DS@GT at CheckThat! 2025: Detecting Subjectivity via Transfer-Learning and Corrective Data Augmentation

Notebook for the CheckThat! Lab at CLEF 2025

Maximilian Heil[1,*], Dionne Bang[1]

[1]Georgia Institute of Technology, North Ave NW, Atlanta, GA 30332

Abstract

This paper presents our submission to Task 1, Subjectivity Detection, of the CheckThat! Lab at CLEF 2025. We investigate the effectiveness of transfer-learning and stylistic data augmentation to improve classification of subjective and objective sentences in English news text. Our approach contrasts fine-tuning of pre-trained encoders and transfer-learning of fine-tuned transformer on related tasks. We also introduce a controlled augmentation pipeline using GPT-4o to generate paraphrases in predefined subjectivity styles. To ensure label and style consistency, we employ the same model to correct and refine the generated samples. Results show that transfer-learning of specified encoders outperforms fine-tuning general-purpose ones, and that carefully curated augmentation significantly enhances model robustness, especially in detecting subjective content. Our official submission placed us 16^{th} of 24 participants. Overall, our findings underscore the value of combining encoder specialization with label-consistent augmentation for improved subjectivity detection. Our code is available at https://github.com/dsgt-arc/checkthat-2025-subject.

Keywords

Subjectivity Detection, Transfer Learning, Transformer, Data Generation, GPT, Fine Tuning, CEUR-WS

1. Introduction

Given the great risk of misinformation globally[1], the need for automatic fact-check systems is vital. Similar to machine learning pipelines, an automatic fact-checking system also includes more than just a classifier: retrieval to equip the system with evidence for the fact-check, data preparation to comply with the format requirements of the system, training or fine-tuning to enhance classification performance, and much more. For example, an objective sentence can directly be fact-checked, a subjective sentence needs further data augmentation before it can be passed forward into a fact-checking system. The subjective sentence must be stripped from emotions, opinions or personal interpretations so that the fact-checking system can subsequently focus on the factual verification only. This motivates the CheckThat! Lab of CLEF 2025[2], where Task 1 focuses on identifying subjective and objective sentences in news paper articles[3]. Task 1 of CheckThat! has evolved over recent years to address subjectivity detection in multilingual and monolingual contexts. Previous editions in 2023 and 2024 have established strong baselines using transformer-based models and explored both traditional and generative approaches [4, 5]. Participating teams have applied lexicon-based classifiers, fine-tuned encoders, and increasingly, synthetic data generation to boost performance under limited data settings. In this paper, we present our contribution to the 2025 English monolingual task. Our approach explores three key research areas focusing on transfer-learning, data-augmentation and the ability of a generative model to correct and refine itself (self-correction). We evaluate both general pre-trained encoders and compare them with encoders that have already been fine-tuned on related tasks (specified encoders). In addition, we investigate the role of data augmentation through stylistic paraphrasing via a large language model (LLM). Furthermore, we introduce a correction pipeline using the same LLM to align generated paraphrases with their intended labels and stylistic attributes. The impact of each

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

mheil7@gatech.edu (M. Heil); dbang6@gatech.edu (D. Bang)

0009-0002-6459-6459 (M. Heil); 0000-0002-9165-8718 (D. Bang)

component is assessed through detailed ablation experiments. Overall, our approach to the competition ranks us 16 out of 24.

The paper is structured as follows: Section 2 highlights related work, Section 3 presents our methodology, Section 4 describes the dataset, Section 5 and Section 6 shows and discusses the results, Section 7 highlights future research avenues, and Section 8 concludes.

2. Related Work

Subjectivity detection has a long history across contexts[6], domains[7], and languages[8]. More specifically, subjectivity detection in news paper articles has a three-year history with CheckThat! at CLEF. Most of the results [4, 5] have been driven via the advent of transformer architectures [9] and the introduction of BERT-like encoder models[10] for natural language processing (NLP). NLP systems now easily expand across domains or languages with high accuracy and robustness. In addition, generative models have made substantial contributions due to their outstanding zero-shot and few-shot capability, as well as a significant long context window[11].

Last year's winner of the monolingual English task, Team DWReCo, employed an LLM to generate subjective training examples with subjectivity styles (e.g. partisan, exaggerated, emotional) given expert knowledge. Synthetic data has been used to balance the data set and enrich the fine-tuning of a *RoBERTa-base*[12] model with a classification head.[13] Our data augmentation approach is influenced by Team DWReCo, but we are less interested in data sampling. In contrast, we use stylistic data augmentation for contrastive learning[14]. The 2023 monolingual English task winner, Team HYBRINFOX, builds an ensemble of a fine-tuned *RoBERTa-base* and a *DistilBERT-base-nli-mean-tokens* (sentence transformer[15]) to capture the synthetic as well as semantic meaning of the sentences. This is completed with a self-designed expert system that classifies given NER and lexicon-based methods[16]. Similarly, we also fine-tune general-purpose encoders on the data set and, in addition, investigate the capability of transfer learning in subjectivity detection.

3. Methodology

In the course of the competition, we explored the potential of specialized encoders and data augmentation. As shown in Table 1, we contrasted the fine-tuning of general-purpose encoders with transfer-learning of specialized encoders, fine-tuning both on the original data set. In a second step, we explored data augmentation and investigated its added benefit. Finally, we also added a self-corrective data alignment procedure to the data augmentation to ensure that generated paraphrases match their intended labels and styles, using GPT-4o to identify and rewrite those it considers inconsistent.

Table 1
Overview of Explored Methods

Domain	Model	Dataset
General	RoBERTa-base	Original Train
	MiniLM-L12-v2	Original Train
	ModernBERT	Original Train
Transfer	Sentiment-Analysis-BERT	Original Train
	Emotion-English-DistilroBERTa-base	Original Train
	Emotion-English-RoBERTAa-base	Original Train
Transfer	Sentiment-Analysis-BERT	Augmented Train
	Emotion-English-DistilRoBERTa-base	Augmented Train
Transfer	Sentiment-Analysis-BERT	Self-Corrected Augmented Train
	Emotion-English-DistilRoBERTa-base	Self-Corrected Augmented Train

First, our approach contrasted general-purpose encoders and specialized encoders by evaluating

their respective capabilities to distinguish objective from subjective newspaper sentences. Initially, we assessed the performance of general-purpose encoders including *RoBERTa-base*, *MiniLM-L12-v2*, and *ModernBERT*[17]). This evaluation focused on comparing their capabilities in token-level and sentence-level semantic relationships, as well as contextual understanding. We then employed the transfer-learning capabilities of encoders *Sentiment-Analysis-BERT*[18], *Emotion-English-DistilroBERTa-base*, and *Emotion-English-RoBERTa-large* [19], which were already fine-tuned on related tasks with domain-specific datasets for sentiment analysis and emotion recognition. These models are better equipped to detect emotional tone and subjective language, improving their ability to distinguish between subjective and objective statements.

Hypothesis H1: Transfer-learning with specialized encoders will result in greater sensitivity in distinguishing between subjective and objective language expressions.

Second, we investigated the added benefit of data augmentation. Given the small size of the original dataset, we pursued a strategy to synthetically augment and expand the training data using GPT-4o[20]. Inspired by ClaimDecomp[21], which decomposed complex political claims into literal and implied sub-questions to improve fact verification, we hypothesized that generating stylistic paraphrases of labeled sentences could improve classification performance by increasing training diversity.

ClaimDecomp evaluated sub-question generation using encoder-based language models but found that while literal questions were tractable, implied ones remained challenging due to the models' limited reasoning and contextual capabilities. They also noted that larger language models like GPT could be more suitable for handling these implicit inferences. In parallel, the CLEF CheckThat! 2024 overview[5] found that transformer-based classifiers augmented with domain-informed synthetic data outperformed baselines and that models consistently struggled more with identifying subjective sentences across languages.

Drawing on these findings, we adopted a generation approach similar in spirit to CLaC-2[22], who used zero-shot GPT-3 to generate two paraphrases per sentence and labeled them via majority vote. However, rather than perform on-the-fly classification, we used GPT-4o in a few-shot setup to generate paraphrases with controlled styles, both subjective and objective, based on the original sentence content. Unlike DWReCo[13], who used a zero-shot prompt method to generate new subjective examples based on a subjectivity style checklist, we generated both subjective and objective sentences to augment each data point. DWReCo also found that augmentation with paraphrased sentences produced lower-diversity samples and required post-hoc filtering. We did not apply such filtering in our submitted results due to time constraints.

Our generation approach was guided by the hypothesis that providing both subjective and objective perspectives per original content could improve a model's ability to distinguish stylistic cues in spirit of contrastive learning. For each subjective sentence, we generated two or six objective paraphrases. Conversely, for each objective sentence, we generated two or six subjective variants, explicitly styled as *propaganda*, *exaggerated*, *emotional*, *derogatory*, *partisan*, or *prejudiced* (categories aligned with DWReCo's most effective style prompts).

- Original (SUBJ): "Gone are the days when they led the world in recession-busting."
- Generated (OBJ): "The era in which they were at the forefront of overcoming economic downturns has ended."

- Original (OBJ): "The trend is expected to reverse as soon as next month."
- Generated (SUBJ): "A promising turnaround is on the horizon, with expectations for change as early as next month."

This resulted in two augmented datasets:

- **Balanced-2**: Each sentence augmented with 2 paraphrases in the opposite style

- **Balanced-6**: Each sentence augmented with 6 paraphrases covering a wider stylistic range

To assess the impact of this augmentation, we selected two encoder models for fine-tuning: *Sentiment-Analysis-BERT* and *Emotion-English-DistilRoBERTa-base*. These were chosen based on their stronger baseline performance compared to general-purpose encoders, which underperformed consistently in earlier trials and was excluded from further experiments.

Hypothesis H2: Fine-tuning with augmented train data will improve the classification performance over the original data set.

After submission, in attempt to further improve the quality of our augmented datasets, we developed a second-stage validation and correction pipeline using GPT-4o. Our goal was to ensure that each generated sentence not only aligned with its assigned label (**SUBJ** or **OBJ**), but also reflected the appropriate stylistic intent. After visual inspection of the generated samples we recognized that some synthetic data samples were misleading and could degrade the performance when used in fine-tuning the classifier. Therefore, we implemented an automated revision program that iterates over each sentence in the augmented training set. For each sample, GPT-4o was prompted with the original sentence, its intended label, and associated style (e.g., *partisan*, *emotional*). The model was instructed to do nothing if the sentence already matched the label and style. Otherwise, it rewrote the sentence to meet the intended classification and style requirements, preserving the subject matter and keeping outputs under 25 words. This method ensured consistency in label-style alignment and improved stylistic clarity without introducing content drift. The system was built in Python using LangChain's `ChatOpenAI`[23] wrapper and Polars[24] for input/output. The prompt template emphasized minimal intervention:

Correction Data Augmentation Prompt:
```
You are an expert in rewriting sentences to match specific subjectivity and style requirements.

Instructions:
- You will be given a sentence and its intended label ("SUBJ" or "OBJ") and style.
- If the sentence already matches the label and style, return it unchanged.
- If it does NOT match, rewrite the sentence so it reflects the correct label and style.
- Always preserve the subject matter.
- Only apply style if the label is SUBJ. For OBJ, remove all subjective language and opinion.
- Keep the rewritten sentence **under 25 words**.

Now perform the task:
Label: {label}
Style: {style}
Sentence: "{sentence}"

Response:
```

The model operated asynchronously with a concurrency cap to process thousands of samples efficiently, with built-in error handling and whitespace normalization. While the correction step added complexity, the pipeline remained efficient and reproducible. It was run locally using the OpenAI API through LangChain, with GPT-4o rewriting only when it detected a mismatch between a sentence's content and its assigned label or style. The process completed in a few minutes in practice and outputs were saved as .tsv files with no manual edits. The full pipeline also ran on Georgia Tech's PACE cluster, ensuring

consistent results under controlled API and compute conditions.

This process yielded two corrected datasets:

- **Corrected Balanced-2**: Based on the original Balanced-2 augmentation, where each sentence had two style-flipped paraphrases.
- **Corrected Balanced-6**: Based on Balanced-6, with six paraphrases per sentence spanning multiple style prompts.

Table 2
Examples of generated sentences deemed mislabeled by GPT-4o

Initial Generation (Label)	Corrected Version (Rationale)
The plight of Serbia's LGBTQ+ community remains largely unaddressed, leaving them in a void of neglect. (SUBJ, exaggerated)	Serbia's LGBTQ+ community is shockingly ignored, casting them into an abyss of utter neglect! (increased rhetorical intensity for the exaggerated category)
A promising turnaround is on the horizon, with expectations for change as early as next month. (SUBJ, propaganda)	A glorious transformation awaits us, with change destined to arrive as soon as next month! (intensified tone to better fit propaganda category)
He expressed that a new variant emerging this fall would not come as a shock to him. (SUBJ, propaganda)	The emergence of a new variant this fall is inevitable and will not surprise the vigilant. (rewritten to sound more declarative and assertive to better fit propaganda category)

We then fine-tuned *Sentiment-Analysis-BERT* and *Emotion-English-DistilRoBERTa-base* on these cleaned and enhanced datasets. As can be seen in Table 4, this refinement led to improved macro F1 scores and more stable class-wise performance, particularly in subjective detection.

Hypothesis H3: Self-corrected data-augmentation increases the quality of synthetic data and therefore the fine-tuned classifier performance.

All results of these described methods can be found in Section 5.

3.1. Evaluation

Our models were fine-tuned and evaluated on the macro-averaged F1-measure (macro F1):

$$\text{Macro-F1} = \frac{1}{N}\sum_{i=1}^{N} F1_i \tag{1}$$

where $F1_i$ is the the class-wise F1 score:

$$F1_i = 2 \times \frac{P_i \times R_i}{P_i + R_i} \tag{2}$$

where R_i is the recall of class i and P_i is the precision of class i.

Training was performed on a Tesla V100-16GB on the Phoenix cluster of Georgia Tech's Partnership for an Advanced Computing Environment[25] or locally on an Apple M3 Pro GPU-36GB and Metal Performance Shaders.

4. Data

We used the English dataset provided for Task 1 of CheckThat! 2025, which consisted of labeled sentences from news articles. Each sentence was annotated as either **objective (OBJ)** or **subjective**

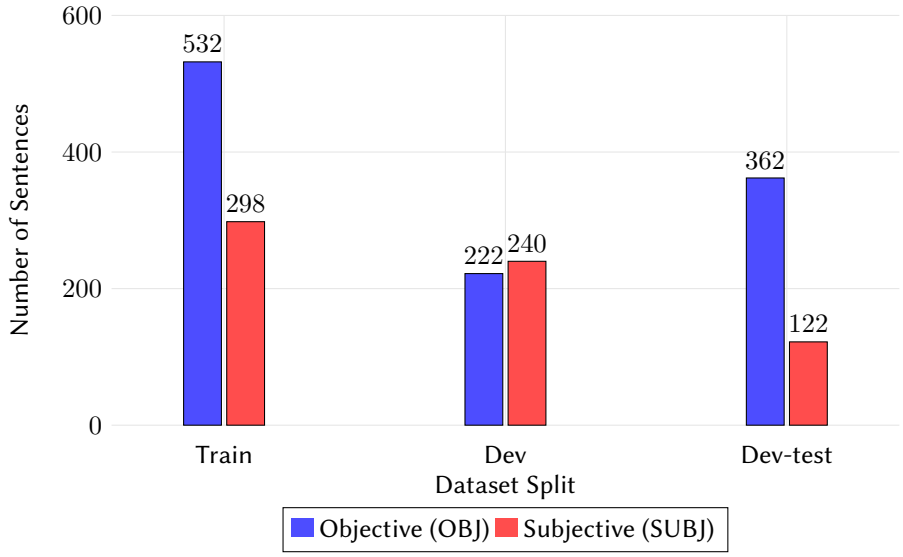

Figure 1: Class distribution across the English dataset splits

(SUBJ). The dataset was divided into training, development (dev), and test (test) sets. The distribution of class labels across these splits is shown in Figure 4.

The train dataset had substantially more objective sentences (523) instead of subjective sentences (298), but this difference was insufficient to be determined as a significant class imbalance. This did not align with the dev set, which we used as our validation set. Here, the number of objective (222) and subjective (240) examples was more balanced. Below are representative examples from each class:

- **Subjective:** "Gone are the days when they led the world in recession-busting."
- **Objective:** "The trend is expected to reverse as soon as next month."

The subjective sentence showed a personal interpretation or opinion, reflecting the speaker's sentiment toward a past event. In contrast, the objective sentence presented factual information or predictions without personal bias. This distinction illustrates that subjective statements often involve emotional or evaluative language, while objective statements rely on logical reasoning.

5. Results

Table 3 highlights our main results for the general-purpose encoders and domain-specific encoders that we used. *RoBERTa-base* achieved 0.70 macro-F1 on train, 0.75 macro-F1 on validation, and 0.65 macro-F1 on test after fine-tuning for 3 epochs with a learning rate α of 1e-4 on the original train data set. During validation, the *RoBERTa-base* model demonstrates a balanced ability to identify objective and subjective sentences. However, this dropped during test, where the subjective F1 (0.58) was significantly lower than objective F1 (0.72).

MiniLM-L6-v2 showed weaker generalization across splits, achieving 0.81 macro-F1 on train, 0.69 on validation, and 0.64 on test after fine-tuning for 3 epochs with a learning rate of α 1e-4. It had difficulties with the recognition of objective sentences during validation (only 0.58 F1), though test performance saw a drop in subjective F1 (0.51) compared to objective F1 (0.76), indicating a slight bias toward objective language cues.

ModernBERT-large struggles with subjective classification, achieving 0.41 macro-F1 on train and 0.41 on test, despite showing high objective F1 on test (0.84). After being fine-tuned for 2 epochs with α = 2e-5, it completely failed to capture subjective expressions during validation and test, suggesting overfit and limited adaptability to subjective nuances in the training data.

Table 3
Results After Fine-Tuning with Original Train Data

Model	Train	Validation F1			Test F1		
	Macro F1	Obj.	Subj.	Macro	Obj.	Subj.	Macro
ModernBERT-large	0.41	0.64	0.00	0.32	0.84	0.00	0.42
RoBERTa-base	0.79	0.77	0.72	0.75	0.72	0.58	0.65
MiniLM-L6-v2	0.81	0.58	0.75	0.69	0.76	0.51	0.64
Emotion-english-RoBERTa-large	0.97	0.73	0.77	0.77	0.76	0.59	0.67
Emotion-english-DistilRoBERTa-base*	0.90	0.78	0.75	0.77	0.80	0.57	0.68
Sentiment-Analysis-BERT	0.87	0.71	0.53	0.64	0.77	0.58	0.67

Among the transfer-learning models, *Sentiment-Analysis-BERT* showed strong performance, attaining 0.87 macro-F1 on train, 0.64 on validation, and 0.67 on test with 4 epochs and α 2e-5. Although validation subjective F1 (0.53) was notably lower than objective F1 (0.71), it generalized better on test with balanced F1 scores (0.58 and 0.77).

Emotion-English-DistilRoBERTa-base achieved the best performance among all models with 0.90 macro-F1 on train and 0.77 macro-F1 on validation. It also showed strong and balanced validation scores for both objective (0.78) and subjective (0.75) classification. After being fine-tuned for 6 epochs with α 2e-4, it sustained decent test performance (0.68 macro-F1), although subjective F1 dropped to 0.57. This model was used for submission and placed us 16^{th} out of 24.

Emotion-English-RoBERTa-large showed similar performance to the distilled model before: 0.67 macro-F1 on test after 7 epochs with α 2e-5. It had similar validation results (0.77 macro-F1), and a very high train performance (macro-F1: 0.97).

Table 4
Results After Fine-Tuning with Augmented Train Data

Model	Dataset	Train	Validation F1		
		Macro F1	Obj.	Subj.	Macro
Emotion-English-DistilRoBERTa-base	Balanced-2	0.95	0.62	0.68	0.68
	Balanced-6	0.98	0.57	0.70	0.64
	Corrected Balanced-2	0.99	0.71	0.67	0.74
	Corrected Balanced-6	0.99	0.71	0.70	0.71
Sentiment-Analysis-BERT	Balanced-2	0.91	0.64	0.65	0.67
	Balanced-6	0.99	0.54	0.68	0.62
	Corrected Balanced-2	0.99	0.68	0.36	0.75
	Corrected Balanced-6	0.99	0.67	0.73	0.74

Table 4 reports validation results from fine-tuning *Emotion-English-DistilRoBERTa-base* and *Sentiment-Analysis-BERT* on four augmented versions of the training data. Both models achieve high training macro-F1 scores (above 0.90), indicating strong fit to the training data across all configurations.

For *Emotion-English-DistilRoBERTa-base*, the highest validation macro-F1 was achieved with the Corrected Balanced-2 dataset (0.74), where objective and subjective F1 scores were 0.71 and 0.67, respectively. Corrected Balanced-6 performed slightly worse, but remained strong (0.71 macro-F1). The Balanced-2 dataset without editing resulted in a lower macro-F1 (0.68), while Balanced-6 showed the weakest performance overall (0.64), with an especially low objective F1 (0.57).

Sentiment-Analysis-BERT showed greater variability. Corrected Balanced-6 yielded the best overall macro-F1 (0.74) with balanced class performance. Corrected Balanced-2 produced a slightly higher macro-F1 (0.75), but this was driven by a strong objective F1 (0.68) and a low subjective F1 (0.36), indicating an imbalance in class predictions. The Balanced-2 dataset without editing led to more balanced class scores (0.64 and 0.65) and a moderate macro-F1 of 0.672. Balanced-6 performed the worst (0.63 macro-F1), with notably lower objective performance (0.54).

In general, editing improved consistency in subjective classification. *Emotion-English-DistilRoBERTa-*

base performed reliably across both dataset sizes when editing was applied. In contrast, *Sentiment-Analysis-BERT* was more sensitive to augmentation type and prone to overfitting, particularly favoring objective predictions when trained on the Corrected Balanced-2 dataset.

6. Discussion

Our results highlight the benefits and limitations of transfer-learning of already specified models and data augmentation for subjectivity detection in news text. Among models trained only on the original data, already fine-tuned encoders on related tasks, like *Sentiment-Analysis-BERT* and *Emotion-English-DistilRoBERTa-base*, outperformed general pre-trained models in macro-F1 on validation and test set. In addition, they achieve a better performance over the general-purpose encoders on the test set after fine-tuning. This suggests that pretraining on sentiment and emotion-related corpora improves sensitivity to subjective linguistic cues and confirms **Hypothesis H1**.

Gains from augmentation were not uniform. While adding more paraphrases (Balanced-6) increased training scores, it did not always translate to better validation performance. In some cases, especially without correction, augmentation degraded performance due to inconsistencies between the generated sentence and its intended label. Clearly, the model performance suffered due to overfit on noise, which confirms past findings that LLM-generated data can introduce noise without sufficient validation. Therefore, **Hypothesis H2** needs to be rejected as fine-tuning with augmented train data did not improve performance.

Applying a correction pipeline improved validation macro-F1 for both models. *Emotion-English-DistilRoBERTa-base* showed consistent gains across datasets, with Corrected Balanced-2 and Corrected Balanced-6 outperforming their uncorrected counterparts. *Sentiment-Analysis-BERT* exhibited more volatile behavior. Its macro-F1 improved in Corrected Balanced-6 with balanced class performance, but Corrected Balanced-2 showed a misleading macro-F1 gain driven by high objective performance, while subjective performance dropped substantially. Among all configurations, Corrected Balanced-6 yielded the most stable performance across classes for both models. This suggests that correcting label and style alignment in synthetic data improves generalization and robustness, confirming **Hypothesis H3**.

Importantly, these improvements were not just a result of adding more data, but of the generative model self-refining the augmented data to ensure semantic and stylistic alignment. The second-stage editing process using GPT-4o, which rewrote misaligned samples while preserving subject matter and label intent, played a key role in reducing label noise and improving model reliability. This distinction between simple augmentation and corrected augmentation was critical to achieving consistent gains.

Although the models fine-tuned on Corrected Balanced-2 and Corrected Balanced-6 performed best overall, we were not able to submit them to the shared task because results were finalized after the submission deadline. We submitted results with *Emotion-English-DistilRoBERTa-base* fine-tuned on the original train dataset which resulted in a 0.68 test macro-F1 and placed us on rank 16 out of 24 in the monolingual-english competition. Therefore, our submission is significantly better than the organizers baseline (test macro-F1: 0.54) but leaves room for improvement due to the top competitor achieving a 0.81 macro-F1 on the test set. Overall, these findings show that combining transfer-learning with encoders with carefully curated synthetic examples can improve performance in low-resource tasks like subjectivity detection, as long as the synthetic data is reliable and label-consistent.

7. Future Work

The research can be extended by applying our approach to multilingual settings, leveraging languages such as Arabic, Bulgarian, German, Italian, Spanish, and French included in Task 1. Our approach can further be improved by incorporating more data, e.g. the dev dataset, into the model fine-tuning for submission. More labeled data can improve the quality of the fine-tuned classifier. Also, the contrastive learning approach can be enhanced by incorporating more specialized loss-functions for the

model. Furthermore, we have observed models with varying qualities. For example, *Emotion-english-RoBERTa-large* has demonstrated a greater capability of correctly identifying objective sentences while *Emotion-english-DistilRoBERTa-base* has superior performance for subjective sentences. An ensembling approach of these models could further enhance the classification performance and is an open research avenue. Our contribution to subjectivity detection may be integrated into the broader fact checking framework via misinformation analysis, bias assessment, and claim verification workflows .

8. Conclusions

In this paper, we explored subjectivity detection in news text through the lens of transfer-learning and data augmentation. Our findings highlight three key insights: First, domain-specific encoder models already fine-tuned on sentiment or emotion datasets consistently outperformed general pre-trained encoders, supporting our hypothesis that transfer learning can enhance sensitivity to subjective language. Second, while naive data augmentation introduces inconsistencies that occasionally degraded performance, a "self-corrective" pipeline using GPT-4o significantly improved the quality and label alignment of synthetic examples. This allowed models to generalize better and increased macro-F1 scores, particularly for the harder-to-detect subjective class. Third, although our final submission was constrained to uncorrected data due to time limits, our post-submission results demonstrate the value of integrating high-quality synthetic training data. Overall, our approach underscores the importance of not only augmenting data, but shows the ability of a generative model to check itself for semantic and stylistic fidelity. Combining specialized encoders with refined augmentation holds promise for improving low-resource NLP tasks like subjectivity detection.

Acknowledgements

We thank the DS@GT CLEF team for providing valuable comments and suggestions. This research was supported in part through research cyberinfrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI-GPT-4o in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Elsner, G. Atkinson, S. Zahidi, Global Risks Report 2025, Technical Report, World Economic Forum, Geneva, 2025.

[2] F. Alam, J. M. Struß, T. Chakraborty, S. Dietze, S. Hafid, K. Korre, A. Muti, P. Nakov, F. Ruggeri, S. Schellhammer, V. Setty, M. Sundriyal, K. Todorov, V. V., The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 467–478.

[3] F. Ruggeri, A. Muti, K. Korre, J. M. Struß, M. Siegel, M. Wiegand, F. Alam, R. Biswas, W. Zaghouani, M. Nawrocka, B. Ivasiuk, G. Razvan, A. Mihail, Overview of the CLEF-2025 CheckThat! lab task 1 on subjectivity in news article, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 2025.

[4] A. Galassi, F. Ruggeri, A. Barrón-Cedeño, F. Alam, T. Caselli, M. Kutlu, J. M. Struß, F. Antici, M. Hasanain, J. Köhler, K. Korre, F. Leistra, A. Muti, M. Siegel, M. D. Türkmen, M. Wiegand, W. Zaghouani, Notebook for the CheckThat! Lab at CLEF 2023, in: Proceedings of the Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (CLEF), 2023.
[5] J. M. Struß, F. Ruggeri, D. Dimitrov, A. Galassi, G. Pachov, I. Koychev, P. Nakov, M. Siegel, M. Wiegand, M. Hasanain, R. Suwaileh, W. Zaghouani, Notebook for the CheckThat! Lab at CLEF 2024, in: Proceedings of the Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum (CLEF), 2024.
[6] J. Wiebe, E. Riloff, Creating subjective and objective sentence classifiers from unannotated texts, in: A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg, 2005, pp. 486–497.
[7] L. L. Vieira, C. L. M. Jeronimo, C. E. C. Campelo, L. B. Marinho, Analysis of the subjectivity level in fake news fragments, in: Proceedings of the Brazilian Symposium on Multimedia and the Web, WebMedia '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 233–240.
[8] R. Mihalcea, C. Banea, J. Wiebe, Learning multilingual subjective language via cross-lingual projections, in: A. Zaenen, A. van den Bosch (Eds.), Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 976–983.
[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[10] J. Devlin, M.-W. 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, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.
[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners (2019).
[12] 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).
[13] I. B. Schlicht, L. Khellaf, D. Altiok, DWReCO at CheckThat! 2023: Enhancing Subjectivity Detection through Style-based Data Sampling, in: Proceedings of the Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (CLEF), 2023.
[14] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, Simclr: A simple framework for contrastive learning of visual representations, in: Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR, 2020.
[15] 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.
[16] M. Casanova, J. Chanson, B. Icard, G. Faye, G. Gadek, G. Gravier, P. Égré, Notebook for the HYBRINFOX Team at CheckThat! 2024 - Task 2, in: Proceedings of the Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum (CLEF), 2023.
[17] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, I. Poli, Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024.
[18] M. Ange, Semtiment-analysis-bert, https://huggingface.co/MarieAngeA13/Sentiment-Analysis-BERT, 2023.
[19] J. Hartmann, Emotion-english-distilroberta-base, https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/, 2022.
[20] OpenAI, Gpt-4 technical report, 2024.
[21] J. Chen, A. Sriram, E. Choi, G. Durrett, Generating literal and implied subquestions to fact-check complex claims, 2022.

[22] A. Awadallah, A. Trabelsi, Clac at checkthat! 2024: Sentence-level subjectivity classification using zero-shot gpt-3 and majority voting, in: CLEF 2024 Working Notes, 2024.
[23] L. AI, L. Contributors, Langchain: A python package for natural language processing, https://github.com/langchain-ai/langchain, 2025.
[24] P. BV, P. Contributors, Polars: A fast dataframe library implemented in rust, https://github.com/pola-rs/polars, 2025.
[25] PACE, Partnership for an Advanced Computing Environment (PACE), 2017. URL: http://www.pace.gatech.edu.