FACTA UNIVERSITATIS
Series: **Electronics and Energetics** Vol. 38, Nº 4, December 2025, pp. 697 - 714
https://doi.org/10.2298/FUEE2504697K

Original scientific paper

A STRUCTURE BASED ON TROCR TRANSFORMER AND LARGE LANGUAGE MODEL FOR CLASSIFICATION OF HANDWRITTEN TEXTS

Hossein KardanMoghaddam, Adel Akbarimajd, Mohammad Ranjbarpour, Mahdi Nooshyar, Shahram Jamali

Department of Electrical and Computer Engineering, University of Mohaghegh Ardabili, Ardabili, Iran

ORCID iDs:	Hossein KardanMoghaddam	https://orcid.org/0000-0002-9304-5093
	Adel Akbarimajd	https://orcid.org/0000-0002-7019-9655
	Mohammad Ranjbarpour	https://orcid.org/0009-0004-4522-2421
	Mahdi Nooshyar	https://orcid.org/0000-0002-6786-7763
	Shahram Jamali	https://orcid.org/0000-0003-2764-6373

Abstract. *Processing handwritten texts and classification and their content analysis are among the most important problems in the realm of text analysis. Microsoft has presented pre-trained TrOCR models for printed and hand-written texts. These models due to prior pre-training are better starting point for image processing. For using TrOCR with the aim of detecting printed and handwritten texts, we can use fine-tuning technique on pre-trained model using different datasets. This process helps the model to learn better the specific features of image processing and hand-written or semi-handwritten texts. TrOCR uses transformer models for OCR and its fine-tuning on special datasets especially, hand-written datasets is a common task. TrOCR model from Microsoft extracts text from these images, and in this research a structure based on TrOCR and LLM has been proposed whose aim is extraction of hand-written texts from existing images in a dataset (English handwritten line dataset) and converting them to text data and then this data has been given to LLM as an input so that the extracted texts can be classified (using BART model) based on different subjects and contents.*

Key words: *Large Language Model, Fine-tuning, TrOCR, Neural Network, Deep Learning*

1. Introduction

Natural language processing is one of the significant subfields in artificial intelligence and also in linguistic field. Handwritten Recognition (HWR) is the ability of computer for receiving and interpreting conceivable handwritten input from sources like paper

Received December 28, 2024; revised April 03, 2025 and August 07, 2025; accepted August 10, 2025
Corresponding author: Hossein KardanMoghaddam
Department of Electrical and Computer Engineering, University of Mohaghegh Ardabili, Ardabili, Iran
E-mail: kardanmoghaddam@uma.ac.ir

documents, image, touch screen and other devices. Text detection is an old research problem for digitalization of documents. Natural language processing is a field of artificial intelligence uses linguistic and computational techniques for helping computers in perceiving and producing human languages in written and spoken forms [1]. Among the most important issues in the realm of natural language processing are automatic translation, question and answering systems, retrieving and searching information, text summarization and emotion analysis [2][3]. In recent years, deep learning models have been presented for detecting and recognizing text contents that compared with traditional models and machine learning have higher power meaning modeling and word sequence [1][4][5]. Deep learning models can extract hidden and inherent details of meaning from a sequence of deep neural network layers and during decision making act on the basis of text meaning [6][7]. Deep models using deep neural structures and layers inside them can learn automatically desirable features in applied realm and usually obtain better results. In natural language processing, like transformer based models and deep learning techniques such as BERT and GPT-3 the capacity of condense text automatically has been improved considerably [8][9][10]. However, the challenges of perceiving complex occasions of text and correcting biases in textual data are continuous [11-15]. Nowadays because of high density of scanned documents and information, one of the main functions of natural language processing is retrieving these data to the computer documents and classifying this information and searching among them. Companies and organizations using text classification can automatically organize and order any kind of text such as e-mail, legal documents, social media, chat bots, surveys and other cases in a fast and economical way. Text classification is one of machine learning methods allocates a collection of pre-defined classes to the text. Text classification is one of the fundamental duties in natural language process with widespread usages like sentiment analysis, topic labeling and spam detection. Machine learning-based text classification learns to classify the text based on past observations. Using pre-labeled samples as training data, machine learning algorithms can learn different relations among the texts and for each unique input, they allocate a special label. The "label" is the class or desired set that each input text can be placed in a special set. Converting text data into quantitative data for obtaining practical insight and scientific and business decisions is very useful. This paper presents a new Transformer-based structure and integrates it with large language models (combining TrOCR with an LLM) to improve the accuracy of handwritten text recognition and classification, yielding better results than traditional methods.

2. Research Background

In this section, we will examine some proposed research of different researchers based on deep learning presented for detecting and evaluating in texts: In research (Dhivyaa et al.) [16], fine-tuned Convolutional Neural Network has been proposed for detecting Tmail handwritten text. The proposed approach includes using pre-trained Convolutional Neural Network models and their exact adjustment on collection of Tmail handwritten characters. In research (Daniel Parres et al) [17], pre-trained visual encoder-decoder transformers have been used for detecting handwritten text in historical documents. In research (Mehar Prateek Kalra et al) [18] examined how to combine Handwritten Text Recognition (HTR) systems with LLM. Given the impressive capabilities of LLMs in language understanding and processing, this integration could be especially beneficial in fields that demand high accuracy in text recognition. In research (Jan Kohút et al) [19], the effect of fine-tuning methods on handwriting recognition is investigated. The research emphasizes that fine-tuning is a simple and effective method for domain adaptation in handwritten text recognition and can significantly improve handwriting recognition systems' performance. Kumar and Raman [20] has used BERT model as a placement method and in next layers a bi-channel structure that the first channel contains convolution layers and the second channel contains BiLSTM. In their proposed method, convolution layer has been used for modeling local features of text, and BiLSTM has been used for modeling word sequence. Badpeima et al [21] have used combined Long-Short Term Memory recurrent network for extracting features and support vector machines for classifying Persian texts in three groups of positive, negative and neutral. In this model, after pre-processing and extracting word bank, the raw text has been used as input for recurrent networks. In the following, vectors resulted from the network output has been averaged and finally have been classified by support vector machine. Onan [22] analyzed review database containing 66000 comments using machine learning methods, Ensemble Learning and deep learning methods. In research [23], a multimodal fusion neural network with dual-attention based on textual double-embedding networks for rumor detection has been proposed consider both images and texts and comparing experiments have been carried out on two sets of Weibo and Twitter. In the research (Majma et al) [24] using blocking criteria of text and cosine similarity, copying in scientific texts has been detected. In the research (Behzadidoost et al) [25] a deep learning model based on granular computation has been proposed for text classification. In the research (Feizi-Derakhshi et al) [7] it has been proposed that convolutional neural network and bidirectional long short term memory neural network with attention layer used for classification of Persian texts.

3. Modern Deep Learning Methods in the Realm of Text Recognition

Modern Deep learning methods in the realm of text recognition (OCR-optical character recognition) because of processing ability and detecting convoluted texts in images or videos are very important. The first method is CRNN (convolutional recurrent neural network) is a fusion of convolutional neural network for extracting visual features and recurrent neural networks for modeling time sequencing. Its usage is detecting sequence texts like handwritten or printed ones in the images [26]. The second method is transformer-based OCR (TrOCR) from Microsoft uses Vision Transformer (ViT) and linguistic transformer for text recognition and its main usage is detecting printed and handwritten texts with high accuracy [27]. The third method is EAST (Efficient and Accurate Scene Text Detector) which is an efficient model detecting textual areas in images, without requiring exact classification. Its usage fused with OCR for text recognition in convoluted images, is like scanned documents [28]. The fourth method, is the CTC (Connectionist Temporal Classification) which is a deep learning algorithm designed for modeling issues related to sequence data without requiring exact alignment between input and output. This structure is usually applied in cases like speech recognition, text recognition (OCR) and sequence analysis [29]. The fifth method, is the pre-trained models that in this method deep learning models are used previously trained on big data sets and ready for direct usage and some cases are:

- EasyOCR: A library with support for multiple languages [30].
- PaddleOCR: Comprehensive framework with the capability of text recognition in multilingual images [31].
- Keras OCR: open text framework for teaching and using OCR models [32].

The sixth method is Vision-Language Models for OCR that this method is set of deep learning methods fusing the computer visual capabilities and natural language processing for recognition, interpretation and comprehension of text from images. These models are more advanced than traditional OCR Methods and are used in complex cases such as text recognition in multi-page documents, chart analysis or images with disordered or sparse texts [33]. These models comprise two main parts:

- Vision Module: this module is responsible for analysis of visual aspects of an image. Most of the time, computer vision advanced structures like Vision Transformer (ViT), ResNet or CNN are used for extracting visual features.
- Language Module: this part processes extracted text features and interprets their language content. Language structures like Transformer or BERT is used for text production by language structure.

Among the prominent cases of Vision-Language Models for OCR we can name Donut (Document Understanding Transformer) [33] and TrOCR [27] and LayoutLM [34].

4. The Advantages of Pre-trained Models in TrOCR

The pre-trained models in TrOCR propose some key advantages for optical character recognition that considerably improves performance compared with traditional methods:

4-1. Lowering need for labeled data: pre-trained models are trained with big and general data (labeled and without label). This method lowers the cost and needing to get high density of labeled data and also provides high accuracy level. Pre-trained level data include large sets of handwritten or printed texts made with artificial data production tools.

4-2. Better Generalization of TrOCR: It uses extracted general features in pre-trained stage and can be adjusted for more special data. This feature causes that the model can act with higher accuracy on manifold languages, fonts and text structures.

4-3. Increases speed of fine-tuning: Using pre-trained models lowers time and computation resources required for adjustment with more special tasks as the model has previously learnt general knowledge of the text and visual features.

4-4. Advanced Performance in Multilingual texts Recognition: Transformer based structure in TrOCR like using models like BEiT and RoBERTa, improves recognition of different lingual structures and makes it suitable for multilingual usages.

4-5. Higher Quality in Convoluted Text Recognition: the structure of Vision Encoder and Language Decoder in TrOCR provides the possibility of exact extraction of visual features (like word form and arrangement) and their combination with lingual information. This feature is very effective especially in cases with noised or low quality images.

4-6. Transfer Learning: Using known models like BEiT and RoBERTa for Encoder and Decoder parts contributes on transferring knowledge from similar ranges and improves the performance.

4-7. Compatibility with Modern Structure: Using Structural benefits of Transformer like Self-Attention and Multi Head Attention, increases precision and speed of processing.

These benefits have caused that TrOCR be considered as one of the best existing models in printed, handwritten and convoluted text recognition tasks [27].

5. LARGE LANGUAGE MODELS

Large language models (LLM) are one of the latest advancements in artificial intelligence having the required ability in producing a text with high quality and precision. These models use convoluted structures and deep neural networks that enables them to produce a coherent text compatible with a realm of subjects. Using these models considered as deep learning algorithms, machines can understand the human's talking and produce something like that. In this way, they can arrange a dialogue with a person and this dialogue can be very beneficial. These models can be considered as a very smart and talented person that spends all of his time on studying and learning from different sources and now s/he can answer the person's questions in different realms and in many cases helps the person. Gradually these models learn new information from this process of asking and answering, and gives better answers to the person. The basic structure of LLM has been made based on transformers and this advancement was proposed in 2017 and due to its unique efficiency, it has been converted to natural language processing milestone at high speed [35]. Transformer model is considered as one of the most important elements of LLM that can be trained with high amount of manifold data. Using these large datasets causes that we call them large data. Some of the key usages of LLM Models include automatic content producing, machine translation, natural language processing and answering systems to questions. LLM can analyze huge amount of data, learn lingual patterns enabling them to produce logical and meaningful sentences often cannot be recognized from the handwritten text [36]. Some of the significant cases of LLM are GPT [37] developed by OpenAI and BERT [38] and T5 [39] developed by google. These models have learnt convoluted lingual patterns through analysis of large amount of data that enables them to produce sentences which are both logical and rich-content and their output differentiation challenges handwritten text. The main benefit of these models is their unique capability in comprehending and producing a high quality text in multiple languages and also presenting exact answering to questions with different complexities. Therefore, LLM have widespread applications in the realms like machine translation, chat robots, automatic content producing and even advice systems. According to these capabilities, LLMs are not only powerful devices for natural language processing, but they have been considered as key elements in developing AI-based technologies. For using a LLM for different tasks, a common approach, fine-tuning a pre-trained model on specific task data has been proposed [36][38]. This fine-tuning is vital for optimizing the performance of LLM in special programs. The exact fine-tuning a lingual model can be computationally intensive, usually requires updating all parameters in pre-trained model and the fine-tuned model can have parameters the same as the key model [40].

5.1 Fine-tuning Language Models

Training a LLM at first requires time and considerable financial resources. Using thousands of GPU can last several days [41] and requires considerable financial investment [42]. Fine-tuning pre-trained models has been emerged as an efficient method for obtaining LLM benefits. For starting fine-tuning process, at first a pre-trained Language model like BERT, GPT or LLaMA is chosen, then data related to task or required range are gathered. These data can include texts, questions and answers or any other kind of information and then model parameters are fine-tuned to be compatible with new data. This includes choosing learning rate, the number of training courses and other fine-tunings and after that the model is trained using special data. This stage includes several training courses so that the model can be compatible with data so well and after that the model performance on experimental data is evaluated and when required, more fine-tuning is performed [43]. Fine-tuning is the process of compatibility of a pre-trained model with a special task by training it on data related to task, and finally improves it. This approach has been accepted widely, as this allows the researchers to use pre-trained models with general aim and design them for meeting special needs. Many organizations like Meta (Facebook) with their LLaMA [44] make their pre-trained models accessible to the public. This pre-trained models accessible to the public can be fine-tuned with different downstream tasks and using fine-tuning convert them to the most practical way for using LLM benefits. However, the complete fine-tuning of LLM is computationally expensive, as it needs updating all model parameters.

6. Used Dataset in this Research

The dataset used in this study is known as the English Handwritten Line Dataset. It comprises approximately 400 images of handwritten lines in English, which are utilized to train and evaluate machine learning and computer vision models related to Handwritten Text Recognition. This dataset features images of handwritten text created by different individuals, each accompanied by labels corresponding to the content of the images. The dataset is particularly effective for training deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) like LSTM (Long Short-Term Memory), to transform handwritten text into digital text. By leveraging this dataset, models can be developed to comprehend handwritten text and apply it in Natural Language Processing (NLP) tasks such as translation, summarization, and text analysis. Furthermore, this dataset serves to enhance Optical Character Recognition (OCR) systems that convert handwritten text into digital formats. It is widely regarded as a standard data source for both training and evaluating various machine learning and deep learning models within the field of handwritten text recognition. Each image in the dataset is paired with a text file that contains the corresponding text, which can be used to train models effectively. The diversity of examples within this dataset aids in improving the accuracy and efficiency of handwritten text recognition systems. An example of these images can be seen in Figure 1.

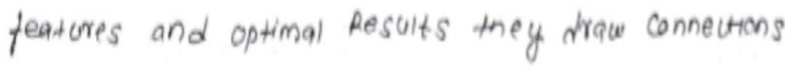

Fig. 1 an Example Image of the English Handwritten Line Dataset

This dataset can be combined with the IAM dataset [45] to create a real-time handwriting recognition system. To use this dataset, you can visit the Kaggle site or GitHub by following the links [46] and [47].

7-The Proposed Structure in this Research

The methods proposed so far about printed and handwritten text content typically have used traditional methods and LLM and TrOCR has not been used in combination. Meanwhile by developing Deep learning methods and using pre-trained models and also using LLMs speed and productivity have been progressed in many cases. Nowadays most of the texts are obtained in film-taking mode that are required to be processed and even corrected before entering LLM so that they can be categorized correctly. In this research a structure based on TrOCR and LLM has been proposed that the main aim of this research is extracting text from existing images in an English handwritten line dataset by using pre-trained models. Table 1 outlines the pre-trained models utilized in this study along with their respective characteristics.

Table 1 Pre-Trained Models TrOCR

Pre-trained Models	Features
Microsoft/TrOCR-base-handwritten	Based on “Base” Transformer structure especially has been designed for handwritten text and optimized for recognition of text in images containing handwritten texts
Microsoft/TrOCR-small-handwritten	The smaller version of TrOCR(Small)with less computational amount is suitable for systems with limited computational resources with acceptable performance in handwritten recognition
Microsoft/TrOCR-large-handwritten	The large version with high capacity for better learning of convoluted patterns, is suitable for cumbersome tasks requiring high precision like processing sensitive or complicated documents, suitable for use in research projects, reading handwritten texts with complicated details and organizational and industrial usages requiring high precision

These pre-trained models by Microsoft have been trained using general OCR data and handwritten texts and after that the obtained information has been saved on a CSV file then this CSV file contains obtained texts from images is given to LLM as an input so that the extracted texts are categorized into different subjects (“Health”, “Technology”, “Finance”, “Education”, “Sports”, “Others”) based on the content. The order of proposed stages and structure used has been illustrated in Figure 2.

As we can see in figure 2, after uploading OCR model and entering dataset images and performing elementary steps, data are given to the transformer model as input. Transformer-based models include two parts: encoder, decoder. The encoder part processes the input and converts it to a rich vector representation. In this case, the input is given to the model as a text or image and at first the words or the elements are converted to the embedding vectors. Vectors cross self-attention layers where the model learns the relation between each word and element with other words and elements and then encoder output is a compressed representation of total input sent to decoder. The next part is decoder uses output vector of encoder and converts it to final output like text, translation or any other format. Unlike RNNs that process data in ordinal manner, transformers can process data in parallel manner leading to high speed computations. Transformers due to using self-Attention and parallel processing can model long-term relationships in ordinal data so well. This capability has made them more efficient for many Deep learning tasks [35]. Transformer-based Optical Character is a modern method for recognition of optical characters based in Transformer models. This model has been developed by Microsoft research team and uses transformer structure for extracting text from the images [27]. The architecture of this method can be observed in Figure 3.

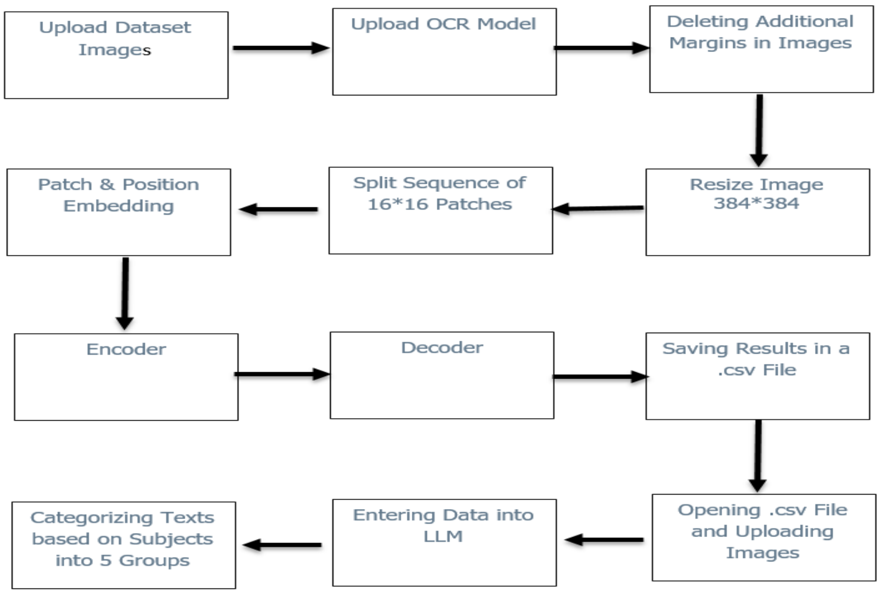

Fig. 2 Proposed Structure

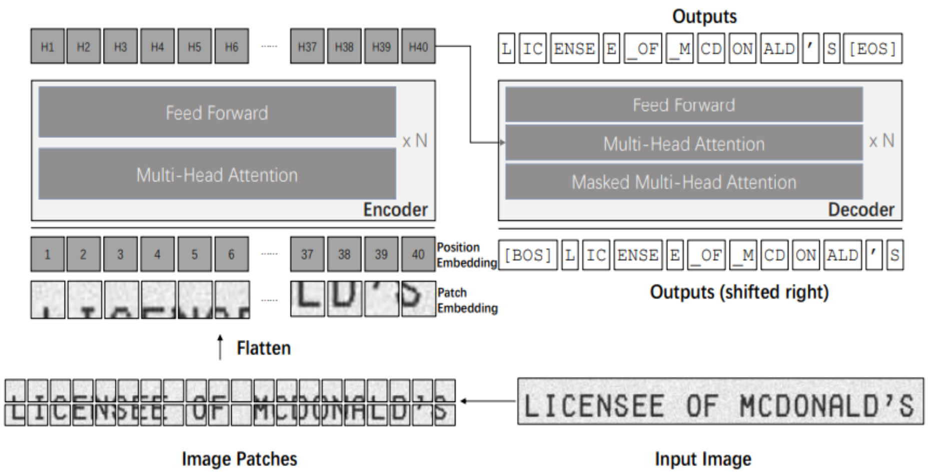

Fig. 3 The Architecture of TrOCR [27]

The TrOCR is easy but effective and it can be fine-tuned with artificial data in pre-trained large scale with labeled dataset by human. In research (Li et al) [27] is was illustrated that TrOCR model has performed better in printed and handwritten text recognition tasks than advanced models. Microsoft has presented pre-trained TrOCR models for printed and handwritten texts. These models due to early pre-training, are better starting point for processing the images. In this research Transformer OCR (TrOCR) has been used for recognition of handwritten texts from images developed by Microsoft uploaded in two parts:

- TrOCR Processor: The task of input processing (images) and transforming them into a suitable format for the model.
- Vision Encoder Decoder Model: Deep learning model that recognizes the handwritten text from the images and produces a text according to them.

After using TrOCR for converting image to text, the obtained texts have been given to a LLM as an input. The basic architecture of LLMs has been made on the basis of transformers, this progression was introduced in 2017 and because of its unique efficiency and it has been turned into a milestone of natural language processing [35]. Large language model of Facebook/BART-large-cnn used in this research is one of the advanced models of natural language processing developed by Facebook. This model has been made based on BART (Bidirectional Encoder Representations from Transformers) and is used for tasks like text summarization, machine translation and natural language comprehension. The BART model has been trained by a self-supervised pre-training. In this method, the model is trained on a large dataset from unlabeled texts. This process helps the model to comprehend the language structure and relation between the words. BART is a combination of bidirectional models (like BERT) and self-regression (like GPT). BART uses attention mechanism for text processing and helps it to comprehend the relation between the words better [48] [49]. The input texts to LLM has been categorized in pre-determined categories (like (“Health”, “Technology”, “Finance”, “Education”, “Sports”, “Others”).

8. RESULTS

In the classification discussion a dataset using classification methods, is the aim of achieving the highest possible precision and accuracy in classification and recognition of sets. In some problems, the correct recognition of samples related to one of the sets is very important for us. Confusion table or matrix represents the results of classification based on real existing information. Now on the basis of these values we can classify different evaluating criteria and define precision measurement. The confusion matrix is one of the most important tools for evaluating multi-class machine learning models. It serves as a method to assess the performance of classification models by comparing their predictions with the actual values. The matrix highlights where the model has made errors, offering insights into how its performance can be improved. Additionally, by calculating metrics such as Precision, Recall, and F1-score, we can obtain a comprehensive evaluation of the model's performance for each class. Here we examine three cases based on pre-trained model of TrOCR and LLM. The confusion matrix has been used for a case of pre-trained TrOCR model Microsoft/trocr-base-handwritten and LLM also has been used for text classification as Facebook/BART-large-cnn. The obtained results have been presented in table 2.

Table 2 Confusion Matrix for the case where the microsoft/trocr-base-handwritten model and LLM are used

		Predicted class					
		Technology	Health	Education	Finance	Sports	Others
Actual class	Technology	41	5	2	1	1	39
	Health	2	1	0	0	0	12
	Education	0	0	0	0	0	6
	Finance	0	0	0	0	0	3
	Sports	0	0	0	0	1	1
	Others	22	20	4	0	2	173

Now we obtain the values of precision, recall, accuracy, f1-score for the above table.

$$Precision_i = \frac{M_{ii}}{\sum_j M_{ji}} \tag{1}$$

$$Recall_i = \frac{M_{ii}}{\sum_j M_{ij}} \tag{2}$$

$$f1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

That the results obtained from table 2 has been conformed to table 3.

Table 3 The results of text classification by the model Microsoft/trocr-base-handwritten and LLM

	Technology	Health	Education	Finance	Sports	Others
Precision	0.63	0.0384	0	0	0.25	0.739
Recall	0.460	0.066	0	0	0.5	0.782
f1-score	0.5317	0.04855	0	0	0.3333	0.7598
Accuracy			0.6428			

In the second case, for the case of pre-trained model of TrOCR and LLM has been according to table 4.

Table 4 Pre-trained TrOCR and LLM model used in the second case

Pre-trained model TrOCR	LLM used
microsoft/trocr-large-handwritten	Facebook/BART-large-cnn

The confusion matrix has been obtained as table 5.

Table 5 Confusion Matrix for the case where the microsoft/trocr-large-handwritten model and LLM are used

		Predicted class					
		Technology	Health	Education	Finance	Sports	Others
Actual class	Technology	52	4	2	0	0	31
	Health	3	2	0	0	0	10
	Education	0	1	0	0	0	5
	Finance	0	0	0	2	0	1
	Sports	0	0	0	0	1	1
	Others	36	12	4	1	3	165

Now we obtain the values of precision, recall, accuracy, f1-score for table 5 that the results have been presented in table 6.

Table 6 The results of text classification by the model Microsoft/trocr-large-handwritten and LLM

	Technology	Health	Education	Finance	Sports	Others
Precision	0.571	0.105	0	0.666	0.25	0.774
Recall	0.584	0.133	0	0.666	0.5	0.7466
f1-score	0.5774	0.1173	0	0.666	0.3333	0.760
Accuracy			0.6529			

In the third case, for the case of pre-trained model of TrOCR and LLM has been according to table 7.

Table 7 Pre-trained TrOCR and LLM model used in the third case

Pre-trained model TrOCR	LLM used
microsoft/trocr-small-handwritten	Facebook/BART-large-cnn

The confusion matrix has been obtained as table 8.

Table 8 Confusion Matrix for the case where the microsoft/trocr-small-handwritten model and LLM are used

		Predicted class					
		Technology	Health	Education	Finance	Sports	Others
Actual class	Technology	28	3	6	1	3	48
	Health	1	2	0	0	0	12
	Education	2	1	0	0	0	3
	Finance	0	0	0	0	0	3
	Sports	0	0	0	0	0	2
	Others	22	13	8	5	5	168

Now we obtain the values of precision, recall, accuracy, f1-score for table 8 that the results have been presented in table 9.

Table 9 The results of text classification by the model Microsoft/trocr-small-handwritten and LLM

	Technology	Health	Education	Finance	Sports	Others
Precision	0.5283	0.105	0	0	0	0.7118
Recall	0.314	0.133	0	0	0	0.7601
f1-score	0.3938	0.1173	0	0	0	0.7351
Accuracy			0.5823			

9. Discussion

Handwritten text recognition and classification has long been a significant challenge in the fields of image processing and natural language processing (NLP). This difficulty arises from the variety of writing styles, the differences in letter shapes, and the noise present in images. Handwritten text recognition has been progressed considerably in recent years and this issue is mainly because of combining Deep learning techniques. Modern Deep learning methods like CRNN, TrOCR, EAST, and CTC in combination with pre-made models and transformer architecture, have provided high precision and efficiency in recognition of texts from images. Among the famous models based on transform we can name GPT (Generative Pre-trained Transformer) [50] for producing text, model of BERT (Bidirectional Encoder Representations from Transformers) [38] for language comprehension and T5 model [39] for transforming text into text (for translation, summary and …) and model of Vision Transformer (ViT) [51] for image processing and model of TrOCR [27] for handwritten text recognition from the images. This paper introduces a new framework built on advanced models designed to enhance accuracy and efficiency in the field. The importance of this research lies in its integration of two cutting-edge technologies from natural language processing and computer vision. By combining TrOCR with large language models, we can achieve greater accuracy and efficiency in classifying handwritten text. This integration streamlines both the complex pre-processing and post-processing steps, minimizing the need for separate models. In this research, pre-trained model of TrOCR has been used for image processing and text extraction. TrOCR can be combined with advanced tools like language recognition models, image processing tools or large language modes to present high performance. The TrOCR model is based on Transformer architecture and is specifically designed for recognizing text from images. By integrating image processing with natural language processing capabilities, this model offers enhanced performance in recognizing handwritten text. The application of this model in this paper reflects the latest technological advancements in this field. TrOCR is usually pre-trained on large scale data (like digital documents or images containing text) and also there are different versions of TrOCR for recognition of printed and handwritten texts that in this research as the used information bank was in handwritten format we just have used handwritten versions of TrOCR. This research uses Large Language Models to improve text classification. These models can help increase the accuracy of handwritten text recognition and classification by better understanding the meaning of the text and its context. Combining TrOCR with LLM is an innovative approach that can provide better results than traditional methods. Overall, the use of this structure can lead to significant improvements in handwritten text recognition and classification systems. In this study, the initial stage involved using a database that includes a text file (in .txt format) which can be downloaded alongside the image database. This text file contains the content for all the images. To classify the texts, we first performed text classification into six categories: "Technology," "Health," "Education," "Finance," "Sports," and "Others." This classification was carried out separately on the text file using a large language model (LLM). The results of this classification served as the evaluation criterion for categorizing images in the subsequent stages. In the next step, the TrOCR model was utilized to load images and extract text from the English handwritten line dataset. These images were then converted into text data. Following this, the extracted text data was analyzed for classification using the BART model, which is based on a large language model (LLM). The output of this classification process was saved as a .CSV file. The output file from cases where images were converted to text using the TrOCR model and subsequently classified with LLM was compared to instances where text files (.txt) related to the dataset were classified using LLM (first stage). The level of agreement in text classification between the two methods was then evaluated. When comparing the results of two methods—the first using the TrOCR model to convert images to text and subsequently classifying that text with a language model (LLM), and the second using a text file with a database for classification with the LLM—there is a noticeable difference of about 30 percent. This discrepancy is expected, as the database is based on handwritten text and the output from TrOCR remains unaltered. In this research the extracted texts from images by TrOCR haven't been corrected and upgraded and this has lowered the text recognition and consequently decreased the output of classification of texts. If the obtained texts from images are corrected and upgraded before entering LLM, it will have better result in text classification and this can be done by fine-tuned models of LLM and other new methods like VAE and other techniques. In many cases, when reading texts indie images and transformed into text (.txt), the correction is required as some characters or words may lose their forms or even some of the letters or words may be deleted to added or deformed and require correction that these are performed by generative models in NLP and these cases are suggested as future researches in this realm along with combination with other image processing techniques. One of the techniques suggested as an efficient research in the future is using Variational Autoencoder(VAE) for making deformed characters and words in the text. VAE is a kind of neural network used for making new data similar to training data. By using VAE we can use text for making new images in order to improve the performance of TrOCR. We can also use VAE for decreasing the dimensions of the images and upgrading data quality before entering TrOCR model and in some cases, VAE can be used with TrOCR as one part of combinational architecture, in a manner that extracted features of VAE in TrOCR are used. This paper signifies a major breakthrough in the fields of image and natural language processing. The employment of Transformer-based architectures, combined with the integration of large language models, represents an important advancement toward improving AI systems for more complex applications. This paper can serve as a reference for future research in handwritten text recognition and the integration of image processing models with large language models. Additionally, the methods presented may inspire developers to create similar systems in other domains.

10. CONCLUSIONS

In their study, Chae and Davidson (2023) [52] highlight that large language models (LLMs) excel in text classification tasks, significantly outperforming traditional methods. They note that these models can achieve high accuracy with minimal training examples and instructions, making them valuable tools for sociologists. Furthermore, the study suggests that fine-tuning smaller models provides an optimal solution for researchers. Wang et al. (2023) [53] analyzed the performance of GPT models in text classification. The researchers compared the capabilities of zero-shot LLMs with other advanced text classification methods, including traditional machine learning and deep learning techniques. The experimental results demonstrated that LLMs can effectively function as zero-shot text classifiers, showing strong performance in three out of four datasets examined in the study. In research conducted by Zhong et al. (2024) [54], a semi-supervised learning framework was introduced specifically for text classification tasks. This study solely focuses on text classification and does not involve image processing or image-to-text conversion. The framework was tested on two datasets: Reuters 20 Newsgroups and Web of Science, achieving accuracy rates of 95.41% and 82.43%, respectively. In 2024, Fabio Dennstädt and his colleagues [55] developed a general-purpose text classification framework utilizing LLMs for classifying oncological trials, achieving an overall accuracy of over 94%. Guo et al. (2024) [56] assessed the effectiveness of LLMs in text classification tasks using methods such as zero-shot classification, data annotation, and data augmentation. They found that LLMs, particularly GPT-4, demonstrated improved performance when combined with human-annotated data. The study indicates that data augmentation using LLMs yields better results than training with human-annotated data alone. Liu and Shi (2024) [57] introduced the PoliPrompt framework, a high-performance and cost-effective text classification system based on LLMs specifically for political science applications. This framework employs in-context learning, automatic prompt generation, and consensus mechanisms. It significantly reduces economic costs by 78% compared to traditional human labeling methods and achieves an improvement of 0.36 in the F1 score for zero-shot classification tasks. In research conducted by Mohajeri et al. (2024) [58], the authors proposed a method called the Code Completion Prompt (CoCoP), which leverages the capabilities of LLMs for text classification. This innovative approach transforms the text classification problem into a code completion task, resulting in improved accuracy of nearly 20% on the SST-2 dataset. The method was tested on several datasets, including SST-2, CoLA, MRPC, and SNLI, achieving the following best results: 93.5 ± 0.5, 78.5 ± 1.2, 77.4 ± 0.8, and 67.7 ± 2.1, respectively. In the study by Zhang et al. (2024) [59], text classification was performed on four datasets: SST-2, AG, Ohsumed, and MR. The researchers achieved accuracy rates of 98.68%, 97.61%, 77.41%, and 94.27%, respectively. They proposed the RGPT framework, which performed 1.36% better than SOTA PLM8 and SOTA LLM7 across four criteria on average. Yin et al. (2024) [60] proposed the CrisisSense-LLM method for text classification. This approach utilizes a pre-trained large language model that has been specifically fine-tuned for text classification tasks. The model incorporates Low-Rank Adaptation (LORA), which introduces additional trainable parameters and enhances its performance. In another study, Di Palo et al. (2024) [61] presented the Performance-Guided Knowledge Distillation (PGKD) method for text classification. This method completed text classification tasks up to 130 times faster and 25 times cheaper than LLMs. The study demonstrates that PGKD outperforms traditional BERT-based models in multi-class classification. Recent approaches in text classification, such as CoCoP (which transforms text classification into a code completion task) [58], RGPT (increasing accuracy over SOTA models) [59], CrisisSense-LLM (which utilizes LORA to improve text classification) [60], and PGKD (which reduces computational costs by up to 130 times) [61], illustrate the rapid development of innovative and adaptable methods using large language models (LLMs). Recent research reviews indicate that LLMs can serve as accurate, cost-effective, and reliable tools for classifying specialized texts. However, a review of past studies shows that the input data has typically been structured text, with no research conducted on using unstructured text data, such as images, as input. This highlights the necessity to explore the utilization of unstructured input data, specifically images, in text classification. Recent studies, such as those by Chae & Davidson (2023) [52] and Wang et al. (2023)[53], have focused on digital and structured text classification, demonstrating the effectiveness of large language models (LLMs) in zero-shot or few-shot formats with accuracies reaching or exceeding 90%. Semi-supervised approaches, like the one conducted by Zhong et al. (2024) [54], and techniques based on specific prompt designs, such as CoCoP [58], have also achieved over 90% accuracy in controlled environments. Additionally, research by Fabio et al. (2024) [55], Liu & Shi (2024)[57], and Zhang et al. (2024) [59] has successfully optimized LLMs for more precise and domain-specific classification by creating specialized frameworks tailored to specific fields, including medical, social, scientific, and political domains. The proposed method in this research involves extracting handwritten texts from images within a database dataset and converting them into text data. This extracted text data is then fed into a Large Language Model (LLM) for categorization based on content using the BART model. The research presents a framework for text classification through the integration of the TrOCR model and large language models. The results indicate that among the pre-trained TrOCR models, the large version performs the best in text classification, achieving an accuracy of nearly 65%. The method proposed in this study demonstrates superiority when working with unstructured and image-based data, unlike other methods that require textual data. This study stands out from most others as it focuses on image data rather than just digital text. By combining TrOCR with linguistic models, it allows for the processing of real, unstructured documents. This capability has the potential to lead to the development of systems that can automate the analysis of forms, contracts, and manuscripts. However, the classification performance in this study still requires optimization when compared to the accuracy of fine-tuned models designed for structured texts. Future studies could enhance performance by utilizing more advanced OCR models, fine-tuning the language model used, or applying data augmentation techniques. Although the method presented in this study has lower accuracy compared to purely textual methods, it excels at processing more complex data. It offers a significant advantage in applications where data is available in image form. The innovation of this method lies in its use of advanced Optical Character Recognition (OCR) models alongside large language models, making it highly suitable for real-world projects that involve scanned or unstructured data. This approach is particularly valuable in environments with image data, especially when information is stored in handwritten formats within forms or archives. However, for applications where data is available in structured text form, other methods such as RGPT, CoCoP, or PoliPrompt are more effective, offering higher accuracy at a lower cost.

REFERENCES

[1] D. W. Otter, J. R. Medina and J. K. Kalita, "A Survey of the Usages of Deep Learning for Natural Language Processing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 604-624, Feb. 2021.

[2] A. Khosravi and H. Abdolhosseini, "Personality in Social Networks Using Thematic Modelling of User Feedback", *Soft Comput. J.*, vol. 11, no. 2, pp. 51-60, 2023.

[3] F. Pourgholamali, M. Kahani and E. Asgarian, "Exploiting Big Data Technology for Opinion Mining", *Soft Comput. J.*, vol. 9, no. 1, pp. 26-39.

[4] F. A. Acheampong, C. Wenyu and H. NunooMensah, "Text-based Emotion Detection: Advances, Challenges, and Opportunities", Eng. Rep., vol. 2, no. 7, p. e12189, 2020.

[5] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel and H. Prendinger, "Deep Learning for Affective Computing: Text-Based Emotion Recognition in Decision Support," Decis. Support Syst., vol. 115, pp. 24-35, Nov. 2018.

[6] F. Zare Mehrjardi, M. Yazdian-Dehkordi and A. Latif, "Evaluating Classical Machine Learning and Deep-Learning Methods in Sentiment Analysis of Persian Telegram Message", *Soft Comput. J.*, vol. 11, no. 1, pp. 88-105, 2022.

[7] M. Feizi-Derakhshi, Z. Mottaghinia and M. Asgari-Chenaghlu, "Persian Text Classification Based on Deep Neural Networks", *Soft Comput. J.*, vol. 11, no. 1, pp. 120-139.

[8] S. Freyberg and H. Hauser, "The Morphological Paradigm in Robotics", *Stud. Hist. Philos. Sci.*, vol. 100, pp. 1-11, 2023.

[9] A. Ganesh, A. Jaya and C. Sunitha, "An Overview of Semantic Based Document Summarization in Different Languages", *ECS Trans*, vol. 107, no. 1, pp. 6007-6017, 2022.

[10] L. Geiszler, "Imitation in Automata and Robots: A Philosophical Case Study on Kempelen", *Stud. Hist. Philos. Sci.*, vol.100, pp. 22-31, Aug. 2023.

[11] V. K. Finn, "Exact Epistemology and Artificial Intelligence", *Autom. Document. Math. Linguist.*, vol. 54, pp.140-173, 2020.

[12] F. Ansari, "Knowledge Management 4.0: Theoretical and Practical Considerations in Cyber Physical Production Systems", *IFAC-PapersOnLine*, vol. 52, no. 13, pp. 1597-1602, 2019.

[13] E. Baralis, L. Cagliero, S. Jabeen, A. Fiori and S. Shah, "Combining Semantics and Social Knowledge for News Article Summarization", in *Data Mining and Analysis in the Engineering Field*, IGI Global, 2014, pp. 209-230.

[14] L. Waardenburg and M. Huysman, "From Coexistence to Co-creation: Blurring Boundaries in the Age of AI", *Inf. Organ.*, vol.32, no. 4, p. 100432, Dec 2022.

[15] S. Gupta, S. Modgil, A. Kumar, U. Sivarajah and Z. Irani, "Artificial Intelligence and Cloud-Based Collaborative Platforms for Managing Disaster, Extreme Weather and Emergency Operations", *Int. J. Prod. Econ.*, vol. 254, p. 108642, Dec 2022.

[16] C. R. Dhivyaa, K. Nithya, R. Dharshini, R. Sudhakar., K. Sathis Kumar and T. Janani. "Fine-Tuned Convolutional Neural Networks for Tamil Handwritten Text Recognition.", In Proceedings of the 8th International Conference on Communication and Electronics Systems (ICCES), 2023, pp. 887-893.

[17] D. Parres and R. Paredes, "Fine-Tuning Vision Encoder–Decoder Transformers for Handwriting Text Recognition on Historical Documents" In Proceedings of International Conference on Document Analysis and Recognition. 2023, pp. 253-268.

[18] M. P. Kalra, A. Kushwaha and P. P. Vuppuluri, "LLM Powered HTR: Integrating Handwritten Text Recognition System with Large Language Model," In Proceedings of IEEE Students Conference on Engineering and Systems (SCES), Prayagraj, India, 2024, pp. 1-6.

[19] J. Kohút and M. Hradiš. "Fine-tuning is a Surprisingly Effective Domain Adaptation Baseline In Handwriting Recognition", In Proceedings of International Conference on Document Analysis and Recognition, arXiv:2302.06308, 2023.

[20] P. Kumar and B. Raman, "A BERT Based Dualchannel Explainable Text Emotion Recognition System," *Neural Netw.*, vol. 150, pp. 392-407, June 2022.

[21] M. Badpeima, H. Shirazi and S.S. Sadidpur, "Determining the Polarity of Persian Texts Using LSTM Recurrent Networks", In Proceedings of the 3rd International Conference on Electrical, Electronic, and Computer Engineering, Norway, 2016 [In Persian].

[22] A. Onan, "Sentiment Analysis on Massive Open Online Course Evaluations: A Text Mining and Deep Learning Approach", *Comput. Appl. Eng. Educ.*, vol. 29, no. 3, pp. 572-589, 2021.

[23] H. Han, Z. Ke, X. Nie, L. Dai and W. Slamu, "Multimodal Fusion with Dual-Attention Based on Textual Double-Embedding Networks for Rumor Detection," *Appl. Sci.*, vol. 13, no. 8, p. 4886, 2023.

[24] N. Majma and S.Bashtin, "Detection of Plagiarism in Scientific Texts Based on Text Blocking and Cosine Similarity Criteria", *Soft Comput. J.*, vol. 11, no. 1, pp.60-71, 2022.

[25] R. Behzadidoost, F. Mahan and H. Izadkhah, "Granular Computing-Based Deep Learning for Text Classification," *Inf. Sci.*, vol. 652, p. 119746, 2024.
[26] B. Shi, X. Bai and C. Yao, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298-2304, 2017.
[27] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li and Wei, F., " TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models", In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 11, pp. 13094-13102, 2023.
[28] X. Zhou, C. Yao, H. Wen, et al., " EAST: An Efficient and Accurate Scene Text Detector", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp.5551-5560.
[29] H. Sak, A. Senior and F. Beaufays, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition",, arXiv:1402.1128, 2014.
[30] [Online]. Available: https://github.com/JaidedAI/EasyOCR
[31] [Online]. Available: https://github.com/PaddlePaddle/PaddleOCR
[32] [Online]. Available: https://github.com/faustomorales/keras-ocr
[33] G. Kim, T. Hong, M. Yim, J. Y. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han and S. Park. "Ocr-Free Document Understanding Transformer", In Proceedings of European Conference on Computer Vision, 2022, pp. 498-517.
[34] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei and M. Zhou, "Layoutlm: Pre-training of text and layout for document image understanding." Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 2020, pp. 1192-1200.
[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez and I. Polosukhin, " Attention is All You Need", Adv. Neural Inf. Process. Syst., vol. 30, pp. 1-15, 2017.
[36] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell and S. Agarwal, " Language Models Are Few-Shot Learners", Adv. Neural Inf. Process. Syst., vol. 33, pp. 1877-1901, 2020.
[37] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat and R. Avila, "Gpt-4 Technical Report", arXiv preprint arXiv:2303.08774, 2023.
[38] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding", In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171-4186.
[39] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena and P. J. Liu, " Exploring the Limits of Transfer Learning with a Unified Text-To-Text Transformer", *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1-67, 2020.
[40] K. Lv, Y. Yang, T. Liu, Q. Gao, Q. Guo, and X. Qiu, "Full Parameter Fine-Tuning for Large Language Models with Limited Resources", arXiv preprint arXiv: 2306.09782, 2023.
[41] D. Narayanan, M. Shoeybi, J. Casper, P. LeGresley, M. Patwary, V. Korthikanti, D. Vainbrand, P. Kashinkunti, J. Bernauer, B. Catanzaro and A. Phanishayee, "Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM", In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2021, pp. 1-15.
[42] O. Sharir, B. Peleg, and Y. Shoham, "The Cost of Training NLP Models: A Concise Overview," arXiv preprint arXiv: 2004.08900, 2020.
[43] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi and N. Smith, "Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, And Early Stopping", arXiv preprint arXiv:2002.06305, 2020.
[44] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov et al. "Llama 2: Open Foundation and Fine-Tuned Chat Models", arXiv preprint arXiv:2307.09288, 2023.
[45] Research Group on Computer Vision and Artificial Intelligence — Computer Vision and Artificial Intelligence (https://fki.tic.heia-fr.ch/databases).
[46] Handwritten Line Text Recognition using Deep Learning with Tensorflow. [Online]. Available: https://github.com/sushant097/Handwritten-Line-Text-Recognition-using-Deep-Learning-with-Tensorflow
[47] English handwritten line dataset. [Online]. Available: https://www.kaggle.com/datasets/sushant097/english-handwritten-line-dataset
[48] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov and L. Zettlemoyer, "BART: Denoising Sequence-To-Sequence Pre-Training for Natural Language Generation, Translation, And Comprehension", arXiv preprint arXiv:1910.13461, 2019.

[49] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, "Roberta: A Robustly Optimized BERT Pretraining Approach", arXiv preprint arXiv:1907.11692, 2019.

[50] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever. "Language Models are Unsupervised Multitask Learners" *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[51] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", arXiv preprint arXiv:2010.11929, 2020.

[52] Y.-M, Chae and T. Davidson, "Large Language Models for Text Classification: From Zero-Shot Learning to Fine-Tuning", SocArXiv: 10.31235/osf.io/sthwk, Aug. 2023.

[53] Z. Wang, Y. Pang and Y. Lin. "Large Language Models Are Zero-Shot Text Classifiers." arXiv preprint arXiv:2312.01044, 2023.

[54] S. Zhong, J. Zeng, Y. Yu and B. Lin, "Clustering Algorithms and RAG Enhancing Semi-Supervised Text Classification with Large LLMs", *Int. J. Data Sci. Anal.*, vol. 2025, pp. 1-22, 2025.

[55] F. Dennstaedt, P. Windisch, I. Filchenko, J. Zink, P. M. Putora, A. Shaheen, R. Gaio, N. Cihoric, M. Wosny, S. Aeppli, et al., " Application of a General Large Language Model-Based Classification System to Retrieve Information about Oncological Trials ", medRxiv: 10.1159/000546946, 2024.

[56] Y. Guo, A. Ovadje, M. A. Al-Garadi and A. Sarker, "Evaluating Large Language Models for Health-Related Text Classification Tasks with Public Social Media Data", J. Am. Med. Inform. Assoc., vol. 31, no. 10, pp. 2181-2189, 2024.

[57] M. Liu and G. Shi, "Enhancing LLM-Based Text Classification in Political Science: Automatic Prompt Optimization and Dynamic Exemplar Selection for Few-Shot Learning", arXiv preprint arXiv:2409.01466, 2024.

[58] M. M. Mohajeri, M. J. Dousti and M. N. Ahmadabadi, "CoCoP: Enhancing Text Classification with LLM through Code Completion Prompt", arXiv preprint arXiv:2411.08979, 2024.

[59] Y. Zhang, M. Wang, Q. Li, P. Tiwari and J. Qin, "Pushing the Limit of LLM Capacity for Text Classification", In Proceedings of the ACM on Web Conference, 2025, pp. 1524-1528.

[60] Yin, Kai, Chengkai Liu, Ali Mostafavi, and Xia Hu., "CrisisSense-LLM: Instruction Fine-Tuned Large Language Model for Multi-Label Social Media Text Classification in Disaster Informatics", arXiv preprint arXiv:2406.15477, 2024.

[61] F. Di Palo, P. Singhi and B. Fadlallah. "Performance-Guided LLM Knowledge Distillation for Efficient Text Classification at Scale", arXiv preprint arXiv:2411.05045, 2024.