CheckMates At CheckThat! 2025: Transformer-Based Models For Subjectivity Classification

Notebook for the CheckThat! Lab at CLEF 2025

Karthik V[1], R Padmashri[1], V Srikumar[1] and Durairaj Thenmozhi[1]

[1]*Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam, Tamil Nadu, India, 603110*

Abstract

The recent spike in popularity of social networks has led to the need for reliable data to make informed decisions, as well as to ensure that automated systems and social research are based on objective and unbiased data. This necessitates the need to filter out unverified subjective data. The aim of this research is to train an AI model to classify a given sentence as subjective or objective. We explored various models such as logistic regression, Support Vector Machine (SVM), BERT, Sentence-BERT, and DistilBERT. Our evaluation showed that DistilBERT outperformed the other models, making it the best choice for the given task. In the CheckThat! 2025 Lab, under monolingual English setting we ranked 15th with an F1 score of 0.7009.

Keywords

Subjectivity, Classification, BERT, Transformers, SBERT, distilBERT, Deep Learning

1. Introduction

Classifying content in news articles as subjective or objective is critical in various natural language processing tasks, especially those involving information credibility, sentiment evaluation, and content classification. Objective sentences give information that is verifiable, measurable, and independent of personal opinions, often reflecting facts or universally accepted truths. On the other hand, subjective sentences are influenced by personal views, opinions, and emotions and their interpretation may change depending on individual or contextual factors [1, 2]. The goal of the subjectivity classification task in the CLEF CheckThat! Lab Task 1 [3, 4, 5] is to classify a sentence as either subjective or objective, allowing systems to differentiate between sentences containing opinions and those stating facts. This task focuses on sentence-level classification and uses data annotated with language-agnostic prescriptive guidelines to achieve consistency in linguistic contexts. Existing approaches vary from syntactic techniques such as keyword spotting [6] to strong semantic approaches like statistical models, annotation-guided heuristics and transformer-based deep learning models [1, 2]. Among them, fine-tuned transformer models like BERT and RoBERTa have shown improved performance with high F1 scores on subjectivity classification tasks.

In our submission for CLEF 2025 [3, 4, 5], we take different transformer models like **BERT**[7], **Sentence-BERT**[8] and **DistilBERT**[9]—a distilled, lighter and faster form of BERT—as our core model. These models use self-attention mechanisms to understand the context deeply and process text bidirectionally unlike traditional rule-based models. While BERT and SBERT provided moderate results, we decided to go with DistilBERT which maintains over 97% of BERT's performance while being 40% smaller and 60% faster [10]. This makes it a suitable choice for real-world deployment. Our aim is to contribute to the ongoing research to develop scalable, accurate, and interpretable models for subjectivity classification, specifically in English news articles. We first cleaned the training data and split 20% of it into a validation set. The sentences were then tokenized using the DistilBERT tokenizer and we fine-tuned DistilBERT for binary subjectivity classification. Model performance was evaluated by comparing predictions on the validation set against the true labels, using macro-averaged F1 score.

The remainder of the paper is structured as follows: Section 2 reviews related work on subjectivity classification; Section 3 describes the datasets used; Section 4 explains the methodology for the task; Section 5 presents and analyzes the results; and Section 6 concludes the article.

2. Related Works

Balahur Alexandra et al. [11] describe the importance of subjectivity detection as a preprocessing step for sentiment analysis. Since objective sentences just state facts without emotion, they're usually left out of sentiment analysis. Subjectivity detection is also essential to ensure that unbiased, factual, and reliable data is available for research works and model training.

Subjective sentences need not always be purely opinion-based and can sometimes incorporate facts in them. [12]. While traditional algorithms find identification of such sentences challenging, pretrained models like BERT, which have contextual understanding, can identify them with proper training. C. Zhu, and Y. Yu [13] trained various machine learning algorithms on a dataset of 1000 articles for the classification task, and concluded that KNN performed best with an accuracy of 84%. This suggests that the sentences belonging to the same class share structural similarities and are clustered closely in the KNN space. F. Antici et al.[14] observed that BERT-based models outperformed other models like SVM and LR in the subjectivity classification task.

Georgi Pachov et al. [15] who participated in CLEF 2023 used a multi-model approach for Subjectivity Detection and achieved a Macro F1 score of 0.77, securing second place in the English classification task. They trained a sentence-embedding encoder model (Sentence BERT), a sample-efficient few-shot learning model (SetFit) and a multilingual transformer (XLM-RoBERTa), and the three approaches were then combined into a simple majority voting ensemble to classify sentences for the given task. Similarly, Samir Rustamov [16] also used a hybrid system that combines HMMs (Hidden Markov Model), FCS (Fuzzy Control System), and ANFIS (Adaptive Neuro-Fuzzy Inference System) and obtained . All three models evaluate subjectivity scores individually, and the outcomes of the three are then directed into a decision-making block, which returns the result. This approach achieved 92% accuracy when trained on a dataset consisting of 5000 subjective and 5000 objective sentences. While multi-model systems provide good results, its complexity increases risk of overfitting especially when the dataset is small.

Samuel Akpatsa et al. [17] approached a sentiment classification task using various methods. Their results show the superiority of BERT-based models over traditional machine learning and deep learning models for the given classification task due to its bi-directional context awareness in text sequences. H. Huo and M. Iwaihara [18] noted that the BERT-base model produced good results for the subjectivity detection task with minimal fine-tuning. They further improved on their model by experimenting with various fine-tuning strategies, demonstrating the potential offered by pre-trained transformers in NLP tasks. Berfu Büyüköz et al. [19] compared the performance of bi-LSTM network(ELMo) and distilBERT. Their research suggested that distilBERT generalized better when compared to ELMo in a cross-context environment. This is due to distilBERT's deep contextual awareness. distilBERT is also 30% smaller and 83% faster than ELMo, making it easier to train and fine-tune.

3. Dataset Description

The model is trained on the dataset which is divided into three parts: training, validation, and test. The training and validation datasets each contain a id, the news article sentence, its label, and the corresponding solved conflict. The validation dataset is derived by taking 20% from the training dataset in random order. The test dataset consists of sentence id and a sentence. Refer to Table 1, Table 2 for detailed dataset statistics. For the competition, in the training data set, it is observed that 532 sentences are **Objective** and 298 sentences are **Subjective**. Based on this, we note that the training phase will be biased towards Objective sentences more. To further enhance the training set post competition results, we included the dev_en.tsv provided - - in our training data. To overcome the Objective bias, we also upsampled the dataset to hold 532 records each of Subjective and Objective sentences.

Table 1
Dataset statistics for the competition

Split	English	SUBJ	OBJ
Training Set	830	298	532
Validation Set	166	62	104
Test Set	300	85	215
Total	1296	445	851

English, SUBJ and OBJ columns indicate the number of sentences.

Table 2
Dataset statistics post competition

Split	English	SUBJ	OBJ
Training Set	1064	532	532
Validation Set	212	104	108
Test Set	300	85	215
Total	1364	617	747

English, SUBJ and OBJ columns indicate the number of sentences.

4. Methodology

Before evaluating model performance, it is essential to detail the steps taken to prepare the data and design the experimental pipeline. This section outlines the full methodology, starting with preprocessing of the input text, followed by an overview of the transformer models used—BERT, SBERT, and DistilBERT—and concluding with the performance metrics applied to compare these models. Each model was tested using consistent training parameters to ensure a fair and interpretable comparison.

4.1. Data Preprocessing

The preprocessing step consists of several phases to clean and standardize the text. Sentences were converted to lowercase and spaces were trimmed from the start and end to normalize the text. Data cleaning normalized the quotes, eliminated text in square brackets, removed special characters (except few punctuations), and condensed multiple spaces to one. Once the text was normalized, the corpus was vectorized within a TF-IDF vectorizer of 4-ngrams, with a cap of 10,000 features, and minimum document frequency of 2. The use of vectorization allowed for text to matrix with numeric representation conversion that could be used for machine learning classifiers input.

4.2. Transformer Models Tested

4.2.1. Bidirectional Encoder Representations from Transformers (BERT)

BERT [7] is an encoder-only transformer model which is pre-trained on MLM(Masked Language Model) and NSP(Next Sentence Prediction) with a large corpus of unlabeled data. BERT uses self supervised learning methods to learn how to represent text as a sequence of vectors. BERT(base) has a total of 12 transformer layers with a parameter size of 110M and hidden size 768. The key difference between BERT and other models is its unique self-attention mechanism. BERT has a unique way of focusing on the context of the sentence by assigning attention scores to the words. BERT(base) has 12 self attention heads per layer to capture various relationships between the words and understand the context about each word.

To tune a BERT model for the given task we first tokenize the data. To do this a suitable tokenizer is instantiated with parameters `padding="max_length"`, `truncation=True`, `max_length=128` (to

ensure length of sequences are as expected by BERT). This tokenizer is applied to the training data which generates corresponding `input_ids`, `attention_masks` and `labels`.

We loaded the pre-trained `BertForSequenceClassification` model of Hugging Face's transformer module for this task. The parameter `num_labels=2` is set to ensure binary classification of sentences. Training is done with a batch size of 8, `num_train_epochs = 3`, and a learning rate of 2×10^{-5}. This ensures that the model is simple and fast, while reducing the risk of overfitting and bias.

4.2.2. Sentence-BERT (SBERT)

SBERT [8] is a framework that finetunes pretrained transformer models (like MiniLM/BERT) to compute sentence embeddings efficiently. SBERT is a partially self-supervised transformer model, pre-trained using a Siamese/triplet architecture. SBERT uses cosine similarity to analyse semantic equivalence of fixed length vectors, drastically decreasing its computational time and complexity. The input data is Tokenized using a WordPiece tokenization strategy. The input is split into subwords and `[CLS]` and `[SEP]` tokens are added for each sentence. For example:
```
input = "Tokenization with SBERT"
tokenized = [CLS],"Tokenization","with","SBERT", [SEP]
```
The tokenizer also generates Attention masks for the tokens to differentiate padding tokens from the rest. SBERT adapts the architecture of pretrained transformer models like MiniLM and replaces the output layer to produce sentence embeddings. SBERT has 6 transformer layers and 384-dimensional hidden states with 12 self attention heads per layer. The tokenized data is passed through the 6 transformer layers, each layer adding more contextual information to the embedding. The final token level embedding which contains rich contextual data is then passed to the pooling layer, where a suitable pooling strategy is applied to convert the token embeddings into fixed length sentence embeddings.

We finetuned our SBERT model by adjusting neuron dropout rate `0.1` (to improve generalization). We trained the model with, `num_train_epochs = 3`, and a learning rate of 2×10^{-5} (to reduce risk of overfitting). This keeps the training process simple and quick while maintaining a good accuracy score.

4.2.3. Distilled BERT (distilBERT)

DistilBERT [9] is a trim and efficient version of BERT achieved through the process of knowledge distillation. When creating DistilBERT, the original model's 12 transformer layers are reduced down to 6 transformer layers but are left with basically the same hidden dimension (768) and number of attention heads (12). The end result is a model 40% less in size and 60% faster compared to original BERT but with roughly 97% of the performance of original BERT. The distillation process helps to really balance performance and efficiency by removing the size constraints from a much larger model.

In the distillation process, the compressed distilBERT - student model learns the predictions made by the larger and more robust BERT - teacher model. In this distillation step, the student model will try to output an approximate version of the teacher model's output distributions, to help the distilBERT model learn important patterns while filtering out the less important information so that like BERT, the distilBERT can perform similar tasks but with fewer parameters and thus more computationally efficient inference and faster times. During fine-tuning the distilBERT model is trained on certain datasets to fit to the tasks associated with them, and because it has a smaller architecture it is able to fine-tune pretty quickly and uses less resources than BERT.

In our testing, the data is tokenized using the DistilBERT tokenizer, then formatted into PyTorch datasets for training and evaluation. The model is trained for 3 epochs with a learning rate of 2×10^{-5} and a batch size of 8, and its performance is evaluated using the macro F1-score. Post competition results, followed the same procedure for dataset training with the upsampled dataset, followed by assigning weights of **0.8** and **1.2** respectively to solved_conflict **FALSE** and **TRUE**. The model is trained for 4 epochs with a learning rate of 1×10^{-5} and a batch size of 16, and its performance was once again evaluated using the macro F1-score.

4.3. Performance Metrics

The Macro Averaged F1 Score is a measure of binary classification model performance across multiple classes that treats all classes equally. Precision is the number of true positive predictions divided by the total number of positive predictions, while Recall is the number of true positives divided by the total number of actual positives. The F1 score serves to combine both Precision and Recall into a single score. The macro F1 Score is calculated by obtaining the F1 score for each class (objective and subjective) first, noting that the F1 score is the harmonic mean of Precision and Recall.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

For binary classification task, the Macro F1 Score is calculated by averaging the F1 scores of both classes as it is calculated using the arithmetic mean of the F1 score of the "objective" class and the F1 score of the "subjective" class. The Macro F1 score treats both classes equally and is therefore not influenced by class imbalance and represents how well the model performs in a way that it is not biased towards the more frequent class.

$$\text{Macro F1} = \frac{1}{2}\left(\text{F1}_{\text{objective}} + \text{F1}_{\text{subjective}}\right)$$

5. Result Analysis

We tested five models: Logistic Regression, SVM, BERT, SBERT and DistilBERT and each model was evaluated using macro-averaged F1 scores on the validation and test sets. Table 3 reports the results.

Table 3
Macro F1 Scores for Different Models on Validation and Test Sets

Model	Validation Set F1 (competition)	Test Set F1 (competition)
Logistic Regression	0.6389	0.5573
SVM	0.6307	0.5342
BERT	0.8727	0.6725
SBERT	0.8175	0.6991
DistilBERT	0.8593	0.7009

Logistic Regression and SVM provide lower F1 scores on both validation and test sets, due to their limited ability to generalize. These models rely on TF-IDF features and fail to understand nuanced language patterns needed for subjectivity detection. In contrast, the transformer-based models performed significantly better, with validation scores above 0.81 and test scores above 0.67. They show much stronger generalization as they are built to understand words in context, which helps them recognize subtle patterns in subjective text. Among the five models evaluated, DistilBERT achieved the highest macro-averaged F1 score on the test set (0.7009). The drop in test score compared to validation score is due to overfitting caused in the model - it trained on patterns over generalization due to the small size of the training dataset. Upon upsampling and fine-tuning, we were able to achieve a test set F1 score of **0.7174** and validation set F1 score of **0.7662**. The improved results of DistilBERT over other models can be attributed to its distillation process and optimized token-level understanding. While BERT performed well on the validation set due to its powerful contextual understanding, its large capacity also makes it prone to overfitting. However, DistilBERT with fewer parameters and faster inference than BERT, achieves greater efficiency.
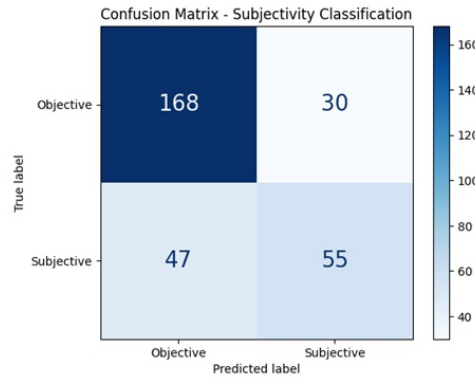
Figure 1: Confusion Matrix for Subjectivity Classification with DistilBERT

We further analysed the performance of DistilBERT model with the help of confusion matrix and ROC curve along with the performance metric of macro-averaged F1 score. From the confusion matrix depicted in Figure 1 we see that the model performs better at identifying **Objective** sentences (168 correctly predicted - 84%) than **Subjective** ones (55 correctly predicted - 59%). However, it misclassifies:

- 30 Objective samples as Subjective (false positives),
- 47 Subjective samples as Objective (false negatives).

This indicates a bias toward predicting Objective, likely due to class imbalance or weaker feature signals for subjectivity. Improving recall on the Subjective class may involve balancing data or enhancing feature representation.
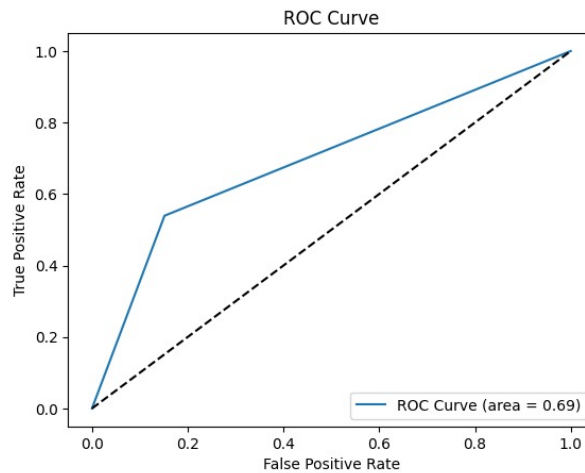
Figure 2: ROC Curve showing DistilBERT performance

In Figure 2 AUC = 0.69 shows that the model has a 69% chance of ranking a randomly chosen positive example higher than a negative one. The curve shape follows a steep early rise indicating decent sensitivity at low FPRs, but it flattens quickly — suggesting limited discriminative power at higher thresholds. DistilBERT is able to capture some subjectivity signals from the data, hence its performance is better than random guessing. However, the boundary between subjective and objective samples may be fuzzy, or the model might require better fine-tuning, more training data, or stronger feature representations (e.g., context window, additional metadata).

6. Conclusion

This study evaluates multiple machine learning and pretrained transformer-based algorithms for the given subjectivity classification task. We observed that BERT, Sentence-BERT and Distilled-BERT show exceptionally good performance when compared with Logistic Regression and Support Vector Machine (SVM) due to its ability to understand the context of the data. From the tested models, we picked DistilBERT as the optimal choice since it demonstrates a reasonable tradeoff between computational efficiency and model performance with minimal tuning. Our model predicted the given test data with a macro F1 score of 0.7009 in the competition results and 0.7174 post competition upon further enhancement. With better accounting for the domain-specific fine-tuning, the accuracy of the model can be increased further.

Declaration on Generative AI

During the preparation of this work, the author(s) used QuillBot in order to: Grammar and spelling check, Plagiarism detection. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] 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, Overview of the clef-2023 checkthat! lab: Task 2 on subjectivity in news articles, in: CLEF-WN, volume 3497 of *CEUR-WS*, CEUR-WS, Thessaloniki, Greece, 2023. URL: http://ceur-ws.org/Vol-3497/paper-35.pdf, notebook for the CheckThat! Lab at CLEF 2023.

[2] J. M. Struß, F. Ruggeri, A. Barrón-Cedeno, F. Alam, D. Dimitrov, A. Galassi, G. Pachov, I. Koychev, P. Nakov, M. Siegel, M. Wiegand, M. Hasanain, R. Suwaileh, W. Zaghouani, Overview of the clef-2024 checkthat! lab task 2 on subjectivity in news articles, in: CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-XXXX/, notebook for the CheckThat! Lab at CLEF 2024.

[3] 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 Int. Publ., Cham, 2025, pp. 467–478.

[4] 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., Overview of the clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Jose, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), CEUR Workshop Proceedings, 2025. To appear.

[5] F. Ruggeri, A. Muti, K. Korre, J. M. Struß, M. Siegel, M. Wiegand, F. Alam, R. Biswas, W. Zaghouani, M. Nawrocka, B. Ivasiuk, R. Gaina, A. Mihail, Overview of the clef-2025 checkthat! lab task 1: Subjectivity in news articles, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025. To appear.

[6] J. Wiebe, E. Riloff, Creating subjective and objective sentence classifiers from unannotated texts, in: P. Sojka, I. Kopeček, K. Pala (Eds.), Text, Speech and Dialogue. Proceedings of the 8th International Conference, TSD 2005, Karlovy Vary, Czech Republic, September 12–15, 2005, volume 3658 of *Lecture Notes in Computer Science*, Springer, 2005, pp. 486–497. URL: https://link.springer.com/chapter/10.1007/978-3-540-30586-6_51. doi:10.1007/978-3-540-30586-6_51.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: 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. URL: https://aclanthology.org/N19-1423.

[8] 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410.

[9] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108, 2019. URL: https://arxiv.org/abs/1910.01108.

[10] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint, 2019. URL: https://arxiv.org/abs/1910.01108, arXiv:1910.01108.

[11] A. Balahur, R. Mihalcea, A. Montoyo, Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications, Comput. Speech Lang. 28 (2014) 1–6. URL: https://www.sciencedirect.com/science/article/pii/S0885230813000697.

[12] B. Saberi, S. Saad, Sentiment analysis or opinion mining: A review, Int. J. Adv. Sci. Eng. Inf. Technol. 7 (2017) 1660–1666. URL: https://www.researchgate.net/profile/Saidah-Saad/publication/320748824_Sentiment_Analysis_or_Opinion_Mining_A_Review/links/5a8f629ca6fdccecfffffdf0a/Sentiment-Analysis-or-Opinion-Mining-A-Review.pdf.

[13] C. Zhu, Y. Yu, Subjective or objective: A study of classifying news content based on machine learning algorithms: – an example of sports articles, J. Innov. Dev. 6 (2024) 53–59. URL: https://doi.org/10.54097/ad878g55. doi:10.54097/ad878g55.

[14] F. Antici, et al., A corpus for sentence-level subjectivity detection on english news articles, arXiv preprint, 2023. URL: https://arxiv.org/abs/2305.18034, arXiv:2305.18034.

[15] G. Pachov, et al., Gpachov at checkthat! 2023: a diverse multi-approach ensemble for subjectivity detection in news articles, arXiv preprint, 2023. URL: https://arxiv.org/abs/2309.06844, arXiv:2309.06844.

[16] S. Rustamov, A hybrid system for subjectivity analysis, Adv. Fuzzy Syst. 2018 (2018) 2371621. URL: https://onlinelibrary.wiley.com/doi/full/10.1155/2018/2371621.

[17] S. Akpatsa, et al., Online news sentiment classification using distilbert, J. Quantum Comput. 4 (2022) 1. URL: https://www.researchgate.net/profile/Prince-Addo/publication/362675821_Online_News_Sentiment_Classification_Using_DistilBERT/links/62f7c74cc6f6732999c99fff/Online-News-Sentiment-Classification-Using-DistilBERT.pdf.

[18] H. Huo, M. Iwaihara, Utilizing bert pretrained models with various fine-tune methods for subjectivity detection, in: Web and Big Data: 4th International Joint Conference, APWeb-WAIM 2020, Tianjin, China, September 18-20, 2020, Proceedings, Part II, Springer Int. Publ., 2020. URL: https://db-event.jpn.org/deim2020/post/proceedings/papers/G1-1.pdf.

[19] B. Büyüköz, A. Hürriyetoğlu, A. Özgür, Analyzing elmo and distilbert on socio-political news classification, in: Proc. Workshop AESPN, 2020. URL: https://aclanthology.org/2020.aespen-1.4/.