Gpachov at CheckThat! 2023: A Diverse Multi-Approach Ensemble for Subjectivity Detection in News Articles

Notebook for the CheckThat Lab at CLEF 2023

Georgi Pachov[1,*], Dimitar Dimitrov[1], Ivan Koychev[1] and Preslav Nakov[2]

[1]*Sofia University "St. Kliment Ohridski", Bulgaria*
[2]*Mohamed bin Zayed University of Artificial Intelligence, UAE*

Abstract
The wide-spread use of social networks has given rise to subjective, misleading, and even false information on the Internet. Thus, subjectivity detection can play an important role in ensuring the objectiveness and the quality of a piece of information. This paper presents the solution built by the Gpachov team for the CLEF-2023 CheckThat! lab Task 2 on subjectivity detection. Three different research directions are explored. The first one is based on fine-tuning a sentence embeddings encoder model and dimensionality reduction. The second one explores a sample-efficient few-shot learning model. The third one evaluates fine-tuning a multilingual transformer on an altered dataset, using data from multiple languages. Finally, the three approaches are combined in a simple majority voting ensemble, resulting in 0.77 macro F1 on the test set and achieving 2nd place on the English subtask.

Keywords
Subjectivity detection, Sentence Embeddings, Few-shot learning, Transformer, Ensemble, Natural Language Processing, Deep Learning

1. Introduction

Subjectivity is a feature of language and a form of bias, in which whenever a person is sharing information, it comes out skewed by the speaker's own personal preferences, beliefs and views. In today's interconnected world, where opinions and biases travel fast and far, subjectivity detection can be a very important piece in order to ensure information reporting is done in a clear, objective and unbiased fashion.

Specifically, subjectivity in news and media articles can be nuanced, subtle and difficult to identify. Detection of subjectivity in such texts can play an important role in identifying potentially misleading or malicious texts and in detecting fake news online.

In Task 2 of CheckThat! Lab at CLEF 2023 [1, 2], systems are required to distinguish whether a sentence from a news article expresses the subjective view of the author or presents an objective

view on the covered topic instead. This is a binary classification task in which systems have to identify whether a text (a sentence or a paragraph) is subjective or objective.

This paper explores the effects of fine-tuning a large pre-trained language model on the subjectivity task. Additionally, we have examined the angles of few-shot-learning and fine-tuning a sentence embedding model. Finally, an ensemble method is proposed to unify all three into one solution.

2. Related Work

While sentiment analysis can be regarded as a classic NLP task with lots of research already available on the subject, subjectivity classification is deemed to be a slightly less popular research topic. [3] looks at the subjectivity detection task as a way to improve sentiment analysis classifiers by excluding neutral (objective) sentences. They offer a broad survey on published subjectivity detection methods, categorizing them into syntactic (keyword-spotting, lexical affinity, statistical methods), semantic (parse trees, convolutional neural networks, extreme learning machines) and multi-modal (BiLSTM, multiple-kernel learning).

In [4], authors explore multi-task learning with hard parameter sharing via Neural Tensor Network. They demonstrate that using a single network with shared layers while learning on two semantically related datasets can improve performance on both datasets.

In [5], authors compare Word2Vec and BERT embedding models in the context of subjectivity detection. With additional classification models to process the embedding outputs, authors demonstrate the superiority of BERT embeddings in high-resource settings, while showing Word2Vec embeddings can be more efficient in low-resource settings. In the current challenge, pairing an embedding encoder with various classification models is also explored.

In [6], authors compare the performance of pure transformer models against a variety of more specialized methods for short text classification. Their results show superior performance of the transformer models and part of the research performed in this paper is influenced by their findings.

3. Data and Baseline Solution

For the subjectivity detection task, datasets in 6 different languages were provided - Arabic, Dutch, English, German, Italian and Turkish. A total of 7 datasets were available - one for each language, and one for the multilingual version of the task. The English dataset contained a total of 1019 examples. 800 of the provided examples were labeled as training, the other 219 as validation. A baseline solution[1] is provided by competition organizers, which consists of a sentence encoder model, producing sentence embeddings, which are then classified with Logistic Regression.

An interesting insight was that most of the sequences were relatively short. For the English dataset, the average number of words in a sequence was 23, while 90% of sequences consisted of 40 words or less.

---

[1]https://gitlab.com/checkthat_lab/clef2023-checkthat-lab/-/tree/main/task2/baseline

The English training set is imbalanced, with 64% of samples labeled objective and 36% - subjective. This imbalance is not present in the validation set.

4. Experiments and Evaluation

Three research directions were explored, each of them resulting in a separate solution. The final program used for submission is a simple majority voting ensemble of the three solutions. The first research direction explores what is achievable using sentence embeddings. The second one looks at a few-shot-learning model and dual-stage fine-tuning. The third is based on fine-tuning a pre-trained transformer model, also utilizing training data from the other languages available for the task.

All evaluations of experiments are done on the English validation set, provided by the organizers. All research directions will be described in more detail in the next subsections.

4.1. Sentence Embeddings

Multiple experiments with sentence embeddings were conducted. All of them were based on using a pre-trained sentence embedding encoder model [7] . The following ideas were explored:

- Using more powerful classifiers on top of sentence embeddings output
- Using dimensionality reduction
- Fine-tuning the sentence embeddings encoder in the context of subjectivity detection

Initially, the baseline solution provided by organizers used sentence embeddings and a simple Logistic Regression on top to produce classification outputs. More complex classifiers were tested. Multiple different classifiers yielded improvements, measured on the validation set. The most performant was LogisticRegression (from sklearn), using ElasticNet penalty, balanced class weights, 'saga' solver and 0.5 as regularization constant.

A potential place for improvement was related to sentence embeddings dimensionality. Due to only having 800 training examples, embeddings of dimensionality 384 could prove challenging for a classifier. Using dimensionality reduction, information from embedding vectors can be further compressed in a way that could make it easier for classifiers to find a proper decision boundary.

Best performance was achieved using PCA with 110 remaining components (out of 384), which explained 92.5% of total variance. Experiments with different classifiers and dimensionality reduction are summarized in the first column of Table 1.

While dimensionality reduction can help classifiers, the sentence embeddings themselves were generally created for a very broad category of NLP tasks. Fine-tuning the embeddings encoder for subjectivity detection proved to be helpful for all of the tested classifiers.

Embeddings were fine-tuned using cosine similarity loss in a contrastive learning manner. The new similarity label of a pair of sentences consisted of two equally weighted components - their original similarity and their label-based similarity. Label-based similarity is defined as 1 if the two sentences are from the same class and 0 otherwise. This can be summarized with the following equation:

Table 1

Classifiers and dimensionality reduction with original vs fine-tuned embeddings encoder (Macro F1)

Classifier\Embedding	Original Sentence Embeddings	Fine-Tuned Embeddings (N=100)
Baseline (SBERT + LR)	0.74	0.76
SVM	0.76	0.78
ElasticNet	0.77	0.8
PCA + ElasticNet	**0.78**	**0.81**

Table 2

Few-shot learning experiments with SetFit

Fine-tuning regime	Number of samples	Macro F1
Single stage	10	0.79
Single stage	20	0.8
Dual stage	20	**0.81**
Dual stage	64	0.8
Dual stage	100	0.79

$$New\_Similarity\_Label(A, B) = 0.5 * Similarity(A, B) + 0.5 * (class(A) == class(B))$$
(1)

To generate training samples, N objective and N subjective sentences were randomly selected. Training pairs were generated - each sentence was paired with every other sentence and a similarity label was generated using Formula 1. In total, 2N*(2N-1) training pairs were generated.

All of the tested classifiers performed better using the fine-tuned embeddings (with N=100) instead of original embeddings. All experiments with classifiers, dimensionality reduction and fine-tuned sentence embeddings are summarized in Table 1. For both the original and the fine-tuned embeddings, best performance was achieved through PCA and Linear Regression with elastic net penalty. All macro F1 scores are measured on the (English) validation set.

4.2. Few-Shot Learning

The second research direction explored what can be achieved with few-shot learning. Experiments are based on the SetFit model from HuggingFace [8]. While conceptually similar to the idea of fine-tuning sentence embeddings, the SetFit model has numerous advantages, including faster training, better sample efficiency and a dual-stage fine-tuning mechanism. In the first stage, the classification head is frozen and embeddings are fine-tuned. Vice versa in the second stage.

Experiments and results for this approach are outlined in Table 2. Results are similar to fine-tuning sentence embeddings encoder, but achieved while using a lot less information (samples) from the task-specific dataset, which showed promise of low variance and good generalization capabilities.

Table 3
Transformer experiments with English dataset

Transformer Model	Architecture	Macro F1
BERT	base	0.77
BERT	large	0.78
RoBERTa	base	0.81
DeBERTa-v2	large	**0.82**

Table 4
Transformer experiments with multilingual datasets

Model	Architecture	Training Set	F1 macro
bert-multilingual	base	All data	0.81
xlm-roberta	large	All data	0.81
mdeberta-v3	base	All data	0.82
xlm-roberta	base	All data	0.83
xlm-roberta	base	English, Arabic and Turkish	0.82
xlm-roberta	base	English, German translated to English	**0.84**

4.3. Transformer Fine-Tuning

The third research direction was based on the idea of fine-tuning a transformer model to the dataset provided. Lots of transformer models are available in the HuggingFace hub [9]. BERT [10], RoBERTa [11] and DeBERTa [12] were used throughout the experiments. Limitations in hardware capabilities restricted the experiments to only 'base' and 'large' variants of the models. Xlarge and xxlarge models were not tested. Initially, experiments were performed using the English dataset only. Later on, additional experiments were performed including data from other languages.

Table 3 summarizes the experiments and results when using the English dataset. Fine-tuning a DeBERTa-v2-large achieved the best performance on the validation set.

Additional experiments were performed using data from other languages. Multiple training datasets were created. The first one consisted of all the available data for all languages. The second one consisted of English, Arabic and Turkish - the languages who had published baseline solutions with higher F1 than English. The third one contained training samples from English and German translated to English using a neural model [13]. While creating the datasets, English validation samples were always manually excluded as a final step to prevent data leakage.

For all three datasets, multilingual transformer models were used - bert-base-multilingual [10], mDeBERTa [14] and xlm-roberta-base [15].

Table 4 summarizes the results. An xlm-roberta-base model was fine-tuned using all available data, achieving a macro F1 score of 0.83, which showed slight improvement over using English-only dataset. The best results were obtained by using English and German translated to English, fitted to xlm-roberta-base model.
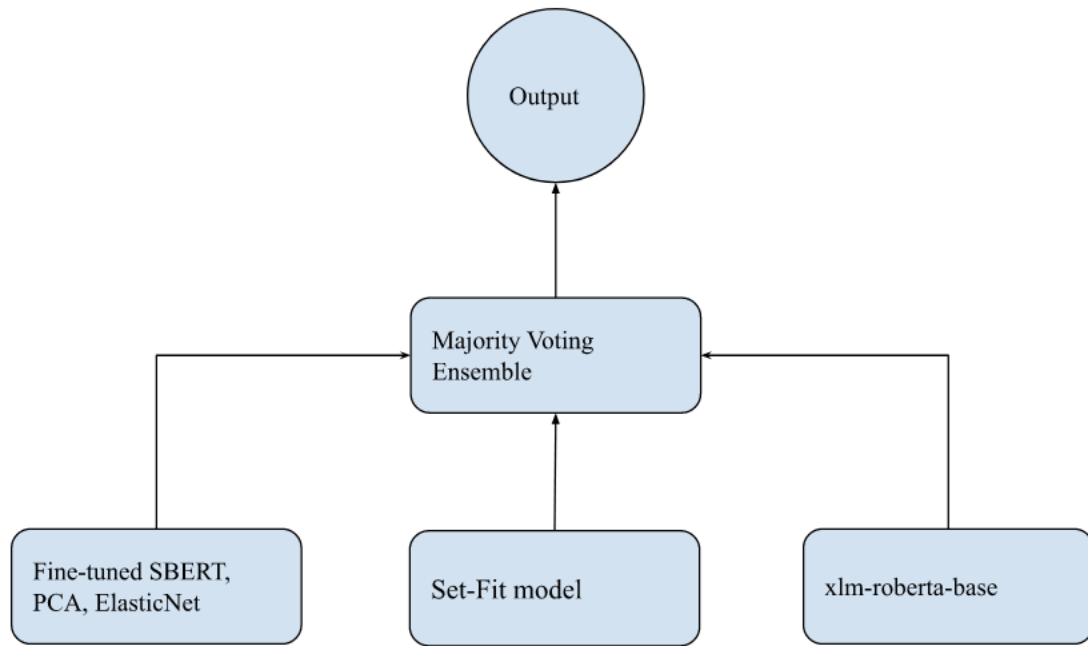
Figure 1: Schema of the ensemble

4.4. Ensemble

Best performing solutions from all three research directions were chosen and formed into an ensemble (Figure 1). The final solution was a simple majority voting ensemble from the following solutions:

- A fine-tuned sentence embeddings encoder, producing improved sentence embeddings (in the context of this task). The encoder output passes through dimensionality reduction (384 dimensions reduced to 110). The reduced embeddings are then classified through LogisticRegression with equally weighted L1 and L2 penalties, 'saga' solver, balanced class weights and 0.5 regularization constant.
- A few-shot learning SetFit model trained using dual-stage fine-tuning procedure with N=20 sentence pairs.
- An xlm-roberta-base model, fine-tuned on English and German translated to English.

The final submitted ensemble was able to achieve 0.85 macro F1 on English validation set.

5. Conclusions

The final solution achieved macro F1 of 0.85 on validation set and 0.77 on test set. This indicates generalization issues, likely stemming from overfitting the solution to the validation set.

Each of the three separate methods was analyzed on the test set, to identify possible causes for the significantly lower performance on the test set. All components of the solution perform

Table 5
Test scores of individual solutions and ensemble

Solution	Val F1 score	Test F1 score
Fine-tuned S-BERT, PCA, ElasticNet	0.81	0.7
SetFit model	0.81	0.76
xlm-roberta-base	0.84	**0.79**
Majority voting ensemble	**0.85**	0.77

Table 6
Per-class precision, recall and f1-scores of the final solution

Class	Precision	Recall	F1
subjective	0.83	0.71	0.77
objective	0.73	0.84	0.78

worse on the test set than on the validation set, indicating a more difficult test set. The first method however incurs a bigger loss of performance than the others. This is likely caused by an issue in the fine-tuning procedure, resulting in the encoder model producing highly specialized embeddings, which lose a bigger part of their pretrained semantics than optimal and thus generalize poorly. Less examples used for contrastive learning, smaller learning rate, or using a holdout set to check for generalization issues could have mitigated this problem.

Results also indicate that the best performer is an xlm-roberta-base model trained on English and translated German. All test and validation F1 scores are summarized in Table 5.

Table 6 indicates that the solution is biased to expect more frequent 'objective' examples, hence the higher recall but lowered precision. This is likely stemming from the initial class imbalance in the English training set, where 64% of examples are labeled objective. While this imbalance doesn't seem to have a major impact on final F1 scores, a more balanced solution could have been achieved by using any of the well-known class-imbalance techniques, e.g. sample weighing or choosing a different threshold.

Overall, xlm-roberta model trained on English and translated German seem to be the best performer. The few-shot-learning approach yielded decent test results. The fine-tuned SBERT encoder method seems to not be general enough and is reducing the performance of the full solution. Submitting predictions from only the transformer model would have resulted in 0.79 macro F1, which could have won the English subtask challenge.

6. Future Work

As indicated by Table 5, the transformer-based solution proved to be the most effective and robust out of the ones used in the ensemble. Further experiments can be conducted with newer and more promising versions of existing transformer models. For example, DeBERTa-V3 [14] is reported to improve performance of the original DeBERTa model with 1.37% on the GLUE benchmark. This can prove relevant to the task of subjectivity classification as well.

Additionally, due to resource constraints, transformer models with 'xlarge' and 'xxlarge' architectures were not used for this research. For the same reason, very little hyperparameter tuning was performed. Using a larger model and exploring bigger hyperparameter space can potentially improve the results.

Finally, the transformed-based approach used for subtask 2A (English) can also be attempted and used for the other 5 languages in the task - Arabic, Dutch, German, Italian and Turkish.

Acknowledgments

References

[1] A. Barrón-Cedeño, F. Alam, A. Galassi, G. Da San Martino, P. Nakov, , T. Elsayed, D. Azizov, T. Caselli, G. Cheema, F. Haouari, M. Hasanain, M. Kutlu, C. Li, F. Ruggeri, J. M. Struß, W. Zaghouani, Overview of the CLEF–2023 CheckThat! Lab checkworthiness, subjectivity, political bias, factuality, and authority of news articles and their source, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), 2023.

[2] A. Galassi, F. Ruggeri, A. Barrón-Cedeño, F. Alam, T. Caselli, M. Kutlu, J. Struss, F. Antici, M. Hasanain, J. Köhler, K. Korre, F. Leistra, A. Muti, M. Siegel, T. Mehmet Deniz, M. Wiegand, W. Zaghouani, Overview of the CLEF-2023 CheckThat! lab task 2 on subjectivity in news articles, in: M. Aliannejadi, G. Faggioli, N. Ferro, Vlachos, Michalis (Eds.), Working Notes of CLEF 2023–Conference and Labs of the Evaluation Forum, CLEF 2023, Thessaloniki, Greece, 2023.

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

[4] R. Satapathy, S. Pardeshi, E. Cambria, Polarity and subjectivity detection with multitask learning and bert embedding, 2022. arXiv:2201.05363.

[5] W. Chong, H. Ng, T. T. V. Yap, W. Soo, V. T. Goh, D. Cher, Objectivity and Subjectivity Classification with BERT for Bahasa Melayu, 2022, pp. 246–257. doi:10.2991/978-94-6463-094-7_20.

[6] F. Karl, A. Scherp, Transformers are short text classifiers: A study of inductive short text classifiers on benchmarks and real-world datasets, 2022. arXiv:2211.16878.

[7] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. arXiv:1908.10084.

[8] 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`.

[9] 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. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. 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. doi:`10.18653/v1/2020.emnlp-demos.6`.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. `arXiv:1810.04805`.

[11] 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`.

[12] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, 2021. `arXiv:2006.03654`.

[13] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, A. Fan, Multilingual translation with extensible multilingual pretraining and finetuning, 2020. `arXiv:2008.00401`.

[14] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2023. `arXiv:2111.09543`.

[15] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. doi:`10.18653/v1/2020.acl-main.747`.