JK_PCIC_UNAM at CheckThat! 2024: Analysis of Subjectivity in News Sentences Using Transformers-Based Models

Notebook for the CheckThat! Lab at CLEF 2024

Karla Salas-Jimenez[1,†], Iván Díaz[1,†], Helena Gómez-Adorno[2], Gemma Bel-Enguix[3,4] and Gerardo Sierra[3]

[1]*Posgrado en Ciencias e Ingeniería de la Computación, Universidad Nacional Autónoma de México, Ciudad de México 04510, México.*

[2]*Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Ciudad de México 04510, México.*

[3]*Instituto de Ingeniería, Universidad Nacional Autónoma de México, Ciudad de México 04510, México.*

[4]*Departament de Filologia Catalana i Lingüística General, Universitat de Barcelona, Barcelona, España.*

Abstract

Recognizing subjectivity in online content is essential for understanding public opinion, detecting bias, and managing misinformation. This year's CheckThat! 2024 Task 2 emphasized the identification of subjective and objective news sentences. Transformer models, particularly BERT, have demonstrated high efficacy for this task. In our study, we trained and evaluated our methodologies on the English and Italian sub-tasks of the challenge. A thorough data analysis was conducted, emphasizing the importance of extracting relevant features for accurate classification. Although traditional machine learning algorithms were utilized for this task, the BERT models significantly outperformed them, demonstrating superior performance. Specifically, our BERT-based classifiers achieved a macro F1 score of 0.82 on the English development dataset and 0.81 on the Italian development dataset. These results underscore the effectiveness of transformer models in distinguishing subjective content.

Keywords

Subjectivity, News sentences, Transformer Models, BERT,

1. Introduction

A subjective sentence expresses the position, attitude, or feelings of its author [1]. The detection of subjectivity is a challenging task for computers due to the intricate nature of human language. Subjective statements rely on personal opinions and emotions, which are difficult to quantify and interpret accurately. Context and cultural references further complicate the task, as words can have different meanings in different situations. This complexity requires the application of advanced natural language processing techniques, which still struggle to reliably distinguish between subjective and objective content.

Detecting subjectivity in news articles is essential for numerous applications, including sentiment analysis, opinion mining, fact-checking, understanding public opinion, identifying bias, and combating misinformation. In the realm of journalism, where articles are widely disseminated and opinions are often intertwined with facts, differentiating between subjective and objective tones is a critical task.

The 2024 edition of CheckThat! shared task [2] included 6 subtasks. Subtask 2 focused on evaluating whether a sentence within a news article is presented with an objective or subjective tone. This task seeks to address the challenge of discerning whether online news articles are composed of subjective opinions

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

[†]These authors contributed equally.

karla_dsj@ciencias.unam.mx (K. Salas-Jimenez); diazrysivan@gmail.com (I. Díaz); helena.gomez@iimas.unam.mx (H. Gómez-Adorno); gbele@iingen.unam.mx (G. Bel-Enguix); gsierram@iingen.unam.mx (G. Sierra)

https://github.com/KarlaDSJ (K. Salas-Jimenez); https://github.com/JuanIvanDiazReyes (I. Díaz); https://helenagomez-adorno.github.io (H. Gómez-Adorno)

or objective statements. This diversity highlights the importance of language-agnostic approaches to subjectivity detection, enabling broader applicability across different linguistic contexts. The official evaluation metric for the shared task is the macro-averaged F1 score between the two classes (subjective and objective).

Datasets were offered in five languages: Arabic, Bulgarian, English, German, and Italian, in addition to a Multilingual, mixing all the above languages. We participated in the English and Italian subtasks.

The sentences from each dataset are from news articles dealing with controversial topics such as political issues, COVID-19, civil rights, and economics. In addition to annotating the data, the organizers developed a set of guidelines [1] that can be applied to any language to generate corpora in other languages, trying to assist in any disagreements that may arise between annotators.

Among the guidelines [1], there are the following cases:

- Sentence is subjective if it contains:
  - Speculations that draw conclusions that are considered opinions
  - Sarcastic or ironic expressions
  - Exhortations or personal auspices
  - Discriminating or downgrading expressions
  - Rhetorical figures explicitly made by its author to convey their opinion
  - A conclusion made by its author that is drawn despite insufficient factual information.
  - Intensifiers that can be attributed to its author to express their opinion
- Sentence is objective when it:
  - Describes the personal feelings, emotions, or moods of its author without conveying opinions on other matters
  - Expresses an opinion, claim, emotion, or a point of view that is explicitly attributable to a third-party
  - Presence of quotation marks, when used to quote a third person

We consider these guidelines to select the characteristics that can help us to determine whether a sentence is subjective or not, as mentioned in section 3.1.

We first performed handcrafted feature extraction with Machine Learning Models to train our subjectivity detection models. Additionally, we fine-tuned BERT-based models for comparison purposes. During the development phase, our best-performing models were those based on BERT.

The remainder of the paper is organized as follows. Section 2 discusses related work, introducing transformer-based models and recent applications in subjectivity classification. Section 3 describes the methodology, detailing the dataset provided and used throughout the competition, as well as the models employed. The results of the experiments performed and an analysis are presented in Section 4. The paper concludes with a discussion of the findings and potential future work.

2. Related Work

In the last few years, special attention has been paid to detecting subjectivity in texts. Recently, transformer models have been applied to text classification tasks involving subjectivity detection. For example Timo Spinde's work [3], and the Python package DBias [4]. These works use the Bias Annotations By Experts (BABE) corpus, which consists of 3,700 sentences on news with controversial topics extracted from 14 US news platforms from January 2017 to June 2020. In both cases, the authors attacked the problem of classifying whether a sentence is subjective or not using attention-based models such as RoBERTa (F1 0.804) and DistilBERT (F1 0.75), which obtained the best results.

In previous years, the work of DWReCO at CheckThat! 2023 [5] utilized BERT-based models fine-tuned on the competition dataset, augmented with the help of ChatGPT. To encode texts and train subjectivity classifiers, language-specific transformers were employed: 'Roberta-base' for English,

'German BERT' for German, and 'BERTurk' for Turkish, as these models have demonstrated strong performance on the tasks in their respective languages.

These approaches highlight the flexibility and effectiveness of transformer models in handling various NLP tasks. By adapting pre-trained models through fine-tuning on task-specific data, they achieve state-of-the-art results in subjectivity classification.

3. Methodology

3.1. Analysis of the dataset

Since we only participated in the English and Italian language subtasks , we only perform the analysis for these languages.

The first thing observed is that the training dataset provided is unbalanced. The number of subjective sentences is very low compared to the objective sentences. This can be seen in Table 1.

Table 1
Data Split and Distribution

	English			Italian		
	Total	OBJ	SUBJ	Total	OBJ	SUBJ
Train	830	532	298	1613	1231	232
Dev	219	106	113	227	167	60
dev-test	243	116	127	440	323	117

We decided not to make any augmentation to the dataset to balance it out due to the fact that the selected models managed to capture the differences despite the difference in the amount of data in each class, as we will see later in the section on the analysis of the model results.

When analyzing the data, we noted that the guide mentioned in the previous section helped in the task. For example, we observe in Table 2 that the presence of quotation marks is certainly higher in objective sentences.

Table 2
Analysis of the Training Dataset Features Considered for Classic Supervised Classifiers.

Feature	**English**		**Italian**	
	OBJ	SUBJ	OBJ	SUBJ
# of words	11,745	7,358	24,138	7,746
# of different lemmas	2,508	1,826	4,722	2,372
% of nouns	47.67	45.31	50.98	45.98
% of adjectives	18.28	20.10	15.38	16.35
% of verbs	24.69	23.13	15.38	16.35
% of adverbs	9.34	11.34	10.47	14.64
# of quotation marks	110	28	211	59

We inspected the vocabulary of each class, where we can observe that although they share many words, they also disagree in many others. We generated word clouds to visualize this in a better way, and we observed that the objective classes tend to discuss statistics, studies, and reports and frequently mention terms like "infected," "schools," and "teachers." In contrast, the subjective classes use words such as "thought," "consider," "indigent," and "perfectly," among others. This indicates a clear difference in the vocabulary used in each of the two classes. It is also important to note that not only does the vocabulary differ, but the number of words differs. The objective class has more words.

The difference between the features of each class is not as much as we expected, as can be observed for English and Italian in Table 2, except for the feature 'quotation marks.' We expected that in the

subjective sentences, the number of adjectives and adverbs would be higher.

since these usually provide more details on characteristics or attributes about nouns, this could, in certain contexts, introduce a kind of subjectivity into the sentence.

Finally, we count the number of quotation mark pairs and see that there is a difference. As expected, objective sentences have a higher number of quotation marks since they indicate the opinion of a third person, which the annotators considered an objective feature.

3.2. Machine Learning Models

The analysis conducted in Section 3.1 indicates that the main features are:

number of quotation marks, the probability of the sentence being positive and negative, making use of the python pysentimiento package [6], the number of nouns, adjectives, verbs and adverbs, divided by the word length of the sentence, plus a multilingual BERT Sentence [7] to generate the sentence embedding and try to capture the semantics of each sentence, also add a bag of words vector as we see that words not sharing is an important feature.

We tested these characteristics with Logistic Regression (LR), Support Vector Classification (SVC), Support Vector Regression (SVR), RandomForest (RF) and Naive Bayes (NB). These methods have been shown in the literature to work well for learning text features.

3.3. Transformer Training

We employed BERT (Bidirectional Encoder Representations from Transformers) as our primary classifier. BERT models are pre-trained on a vast corpus of text and are specifically tailored for sequence classification tasks, making them ideal for our needs. We utilize language-specific transformers [8]: BERT-base-uncased [9] for English and BERT-base-italian-cased-sentiment [10] for Italian. Both models were fine-tuned on the provided dataset to adapt them for the subjectivity detection task. Our primary focus is on tuning the parameters of the supervised classifier. We train the models for 4 epochs, a batch size of 16, and limit the input size to a maximum of 256 tokens. We trained and ran our system on Google Colab. The experiments used GPUs to leverage faster computation and efficiently handle the large-scale computations involved in fine-tuning BERT models. The dataset used to tune the hyperparameters was the training dataset provided by the organizers.

4. Results and Analysis

In order to obtain the results of the classical methods, we apply the 5-fold cross-validation. The results can be seen in tables 4 and 3, which provide the scores in English and Italian respectively, obtained for each of the machine learning models on the dev-test set. These tables show the effectiveness of the machine learning models in capturing relevant text features.

Table 3
Results of the Classical Machine Learning Models on the English Development Dataset.

Model	Accuracy	Precision	Recall	Macro F1
LR	0.700	0.760	0.622	**0.699**
SVC	0.601	0.841	0.291	0.562
SVR	0.682	0.791	0.535	0.659
RF	0.663	0.747	0.536	0.659
NB	0.572	0.604	0.528	0.572

For the transformer case, results are shown in tables 5 and 6, to further evaluate the performance of our models, we examined the results under different settings for both English and Italian dev-test datasets provided by the organizers. These analyses compare the performance metrics of various configurations of batch size and max length parameters.

Table 4
Results of the Classical Machine Learning Models on the Italian Development Dataset.

Model	Accuracy	Precision	Recall	Macro F1
LR	0.718	0.465	0.402	**0.662**
SVC	0.750	0.621	0.154	0.548
SVR	0.743	0.530	0.299	0.610
RF	0.730	0.484	0.265	0.586
NB	0.650	0.301	0.239	0.518

Table 5
Analysis of Results with Different Settings on the English Development Dataset

Settings	Precision	Recall	F1	Macro F1
Batch size = 32, Max len = 128	0.834	0.761	0.796	0.799
Batch size = 16, Max len = 128	0.798	0.769	0.783	0.780
Batch size = 32, Max len = 256	0.834	0.761	0.796	0.799
Batch size = 16, Max len = 256	0.849	0.796	0.821	**0.821**

Table 6
Analysis of Results with Different Settings on the Italian Development Dataset

Settings	Precision	Recall	F1	Macro F1
Batch size = 32, Max len = 128	0.760	0.633	0.690	0.796
Batch size = 16, Max len = 128	0.677	0.700	0.688	0.787
Batch size = 32, Max len = 256	0.730	0.633	0.678	0.786
Batch size = 16, Max len = 256	0.733	0.733	0.733	**0.818**

As we can see in both English and Italian, the models based on transformers are ahead by almost a decimal point, which is still quite a lot. Something that surprised us was the performance of these models in Italian since they work almost as well as in English. Usually, the transformers have a better performance in English because this is the language in which they were trained, but we can observe that the art of transferring these models to other languages, such as Italian, is getting better and better. This model was fine-tuned specifically for the task of sentiment analysis in Italian texts. This may have helped to obtain better results in Italian.

Note also that in both English and Italian, increasing the maximum token size and reducing the number of batches to 16 helped. This may be because fewer tokens cause information to be lost along the sentences, which could contain bias.

Something similar happens with sentence BERT, which we use for the machine learning models that here in English, it performs better because, for Italian, we use a multilingual model, not one focused on this language. Of these, we can also appreciate that logistic regression is the one with the best performance. It is also important to mention that among the features that helped the most were the bags of words, since the words in which they differ are very significant, followed by the embeddings generated by sentences BERT, which helped to increase the results by almost one-tenth.

5. Conclusion and Future Work

In this research, we use transformer-based models. We fine-tuned BERT-based and Italian BERT-base to analyze the subjectivity of newspaper articles. We also employ classical methods to compare performance on this task.

This approach achieved 5th. place in English and 1st. place in Italian, with 0.7079 and a 0.7917 macro F1-score, respectively. In the case of English, the results are above the baseline, with a difference of only

0.04 from the first place. Our findings show that transformer-based models are effective at detecting subjectivity in sentences.

Future experiments on the classical machine learning models include adding features that consider elements of subjectivity more related to semantics, for example, to detect sarcastic or ironic expressions, to detect if in the sentence a conclusion is expressed, to identify intensifiers in a better way, it is not enough to count adjectives and adverbs. For the part of transformers a domain adaptation can be made before making the classification, in addition to an assembly of models, in the state of the art there are already models that detect feelings, sarcasm, hate speech, etc., look for a way to put them together to enrich the subjectivity detection model.

In future work, several strategies can be employed to enhance the performance of our model. One potential improvement involves implementing weight loss management to address class imbalance. This technique adjusts the loss function to assign higher weights to underrepresented classes, thereby improving the model's ability to learn from these examples. Additionally, freezing model parameters can be beneficial. Specifically, freezing the lower layers of the model while training only the upper layers and the classification layer can lead to more efficient and targeted learning. Combined with further hyperparameter tuning and advanced regularization techniques, these approaches hold promise for achieving better performance metrics in subsequent experiments.

Acknowledgments

K. Salas-Jimenez thanks CONAHCYT scholarship program (CVU: 1291359). J. Díaz-Reyes thanks CONAHCYT scholarship program (CVU: 923309).

This research was funded by CONAHCYT (CF-2023-G-64) and PAPIIT project IT100822, IN104424. G.B.E. is supported by a grant for the requalification of the Spanish university system from the Ministry of Universities of the Government of Spain, financed by the European Union, NextGeneration EU (María Zambrano program, Universitat de Barcelona).

References

[1] F. Ruggeri, F. Antici, A. Galassi, K. Korre, A. Muti, A. Barrón-Cedeño, On the definition of prescriptive annotation guidelines for language-agnostic subjectivity detection., Text2Story@ ECIR 3370 (2023) 103–111.

[2] 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.

[3] T. Spinde, M. Plank, J.-D. Krieger, T. Ruas, B. Gipp, A. Aizawa, Neural media bias detection using distant supervision with babe–bias annotations by experts, arXiv preprint arXiv:2209.14557 (2022).

[4] S. Raza, D. J. Reji, C. Ding, Dbias: detecting biases and ensuring fairness in news articles, International Journal of Data Science and Analytics (2022) 1–21.

[5] L. K. Ipek Baris Schlicht, D. Altiok, DWReCO at CheckThat! 2023: Enhancing Subjectivity Detection through Style-based Data Sampling, in: Notebook for the CheckThat! Lab at CLEF 2023, CLEF 2023: Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, 2023.

[6] J. M. Pérez, M. Rajngewerc, J. C. Giudici, D. A. Furman, F. Luque, L. A. Alemany, M. V. Martínez, pysentimiento: A python toolkit for opinion mining and social nlp tasks, 2023. arXiv:2106.09462.

[7] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, CoRR abs/1908.10084 (2019). URL: http://arxiv.org/abs/1908.10084. arXiv:1908.10084.

[8] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the

2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45.

[9] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[10] NeuralyIA, neuraly/bert-base-italian-cased-sentiment, https://huggingface.co/neuraly/bert-base-italian-cased-sentiment, 2021. Accessed: 2024-05-24.