Indigo at CheckThat! 2024: Using Setfit: A Resource Efficient Technique for Subjectivity Detection in News Article

Notebook for the CheckThat! Lab at CLEF 2024

Soumyadeep Sar, Dwaipayan Roy

Indian Institute Of Science Education and Research Kolkata, India

Abstract

The spread of misinformation and biased news across various reliable news outlets has led to serious consequences in our society. It has become crucial to understand the patterns in such misleading news articles, identify key evidence, and learn how to recognize false information. Subjectivity can play a pivotal role in identifying misleading news. In this work, we employed a resource-efficient method called *SetFit* on the English sub-task for Task-2 (subjectivity detection) at CheckThat!. This technique uses a few shot examples from training data and aims to produce results comparable to those of fully fine-tuned models like BERT on the entire dataset. For the selected sample data, we filter out conflict-resolved instances from the dataset and combine them with some other chosen data-points, then train our desired models on this dataset.

Keywords

SetFit, Sentence Transformers, Few shot text classification, Deep Learning, Fine-tuning, Natural Language processing, LLMs,

1. Introduction

In the realm of natural language processing, accurately detecting subjectivity in text is a crucial task for many reasons. It allows us to distinguish between objective information and content skewed by personal biases and opinions. This is particularly important in the digital age, where opinions and biases spread rapidly through media, potentially influencing public perception. Traditional methods for subjectivity detection often rely on large, labelled datasets, which can be expensive and time-consuming to create. Additionally, achieving high performance often involves fine-tuning large language models (LLMs), further increasing computational costs.

So we explore a resource-efficient approach at CLEF 2024 Check that! [1] for the Task-2 (Subjectivity Detection in News Articles) [2]. We only employed our methods on the English language sub-task of the challenge. We leverage *SetFit* [3], a few-shot learning algorithm that utilizes sentence embeddings and contrastive learning to efficiently fine-tune a model even with limited data. This technique enables the model to learn quickly with minimal labelled examples and requires fewer training epochs compared to traditional fine-tuning. Additionally, SetFit can be run on CPUs, eliminating the need for expensive GPUs. This study investigates the effectiveness of SetFit in distinguishing subjective and objective sentences within news articles. Our goal is to provide a robust and scalable solution for subjectivity detection, paving the way for more efficient and accurate identification of subjective content.

2. Related Work

The ability to distinguish between objective and subjective information in text is crucial for various natural language processing tasks. This distinction is particularly important in the digital age, where

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

soumyadeepsar26@gmail.com (S. Sar); dwaipayan.roy@iiserkol.ac.in (D. Roy)

https://github.com/Soumyadeepsar (S. Sar); https://github.com/dwaipayanroy (D. Roy)

0000-0002-5962-5983 (D. Roy)

opinions and biases spread rapidly through media, potentially influencing public perception. There are several works that deal with the problem. The work by Chaturvedi et al. [4] provide a comprehensive review of subjectivity detection methods, categorizing them into three types: hand-crafted, automatic, and multi-modal. Hand-crafted methods, while effective for identifying strong sentiments, struggle with weakly subjective sentences. Automatic methods, such as deep learning, overcome this limitation by creating meta-level feature representations that generalize well across domains and languages. Multi-modal methods further enhance accuracy by incorporating audio and video data with text using multiple kernels. This review highlights the advantages and limitations of each approach, emphasizing the challenges of high-dimensional n-gram features and the temporal nature of sentiments in long texts.

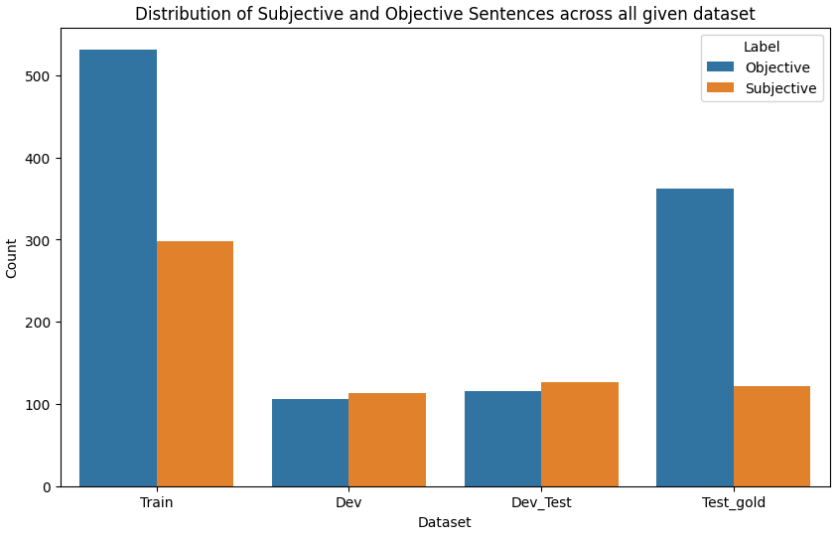

Figure 1: Entire Datasets distribution.

In another work, Pang and Lee [5] introduced a groundbreaking approach to subjectivity detection, focusing on identifying and categorizing subjective portions of a document to determine sentiment polarity. Their method utilizes text-categorization techniques and efficient graph-based algorithms to extract subjective text segments, allowing for the incorporation of cross-sentence contextual constraints. This approach significantly improves sentiment classification accuracy by targeting the relevant subjective content within the text.

Recent advancements have led to the development of resource-efficient methods for several text-processing tasks. Abdedaiem et al. [6] demonstrated the effectiveness of **SetFit**, that offers a highly efficient and prompt-free approach to fine-tuning Sentence Transformers (ST) for few-shot learning scenarios. Its two-stage process begins with contrastively fine-tuning a pre-trained ST model on a limited set of text pairs. This step leverages a Siamese architecture, where the model learns to distinguish similar and dissimilar sentences. The resulting fine-tuned ST then generates rich text embeddings, which are subsequently used to train a separate classification head for the specific task at hand. This elegant framework eliminates the need for handcrafted prompts, enabling accurate classification with minimal labeled data. In our work, we utilize SetFit to tackle the challenge of fake news detection in low-resource languages.

3. Background

3.1. Baseline solution

The baseline solution for subjectivity detection employed a powerful combination of Sentence-BERT [7] and Logistic Regression. Sentence-BERT, a pre-trained sentence encoder, was used to transform each statement into a high-dimensional vector representation, capturing its semantic meaning. This step

Table 1
Conflict resolved Statements in our dataset.

sentence	label	solved conflict
Gone are the days when they led the world in recession-busting	SUBJ	True
Whether this would be all for the best, or otherwise, is not yet the point.	SUBJ	True
Is this a warning of what's to come?	OBJ	True
"That is to say, we have not consciously intended it."	OBJ	True

provided a rich and informative representation of the sentence's content. Subsequently, a Logistic Regression classifier was trained on these sentence embeddings. This classifier learned to distinguish between objective and subjective statements based on the patterns and features present in the embeddings. The complete details and code for this baseline solution are available on the CLEF 2024 CheckThat! Lab GitLab repository[1], providing a valuable resource for researchers and practitioners interested in replicating or building upon this approach.

3.2. Dataset Distribution

We conducted a thorough analysis of the English dataset provided for the subjectivity detection task. The dataset is structured in a tab-separated format, conveniently pre-split into training, development, and development-test sets.

A noteworthy observation is the class imbalance within the training data: subjective sentences constitute only 35.9% (298) of the training set, while objective sentences account for 64.1% (532). This imbalance is not present in the development and development-test sets, which exhibit a more balanced distribution. While development and development-test sets show balanced distributions, the final test set where model performance is measured is skewed towards objective sentences (74.8%, 362) with only 25.2% (122) subjective sentences. This imbalance in the test set should be considered when interpreting and analyzing the model's performance. Figure 1 visually depicts the distribution of subjective and objective sentences across the four dataset splits. This figure provides a clear overview of the class distribution within each set and highlights the potential challenges associated with the imbalanced test set.

3.3. Schema for the Refined Dataset in Our Approach

As SetFit leverages contrastive learning, selecting a set of high-quality examples is crucial for its success. These examples will define the decision boundaries the model learns for the classification task. Given the limited number of examples used in few-shot learning, careful selection is essential.

In this work, we took an innovative approach by focusing on instances from the training set where annotator conflicts were resolved. This marks the first time in the competition that the 'solved-conflict' feature of the dataset has been utilized. We specifically chose 69 statements from the training split where *solved-conflict* had a value of *True*. These statements represent cases where human annotators initially disagreed on the subjectivity label, but ultimately reached a consensus. This consensus can be viewed as a strong indicator of the statement's true subjectivity or objectivity, making them valuable examples for contrastive learning. By focusing on these resolved-conflict instances, we aim to provide the model with clear and unambiguous examples, potentially leading to more robust and accurate decision boundaries for subjectivity classification. Some examples of such instances are shown in Table 1.

While resolved-conflict instances offer valuable training examples, relying solely on them could potentially lead to overfitting, as they represent a specific subset of the data. To address this, we incorporated a filtered selection of additional sentences from the training set. We identified the average number of words in each sentence across the entire training dataset as 22.84. This served as a baseline

[1]Subjectivity detection baseline

Table 2
Improvement in performance using our technique of data sampling.

Sampled Training dataset	Macro-avg F1	SUBJ F1
Randomly sampled dataset	0.73	0.72
Dataset designed according to our schema	**0.75**	**0.78**

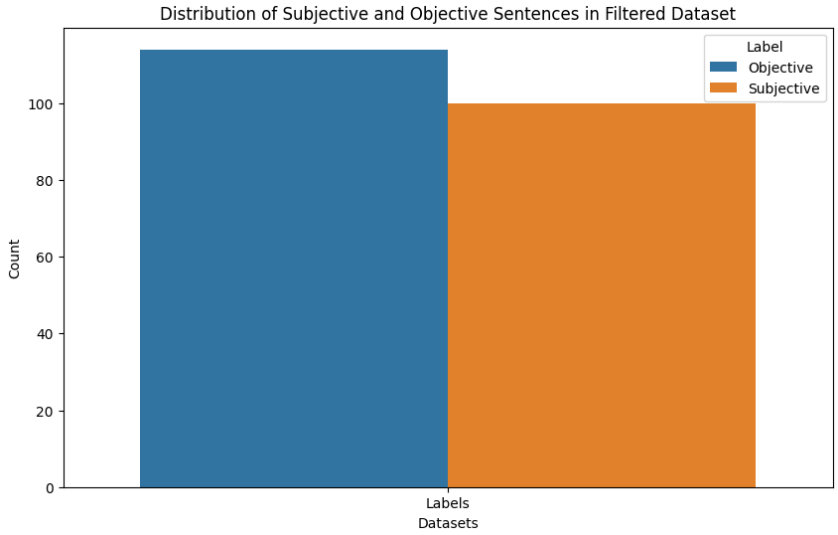

Figure 2: Filtered Dataset distribution.

for filtering sentences based on word count. We aimed to include sentences with a moderate length, avoiding excessively long or short ones. Therefore, we implemented a filtering process where we gradually increased the word count threshold, starting from 24 words. We observed that including sentences with more than 32 words resulted in a significant performance drop due to the limited number of available examples. Hence, we created our sampled dataset accordingly and removed those data points for whom the conflict was resolved since we are already considering them in the conflict-resolved sampled data. This sampled data contains 145 statements. Ultimately, combining both the sampled datasets (69 and 145), we finally get 214 sentences from the training data for contrastive learning. This represents a substantial reduction compared to the full training set, highlighting the efficiency of the approach. Notably, the filtered dataset maintained a balanced distribution with 114 subjective and 100 objective sentences, ensuring a representative sample for model training. Figure 2 visually depicts the distribution of subjective and objective sentences within the filtered dataset.

In Table 2 we demonstrate the improvement in performance of a Setfit Model with SVM classifier head when it is trained on a randomly sampled dataset (using default seed value of 42 for reproducibility) and on our specially designed sampled dataset. In both cases, they were trained for 1 epoch with the same hyperparameters. We clearly observe an improvement in performance of the Setfit Model when it is being trained on our dataset.

3.4. Working of SetFit

SetFit leverages a streamlined two-stage procedure to enhance sentence transformers for classification tasks. This approach offers significant efficiency gains while maintaining accuracy, particularly in scenarios with limited labelled data. The stages of SetFit are described below:

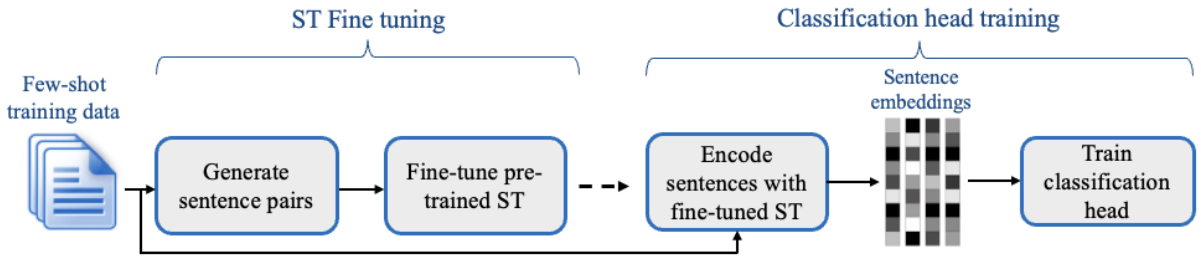

Figure 3: 2 step working phases in SetFit (https://huggingface.co/blog/assets/103_setfit/setfit_diagram_process.png).

3.4.1. Sentence Transformer Fine-tuning

The first stage commences with the provided few-shot training data. SetFit constructs sentence pairs from this training data, enabling the model to grasp the relationships and context within the text. Specifically the following pairs are strategically formed:

- **Positive pairs**: Sentences belonging to the same class are coupled together, representing examples of similar meaning and sentiment.
- **Negative pairs**: Sentences from different classes are paired together, showcasing contrasting meanings and sentiments.

The core objective of contrastive learning in this stage is to:

- **Minimize the distance**: between the embeddings generated for positive pairs, ensuring that sentences with similar meanings have closely aligned representations.
- **Maximize the distance**: between the embeddings generated for negative pairs, creating a clear differentiation between sentences with contrasting meaning and sentiment.

To achieve this, an appropriate loss function, such as the cosine similarity function, is employed to measure the semantic similarity between the sentences within each pair. The model is then fine-tuned to generate embeddings that effectively capture these semantic relationships, laying the foundation for accurate classification.

3.4.2. Classification Head Training

The fine-tuned sentence transformer from the first stage plays a crucial role in the second stage. It encodes each sentence, capturing the essence of the text data in a format suitable for classification. These sentence embeddings can then be utilized to train a variety of classical machine learning models for subjectivity classification such as: **Support Vector Machines (SVMs)** or **Logistic Regression**. Alternatively, *Differentiable Linear Neural Layer*, a neural network layer can be employed for classification. This layer can be trained with its own hyperparameters, potentially different from those used in the first stage. This flexibility allows for customization and fine-tuning of the classification head without retraining the entire model as a single end-to-end system.

By adopting this two-stage approach, SetFit effectively leverages the power of contrastive learning to enhance sentence transformers for subjectivity classification. It achieves this with remarkable efficiency, requiring significantly less data and computational resources compared to traditional fine-tuning methods. A simplified diagram for the entire process is described in the Fig. 3

4. Experiment and Evaluation

4.1. Model and Hyperparameters used

The field of pre-trained sentence transformers offers a wealth of options for various tasks. To select the most suitable model for our subjectivity classification task, we explored the extensive collection

Table 3
Performance of different Classification heads tried for SetFit.

Classification head	Macro-avg F1 score	SUBJ F1 score
Random Forest head	0.72	0.78
SVM Classifier head	0.74	0.78
Linear Differentiable head	**0.74**	**0.79**

Table 4
Hyperparameters used for SetFit.

Hyperparameter	Values
batch size	8
number of epochs (for both the training phases)	1
learning rate (head and body training)	2e-05
seed	42
loss function	Cosine Similarity function
number of iterations	20

available on the sentence bert (SBERT) leaderboard [2]. These pre-trained models have been trained on a massive dataset exceeding one billion training pairs, making them well-equipped for general-purpose use. While all models offer robust performance, the following two models stand out to be significantly effective: i) **all-mpnet-base-v2 model** and ii) **all-MiniLM-L6-v2 model**.

all-MiniLM-L6-v2 model is known for its faster speed and good quality results. On the other hand all-mpnet-base-v2 model is the best model in the entire leaderboard of SBERT, producing the best quality embeddings for a variety of NLP tasks. Since all-mpnet-base-v2 model occupies the topmost position in the leaderboard of SBERT, we consider choosing it for our experiments over the all-MiniLM-L6-v2 model. For the fine-tuning process, we leveraged the filtered dataset of 214 sentences. We specifically trained our model for 1 epoch rather than trying for higher epochs in order to develop resource efficient models, which could perform as good as fully-finetuned LLMs. The cosine similarity was employed to evaluate the semantic similarity between the generated embeddings to provide a reliable measure of their alignment. The seed value was set to 42 to ensure the reproducibility of the results.

For the classification head, multiple scikit-learn based ML classifiers were tried like SVM, Random Forest, etc. We tested their performance using the same hyper-parameters and on the same development-test dataset available before the evaluation cycle of the competition. The performances are listed in the Table 3. The best performing classifier was the linear differentiable layer, which is being trained for classifying the sentence embeddings for 1 epoch. All the hyperparameters used for finetuning the sentence bert and the train the classification head is mentioned in Table 4

4.2. Comparative Performance Analysis

We fully fine-tuned other BERT-based large language models (LLMs) on the entire training dataset of 830 sentences. The models used for comparison are-BERT[8] and RoBERTa[9] . The transfomers[10] library from HuggingFace was being used to get pre-trained models of BERT and RoBERTa. The hyperparameters for training these models were kept the same as those for Setfit for fair comparison. We fine-tuned both the LLMs for 1 epoch and for 4 epochs, respectively. The 1 epoch fine-tuned model's performance will demonstrate the effectiveness of contrastive learning employed in Setfit, which uses nearly 4 times smaller data, yet produces a better result. A comparison is made between their performance on the development test (dev-test.tsv) and the test datasets provided during the evaluation period for the competition. The results are listed in Table 5. We discuss the results obtained in the Result section below.

[2] https://sbert.net/docs/sentence_transformer/pretrained_models.html

Table 5
Comparative study of performances of different models.

	Dev-test		Test	
Models	SUBJ F1	Macro-avg F1	SUBJ F1	Macro-avg F1
BERT(1 epoch)	0.69	0.63	0.35	0.61
RoBERTa (1 epoch)	0.56	0.64	0.23	0.55
Setfit (1 epoch)	**0.79**	**0.74**	**0.47**	**0.64**
BERT (4 epochs)	0.74	0.75	0.57	0.73
RoBERTa (4 epochs)	0.77	0.78	0.54	0.71

5. Result

The performance in Table 5 shows that Setfit performed much better when compared to fine-tuned LLMs, which were trained for 1 epoch. This demonstrates the effectiveness of contrastive learning that allows our model to get well-trained on a 4 times smaller sampled dataset within just 1 epoch. We also notice that the performance of fully fine-tuned models is better than our SetFit approach when they are trained for 4 epochs. This is an anticipated result since they are seeing the entire training dataset and getting a lot of time (epochs) to train. But training on the entire dataset and for a large number of epochs is clearly resource intensive, which from the beginning was our goal to avoid. So considering our approach as a resource-efficient technique the performance was fairly good.

6. Conclusion

In this article, we report the performance of our proposed method for the CLEF 2024 task *CheckThat*. Our approach has successfully outperformed the baseline solution, demonstrating its potential for subjectivity detection. The proposed approach was specifically focused on developing a computationally efficient technique that delivers competitive results compared to existing state-of-the-art models. Large models like GPT, while powerful, contribute significantly to the carbon footprint, posing a threat to our environment. This research delves into the development of lighter and more efficient NLP solutions, potentially paving the way for replacing massive LLMs in various applications in the near future.

References

[1] A. Barrón-Cedeño, F. Alam, T. Chakraborty, T. Elsayed, P. Nakov, P. Przybyła, J. M. Struß, F. Haouari, M. Hasanain, F. Ruggeri, X. Song, R. Suwaileh, The clef-2024 checkthat! lab: Check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 449–458.

[2] J. M. Struß, F. Ruggeri, A. Barrón-Cedeño, F. Alam, D. Dimitrov, A. Galassi, M. Siegel, M. Wiegand, Overview of the CLEF-2024 CheckThat! lab task 2 on subjectivity in news articles, 2024.

[3] L. Tunstall, N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, O. Pereg, Efficient few-shot learning without prompts, 2022. arXiv:2209.11055.

[4] I. Chaturvedi, E. Cambria, R. E. Welsch, F. Herrera, Distinguishing between facts and opinions for sentiment analysis: Survey and challenges, Inf. Fusion 44 (2018) 65–77. URL: https://doi.org/10.1016/j.inffus.2017.12.006. doi:10.1016/J.INFFUS.2017.12.006.

[5] B. Pang, L. Lee, A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, in: D. Scott, W. Daelemans, M. A. Walker (Eds.), Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain, ACL, 2004, pp. 271–278. URL: https://aclanthology.org/P04-1035/. doi:10.3115/1218955.1218990.

[6] A. Abdedaiem, A. H. Dahou, M. A. Chéragui, Fake news detection in low resource languages using setfit framework, Inteligencia Artif. 26 (2023) 178–201. URL: https://doi.org/10.4114/intartif.vol26iss72pp178-201. doi:10.4114/INTARTIF.VOL26ISS72PP178-201.

[7] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.

[8] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/V1/N19-1423.

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.

[10] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.