newsroom environments. This approach represents a significant step forward in bridging the gap between static classification models and the fluid nature of real-world news, ensuring greater accuracy and relevance in classification tasks.

Yet, these findings highlight the need for further investigation into the use of specific criteria in data selection, exploring its potential to enhance overall accuracy under certain conditions.

Moreover, the current semantic enrichment process focuses solely on the entities identified in the texts. Future work could explore the semantic relationships between these entities to further enhance the richness of the documents. At the level of the LLM, it is also recommended to investigate explainability mechanisms, enabling a better understanding of the rationale behind the classifications.

References

[1] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, D. Roth, Recent advances in natural language processing via large pre-trained language models: A survey, ACM Computing Surveys 56 (2023) 1–40.

[2] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, J. Gao, Large language models: A survey, arXiv preprint arXiv:2402.06196 (2024).

[3] T. B. Brown, Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020).

[4] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of naacL-HLT, volume 1, 2019, p. 2.

[5] Y. Liu, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[6] V. Sanh, Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).

[7] M. Lewis, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019).

[8] L. Yang, Y. Huang, C. Tan, S. Wang, News topic classification base on fine-tuning of chatglm3-6b using neftune and lora, in: ACM International Conference Proceeding Series, 2024, p. 521 – 525. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85201301232&doi=10.1145%2f3675249.3675339&partnerID=40&md5=96c212f1b137799db4621a918ddf394f. doi:10.1145/3675249.3675339, cited by: 0.

[9] C. Ye, X. Shi, Optimizing news topic classification with instructional fine-tuning of chatglm3, in: ACM International Conference Proceeding Series, 2024, p. 573 – 577. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85201426681&doi=10.1145%2f3672758.3672851&partnerID=40&md5=4af6a12e72d17b2a62ea4cf7812d078f. doi:10.1145/3672758.3672851, cited by: 0.

[10] N. Nazyrova, S. Chahed, T. Chausalet, M. Dwek, Leveraging large language models for medical text classification: a hospital readmission prediction case, 2024. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85206472619&doi=10.1109%2fICPRS62101.2024.10677826&partnerID=40&md5=57d814e5d4235e5e5f4c820a3601d903. doi:10.1109/ICPRS62101.2024.10677826, cited by: 0.

[11] D. Zhang, R. Mi, P. Zhou, D. Jin, M. Zhang, T. Song, Large model-based data augmentation for imbalanced text classification, in: 2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology, AINIT 2024, 2024, p. 1006 – 1010. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85199212390&doi=10.1109%2fAINIT61980.2024.10581735&partnerID=40&md5=3bb4e1bf49929b06cb1eec09da989df3. doi:10.1109/AINIT61980.2024.10581735, cited by: 0.

[12] A. Edwards, J. Camacho-Collados, Language models for text classification: Is in-context learning enough?, in: 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024 - Main Conference Proceedings, 2024, p. 10058 – 10072. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85195940707&partnerID=40&md5=7696ac972bf51c856dcd1e86d084f407, cited by: 1.

[13] J. Gama, R. Sebastiao, P. P. Rodrigues, On evaluating stream learning algorithms, Machine learning 90 (2013) 317–346.

[14] J. F. C. C. A. P. Tom Ayoola, Shubhi Tyagi, ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking, in: NAACL, 2022.

[15] M. Laurer, Less annotating, more classifying - addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI, 2022. URL: https://osf.io/74b8k.

[16] cross encoder, cross-encoder/nli-deberta-v3-base · hugging face, 2021. URL: https://huggingface.co/cross-encoder/nli-deberta-v3-base.

[17] Z. Wang, A. Bukharin, O. Delalleau, D. Egert, G. Shen, J. Zeng, O. Kuchaiev, Y. Dong, Helpsteer2-preference: Complementing ratings with preferences, 2024. URL: https://arxiv.org/abs/2410.01257. arXiv:2410.01257.

[18] D. Sileo, tasksource: A large collection of NLP tasks with a structured dataset preprocessing framework, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 15655–15684. URL: https://aclanthology.org/2024.lrec-main.1361.

[19] joeddav, joeddav/xlm-roberta-large-xnli · hugging face, 2023. URL: https://huggingface.co/joeddav/xlm-roberta-large-xnli.

[20] M. Laurer, W. van Atteveldt, A. Casas, K. Welbers, Building Efficient Universal Classifiers with Natural Language Inference, 2023. URL: http://arxiv.org/abs/2312.17543. doi:10.48550/arXiv.2312.17543, arXiv:2312.17543 [cs].

[21] cross encoder, cross-encoder/nli-MiniLM2-l6-h768 · hugging face, 2021. URL: https://huggingface.co/cross-encoder/nli-MiniLM2-L6-H768.

[22] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, 2016. arXiv:1509.01626.

[23] Opensnippets, BBC UK news dataset - dataset by opensnippets, 2021. URL: https://data.world/opensnippets/bbc-uk-news-dataset.

[24] H. Unger, CNN Articles - Data Cleaning & Visualization, 2021. URL: https://kaggle.com/code/hadasu92/cnn-articles-data-cleaning-visualization.

[25] F. Wu, Y. Qiao, J.-H. Chen, C. Wu, T. Qi, J. Lian, D. Liu, X. Xie, J. Gao, W. Wu, M. Zhou, MIND: A Large-scale Dataset for News Recommendation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 3597–3606. doi:10.18653/v1/2020.acl-main.331.

[26] R. Misra, News category dataset, 2022. URL: https://www.kaggle.com/datasets/rmisra/news-category-dataset.

[27] T. G. O. Platform, The guardian api, 2024. URL: https://open-platform.theguardian.com.

[28] T. N. Y. Times, The new york times api, 2024. URL: https://developer.nytimes.com.