

# TUDublin at CheckThat! 2023: ChatGPT for Data Augmentation

Elena Shushkevich<sup>1,\*</sup>, John Cardiff<sup>1</sup>

<sup>1</sup>*Technological University Dublin, Ireland*

## Abstract

This work describes the approach of the TUDublin team at the CheckThat! 2023 Task 2: Subjectivity in News Articles. This task is aimed to discern whether a sentence in a news article conveys the subjective perspective of its author or provides an objective viewpoint on the subject being discussed. Our team worked with English and Italian datasets. We applied mBERT, XLM-RoBERTa, SBERT models, and an ensemble of them. To improve the results, we employ ChatGPT for the news generation. Using such AI-generated news, we balanced the datasets and expanded them, which allowed us to increase the results by 9% macro F1-score for English and 3% macro F1-score for Italian (validation datasets).

## Keywords

Subjectivity detection, AI-generated news, ChatGPT, Large Language Models

## 1. Introduction

Subjectivity detection is a task of natural language processing that aims to identify 'factual' or 'neutral' content in textual data [1]. It plays a crucial role in various domains, including sentiment analysis, opinion mining, and information retrieval. The ability to differentiate subjective and objective expressions enables more accurate understanding and analysis of text, allowing researchers, businesses, and organizations to gain valuable insights from large volumes of textual data. The importance of subjectivity detection lies in its practical applications. In today's digital era, where information overload and fake news proliferate, it is crucial to distinguish between subjective opinions and objective facts. Subjectivity detection enables the identification of biased or subjective content, allowing individuals, businesses, and organizations to make informed decisions based on reliable information.

This paper presents the experience of the TUDublin team in participating in the CheckThat!-2023 Task 2, which focused on subjectivity detection in news articles. The primary objective of the task was to determine whether a given message was objective or subjective. The article is structured into six sections to provide a comprehensive overview of the team's approach and findings.

Section 1 serves as the introduction, providing background information on the problem and the significance of subjectivity detection in news articles. In Section 2, we explore relevant

---

*CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece*


\*Corresponding author.

✉ elena.n.shushkevich@gmail.com (E. Shushkevich); john.cardiff@tudublin.ie (J. Cardiff)

🆔 0009-0002-8899-3942 (E. Shushkevich); 0000-0003-1863-4557 (J. Cardiff)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

research conducted in the field, highlighting prior works. Section 3 details the dataset utilized for the study, along with our approach of augmenting the original dataset using ChatGPT 4<sup>1</sup>. Section 4 outlines the preprocessing steps employed and describes the models used for subjectivity detection. Section 5 presents the training phase results, as well as the team’s performance in the challenge, followed by an analysis of the achieved results. Finally, in Section 6, we summarize the conclusions from the work and outline potential avenues for further improvements to the subjectivity detection model.

## 2. Related work

Subjectivity detection is a popular task in natural language processing, where researchers use various methodologies and techniques to differentiate between subjective and objective content. In [2] authors introduced a machine learning-based approach that utilized lexical and syntactic features for sentiment analysis, including subjectivity detection. In [3] authors proposed a bootstrapping algorithm that acquired subjective patterns using syntactic patterns. Recent advancements include the use of deep learning models like convolutional neural network architecture [4] and BERT model [5], showcasing improved performance in subjectivity detection tasks in comparison with classic techniques. These studies highlight the ongoing evolution of subjectivity detection techniques, incorporating lexical, syntactic, and contextual information to accurately identify subjectivity in text.

## 3. Datasets

### 3.1. CheckThat!-2023 dataset

The CheckThat!-2023 Task 2 (Subjectivity detection) [6, 7] introduced datasets in various languages, including Arabic, Dutch, English, German, Italian, and Turkish. However, for our experiments, we focused solely on the English and Italian datasets. Each dataset consists of three labels: `sentence_id` (representing the unique identifier for a sentence within a news article), `sentence` (containing the textual content of the sentence), and `label` (denoting 'OBJ' for objective messages and 'SUBJ' for subjective messages). The statistics of both datasets are presented in the Table 1.

**Table 1**

Statistics of the CheckThat!-2023 datasets for English and Italian, in the training/validation and test sets

Language	Objective messages	Subjective messages	Total	Messages in the Test dataset
English	532/106	298/113	830/219	243
Italian	1231/167	382/60	1613/227	440

The Italian training dataset stands out for its size, being twice as large as the English training dataset. Both datasets are unbalanced: there is a clear discrepancy in the distribution of objective

---

<sup>1</sup><https://chat.openai.com/>

and subjective messages, with a notable overrepresentation of objective messages compared to subjective ones. This discrepancy is particularly pronounced in the case of the Italian training dataset, where the number of objective news instances is three times greater than the number of subjective news instances. This imbalance poses a challenge when training models, as it may impact the overall performance and accuracy of subjectivity detection tasks on these datasets.

### **3.2. Expansion of the datasets using ChatGPT 4**

Considering the potential impact of the inherent dataset imbalance on the classification results, we made a strategic decision to address this issue by balancing and augmenting both the English and Italian datasets using AI-generated messages. Our hypothesis is that AI-generated sentences can closely approximate, though not replicate, the human-generated messages collected by the datasets creators. As it was repeatedly demonstrated [8, 9], a dataset expansion can enhance classification outcomes, and we believe it will prove beneficial in our specific case as well.

To achieve dataset balance, we employed ChatGPT 4, providing it with the training datasets for both English and Italian. We instructed ChatGPT to generate messages that closely resemble the original ones, both in terms of subjectivity and objectivity, but not identical to them. Some examples of the ChatGPT-generated sentences are presented in the Table. 2.

As a result, we expanded the number of messages in the training datasets. For the English training dataset, we expanded it from 830 to 1546 messages (773 objective, 773 subjective). Similarly, the Italian training dataset was expanded from 1613 to 2118 messages (1059 objective, 1059 subjective). The comparison between the original and the expanded datasets is visually depicted in Figure 1.

This approach allows us to address the dataset imbalance and potentially improve the classification results by incorporating a wider range of artificially generated data while maintaining the essence of the original messages.

## **4. Modeling**

This section describes the two stages of our model creation: the preprocessing and the modeling steps.

### **4.1. Preprocessing**

During the preprocessing stage, we meticulously followed a set of steps to ensure the data was appropriately prepared. These steps included:

- Converting all characters to lowercase: To ensure consistency and remove any potential case-related discrepancies, we transformed all characters in the dataset to lowercase.
- Removing non-alphabetic and non-numeric characters: To focus solely on the textual content, we eliminated any characters that did not belong to the English and Italian alphabet or numeric values.
- Removing stopwords, as they do not typically carry significant meaning and can hinder classification accuracy.

**Table 2**

Examples of sentences from the original dataset vs ChatGPT-generated messages

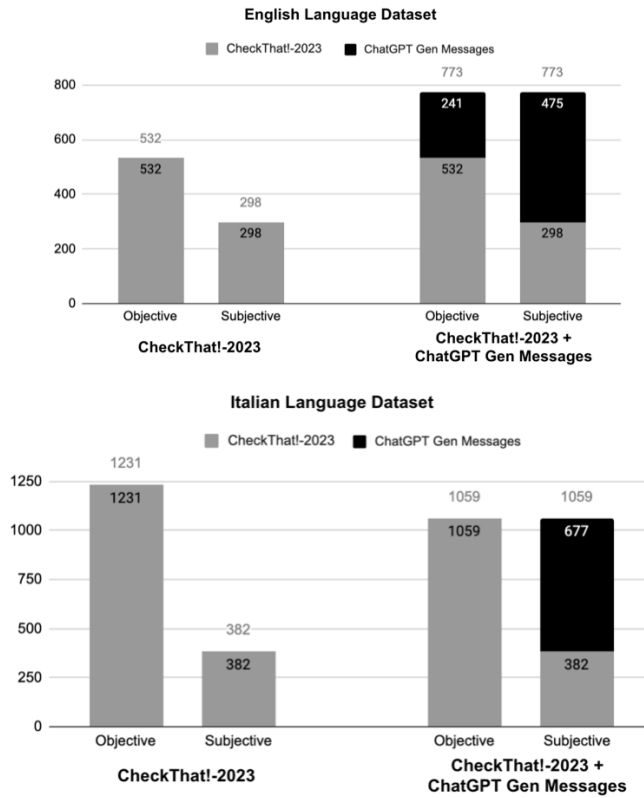
Message	Label	Origin
The former president has continued to deny wrongdoing, accusing James of being politically motivated, CNBC reports.	OBJ	Original Dataset
European Pride Organisers Association’s Latvian President, Kristine Garina, stated that they are contesting the imposed restrictions through legal channels.	OBJ	ChatGPT
Right now, the most common variant is BA.5, at 85%, BA.4.6, which comprised 10.3% of infections, and BA.2.75, which comprised 1.3%.	OBJ	Original Dataset
This mutation, R346T, has appeared in other variants and is linked to immune evasion, allowing the virus to elude antibodies obtained through vaccination and previous infection.	OBJ	ChatGPT
Torrential spending by the many arms of the state left behind excess capacity, a skewed pattern of production and heavy debts.	SUBJ	Original Dataset
The state’s excessive spending left a legacy of unused capacity, uneven production patterns, and significant debt.	SUBJ	ChatGPT
But taking refuge in public credit will cause that same infection to attack business, banking, industry, agriculture, the entire body of private enterprise.	SUBJ	Original Dataset
Responsibilities to keep infections at bay place significant pressure on local government, potentially detracting attention from efforts to boost public investment, even when adequate funding is accessible.	SUBJ	ChatGPT

- Identifying emotional messages: Messages that contained symbols like "!!!," "...," or "???" often indicate an emotional tone. As part of our preprocessing, we marked these messages as "EMOTIONAL" to capture the distinct emotional content for further analysis, as we believe that more emotional messages tend to be more subjective.
- Identifying first-person messages: we marked messages that contained the pronoun "I" (or its equivalent "Io" in the Italian dataset) as "FIRST." This helped us distinguish these messages for further analysis because we expected them to be more subjective in nature.

By implementing these preprocessing steps, we aimed to ensure data uniformity, eliminate noise, and highlight specific characteristics within the dataset that could provide valuable insights during subsequent stages of analysis.

## 4.2. Modeling

For our subjectivity detection task, we employed several models:



**Figure 1:** Comparison between the original and expanded datasets.

- mBERT [5]: The multilingual BERT (mBERT) model played a crucial role in our experiments. With support for 104 languages, mBERT features a 12-layer architecture, 768 hidden units, 12 attention heads, and around 110 million parameters. Trained on a wide range of Wikipedia articles from diverse languages, mBERT excels in capturing linguistic patterns and contextual information. We leveraged its power to analyze subjectivity in text. For the experiments, we used the batch size of 16, 512-token input, and 0.3 dropout.
- SBERT [10]: We also utilized SBERT, a modified version of BERT that incorporates a siamese and triplet network structure. SBERT facilitates the creation of semantically meaningful sentence embeddings, enabling efficient comparisons using cosine similarity. Its streamlined design allows for faster processing without compromising the quality of results. For the experiments, we used the batch size of 16, 128-token input, and 0.5 dropout.
- XLM-RoBERTa [11]: Another essential model in our arsenal was XLM-RoBERTa, a cross-lingual sentence encoder renowned for achieving state-of-the-art performance on various cross-lingual understanding benchmarks. Trained on a vast corpus of filtered Common-Crawl data spanning 100 languages, XLM-RoBERTa provides robust linguistic representations that excel in capturing nuances across different languages. We used the batch size

of 16, 128-token input, and 0.5 dropout.

- An ensemble of these models, combining their predictions to enhance the accuracy and reliability of our subjectivity detection. The ensemble approach is very popular and often allows to improve the results of classification [12, 13]. This ensemble approach allowed us to capture consensus among the models by considering a label as correct only if it was agreed upon by the majority of models.

## 5. Results

The results of experiments achieved on the English validation dataset are presented in the Table 3.

**Table 3**

Results on the English dataset

Model	Accuracy/Macro F1-score
mBERT	0.493/0.492
SBERT	0.470/0.463
XLM-RoBERTa	0.447/0.447
Ensemble of mBERT, SBERT, XLM-RoBERTa models	0.493/0.492

The best and similar results were obtained by the mBERT model and the ensemble of all models, so we continue the experiments with the mBERT model.

The results of the experiments on the Italian validation dataset are presented in the Table 4.

**Table 4**

Results on the Italian dataset

Model	Accuracy/Macro F1-score
mBERT	0.458/0.455
SBERT	0.414/0.407
XLM-RoBERTa	0.401/0.401
Ensemble of mBERT, SBERT, XLM-RoBERTa models	0.458/0.455

Similarly to the English dataset, we observed that on the Italian dataset the MBERT model consistently achieved the most favorable results. Notably, the ensemble approach yielded the same results as the mBERT model alone. Based on these findings, we made the informed decision to proceed with further experiments utilizing the mBERT model.

We conducted the experiments using the mBERT model on the expanded by AI-generated news datasets (both English and Italian), and the comparison of the results is demonstrated in the Table 5.

In comparison to the results obtained using the original datasets, we observed a noticeable improvement in performance when employing the balanced and expanded datasets, which were augmented using AI-generated news. Specifically, the macro F1-score exhibited a significant

**Table 5**

Comparison of mBERT results with and without ChatGPT-generated news articles

Model	Accuracy/Macro F1-score
English	
mBERT	0.493/0.492
mBERT + ChatGPT-generated data	0.580/0.578
Italian	
mBERT	0.458/0.455
mBERT + ChatGPT-generated data	0.520/0.481

increase of 9% for the English dataset and 3% for the Italian dataset. These improvements underscore the effectiveness of dataset expansion in enhancing the subjectivity detection task.

Building upon these findings, we opted to submit our experimental runs utilizing the mBERT model on the expanded datasets. The resulting performance on the test datasets is presented in the Table 6.

**Table 6**

Results on the test datasets for English and Italian

Language	Macro F1-score
English	0.40
Italian	0.46

Regrettably, the achieved results fell short of our expectations, indicating that further improvements are necessary in this area. Such low results could be explained by the not enough finetuned mBERT. The results on the test set using the ChatGPT-generated data should be compared with the results on the test set with the original dataset as the training one. Furthermore, in future endeavors, it is imperative to enhance the fine-tuning process for the primary model employed in binary classification. Nevertheless, despite the unsatisfactory outcome, it is worth highlighting that the innovative approach of augmenting datasets with ChatGPT-generated texts has shown promising results in the improvement of both English and Italian training datasets.

## 6. Conclusion

In this study, we conducted comprehensive experiments on the CheckThat!-2023 Task 2, focusing on subjectivity detection in news articles. Three transformer models, namely mBERT, SBERT, and XLM-RoBERTa, along with an ensemble of these models, were employed to analyze the subjectivity within the dataset. Among these models, mBERT consistently yielded the highest performance.

To address the issue of dataset imbalance, we leveraged ChatGPT to generate additional data, which proved instrumental in achieving a balanced and expanded dataset. The incorporation

of AI-generated news significantly improved the classification results for both the English and Italian datasets, showcasing a noteworthy improvement of 9% and 3% in macro F1-score, respectively.

These findings highlight the potential of employing novel approaches, such as AI-generated news, to enhance subjectivity detection in news articles. In future, we plan to continue research and exploration of AI-generated news to uncover its potential applications and possibilities in the field.

## References

- [1] I. Chaturvedi, E. Ragusa, P. Gastaldo, R. Zunino, E. Cambria, Bayesian network based extreme learning machine for subjectivity detection, *Journal of The Franklin Institute* 355 (2017) 1780–1797. doi:10.1016/J.JFRANKLIN.2017.06.007.
- [2] B. Pang, L. Lee, A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts (2004). doi:10.3115/1218955.1218990.
- [3] E. Riloff, J. Wiebe, Learning extraction patterns for subjective expressions, in: *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 2003, pp. 105–112.
- [4] Y. Kim, Convolutional neural networks for sentence classification, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, p. 1746–1751.
- [5] 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>. doi:10.18653/v1/N19-1423.
- [6] 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.
- [7] 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.
- [8] E. Shushkevich, J. Cardiff, Detecting fake news about covid-19 on small datasets with



- machine learning algorithms, in: 2021 30th Conference of Open Innovations Association FRUCT, 2021, pp. 253–258. doi:10.23919/FRUCT53335.2021.9599970.
- [9] E. Shushkevich, M. Alexandrov, J. Cardiff, Bert-based classifiers for fake news detection on short and long texts with noisy data: A comparative analysis, in: Proceedings of the 25th International Conference on Text, Speech, and Dialogue (TSD 2022), 2022, pp. 263–274. URL: Unavailable. doi:[https://doi.org/10.1007/978-3-031-16270-1\\_22](https://doi.org/10.1007/978-3-031-16270-1_22).
- [10] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. arXiv:1908.10084.
- [11] 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. URL: <https://aclanthology.org/2020.acl-main.747>. doi:10.18653/v1/2020.acl-main.747.
- [12] A. Glazkova, M. Glazkov, T. Trifonov, g2tmn at constraint@AAAI2021: Exploiting CT-BERT and ensembling learning for COVID-19 fake news detection, in: Combating Online Hostile Posts in Regional Languages during Emergency Situation, Springer International Publishing, 2021, pp. 116–127. URL: [https://doi.org/10.1007/978-3-030-73696-5\\_12](https://doi.org/10.1007/978-3-030-73696-5_12). doi:10.1007/978-3-030-73696-5\_12.
- [13] X. Li, Y. Xia, X. Long, Z. Li, S. Li, Exploring text-transformers in aai 2021 shared task: Covid-19 fake news detection in english, 2021. arXiv:2101.02359.