IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 November 2023, accepted 17 December 2023, date of publication 21 December 2023,
date of current version 28 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3345414

RESEARCH ARTICLE

Evaluating the Effectiveness of GPT Large Language Model for News Classification in the IPTC News Ontology

BAHAREH FATEMI, FAZLE RABBI, AND ANDREAS L. OPDAHL
Department of Information Science and Media Studies, University of Bergen, 5007 Bergen, Norway
Corresponding authors: Bahareh Fatemi (Bahareh.Fatemi@uib.no) and Fazle Rabbi (Fazle.Rabbi@uib.no)

This work was supported in part by SFI MediaFutures Partners and the Research Council of Norway under Grant 309339.

ABSTRACT News classification plays a vital role in newsrooms, as it involves the time-consuming task of categorizing news articles and requires domain knowledge. Effective news classification is essential for categorizing and organizing a constant flow of information, serving as the foundation for subsequent tasks, such as news aggregation, monitoring, filtering, and organization. The automation of this process can significantly benefit newsrooms by saving time and resources. In this study, we explore the potential of the GPT large language model in a zero-shot setting for multi-class classification of news articles within the widely accepted International Press Telecommunications Council (IPTC) news ontology. The IPTC news ontology provides a structured framework for categorizing news, facilitating the efficient organization and retrieval of news content. By investigating the effectiveness of the GPT language model in this classification task, we aimed to understand its capabilities and potential applications in the news domain. This study was conducted as part of our ongoing research in the field of automated journalism.

INDEX TERMS IPTC media topics, journalism, large language models, news classification.

I. INTRODUCTION

Large language models (LLMs) have demonstrated their proficiency in addressing a multitude of natural language processing (NLP) tasks such as named entity recognition [24], text summarization [28] and sentiment analysis [12], and researchers from different domains such as computer science, medicine and social science, have recognized the utility of these models and increasingly adopted them in their work [3]. Within the news domain, researchers have also shown notable interest in using Language Models for tasks that used to be performed in conventional ways. For instance, LLMs have been used to develop creativity support tools that help journalists explore angles for reporting on press releases [15]. Yang et al. [27] investigated whether GPT, a large language model, can assess the credibility of news outlets and provided evidence that its credibility ratings align with human expert judgments, indicating its potential use in fact-checking applications. Goyal et al. [6] examined the impact of large language models, particularly GPT-3 [1], on text summarization, showing their superior performance, especially in tasks that require minimal human prompts.

Effective news classification is essential for categorizing and organizing a constant flow of information, serving as the foundation for subsequent tasks, such as news aggregation, monitoring, filtering, and organization. However, multi-class classification of news articles, in which each article is assigned to one category out of several possible categories, is a challenging task. While immensely valuable for news organizations, the process of multi-class classification of news articles is time-consuming and disliked by journalists because of the need to carefully analyze and assign categorical labels to articles, considering the intricate intersections of news topics and the contextual subtleties embedded within them.

The associate editor coordinating the review of this manuscript and approving it for publication was Khursheed Aurangzeb.

LLMs are well known for their ability to capture the complexities of language and context. In our investigation, we explored the potential of the GPT language model [18] as a foundational element to enhance the efficiency of classifying news articles, particularly within the widely accepted and utilized International Press Telecommunications Council (IPTC[1]) news ontology.

IPTC media topics consist of a layered classification system used in the news industry. It encompasses a six-level hierarchical structure, offering a structured approach for categorizing news subjects. This layered approach simplifies the organization and retrieval of news content, making it more efficient for media professionals [20]. For example, Clercq et al. [4] employed the IPTC news media topics standard to improve the granularity and diversity of news article classification to enhance the representation of news content in their study on automated journalism.

The multiple levels of granularity of the IPTC ontology can be intricate and potentially perplexing, even for news journalists. Consequently, classifying content into fine-grained categories within this framework is inherently more challenging, tedious, and cognitively demanding. Furthermore, achieving fine-grained classification often requires a deep understanding of the domain, making it essential to be a domain expert in addition to being a journalist.

In our study, given the extensive size of the hierarchy, we chose to concentrate on the first two levels to align with our primary goal of showcasing LLM's capabilities. In particular, finding a well-balanced dataset that adequately represents each fine-grained category poses a considerable challenge. This decision helps strike a balance between task complexity and data availability for effective evaluation.

In this investigation, our research question revolves around the effectiveness of utilizing GPT pre-trained language models in news classification. By utilizing LLMs, we not only categorize news articles effectively but also streamline content retrieval in today's digital news landscape. This lays the foundation for future applications, such as tracking event development, which aligns with the underlying motivation discussed later. To evaluate our classification results, we employed both supervised and unsupervised machine learning techniques as proxies because of the absence of a high-quality gold standard. The novelty of this work lies in its exploration of the potential benefits of integrating GPT pre-trained language models into news classification tasks, which have not been extensively explored before.

The remainder of this paper is structured as follows. In Section II, we delve into related work concerning news classification, domain, and large language models. Section III is dedicated to our methods. In Section IV, we present the details of our experiments and the obtained results. In Section V, we examine existing limitations and outline potential avenues for future research. Finally, the conclusions are presented in Section VI.

[1]https://cv.iptc.org/newscodes/mediatopic/

II. ERLATED WORKS

A. NEWS CLASSIFICATION

News classification is a challenging task that has been extensively studied by the research community. Traditional news classification methods typically rely on hand-crafted features (e.g., the number of unique words in the news article, the number of times certain keywords appear) and machine learning algorithms such as support vector machines and decision trees. However, the landscape of news classification has evolved with the emergence of deep learning-based methods that leverage word/document embedding techniques [13], [23], such as [10]. Additionally, a notable shift has emerged towards utilizing convolutional neural networks (CNNs) for sentence classification [8]. Luo [11] explored the use of Long Short-Term Memory (LSTM) networks for text representations. While these methods have shown promise, it is important to note that the effectiveness of text classification relies on the model's ability to capture global word co-occurrence information, a strength notably exemplified by transformer-based models, such as BERT [5] and GPT [1]. These Transformers are at the forefront of natural language processing because of their unmatched capability to understand contexts, adapt to specific tasks, and efficiently handle diverse content, making them particularly suitable for news classification tasks.

B. LARGE LANGUAGE MODELS

Large Language Models such as GPT [1], T5 [19], and BERT [5], are versatile natural language processing models known for their ability to capture extensive semantic and syntactic information within text. There are three general ways of using LLMs: fine-tuned, few-shot, and zero-shot settings. Fine-tuning involves adapting a pre-trained language model to excel in specific tasks through additional training. Few-shot learning is a paradigm where models learn new tasks from a very small number of examples, and zero-shot learning tests a model's ability to tackle new tasks it hasn't been explicitly trained for, relying on its inherent understanding of language and context. Zero-shot learning aims to solve unseen tasks without labeled training examples or gradient updates. Recent research indicates that LLMs are superior at zero-shot learning [2], [25]. The zero-shot learning capability of GPT for different tasks is investigated by Qin et al, [17].

III. METHODS

A. DATA ANNOTATION USING GPT API

OpenAI-python package[2] was used to query the API endpoint. Specifically, we chose the gpt-3.5-Turbo model, which is trained up to September 2021. We set the temperature parameter to zero to minimize the randomness of output generation.

In zero-shot settings, formulating effective prompts is challenging. GPT is not specifically trained for the

[2]https://github.com/openai/openai-python

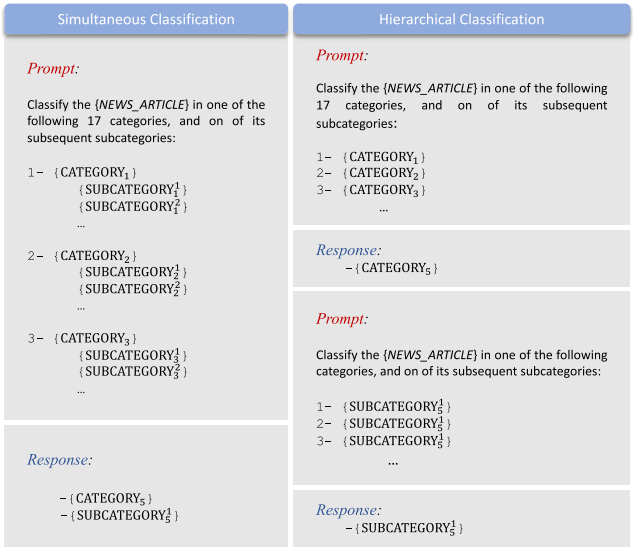


FIGURE 1. Simultaneous and hierarchical classification prompts.

IPTC classification task; therefore using it alone to classify articles into specific categories within this framework could result in unreliable outcomes. The risk of generating hallucinated outcomes makes it necessary to supplement the model with the actual ontology. By providing GPT with the ontology, we can enhance the accuracy and credibility of article classification within the IPTC framework. For this purpose, we explored two prompting strategies, as illustrated in Figure 1:

- **Simultaneous Classification**: In this approach, we provided the model with the entire ontology and tasked it with the simultaneous classification of a news article into one of the Level-1 categories and its corresponding Level-2 subcategories.
- **Hierarchical Classification**: Conversely, in the hierarchical approach, we initially provided the model with Level-1 categories and requested it to classify the news article accordingly. After determining the Level-1 category, we provided the corresponding subcategories belonging to the chosen category and tasked the model with classifying the news article into a specific subcategory.

In the Simultaneous Classification approach, we encountered certain challenges:

- First, the error rate was notably higher, resulting in misclassification. For instance, there were cases where an article was classified into Level-1 Category A and Level-2 Category B, while, in fact, Category B was a subcategory of Level-1 Category C. As an example, the article with title *"Former Bayern and Man Utd star Bastian Schweinsteiger retires from playing football"* was classified in the first-level category of "*sport*", and second-level category of "*retirement*" while "*retirement*" is a subcategory of "*labour*".
- Second, the model exhibited instances of hallucinations, generating categories that did not align with the actual ontology. For instance, an article with the title *"10 of the Best Russia Holiday Destinations - Beyond Moscow and St Petersburg"* was classified into level-1 and level-2 categories "*travel*" and "*Eco-tourism*" which do not exist in the ontology.

Conversely, in the Hierarchical Classification strategy, we achieved significant improvements: 1) the issue of misclassifications was effectively resolved and 2) the problem of hallucination was mitigated to a considerable extent, such that we did not observe any instances of hallucination in the first-level and only a few in the second-level classification. This hierarchical approach demonstrates a more accurate and reliable classification process than the simultaneous strategy.

In Figure 2, we showcase a set of subtractive word cloud visualizations that specifically focus on the "*Education*" category. These visualizations offer a more granular perspective by highlighting distinctive terminologies associated with specific subcategories within "*Education*", namely "*Curriculum*", "*School*" and "*Teachers*". By subtracting the common terms, these word clouds effectively emphasize the unique vocabulary that distinguishes each subcategory.

B. EVALUATION

We employed a dual evaluation approach encompassing classification and clustering techniques to demonstrate the effectiveness of the IPTC labels generated by the GPT, ensuring their suitability for downstream tasks and applications.

- **Classification:** By assessing the performance of a classification model on a labeled dataset, we can indirectly gauge the reliability of the assigned labels. The performance of the model depends on the quality and consistency of the labels. This approach provides a means to evaluate the annotation quality in the absence of a predefined gold standard, ensuring that the assigned labels properly represent the content of the dataset.
  The classification task involved a diverse range of methods, including Multinomial Naive Bayes, Logistic Regression, Support Vector Machines (SVM), and Random Forest, all of which were evaluated using a 5-fold cross-validation setting. In addition to these traditional classification algorithms, we incorporated state-of-the-art deep learning models, namely DistilBERT [21] and BERT [26], based on the complexity of the task and size of the dataset. These models were strategically chosen for their ability to capture intricate contextual information and nuances within the text, which is crucial for achieving accurate classification. For conventional machine learning models, we employed two embedding methods: TF-IDF and Glove [14]. These methods have been used to enhance the performance of traditional models.
- **Clustering:** In addition to classification, we conducted a cluster analysis on the dataset to evaluate the quality of

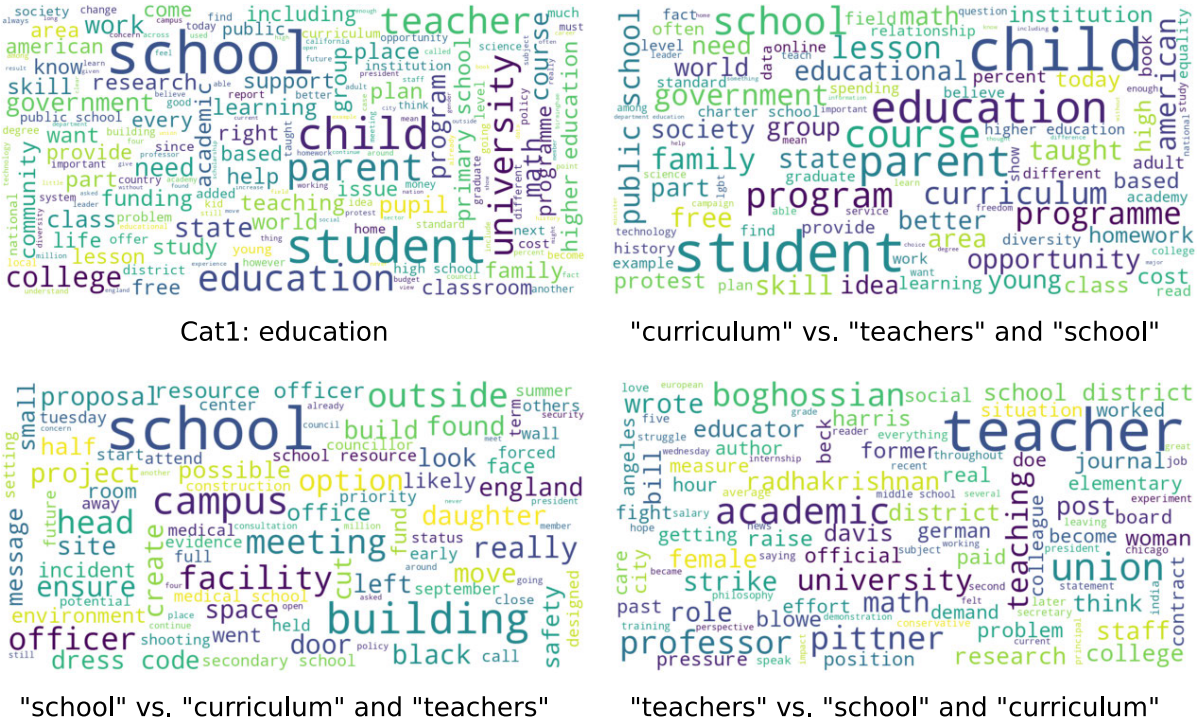


FIGURE 2. Subtractive Word Cloud: Contrasting Themes of "*Curriculum*", "*Teachers*" and "*school*", subcategories of "*Education*".

the annotations in an unsupervised manner. We utilized K-means and agglomerative clustering and reported the *MNI* (Normalized Mutual Information) and *ARI* (Adjusted Rand Index) metrics, which are traditionally used for assessing clustering quality when ground truth labels are available. In our context, we repurposed *MNI* and *ARI* to evaluate the effectiveness of annotations. Higher *MNI* and *ARI* values indicate that the assigned labels accurately capture article groupings, showcasing the higher quality of annotations.

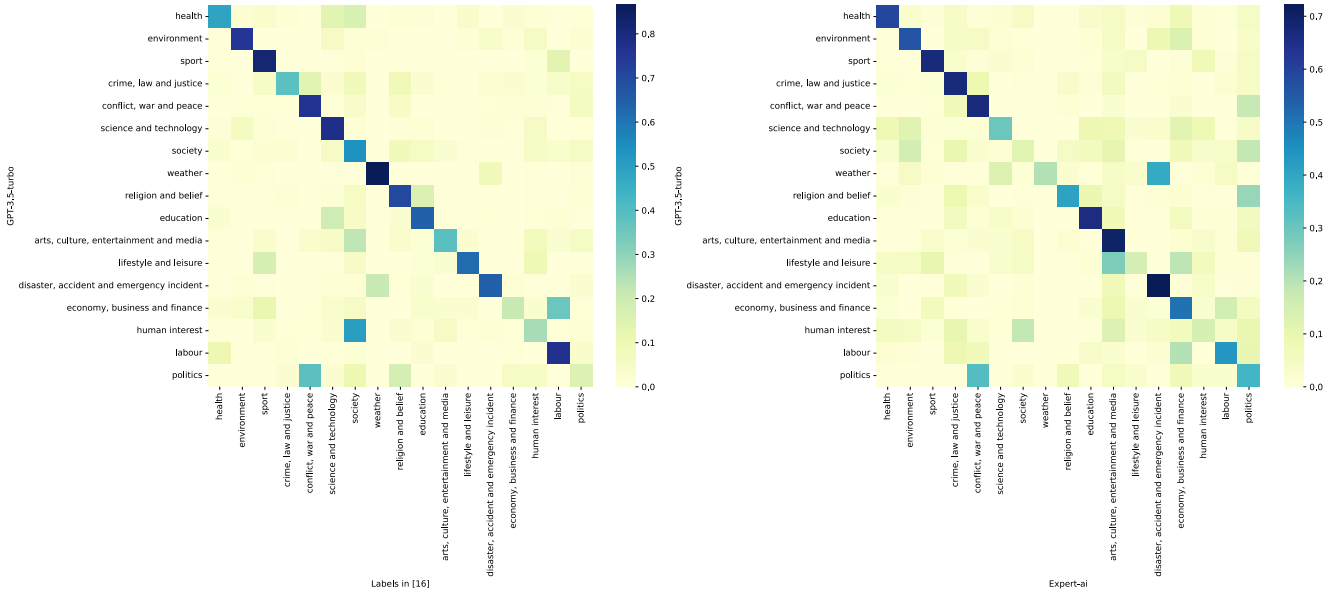
IV. EXPERIMENTS AND RESULTS

Gruppi et al. [7] presented a news dataset for the study of misinformation in news articles. Later on, Petukhova et al. [16] annotated 10,917 news articles from this dataset based on IPTC news codes. We used this dataset as the basis of our study. In addition to evaluating the performance of GPT in multi-class news classification, we compared our results with two other data annotations to which we had access. We conducted a comparison among the three data annotations, which included:

- We used the GPT-3.5 Turbo model to annotate each news article using a hierarchical classification strategy. Owing to limitations in the number of input tokens in the GPT-3.5 model, we had to exclude some articles from consideration.
- We used the annotations presented in [16]. The dataset contained several retired categories in the annotation, making it incompatible with the latest IPTC ontology; therefore, we had to remove these instances.
  Although it is stated that the annotation was conducted manually in a process involving keyword analysis and article reading, it does not appear to represent a set of ground truth labels. We encountered inconsistencies in annotations. For instance, an article titled "*Can Lactobacillus fermentum, a "good" bacteria, address glutathione deficiency?*" which discusses how taking a supplement called Lactobacillus fermentum can help increase the levels of an important antioxidant, is misclassified in the category "*environment*", while GPT more accurately identifies it as a topic related to "*health*". Such observations raise several questions regarding the specifics of their data annotation process and the annotators involved.
- We utilized Expert-ai,[3] a natural language processing platform that provides IPTC news classification models. There were instances of news articles where Expert-ai was unable to process and annotate certain articles effectively, therefore we had to remove them.

A major challenge encountered was obtaining a balanced dataset within the second-level categories, primarily due to the high number of categories at that level. Consequently, some second-level categories had very few articles, which led us to remove them. The final dataset consisted of 4,672 news articles covering 17 first-level categories and

[3]https://www.expert.ai/



(a) (b)

FIGURE 3. The overlap between the labels in [16] 3a, and Expert-ai labels 3b with GPT generated labels.

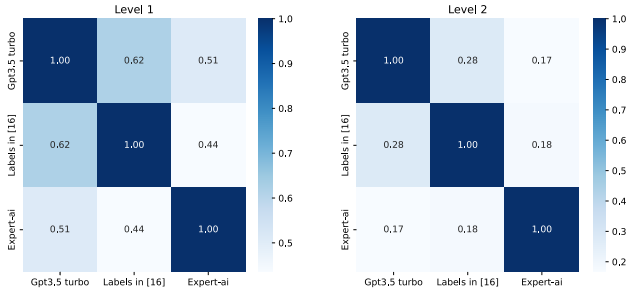


FIGURE 4. Percentage of agreement between the three label sources.

51 second-level categories. The dataset is available at https://zenodo.org/records/10058298.

Figure 3 visually represents the overlap or agreement between different label sources within first-level categories using two heatmaps. As illustrated, the GPT-generated labels are in higher agreement with the labels in [16] rather than the Expert-ai labels. Figure 4 shows the level of agreement between the three labeling mechanisms.

A. CLASSIFICATION RESULTS

We report the *precision*, *recall*, and *F1 scores* for all the models. *Precision* measures the accuracy of a model's positive predictions, *recall* quantifies the coverage capability, and the *F1 score* provides a balanced assessment of precision and recall. Tables 1 and 2 list the classification results.

As shown, GPT performs reasonably well in classifying articles into the Level-1 categories of the IPTC ontology, particularly achieving an average *F1 score* of 80% for BERT-based fine-tuned models. Moreover, it outperforms the baseline in the Level-1 classification task. This observation indicates its ability to handle high-level topic distinctions with a satisfactory level of robustness.

In Level-2 classification, GPT continues to outperform the baselines; however, it is noteworthy that the results showed decreased performance across all three cases. This could be attributed to the inherent complexity of distinguishing between more nuanced and, therefore, more similar sub-classes at Level-2. Additionally, the sharp class imbalance in Level-2 is a contributing factor, with some categories having a significantly larger number of articles, like "*crime*" with 200 articles, in contrast to others such as "*teachers*" with as few as 25 articles. Class imbalance within Level-2 categories is an important factor to consider. When a classification model exhibits a strong bias towards predicting the majority class, in this context, possibly "*crime*" due to the larger number of related articles, it achieves high precision for that class. High precision implies that the model accurately identifies a significant portion of "*crime*" articles. However, this bias negatively impacts recall, especially for minority classes like "*teachers*", as the model might overlook or misclassify a substantial number of "*teachers*" articles. Consequently, the recall of these minority classes tends to be low.

B. CLUSTERING RESULTS

In Tables 3 and 4, two scores *NMI* and *ARI* are reported, which are used to evaluate the performance of clustering algorithms by comparing the clustering results to ground truth labels:

- **NMI** measures the mutual information between the true class labels (if available) and the cluster assignments

TABLE 1. Level-1 classification results.

		TF-IDF			Glove			DistilBert/BertEmbedding		
	Model	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
GPT-3.5 Turbo	Multinomial NB	0.78	0.75	0.75	0.74	0.62	0.64	-	-	-
	Logistic Regression	0.78	0.73	0.75	0.80	0.77	0.78	-	-	-
	SVC Classifier	0.78	0.76	0.76	0.81	0.78	0.78	-	-	-
	Random Forest	0.79	0.72	0.73	0.81	0.75	0.77	-	-	-
	DistilBERT	-	-	-	-	-	-	0.81	0.77	0.78
	BERT	-	-	-	-	-	-	0.82	0.83	0.82
	Avg.	0.78	**0.74**	**0.75**	**0.79**	**0.73**	**0.74**	**0.82**	**0.80**	**0.80**
Labels in [16]	Multinomial NB	0.76	0.69	0.70	0.60	0.51	0.51	-	-	-
	Logistic Regression	0.79	0.69	0.71	0.74	0.70	0.71	-	-	-
	SVC Classifier	0.79	0.75	0.76	0.73	0.71	0.71	-	-	-
	Random Forest	0.79	0.72	0.74	0.73	0.66	0.67	-	-	-
	DistilBERT	-	-	-	-	-	-	0.75	0.74	0.74
	BERT	-	-	-	-	-	-	0.72	0.73	0.70
	Avg.	0.78	0.71	0.73	0.68	0.65	0.65	0.74	0.74	0.72
ExpertAI	Multinomial NB	0.58	0.54	0.52	0.45	0.43	0.42	-	-	-
	Logistic Regression	0.61	0.52	0.52	0.57	0.54	0.54	-	-	-
	SVC Classifier	0.59	0.58	0.57	0.54	0.54	0.53	-	-	-
	Random Forest	0.61	0.55	0.55	0.62	0.55	0.55	-	-	-
	DistilBERT	-	-	-	-	-	-	0.58	0.58	0.58
	BERT	-	-	-	-	-	-	0.57	0.61	0.56
	Avg.	0.60	0.55	0.54	0.55	0.52	0.51	0.58	0.60	0.57

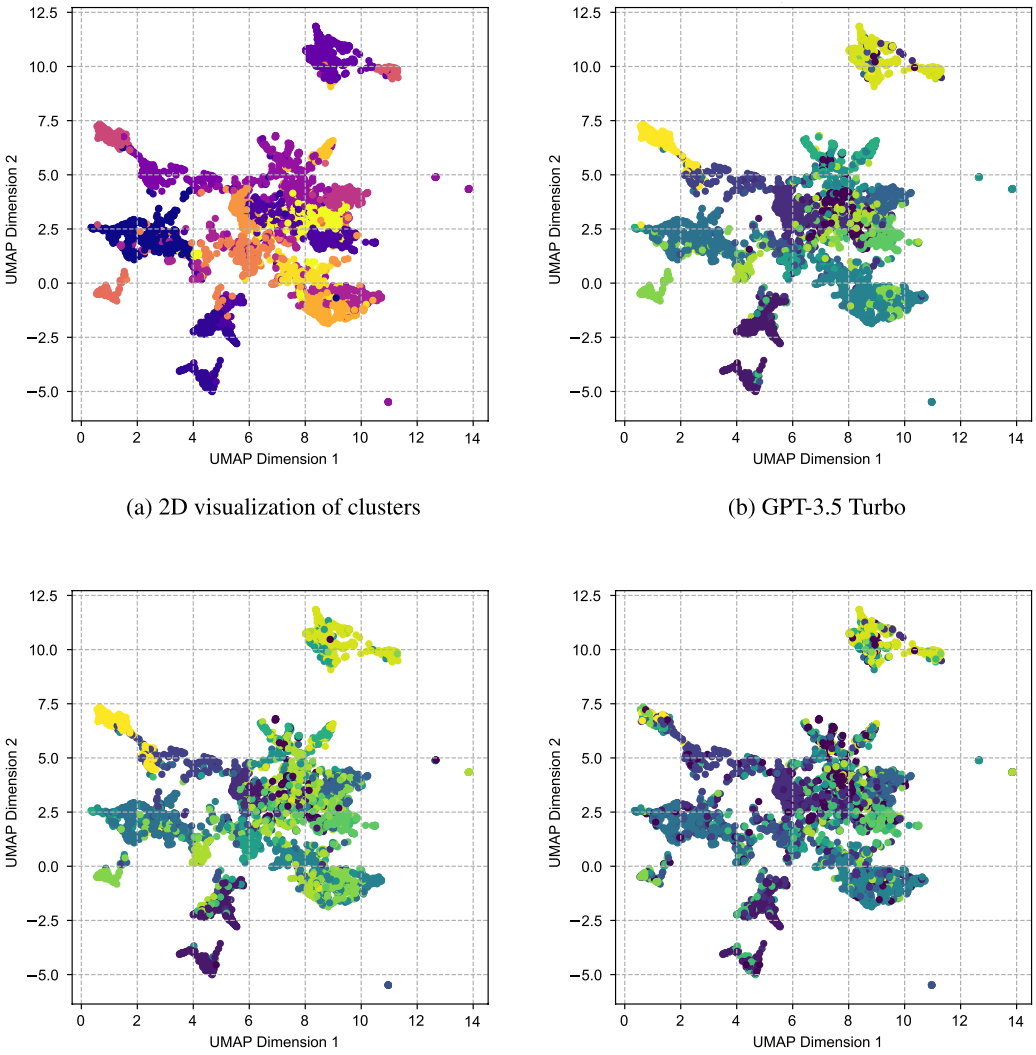
TABLE 2. Level-2 classification results.

	Model	TF-IDF			Glove			DistilBert Embedding		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
GPT-3.5 Turbo	Multinomial NB	0.64	0.52	0.53	0.52	0.41	0.40	-	-	-
	Logistic Regression	0.60	0.50	0.50	0.65	0.59	0.59	-	-	-
	SVC Classifier	0.65	0.59	0.59	0.62	0.59	0.59	-	-	-
	Random Forest	0.62	0.52	0.52	0.63	0.55	0.56	-	-	-
	DistilBERT	-	-	-	-	-	-	0.61	0.60	0.59
	Avg.	**0.63**	**0.53**	**0.54**	**0.61**	**0.54**	**0.54**	**0.61**	**0.60**	**0.59**
Labels in [16]	Multinomial NB	0.49	0.47	0.45	0.40	0.39	0.37	-	-	-
	Logistic Regression	50	0.49	0.47	0.49	0.49	0.47	-	-	-
	SVC Classifier	0.54	0.53	0.52	0.49	0.49	0.47	-	-	-
	Random Forest	0.47	0.48	0.46	0.44	0.43	0.42	-	-	-
	DistilBERT	-	-	-	-	-	-	0.51	0.50	0.49
	Avg.	0.50	0.49	0.48	0.46	0.45	0.43	0.51	0.50	0.49
ExpertAI	Multinomial NB	0.24	0.21	0.20	0.15	0.13	0.12	-	-	-
	Logistic Regression	0.21	0.19	0.18	0.24	0.22	0.22	-	-	-
	SVC Classifier	0.33	0.30	0.29	0.27	0.27	0.26	-	-	-
	Random Forest	0.32	0.24	0.25	0.31	0.24	0.25	-	-	-
	DistilBERT	-	-	-	-	-	-	0.28	0.27	0.27
	Avg.	0.28	0.24	0.23	0.24	0.22	0.21	0.28	0.27	0.27

and is normalized to provide a score between 0 and 1. It quantifies the degree of agreement between the true labels and clusters.

- **ARI** is a clustering evaluation metric that measures the agreement between clustering results and true labels while considering chance correction. It provides a score between -1 and 1, where 1 indicates perfect agreement, 0 implies random chance, and negative values suggest worse than chance agreement.

Higher *ARI* and *NMI* scores for the LLM-generated labels after clustering indicated that the clustering results obtained using this label-set were more consistent with the generated labels. This implies that these labels capture the underlying



(a) 2D visualization of clusters

(b) GPT-3.5 Turbo

(c) Labels in [16]

(d) Expert.ai

FIGURE 5. 2D visualization of the news articles color-coded by different label-sets.

TABLE 3. Level-1 clustering results.

	K-means Clustering		Agglomerative Clustering	
	NMI	ARI	NMI	ARI
GPT-3.5 Turbo	**0.58**	**0.44**	**0.57**	**0.42**
Labels in [16]	0.46	0.29	0.48	0.32
Expert.ai	0.37	0.25	0.36	0.22

TABLE 4. Level-2 clustering results.

	K-means Clustering		Agglomerative Clustering	
	NMI	ARI	NMI	ARI
GPT-3.5 Turbo	**0.55**	**0.27**	**0.54**	**0.26**
Labels in [16]	0.49	0.15	0.50	0.15
Expert.ai	0.39	0.18	0.37	0.17

structure and characteristics of the data better. The results derived from the clustering analysis showcase the satisfactory performance of GPT in the classification task.

To give a visual impression of the difference between the three label-sets, in Figure 5, the 2-dimensional presentations of the articles color-coded with their first-level IPTC labels is presented. Figure 5a shows the output of agglomerative clustering. As illustrated, in 5b the clusters of news articles exhibit more distinct and well-defined formations and look more similar to the actual clusters in 5a, which confirms that the GPT-generated labels are more reliable.

V. DISCUSSION AND FUTURE WORK

Here a few noteworthy points are to be highlighted following the evaluation of the model's performance:

Firstly, the utilization of fine-tuned large language models for various tasks, including news classification, offers several advantages:

- the cost associated with using LLMs such as GPT, in a production environment for processing vast news archives can be prohibitive. However, fine-tuning less expensive open source models such as LLaMA [22] is a cost-efficient alternative that can deliver comparable or even superior results.
- fine-tuning the language models in a few-shot setting reduces the need for prompt engineering. By using null prompts, that contain neither task-specific templates nor training examples, competitive accuracy can be achieved across a wide range of tasks [9]. This approach simplifies the process of fine-tuning and eliminates the need for extensive manual prompt tuning, thereby saving time and effort in the classification process.
- adopting fine-tuned LLMs will allow the creation of a fixed and robust model that remains consistent over time. This consistency is particularly crucial in the sensitive news domain, where reliability and predictability are paramount. Unlike generic LLMs which may undergo changes or updates, a fine-tuned model can maintain its stability, and ensure consistent and reliable news analysis. This feature is essential in scenarios where dynamic model behavior is undesirable, such as news content analysis.

Secondly, the task of analyzing news data extends beyond mere text comprehension and requires a deep understanding of the events being reported and their broader context. Large language models, while being adept at processing text, lack embedded knowledge of the historical and contextual factors shaping news stories. As an example, the GPT classified the article with the title "Syria's Post-War Reconstruction: 600 Establishments Resume Work at Aleppo Industrial City, Sheikh Najjar" into the "*economy, business and finance, economy*" category because it discusses the recovery and development of Aleppo's industrial city, investments, and various business establishments resuming their operations with limited emphasis on the conflict aspect due to mentions of war and sanctions. However, the lack of contextual information prevented it from being categorized under "*conflict, war, and peace, post-war reconstruction*" a more contextually accurate classification. Therefore, there is a growing need to enhance the functionality of LLMs by incorporating background information, thereby enabling them to trace the development of news over time. Contextual awareness not only empowers LLMs to categorize news articles more accurately but also offers a deeper comprehension of how various events interrelate within a larger narrative.

Lastly, our future research will involve a more comprehensive approach to classifying news articles, encompassing other levels of the IPTC ontology. This will allow us to analyze news content more effectively and accurately, enabling us to extract significant information about various events occurring worldwide. This would contribute to enhancing our knowledge and decision-making processes in various domains, such as conflict resolution, disaster management, public health and environmental policies.

VI. CONCLUSION

In conclusion, our study highlights the potential of large language models in news classification within the field of journalism and media research, addressing the need for automation in this task. The desirable performance demonstrated by these models suggests that they can alleviate the workload of journalists and media researchers by automating the time-consuming and labor-intensive classification process. By leveraging the capabilities of large language models, journalists can focus their efforts on more critical aspects of news reporting, such as investigative journalism and in-depth analysis. Moreover, our experiment, conducted in a "zero-shot" setting, emphasizes the versatility of these models, as they achieved impressive results without task-specific training. This underscores the potential usefulness of large language models in improving the efficiency and accuracy in journalism and media research, ultimately enabling journalists to deliver news more effectively to their audiences. As the field continues to evolve, the integration of large language models holds promise for revolutionizing the way news is analyzed and processed, meeting the growing demand for automated solutions in journalism.

REFERENCES

[1] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neur. Inf. Process. Sys.*, vol. 33, 2020, pp. 1877–1901.

[2] T. B. Brown et al., "Language models are few-shot learners," 2020, *arXiv:2005.14165*.

[3] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A survey on evaluation of large language models," 2023, *arXiv:2307.03109*.

[4] O. De Clercq, L. De Bruyne, and V. Hoste, "News topic classification as a first step towards diverse news recommendation," *Comput. Linguistics Netherlands J.*, vol. 10, pp. 37–55, Dec. 2020.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.

[6] T. Goyal, J. J. Li, and G. Durrett, "News summarization and evaluation in the era of GPT-3," 2022, *arXiv:2209.12356*.

[7] M. Gruppi, B. D. Horne, and S. Adali, "NELA-GT-2019: A large multi-labelled news dataset for the study of misinformation in news articles," 2020, *arXiv:2003.08444*.

[8] A. Hassan and A. Mahmood, "Convolutional recurrent deep learning model for sentence classification," *IEEE Access*, vol. 6, pp. 13949–13957, 2018.

[9] R. L. Logan, I. Balavzevi'c, E. Wallace, F. Petroni, S. Singh, and S. Riedel, "Cutting down on prompts and parameters: Simple few-shot learning with language models," in *Proc. Findings Assoc. Comput. Linguistics*, 2021, pp. 2824–2835.

[10] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, vol. 2. Valencia, Spain: Association for Computational Linguistics, 2017, pp. 427–431.

[11] Y. Luo, "Recurrent neural networks for classifying relations in clinical notes," *J. Biomed. Informat.*, vol. 72, pp. 85–95, Aug. 2017.

[12] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 1743–1753, Sep. 2023, doi: 10.1109/TAFFC.2022.3204972.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1–9.
[14] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 1532–1543.
[15] S. Petridis, N. Diakopoulos, K. Crowston, M. Hansen, K. Henderson, S. Jastrzebski, J. V. Nickerson, and L. B. Chilton, "AngleKindling: Supporting journalistic angle ideation with large language models," in *Proc. CHI Conf. Human Factors Comput. Syst.*, Apr. 2023, pp. 1–16.
[16] A. Petukhova and N. Fachada, "MN-DS: A multilabeled news dataset for news articles hierarchical classification," *Data*, vol. 8, no. 5, p. 74, Apr. 2023.
[17] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang, "Is ChatGPT a general-purpose natural language processing task solver?" 2023, *arXiv:2302.06476*.
[18] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, Tech. Rep., 2018. [Online]. Available: https://openai.com/research/language-unsupervised
[19] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, Jan. 2020.
[20] C. Rudnik, T. Ehrhart, O. Ferret, D. Teyssou, R. Troncy, and X. Tannier, "Searching news articles using an event knowledge graph leveraged by Wikidata," in *Proc. Companion World Wide Web Conf.* New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1232–1239.
[21] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.
[22] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.
[23] G. Wang, C. Li, W. Wang, Y. Zhang, D. Shen, X. Zhang, R. Henao, and L. Carin, "Joint embedding of words and labels for text classification," 2018, *arXiv:1805.04174*.
[24] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, and G. Wang, "GPT-NER: Named entity recognition via large language models," 2023, *arXiv:2304.10428*.
[25] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," 2021, *arXiv:2109.01652*.
[26] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2020, pp. 38–45.
[27] K.-C. Yang and F. Menczer, "Large language models can rate news outlet credibility," 2023, *arXiv:2304.00228*.
[28] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto, "Benchmarking large language models for news summarization," 2023, *arXiv:2301.13848*.

BAHAREH FATEMI received the master's degree in information technology engineering (network science) from the University of Tehran, Teheran, Iran. She is currently pursuing the Ph.D. degree with the Department of Information Science and Media Studies, University of Bergen, Norway. Her research interests include machine learning, textual content analysis, and network science.

FAZLE RABBI received the Doctor of Philosophy (Ph.D.) degree in software engineering from the University of Oslo. He is currently an Associate Professor with the University of Bergen, Norway. He has long and varied experience with software development in smaller and larger projects within a large spectrum of domain areas and technological solutions. His research interests include model-based software engineering, data mining, and machine learning, with emphasis on addressing the information science problems in the society. His research portfolio includes software engineering related research: workflow modeling and its verification, metamodeling, building decision support systems, multi-agent systems, and process engineering.

ANDREAS L. OPDAHL received the Ph.D. degree from the Norwegian University of Science and Technology (NTNU), in 1992. He is currently a Professor in information systems development with the University of Bergen, Norway, where he heads the Research Group for Intelligent Information Systems (I2S). His research interests include ontologies and knowledge graphs, enterprise, and IS modeling and their applications to media production. He is the author, the coauthor, or a co-editor of more than a 100 peer-reviewed and widely cited research papers. He is a member of IFIP WG5.8 on Enterprise Interoperability and WG8.1 on Design and Evaluation of Information Systems. He serves as an Associate Editor or renowned international journals and as an organizer of renowned international conferences and workshops.