A REVIEW OF PROMPT-FREE FEW-SHOT TEXT CLASSIFICATION METHODS

Rim Messaoudi, Achraf Louiza and Francois Azelart

Akkodis Research -Akkodis France

ABSTRACT

Text-based comments play a crucial role in providing feedback for various industries. However, effectively filtering and categorizing this feedback based on custom context-specific criteria requires sophisticated language modeling techniques. While traditional approaches have shown effectiveness, they often require a substantial amount of data to compensate for their modeling deficiencies. In this work, we focus on highlighting the performance and limitations of prompt-free few-shot text classification using open-source pre-trained sentence transformers. On the one hand, our research includes a comprehensive study across different benchmark datasets, encompassing 9 dimensions such as sentiment analysis, topic modeling, grammatical acceptance, and emotion classification. Also, we worked at making different experiences to test Prompt-Free Few-Shot Text Classification. On the other hand, we underline prompt-free few-shot classification limitations when the targeted criteria are complex. As an alternative approach, prompting an instruction-fine-tuned language model has demonstrated favorable outcomes, as proven by our application in the specific use case of "Identifying and extracting resolution results and actions from explanatory notes", achieving an accuracy rate of 80%.

KEYWORDS

Language models, Sentence transformers, SetFit, contrastive learning, distillation, intelligence compression, NLP, semantic similarity

1. INTRODUCTION

Automated scoring systems capable of evaluating texts based on custom criteria offer significant advantages in various domains. In fact, the ability to automatically assess and score text provides efficiency, consistency and scalability enabling industries to process and analyze large volumes of text data accurately and quickly. However, traditional objective-specific methods often face limitations in capturing language nuances and thus failing to adapt to abstract classification concepts without a relatively huge amount of training data. These limitations have become gradually accessible since the exploration of general-purpose or foundation models [1]. The term has been popularized by Stanford research team to highlight large models that learn general knowledge from an extensive quantity of data and manage to perform well in downstream tasks without being specifically trained for them.

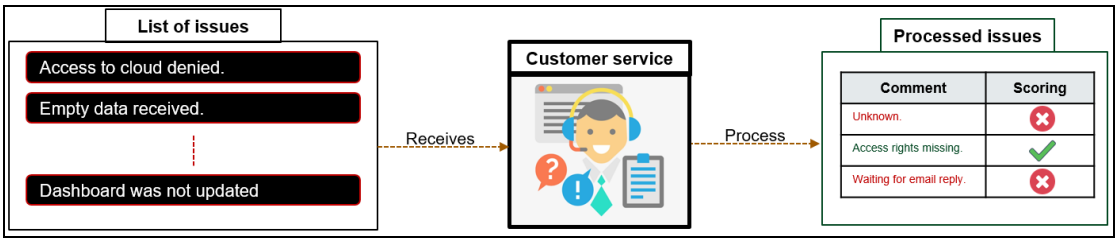

Figure 1. Customer service resolution note scoring

Numerous companies require a precise scoring mechanism for their clients' feedback, or their customer service resolution notes. However, a significant challenge arises due to varying classification objectives across different scenarios. In fact, allocating extensive resources in terms of budget and time for each specific use case becomes impractical. Consequently, leveraging foundation models that are accessible to the open-source community offers a viable solution. For example, efficiently fine-tuning smaller foundation models locally with limited training data proves to be effective. Additionally, the option of using a compact yet efficient language model should not be overlooked, particularly when generating a representative training set poses a challenge. Overall, this work outlines a strategy to leverage these approaches, empowering companies to swiftly align with the latest advancements in the field with minimal investment of effort and time. A part of our approach was presented in this paper [22].

2. RELATED WORK

2.1. General Purpose Language Models

In textual context, we refer to foundation models as language models. They manage to capture abstract concepts of our world using textual data as a comprehensive world projection. This ability has become possible through the Transformers architecture which was originally introduced by Vaswani et al. in 2017 [2] making it possible to capture long-range dependencies and contextual relationships within data using a differentiable and parallelizable modeling. Within the realm of language models, two primary types have proved their efficiency, namely causal language models (CLM) and masked language models (MLM), each with slightly a different generative process inducing distinct characteristics and potential applications. On the one hand, CLM, like GPT (Generative Pretrained Transformer), are autoregressive models since they are trained to predict the next word in a sequence based on the preceding words. On the other hand, MLM, like BERT (Bidirectional Encoder Representations from Transformers), adopt a bidirectional approach by masking random words in a sentence and training the model to predict the missing words using the context around it. With good engineering, both conceptions manage to learn, to some degree, general language knowledge. In fact, Ilya Sustskever, OpenAI's chief scientist stated that the main modification that enabled GPT-4 to reach higher capabilities of reasoning in comparison to GPT-3 is the increase in the base model precision for training data on text generation. Empirically, language model generative precision is positively correlated to its reasoning capabilities and thus at least partially correlated to its downstream task performance. Also we find in [21] authors explained the analysis of the performance of VoIP over the wireless networks.

2.2. From General-Purpose to Downstream Tasks

In the realm of natural language processing, we have witnessed a remarkable shift from the realm of general-purpose language models to their practical deployment in real-world tasks. This transformative journey was pioneered by a series of influential studies that have left a mark on

the landscape of every downstream task. Among these foundational works, Radford and colleagues (2018) introduced the groundbreaking GPT (Generative Pre-trained Transformer) model [3], unveiling its ability to grasp the nuances of human language and produce coherent text. This milestone was further advanced by Devlin et al. (2019), who introduced the widely acclaimed BERT (Bidirectional Encoder Representations from Transformers) model [4], revolutionizing the approach to fine-tuning for task-specific requirements. The subsequent efforts of researchers like Yang et al. (2019) with XLNet [5] and Liu et al. (2019) with RoBERTa [6] further refined pre-training techniques, pushing the boundaries of model adaptability. Collectively, these pioneering studies have paved the way for the integration of foundation models into diverse downstream tasks, demonstrating their remarkable versatility and immense potential for practical applications.

2.2.1. Zero-Shot Learning

Language models have made it possible for natural language processing to achieve astonishing feats that were previously thought to be unattainable. Among these revolutionary developments, zero-shot learning is an absolute testament which allows models to take on tasks for which they have never been formally taught. This method is very promising but lacks two important aspects. First, only the most powerful language models have consistent performance. Secondly, optimized prompt, affined using prompt design and engineering, is necessary. Otherwise, the output is highly impacted by the smallest prompt changes. Even with these limitations, only this approach offers context imputing through prompt, which is a very valuable asset when classifying a complex criterion. Moreover, it is true that the generalization characteristic was mainly consistent in large language model like GPT-4 but thanks to the several advancements in intelligence compression through techniques like quantization, weight punning and knowledge distillation, smaller models have proven to be capable of competitive abstract reasoning and zero-shot learning.

2.2.2. Few-Shot Learning

A better alternative to optimally make use of the current open-source general-purpose models with minimal labeling cost is few-shot learning. It suffices to give a small set of training examples to a language model, as little as 10 examples per class, to achieve high accuracy in a downstream task. One way to tackle this approach is using an instruction followed by a few examples via prompt. Another way to adapt the PLM to a specific task is by fine-tuning the model. Parameter efficient fine-tuning (PEFT) has been a great suggestion to efficiently fine-tune LLMs by fixing the models parameters and only updating adapters which are additive feed forward layers inserted in the architecture. Thanks to this approach larger models have become more accessible, but fine-tuning the language model on text generation and using it in a downstream task via prompt can give answers with relatively high and uncontrolled variance in contrast to a specific classification fine-tuning.

In 2022, Lewis Tunstall et al. propose an efficient few-shot learning method without prompts called SetFit (Sentence Transformer fine-tuning) [7]. The authors make use of Transformer basic architecture and add a pooling layer over words embeddings to create a semantic representation of a sentence. On the one hand, this simple pooling layer added to a language models' architecture gave birth to Sentence Transformers which were first introduced in the article "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks" by Nils Reimer [8]. On the other hand, Lewis Tunstall's contribution resides in the way they advise to fine-tune the ST for a downstream task.

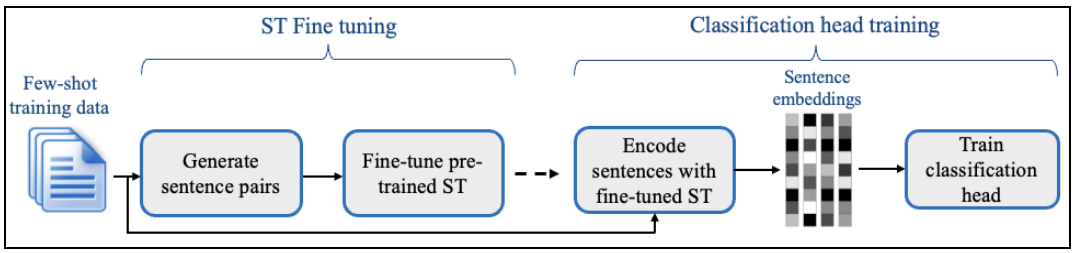

Figure 2. SetFit fine-tuning pipeline for few-shot classification [7]

In fact, by first fine-tuning the ST using a contrastive learning with Siamese network, the ST learns context-specific sentence representations facilitating the training of the classifier head afterwards. (Figure 2) This simple approach has managed to efficiently fine-tune ST for downstream tasks enabling small models to exceed traditional LLMs few-shot performance.

3. BENCHMARK: OPEN-SOURCE ENGLISH SENTENCE TRANSFORMERS

3.1. Studied Uses Cases

To get a grasp of SetFit performance and limitations, we will evaluate several open-source language models on these 4 use-cases: Sentiment analysis, grammatical acceptability, emotion classification and topic modeling as described in the table below.

Table 1. Studied use-cases description

Dataset	Description	Classes
IMDB	The IMDB dataset consists of movie reviews from the Internet Movie Database (IMBD) labeled as positive or negative sentiment based on the overall sentiment. Link: https://huggingface.co/datasets/imdb [9]	0: Negative sentiment 1: Positive sentiment.
CoLA	The CoLA dataset is used to assess the grammatical acceptability of sentences in English. [10] Link: https://huggingface.co/datasets/linxinyuan/cola	0: Grammatically unacceptable 1: Grammatically acceptable.
Emotion	The Emotion dataset consists of English twitter messages with six basic emotions. [11] Link: https://huggingface.co/datasets/dair-ai/emotion	0: Sadness, 1: Joy, 2: Love, 3: Anger, 4: Fear, 5: Surprise
AG News	The AG News dataset is used for news categorization. It contains news articles categorized based on their topics. [12] Link: https://huggingface.co/datasets/ag_news	0: World, 1: Sports, 2: Business and 3: Science/Technology.
Hate speech / Offensive	The hate_speech_offensive dataset is used for hate and offensive speech detection. It contains collected and labeled tweets. [13] Link: https://huggingface.co/datasets/SetFit/hate_speech_offensive	0: Hate speech 1: Offensive language 2: Neither
Enron Spam	The enron_spam dataset is used for spam detection. It contains a mix of non-spam (ham) and spam e-mail messages. [14] Link: https://huggingface.co/datasets/SetFit/enron_spam	0: Ham 1: Spam
SILICONE (dyda_da)	The sequence labelling evaluation benchmark for spoken language (SILICONE) is a collection of resources for analyzing NLU systems. The dataset dyda_da distinguishes the intended communicative purpose or function behind an utterance. [15] Link: https://huggingface.co/datasets/silicone/viewer/dyda_d	0: Commissive 1: Directive 2: Inform 3: Question

Stance Abortion	The tweet_eval_stance_abortion dataset is used to deduce a text stance regarding abortion. It uses labeled tweets. [16] Link: https://huggingface.co/datasets/SetFit/tweet_eval_stance_abortion	0: None 1: Against 2: Favor

3.2. Datasets Label Distribution

The studied datasets are well-known as reference benchmarks for their specific use-case. They are all available through hugging face dataset hub. Since we want to evaluate few-shot classification on these datasets, we need to make sure that we have enough observations in each class inside the training set to be able to study the model efficiency for different training set sizes. Moreover, observing unbalance in test set labels can explain the future potential shift between accuracy and precision/recall.

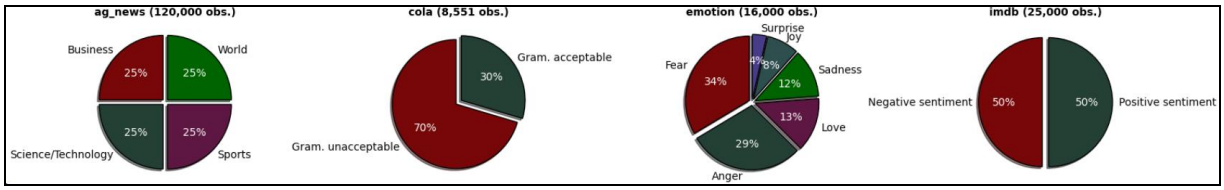

Figure 3. Training datasets label distribution

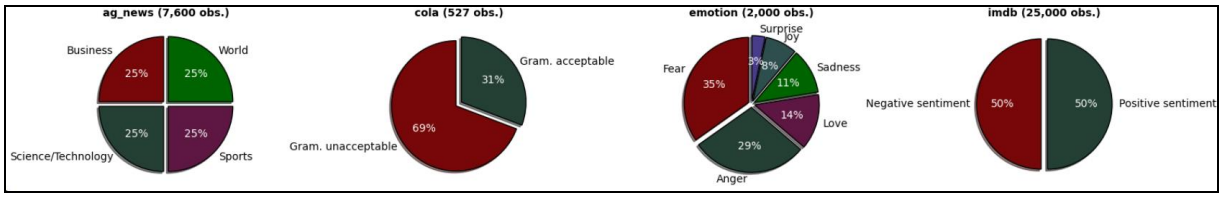

Figure 3. Testing datasets label distribution

First, we have enough samples in the training set to evaluate few-shot classification models. However, AG News and IMDB test datasets are perfectly balanced in contrast to CoLA and Emotion test datasets, so the accuracy measure can be biased in the latter case. Overall, we have enough samples to compare several models' performance.

Explored sentence transformers

In our benchmark study, we leveraged advanced sentence transformers initially trained to excel at recognizing semantic similarity between sentences. These sentence transformers are built upon language models pretrained on extensive text data, allowing them to capture intricate semantic relationships. A pivotal component is the pooling layer, which condenses information within the model's hidden states, producing meaningful sentence embeddings.

To identify the most promising sentence transformers, we initiated our selection process by considering open-source benchmarks available on Hugging Face. We also considered the models' parameter counts, ranging from as few as 22 million to several billion. Our primary focus is on conducting a thorough comparison of the most promising small language models, emphasizing their cost-efficiency and energy consumption. Another crucial factor to consider is the model's licensing. Licenses that allow for potential commercial use are preferred, as they empower the community and small businesses to compete with closed-source models and those operating under proprietary licenses. According to a survey conducted by the Natural Language Processing

(NLP) Foundation in 2022, approximately 70% of NLP practitioners reported using open source LLMs, while around 30% said they utilized closed source models. Another way to look at the distribution of open vs. closed source LLMs is through GitHub repositories. As of March 2023, the top ten open source LLM repositories on GitHub had over 68K stars combined. On the other hand, the top ten closed source LLM repositories had fewer than 10K stars.

In SetFit original article, small models give better performance than larger language models. However, several relatively large language models with several billion parameters have recently managed to reach high quality text generation through chat-based or instruction-based inference. Consequently, models like Llama 2 7B/13B [13], Vigogne 2 7B/13B, Mistral 7B [14] or Zephyr 7B [15] will rather be studied in a prompt-based approach to fully understand the influence of scaling on few-shot classification.

Table 2. Studied sentence transformers description

Sentence transformer	Size	Layers	Model dimension	Description
all-MiniLM-L6-v2	80MB	6	384	MiniLM is a distilled model from the paper "*MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers*". The LM was initially created by Microsoft. The ST was trained on a 1B sentence pairs based on semantic similarity. [16]
multi-qa-distilbert-cos-v1	250MB	6	768	Distilbert is a distilled model from the paper "*DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter".* The ST is trained on a Q&A set of 215M couples of (Question, Answer). [17]
all-mpnet-base-v2	420MB	12	768	The MPNet language model was proposed in "M*PNet: Masked and Permuted Pre-training for Language Understanding*". It inherits MLM and PLM advantages. The ST was trained on 1B pair of sentences using cosine similarity. [18]
bge-base-en-v1.5	440MB	12	768	The BAAI General Embedding (BGE) model is a pre-trained language model developed by the Beijing Academy of Artificial Intelligence (BAAI) and introduced in the article: "*C-Pack: Packaged Resources to Advance General Chinese Embedding*". It is based on the Chinese-RoBERTa model. [19]
gte-large	670MB	12	1024	The General Text Embeddings (GTE) model was introduced in the paper "*Towards General Text Embeddings with Multi-stage Contrastive Learning*". It uses a contrastive learning after three sequential stages in calculating the similarity: local, global and final. [20]
bge-large-en-v1.5	1.34GB	12	1024	This is the large version of bge-base model. This model is ranked first in Hugging Face massive text embedding benchmark (MTEB) leaderboard.

To adapt these models for few-shot classification, we performed an additional round of fine-tuning using contrastive learning techniques. The outcome is sentence transformers with an enhanced ability to distinguish subtle distinctions depending on classification training objective.

Benchmark process

To ensure a comprehensive evaluation, the fine-tuning process involved incrementally increasing the number of labeled examples per class, starting from 10, and gradually scaling up to 30, 50, 70, and, ultimately, 200 examples per class. In addition to **SetFit**, we evaluate a traditional approach consisting of training **a logistic regression over TFIDF** (Term Frequency / Inverse Document Frequency) in a few-shot way. Another traditional method resides in training **convolution neural network** with enough data, in our case a 1000 observation for each class value. While the statistical method serves as a minimum performance reference achievable in a few-shot way with almost no cost, the CNN evaluation offers a maximum reference performance that we hope to be able to surpass with as low as 10-shots instead of thousands.

The evaluation metrics, including accuracy, precision, and recall, were rigorously measured at each fine-tuning stage, utilizing an independent test set previously introduced in the study. The precision and recall were aggregated as non-weighted averages across all classes. This methodological approach gives a global comparison of the open-source models over different use-cases, aiding in the identification of the most effective model for each task. (Figure5) The only issue is the randomness introduced during training set selection at each stage. Ideally, we would have repeated the training set selection, for each number of examples per class several times, and save the mean and standard deviation as robust indicators. This approach is used in a fixed number of examples per class for an advanced comparison later in the study.

Benchmark results

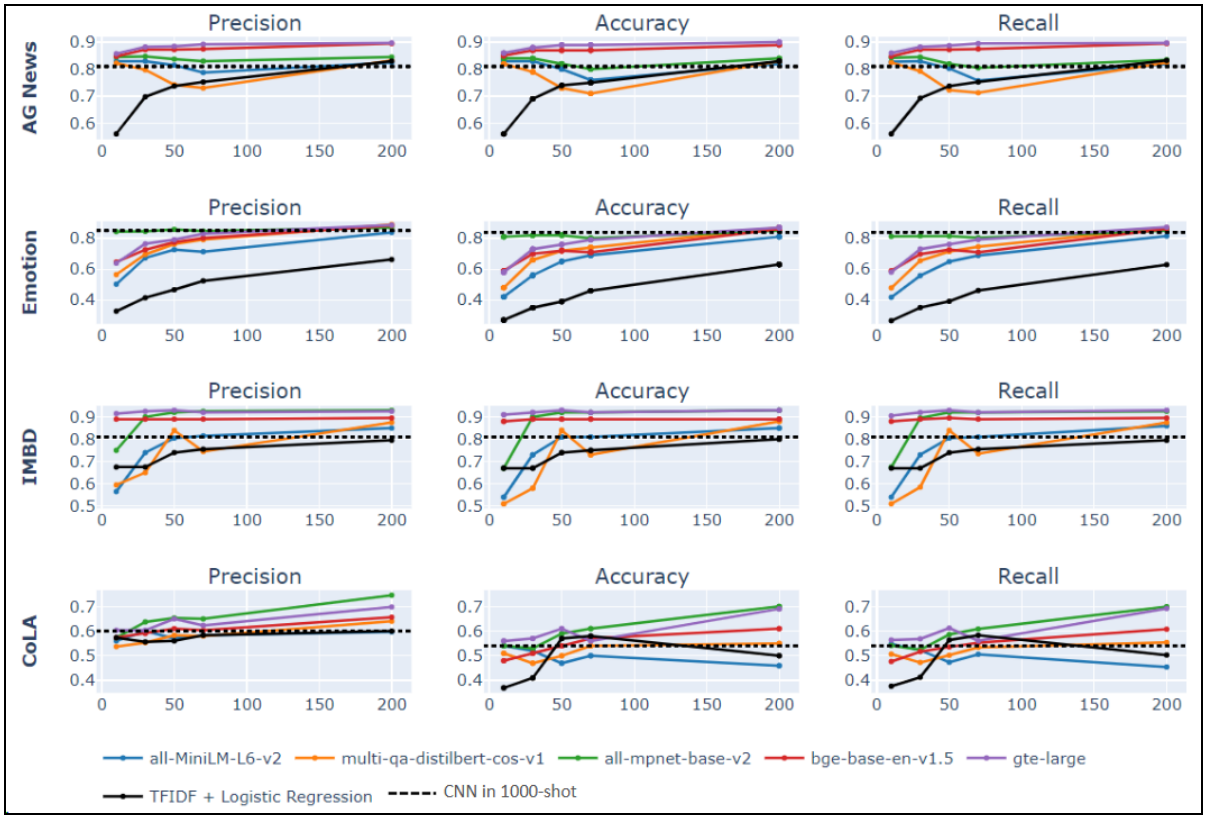

Figure 4. Few-shot classification evaluation of 5 ST on 4 main datasets with training sets with increasing size

The initial conventional method, where logistic regression is applied to TFIDF-encoded data, unfortunately exhibits subpar performance in few-shot classification but, when given 200 examples per class, it manages to reach 80% accuracy in topic modeling and sentiment analysis. Whereas, SetFit models consistently outshine the traditional approach, showcasing remarkable performance even with a minimal dataset of just 10 examples per class. In the realm of Convolutional Neural Network (CNN) training, leveraging 1000 examples per class value has led to an approximately 80% accuracy rate across various use-cases apart from grammatical acceptance where the method failed. Overall, as anticipated, fine-tuning Sentence Transformers outperform traditional methods, delivering superior results with only a minimal number of examples compared to the thousands required by traditional approaches. For the remainder of our analysis, we'll be delving into a more focused comparison of SetFit-based models.

In topic modeling, using AG News dataset, gte-large and bge-base both have between 5% and 10% higher accuracy, precision and recall compared to the rest of the studied models, reaching a 90% score on all metrics starting from as low as 50 examples per class. In fact, in the SetFit's original article, MPNET gave the best results over several benchmarks. In the author's defense, bge and gte weren't publicly available at that time. Overall, all models achieve an impressive minimum accuracy of 80% when categorizing news articles into their respective topics, even with just 10 examples per category, thanks to their ability to harness the semantic understanding acquired during the generative pre-training phase of their language models.

When it comes to emotion classification, MPNET stands out as the sole model achieving an accuracy exceeding 80% with just 10 to 50 examples per category, while all other sentence transformers fall short, unable to exceed a 60% accuracy threshold. Yet, with enough training samples, all the sentence transformers manage to reach the 80% accuracy threshold.

Moving on to sentiment analysis, both bge-base and gte-large consistently achieve a 90% metric from as low as 10 examples per category, maintaining this high accuracy up to 200 examples per category. In contrast, MPNET falls short, unable to exceed 70% accuracy with just 10 examples but gradually reaches the 90% threshold starting at 30 examples per category. On the other hand, distilbert and MiniLM struggle to attain high accuracy unless provided with at least several hundred examples per category. The final use-case under examination, grammatical acceptance, stands out as particularly intricate due to its abstract connotations. Pre-trained language models are renowned for their semantic understanding, but they often encounter challenges when venturing beyond that domain. In practice, even when fine-tuned with 200 examples per category, the smaller models, MiniLM and distilBert, perform worse than a random classifier in this specific scenario. Notably, there appears to be a correlation between model size and performance, with gte-large and MPNET demonstrating the ability to attain a 70% accuracy rate with 200 examples per category.

Limitation analysis

To gain insight into the limitations of sentence transformers across diverse scenarios, it is adequate to examine the distinctions between correctly classified instances and incorrectly classified instances within each respective use-case. To facilitate this examination, we selected the most proficient sentence transformer model for each use-case, employing a dataset consisting of 70 examples per class. Specifically, we used the "gte-large" model for the AG News and MPNET datasets for the remaining three use-cases. Initiating our exploration by examining ST embeddings through density analysis rather than exhaustively scrutinizing individual data points allows us to focus on the broader classification boundaries where the density distributions of two classes overlap. (Figure 5)

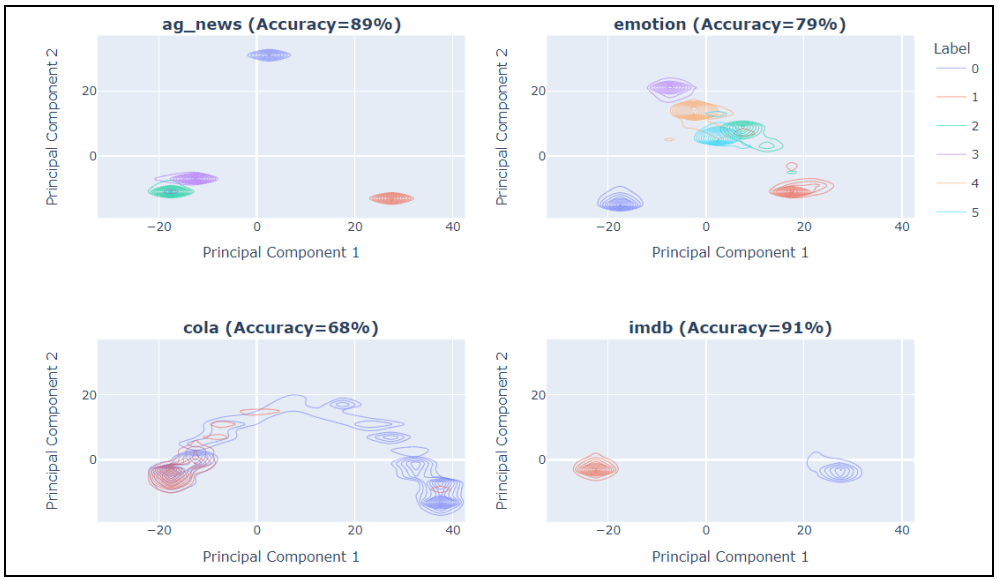

Figure 5. Density contour: PCA 2-dim projection of most efficient fine-tuned ST embeddings

As a result, in AG News, Business & Science/Tech articles often contain topics and terminology that overlap, making it challenging to distinguish between them solely based on the textual content. Likewise, in the Emotion dataset, "Love & Joy" or "Anger & Sadness" can share similar linguistic expressions, leading to overlapping density distributions in the embedding space. This density-based analysis enables us to pinpoint these specific challenges and motivates the development of more robust classification techniques to handle such intricate cases. Moreover, it is worth noting that, in the realm of grammatical acceptance classification, the classes exhibit distinct separation in the projected embedding space, with two nodes situated far apart. Nevertheless, the majority of misclassifications stem from categorizing grammatically unacceptable sentences as acceptable.

Discussion

In our study, we employed relatively small language models and achieved a commendable performance level of up to 80% - 90%. However, our approach does exhibit certain limitations, particularly when dealing with sentences that are ambiguous and fall between two distinct classes. To address this challenge and optimize our methodology, we have identified two potential solutions. The first approach involves incorporating "hard negatives" into the training set, with the hope that these challenging examples will enable our smaller models to uncover more robust underlying criteria for classification, thereby enhancing their performance. Alternatively, we can opt for scalability by employing larger models, such as 1B, 3B or even larger variants, and closely monitoring the impact of scaling on our classification tasks. These strategic considerations provide us with valuable avenues for further refining our approach and tackling the nuances of ambiguous sentence classifications.

When confronted with a complex classification criterion such as grammatical acceptance, prompt-free few-shot classification becomes very impacted by the training set choice. Since the task has not semantic dimensionality, then creating a small yet representative training set can prove to be extremely difficult without the help of a linguistic expert in making meticulous representative choices.

4. SECOND BENCHMARK: RESOLUTION NOTES EVALUATION IN FRENCH

Problem definition

In incident management, the explanatory note left by an agent is the main valuable description of the issue resolution. It assumes a pivotal role not only in ensuring the seamless functionality of technical support processes, but it can also be a valuable repository for future incidents management. However, for an efficient communication of the resolution process, the explanatory note must include on the one hand the result of the agent's intervening and on the other hand, the actions that led to the resolution of the issue. The most problematic notes are those that do not contain any information. One of the main challenges is to automatically detect invalid explanatory notes and thus to increase the quality of the agent's feedback.

Our objective is to create a model that can automatically identify the presence of the resolution result and action in the note left by the agent. This specific task is complex because the classification criteria is not semantically definable: For instance, a resolution result can be expressed in a multitude of different ways since the incidents and their solutions can take several forms in several areas (Cloud, Network, Application, Server, Retail, Customer service, etc.)

Data description

Since we are studying few-shot classification, the training set does not have to be exhaustive. A small representative dataset should be enough in theory. In this retrospective, our focus is aimed at IT-related incidents resolution.

Table 3. Examples of resolution notes labeling

Example	Result presence	Action presence
Incident: Le réseau est très lent. Resolution note: Le problème de latence réseau a été résolu.	Yes	No
Incident: Attaque par déni de service (DDoS) détectée. Resolution note: La menace a été neutralisée, et la stabilité du réseau a été rétablie grâce à l'ajout de règles de filtrage au niveau du routeur.	Yes	Yes (Ajout règles filtrage)
Incident: Erreur de déploiement dans la mise à jour du logiciel client. Resolution note: Rétrogradation à la version précédente	No	Yes (Rétrogradation de version)
Incident: Temps de chargement excessif sur le site web de l'entreprise. Resolution note: Analyse du temps de chargement en cours	No	No

In table 3, we have established a single example for each combination of classes value. However, each case can be expressed in multiple ways making it harder to classify automatically. Using large language models such as ChatGPT, we have been able to generate 50 different examples for each class value, making a total of 100 observations for each target training. Several efforts were dedicated to minimizing semantic confusion in the choice of the training set samples. As an example, for each training observation in a class, a semantically similar one is included in the opposite class to make clear that the semantic meaning of the sentence itself is not the separation criterion we are aiming to distinguish. The training and test generated datasets are publicly available through hugging face dataset under the tag "IT Resolution Notes FR". The dataset has 200 resolution notes with action and result presence labeling. Its purpose is restricted to few-shot

classification of resolution notes in an ideal situation where the resolution notes do not contain any form of noise such as linguistic errors, gibberish, logs, greetings, additional non-informative text, etc. We have first established prompt-free method performance in this ideal situation, then we have identified the main limitations through human-like noise imputation in the resolution notes.

French Sentence Transformers

In our pursuit of training a few-shot classification model, our benchmark has revealed that fine-tuning Sentence Transformers produces cutting-edge results. This chapter focuses on the classification of French texts. Unfortunately, the high-performing language models examined in our initial benchmark lack the capability to handle French texts. Nonetheless, the French community has successfully developed its own set of compact yet powerful language models, achieving state-of-the-art results across various classification scenarios. Additionally, multilingual models demonstrate a commitment to French by incorporating a substantial portion of French text in their training data. A noteworthy example is MiniLM, a compact and effective multilingual distilled model created by Microsoft. In the table below, we have chosen three out of the most efficient small language models of increasing sizes ranging from 20M parameter (MiniLM) to 137M Parameter (FlauBERT).

Table 4. The studied French language models

Language model	Number of parameters	Layers	Model dimension	Description
MiniLM-L12xH384	21M	12	384	MiniLM is a multilingual distilled model from Bert base presented by Microsoft while introducing a new distillation approach through their article: "*MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers*".
Camembert-base	100M	12	768	CamemBERT is a French language model based on RoBERTa and trained on the French section of the multilingual corpus OSCAR. The model was presented in the article: "CamemBERT: a Tasty French Language Model"
Flaubert-base	137M	12	768	FlauBERT is a French language model based on BERT and trained on the new CNRS (French National Center for Scientific Research) corpus. The model was presented in the article: "FlauBERT: Unsupervised Language Model Pre-training for French"

Among the chosen models, we have a BERT architecture, a RoBERTa architecture and a small, distilled model whose teacher is a BERT. Apart from the architecture, the training dataset differs as well. Consequently, their performance will vary, and it is valuable to test each of these models' performance. To do so optimally in the context of few-shot classification, we should fine-tune Sentence Transformers rather than use these language models embeddings directly. Both CamemBERT and MiniLM sentence transformers were fine-tuned on mMARCO dataset which is a question answering dataset based on web searches and other sources. FlauBERT sentence transformer has a different fine-tuning process, starting by 500k sentences of course descriptions

followed by a natural language inference dataset (contradiction, neutral, implication) and finally STS-fr dataset which contains similar and dissimilar sentences.

Public benchmark results: Ideal context with synthetic dataset

In the context of few-shot classification of the result and action presence in a resolution note, we first Benchmark SetFit approach on the generated IT-related resolution notes, using the chosen French language models. This dataset deliberately does not contain any noise and has similar linguistic formulations in both training and test sets. Our objective, through this first step, is to evaluate the performance of a prompt-free approach when dealing with complex criteria in a noise-free, narrow context.

All the three models managed to successfully classify resolution result presence on the 100 test observations with only a single misclassification concerning the incident "Problème de sécurité identifié dans les configurations réseau." with the resolution note "Les règles de pare-feu ont été renforcées." In this case, the result of resolution is not explicitly included in the resolution note, but the models predict otherwise. In fact, the models, inspired by the training examples, interpret a sentence implying an entity positive change as one including the result.

Among the 100 test resolution notes and in the context of action presence detection, FlauBERT and CamemBERT exhibit 3 false negatives, while MiniLM registers 12 misclassifications. Most of these errors are sentences expressed as "Issue has been resolved" instead of "Resolution of the issue". From this small sample, we observe that, without the incident context, the distinction between the result and action can be tricky. Apart from this constraint, the choice of the training set highly impacts the model's performance and must be affined meticulously to embed all the necessary rules and phrasing variations.

Even if the registered performance is very high, considering the similarities with the training set and the absence of noise, the consistence and generalization of prompt-free approach has yet to be proved. In fact, in the following section, we establish that some small noises or variations manage to fool the model's accuracy in this specific use-case.

Private benchmark results: Noisy real-world resolution notes

To assess the model's performance in a real situation, we evaluate it on a confidential dataset of retail/IT incident management englobing 200 manually annotated resolution notes and report the performance in the table below.

Table 5. Private benchmark: Prompt-free few-shot classification of resolution notes using SetFit

Language model	**Resolution result presence**			**Resolution action presence**		
	Accuracy	**Precision$_1$**	**Recall$_1$**	**Accuracy**	**Precision$_1$**	**Recall$_1$**
MiniLM-L12xH384	47%	67%	55%	66%	52%	60%
Camembert-base	44%	70%	41%	**68%**	54%	54%
Flaubert-base	**49%**	80%	40%	64%	49%	79%

By analyzing the model's behavior through experimentation, we have underlined the following implicit rules that we have found to be consistently decisive in the classification inference:

- The model gives high impact to resolution-related words.
    - For example: “functional”, “operational”, “repaired”, “update”, etc.
- The phrasing of the resolution note impacts the predictions.
    - For example, on one hand, if the sentence starts with an action, then the resolution process is predicted as present. On the other hand, if the sentence dictates that an entity has been subject to a positive change then the result of resolution is detected.

Overall, the performance is not efficient enough (less than 70%), especially for the result presence detection. The models struggle with real-world resolution notes classification for several reasons including the following:

- Misspelling words in resolution note.
    - For example, adding or replacing a letter in words with high impact on the predictions can bias the classification. Concretely, if we write the word operational with a spelling mistake in a resolution note, the result presence can be misclassified into absence.
- Radical change in resolution note phrasing
    - For example, if we remove the verb “to be” from the resolution note: “Le serveur est accessible.”, then the result is no longer detected. Moreover, if instead of saying the server has been repaired, we phrase it as Reparation of the server then the result is no longer detected but action is detected instead.
- Non-informative noisy text in resolution note
    - For example, simply adding greetings “Bonjour,” at the beginning of the sentence can bias the prediction.

Most of these constraints are correlated with the training set choice and can potentially be covered by meticulously including them in the training set. However, creating a representative training set can be time-consuming and infeasible in a few-shot setting, especially when the criteria itself is not simple to define or depends on contextual data. Moreover, every modification on the training set can potentially bias other rule-embeddings. SetFit prompt-free few-shot classification is an efficient approach when the criterion has a semantic dimension or a relatively simple rule embedding. When it is not the case, heavy engineering on the data is in order. In our perspective, this task can be better done using a prompt-based few-shot classification approach while communicating the task description with examples as well as the incident and resolution note to classify. Using natural language, it is easier to represent the classification task rather than with a set of labeled examples. Apart from the few given examples in both approaches, a custom description of a complex classification target is priceless and makes the difference in the prompt-free and prompt-based performance comparison. Consequently, we will explore prompting efficient French language model with 7 billion to 13 billion parameters while including post-training quantization, Frantar, Elias, et al., 2022, to reduce required memory to run the model in a local setting.

3. LLM PROMPT-BASED CLASSIFICATION

3.1. Context

In our research, we propose the adoption of prompt-based classification as a strong alternative when dealing with complex classification criteria. While conventional prompt-free few-shot classification methods excel in straightforward scenarios, their effectiveness declines in the face of intricate classification requirements. Prompt-based approaches provide a more flexible and

understandable solution, enabling practitioners to guide the model's attention explicitly to specific aspects of the data. By crafting precise prompts designed for complex classification tasks, we enhance the model's ability to recognize subtle patterns, leading to overall performance improvement. This shift towards prompt-based few-shot classification represents a paradigm change that explores new possibilities for handling diverse and complex criteria, catering to the evolving demands of natural language understanding tasks.

3.1.1. French Language Models for Instruction-Based Inference

The language models, previously discussed, are not capable of precisely and coherently answering questions neither through chat nor instruction. Just a few months ago, only large language models exceeding 70 billion parameters could effectively address questions. However, recent advancements in attention mechanisms, distillation, pruning, and quantization have demonstrated the efficiency of smaller-sized models. Notably, innovations like the Llama architecture and parameter sharing have empowered the community to progressively construct more potent pre-trained models. For example, Sandford successfully trained the Alpaca 7B and 13B language models by leveraging the Llama architecture and employing a dataset of 52,000 Q&A generated by ChatGPT. Similarly, the French community contributed to the development of Vigogne 7B and 13B by training the Alpaca model on the French-translated version of the same 52,000 Q&A dataset used in Alpaca training. Another noteworthy model, Mistral 7B and 13B, has demonstrated state-of-the-art results primarily in English for instruction-based inference, with a recent promising version available in French as well. In this section, we will assess prompt-based classification through instruction-based inference using the most promising French language models, Vigogne and Mistral, both with sizes of 7B and 13B. Additionally, we will evaluate the post-training quantized version with 4-bit precision via GPTQ, as it enables a fourfold reduction in the required GPU memory. For instance, a Vigogne 7B model with a float 16-bit precision has a size of 12GB, while GPTQ quantization compresses the model into 4GB using 4-bit integer precision.

3.1.2. Second Benchmark Results with Prompt-Based Approach

In the pursuit of efficiently extracting valuable information from resolution notes, we introduce a standardized template that facilitates the classification of resolution results and actions. This template is designed in a simple way to enhance the interpretability and efficiency of few-shot classification models, allowing for robust extraction even in scenarios with complex criteria. It is composed of three principal axes. First, we define the instruction followed by some examples and then we ask for an application to a new injectable resolution note. To automatically extract the predictions from the generated answer, we ask the model to return the predictions in Json textual format.

Table 3. Template resolution note result and action presence prompt-based classification and extraction

Request	DEMANDE : Extraire, ou Indiquer [0, none] si non mentionnée, dans la note de résolution le résultat de résolution et le processus (action) de résolution.
Examples (1-shot)	EXEMPLE : Si l'incident est : "Problème de connexion internet" et la note de résolution est : "Suite à la réparation du câble, la connexion a été restaurée.", on obtient : {'resultat_resolution' : [1, "La connexion a été restaurée"], 'action_resolution' : [1, "Réparer le câble"]}
Application	APPLICATION : Si l'incident est "*Inclure l'incident ici*" et la note de résolution est : "*Inclure la note de résolution ici*", on obtient :

This template can seamlessly be applied to new resolution notes, providing a consistent and interpretable approach for result and action extraction. (Table 3)

Table 4. Private benchmark: Prompt-based 5-shot classification of resolution notes

Language model	Resolution result presence			Resolution action presence		
	Accuracy	$Precision_1$	$Recall_1$	Accuracy	$Precision_1$	$Recall_1$
Vigogne-2-7B-Instruct	82%	87%	86%	61%	54%	85%
Vigogne-2-7B-Instruct-GPTQ	76%	97%	66%	72%	66%	96%
Vigogne-2-13B-Instruct-GPTQ	82%	90%	83%	68%	62%	91%
Mistral-7B-Instruct-v0.1-GPTQ	71%	80%	78%	67%	78%	49%

Both Vigogne and Mistral managed to properly classify the result and action presence in the resolution note with an accuracy around 75% for result and 70% for the action presence. The same prompt has been tested on these models in a 5-shot classification approach. Most of the 25% errors made by the models can be considered true. In fact, sometimes the incident is so simple that a direct action applied on the defective object is predicted as both the result and action. It is true that Vigogne-7B or the quantized 13B version are very precise in result presence detection but tend to overfit the action presence detection by extracting an action, sometimes when there is none.

The obtained results can be optimized simply by refining the prompt along with the request and the included examples. Although, our objective was to evaluate the overall performance of prompt-based approaches in French text few-shot classification, and we managed to surpass the performance of prompt-free approaches in this specific context with minimal cost. It is noteworthy to indicate that the Vigogne-2-7B-instruct-GPTQ has a good performance/cost compromise with only 4GB and still the results are almost as precise as the non-quantized version.

Another potential optimization is to distinguish the predictions of the result and action of resolution with two separate prompts. By doing so, the model can have a better understanding of the request as well as a better assimilation of the examples. However, doubling the prompts means doubling the cost of inference. Fortunately, the model has only 4GB and costs tenfold less than GPT APIs.

3.2. Prompt-Based Models Results Comparison

In this segment, attention shifts from traditional classification to the delicate area of information extraction. Prompt-based models perform well, not only in few-shot classification but also in their ability to reason and extract results. Their strength lies in their capability not only to predict outcomes and actions, but also to provide detailed justification for those predictions. This dual functionality of classification and extraction distinguishes prompt-based models and allows for a deeper understanding of the context and nuances of the text and contributes to a more comprehensive understanding of the model's decision-making process.

To explore these features in more detail, we examine and analyze the results of four carefully studied models: Vigogne-2-instruct-7B, Mistral-7B-instruct-GPTQ, Vigogne-2-7B-instruct-

GPTQ, and Vigogne-2-13B-instruct-GPTQ. This study aims to clarify subtle aspects of result extraction and provide valuable insights into the effectiveness and limitations of the model in handling complex linguistic contexts.

Example 1: Result and action are included

- Incident : "Connexion internet ne fonctionne pas."
- Note de résolution : "Suite à l'intervention du technicien SFR, le réseau est rétabli. N'hésitez pas à nous contacter si besoin."

Table 5. First example of resolution result and action extraction through prompting an LLM

	Model			
Extraction	Vigogne-2-7B	Vigogne-2-7B-GPTQ	Vigogne-2-13B-GPTQ	Mistral-2-7B-GPTQ
Result	Le réseau est rétabli	Le réseau est rétabli	Le réseau est rétabli	Le réseau est rétabli
Action	Intervention du technicien SFR	Intervention du technicien SFR	Intervention du technicien SFR	-

In this example, every version of Vigogne managed to properly extract both the result and action of resolution while Mistral failed to extract the action of resolution.

Example 2: Only action included

- Incident : "L'application de Pricing n'est pas accessible"
- Note de résolution : "Merci de contacter l'admin Cloud pour paramétrer les droits d'accès."

Table 6. Second example of resolution result and action extraction through prompting an LLM

	Model			
Extraction	Vigogne-2-7B	Vigogne-2-7B-GPTQ	Vigogne-2-13B-GPTQ	Mistral-2-7B-GPTQ
Result	-	-	-	-
Action	Contacter l'admin cloud pour paramétrer les droits d'accès.	Contacter l'admin cloud pour paramétrer les droits d'accès.	Paramétrer les droits d'accès.	Contacter l'admin cloud pour paramétrer les droits d'accès.

In this second example, all the models manage to recognize that the result of resolution is not included in the resolution note. Similarly, they all managed to recognize the resolution action. What is interesting in this example is that none of the 5-shot examples are like this example and

despite that, the model manages to coherently understand the context and given task. A slight difference is observed in Vigogne-13B-GPTQ extracted action by targeting the technical process while the other models included admin contact directive as well. Both results are coherent, but it all depends on our needs.

Example 3: Only result included

Table 7. Third example of resolution result & action extraction through prompting an LLM

- Incident : "La fonctionnalité de modification ne fonctionne plus sur le site"
- Note de résolution : "Bonjour, Le site n'a plus de bug. Cordialement. Votre support informatique."

	Model			
Extraction	Vigogne-2-7B	Vigogne-2-7B-GPTQ	Vigogne-2-13B-GPTQ	Mistral-2-7B-GPTQ
Result	Le site n'a plus de bug	Le site n'a plus de bug	Le site n'a plus de bug	Le site n'a plus de bug
Action	-	-	-	Résolution du bug

In this example, the detection and extraction of resolution result was successfully done by both Vigogne and Mistral models. The greetings and sign-off phrases were detected as noise and were ignored by the models. In the action detection, however, Mistral fails to recognize the absence of the resolution action and considers implicitly that bug solving is the action that led to clearing bugs from the website. However, our objective is to extract significant information about the process behind the incident resolution. Consequently, the implicit action is not of value in our use-case. In any case, Mistral can also manage to align just by adapting the prompt.

Example 4: Long, detailed resolution note

- Incident : Serveur de développement lent. Problème de réseau.
- Resolution note :

Bonjour,

Afin de remédier aux retards de connectivité dans l'environnement de développement, nous avons optimisé les ressources du serveur en isolant les processus non essentiels, ajusté les paramètres du pare-feu, et réalisé un redémarrage contrôlé du serveur. Ces mesures ont restauré la stabilité du réseau, amélioré les performances du serveur.
Votre support informatique.
Cordialement

Table 8. Forth example of resolution result & action extraction through prompting an LLM

	Model			
Extraction	Vigogne-2-7B	Vigogne-2-7B-GPTQ	Vigogne-2-13B-GPTQ	Mistral-2-7B-GPTQ
Result	Le serveur est maintenant stable	Le serveur est maintenant stable	Stabilité du réseau restaurée	Le réseau est stable. Les performances du serveur sont améliorées
Action	Optimiser les ressources du serveur. Ajuster les paramètres du pare-feu. Redémarrer contrôlé le serveur.	Optimiser les ressources du serveur, ajuster les paramètres du pare-feu, redémarrer le serveur	Optimiser les ressources du serveur, ajuster les paramètres du pare-feu, et réaliser un redémarrage contrôlé du serveur	Isoler les processus non essentiels. Ajuster les paramètres du pare-feu. Redémarrage contrôlé du serveur

Even when presented with a detailed resolution note, the models consistently succeed in accurately extracting the resolution result and action. However, Vigogne-2-7B, whether with or without GPTQ, tends to confuse network stability with server stability. Only Mistral successfully extracts the comprehensive result that the network is stable, and the server performance has been optimized. Vigogne-2-13B-GPTQ, at the very least, can identify the restoration of server stability but falls short in extracting the server performance optimization details from the resolution note. Additionally, Vigogne-2-7B introduces a grammatical error in the extracted actions.
In summary, both Vigogne and Mistral excel in categorizing the presence of resolution results and actions, demonstrating a strong grasp of the given task by accurately extracting and phrasing the results and actions.

3. Conclusion

In this work, we investigated the efficacy and constraints of a state-of-the-art prompt-free few-shot classification approach using foundation models, compared to conventional methods. The fine-tuning of a Sentence Transformers model (SetFit) demonstrated efficiency in addressing semantic criteria like sentiment analysis, emotion classification, and topic modeling. However, SetFit faced challenges in accurately classifying intricate criteria such as grammatical acceptance or resolution note qualification. The complexity lies in the development of a compact yet representative training dataset. Overall, our findings indicate that prompt-free few-shot classification methods outperform prompt-based approaches when the classification criterion has a semantic dimension or is not too complex. This advantage stems from the contrastive fine-tuning which offers sentence representations adequately embedding the separation characteristics. Moreover, in prompt-free approaches, at minimal cost, a modest-sized model, with approximately 400MB, can be easily fine-tuned and seamlessly be integrated into a user-friendly application at a local level. We are excited for the future of both prompt-free and prompt-based few-shot classification as well as several other downstream tasks. One of our prospectives is fine-tuning an instruction-LLM to optimize its downstream task performance potentially using PEFT or QLORA to reduce GPU memory requirements for larger models training and inference.

We are excited for the future of both prompt-free and prompt-based few-shot classification as well as several other downstream tasks. One of our prospectives is fine-tuning an instruction-LLM to optimize its downstream task performance potentially using PEFT or QLORA to reduce GPU memory requirements for larger models training and inference.

REFERENCES

[1] Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." arXiv preprint arXiv:2108.07258 (2021).
[2] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
[3] Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog 1.8 (2019).

[4] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
[5] Yang, Zhilin, et al. "Xlnet: Generalized autoregressive pretraining for language understanding." Advances in neural information processing systems 32 (2019).
[6] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).
[7] Tunstall, Lewis, et al. "Efficient few-shot learning without prompts." arXiv preprint arXiv:2209.11055 (2022).
[8] Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." arXiv preprint arXiv:1908.10084 (2019).
[9] Maas, A. L., et al. (2011). Learning Word Vectors for Sentiment Analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011), June, 142–150. Portland, Oregon, USA: Association for Computational Linguistics. URL: http://www.aclweb.org/anthology/P11-1015
[10] Warstadt, A., et al. (2018). Neural Network Acceptability Judgments. In arXiv preprint arXiv:1805.12471. URL: https://arxiv.org/abs/1805.12471
[11] Saravia, E., et al. (2018). CARER: Contextualized Affect Representations for Emotion Recognition. In Proceedings of EMNLP 2018, Brussels, Belgium, 3687–3697. URL: https://www.aclweb.org/anthology/D18-1404
[12] Zhang, X., et al. (2015). Character-level Convolutional Networks for Text Classification. In NIPS 2015.
[13] Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).
[14] Jiang, Albert Q., et al. "Mistral 7B." arXiv preprint arXiv:2310.06825 (2023).
[15] Tunstall, Lewis, et al. "Zephyr: Direct distillation of lm alignment." arXiv preprint arXiv:2310.16944 (2023).
[16] Wang, Wenhui, et al. "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers." Advances in Neural Information Processing Systems 33 (2020): 5776-5788.
[17] Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108 (2019).
[18] Song, Kaitao, et al. "Mpnet: Masked and permuted pre-training for language understanding." Advances in Neural Information Processing Systems 33 (2020): 16857-16867.
[19] Xiao, Shitao, et al. "C-pack: Packaged resources to advance general chinese embedding." arXiv preprint arXiv:2309.07597 (2023).
[20] Li, Zehan, et al. "Towards general text embeddings with multi-stage contrastive learning." arXiv preprint arXiv:2308.03281 (2023).
[21] A.H. Wheeb, “Performance Analysis of VoIP in Wireless Networks,” International Journal of Computer Networks and Wireless Communications (IJCNWC), vol. 7, no. 4, pp. 1-5, 2017.
[22] Messaoudi, R. et al., Evaluating the Performance and Challenges of Prompt-Free Few-Shot Text Classification, volume 14 number 13, 10th International Conference on Computer Science and Information Technology (CSTY 2024), 2024

AUTHORS

Rim Messaoudi has completed her PhD from the University of Clermont Auvergne, France and the University of Sfax, Tunisia (2021). She has completed her initial education from various reputed educational institutes in Tunisia (ISIMM, EnetCOM and FSEGS). She completed her Master degree in 2017 and her License (Computer Science) in 2014. She has an experience in developing deep learning algorithms and writing research articles. She has been involved in the medical field and the detection of liver cancer lesions from MRI/CT scans images. Her areas of interest are machine learning (ML), artificial intelligence (AI), medical image processing and data classification. She has published several research papers in journals of repute and in refereed international conferences published by Springer. She is also contributing as a reviewer in the editorial boards of a few reputed journals.