

ES-VRAI at CheckThat! 2023: Enhancing Model Performance for Subjectivity Detection through Multilingual Data Aggregation

Hamza Tarik Sadouk^{1,*}, Faouzi Sebbak¹ and Hussem Eddine Zekiri¹

¹*Ecole Militaire Polytechnique, PO Box 17, 16111 Bordj El Bahri, Algiers, Algeria*

Abstract

This paper presents our participation in the CLEF2023 CheckThat! Lab [1], focusing on Task 2, which addresses Subjectivity Detection [2]. Distinguishing subjective and objective content is pivotal in numerous natural language processing tasks. Our work delves into the challenges and techniques associated with the binary classification problem of discerning personal opinions from impartial stances in textual data. The task encompasses six languages: Arabic, Dutch, English, German, Italian, and Turkish. We adopt a multilingual approach, merging diverse datasets into a comprehensive dataset to train a multilingual model. Through fine-tuning pre-trained language models and employing sampling techniques to tackle class imbalance, we optimize the model's performance. Our methodology combines multilingual data aggregation with fine-tuning and class imbalance handling, resulting in a robust subjectivity detection model. By participating in the CheckThat! Lab, we contribute to advancing the understanding of subjectivity detection in different languages, opening avenues for more accurate sentiment analysis and text classification in various applications.

Keywords

Subjectivity Detection, Multilingual, Data aggregation, BERT-Multilingual, XLM-RoBERTa.

1. Introduction

In today's digital age, the vast amount of textual data generated through social media, online forums, news articles, and other sources present a significant challenge for automated systems. A crucial task in natural language processing is to accurately discern whether a comment represents the author's personal opinion or conveys an impartial stance on a discussed topic. This binary classification problem requires sophisticated algorithms capable of analyzing text segments, which may consist of sentences or paragraphs, and accurately classifying them as subjective or objective. Understanding the subjectivity of text is essential for numerous applications, including sentiment analysis, information retrieval, content recommendation, and opinion mining. By developing computational approaches to tackle subjectivity detection, we can unlock valuable insights and improve the overall understanding of textual data.

The utilization of linguistic features, such as syntactic patterns, semantic cues, lexical choices, and stylistic elements, captures the subjective nature of textual content. These features provide

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


*Corresponding author.

✉ sadouk.hamza.tarik@gmail.com (H. T. Sadouk); faouzi.sebbak@gmail.com (F. Sebbak);

houssemzekiri@gmail.com (H. E. Zekiri)



© 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)

valuable clues about the author's emotions, attitudes, and perspectives, enabling a deeper understanding of the text's subjectivity. In addition to linguistic features, contextual information plays a vital role in subjectivity detection. Understanding the surrounding context, including co-occurring words, discourse structure, and dialogue patterns, aids in distinguishing personal opinions from factual statements. By considering the broader context in which the text segment appears, we enhance the system's ability to accurately classify subjective and objective content.

In recent years, language models and transformer architectures have revolutionized the field of natural language processing, offering powerful tools for capturing linguistic features and contextual information in subjectivity detection. These models, such as BERT (Bidirectional Encoder Representations from Transformers) [3], and RoBERTa (Robustly Optimized BERT approach) [4], have demonstrated remarkable success in various language understanding tasks. By leveraging the pre-trained representations learned from vast amounts of text data, language models have the ability to encode rich semantic and syntactic information into their embeddings. This enables them to capture intricate linguistic features that are crucial for identifying subjectivity in textual content. The deep contextual understanding provided by transformer architectures allows for the recognition of subtle nuances and linguistic cues that differentiate subjective expressions from objective statements.

The utilization of language models offers immense potential in uncovering subjective expressions and objective stances in textual data, contributing to the overall understanding and analysis of subjective content. This approach allows us to capture the nuances and intricacies of language usage, enabling our models to better discern subjective elements and provide more precise classifications. By combining these advanced techniques with our comprehensive methodology of multilingual data aggregation, fine-tuning, and class imbalance handling, we create a holistic solution that improves subjectivity detection across languages and facilitates a deeper comprehension of sentiment and perspective in text.

2. Related Work

Subjectivity Detection (SD) is a process aimed at differentiating between objective and subjective information. Historically, two primary methods have been employed: syntactic and semantic.

Semantic approaches tackle subjectivity detection by utilizing statistical or neural text representation techniques [5, 6], necessitating labeled training data. These methods may incorporate domain-specific assumptions or employ guidelines for the annotation process to acquire the required training data [7].

Syntactic methods primarily rely on identifying keywords [8] or employing lexicons [9]. However, these techniques are often specific to particular languages and may lose information during translation. Additionally, lexicon-based approaches require external databases, which can limit their applicability in various scenarios.

While semantic methods offer advantages such as language independence and applicability to multiple languages, they also present challenges. The perception of subjectivity is inherently subjective, leading to interpretation bias, ambiguity in annotation, and difficulties in handling edge cases [10].

Table 1 summarizes the methods used in subjectivity detection.

Table 1
Taxonomy of Subjectivity Detection Methods

Method	Description	Advantage	Disadvantage
Syntactic	Identify keywords or use lexicons to determine subjectivity.	Easy to implement and fast.	Language-specific and can lose information during translation.
Semantic	Use statistical or neural text representation techniques to determine subjectivity.	More accurate than syntactic methods.	Requires labeled training data and can be prone to interpretation bias.

Table 2
Dataset description for subjectivity detection

	Training		Development	
	Subjective	Objective	Subjective	Objective
Arabic	280	905	70	227
Dutch	311	489	93	107
English	298	532	113	106
German	308	492	77	123
Italian	382	1281	60	167
Turkish	378	422	100	100
Total	1957	4072	513	830

In this work, we leverage the power of language models and transformer architectures to address linguistic features and contextual information in subjectivity detection. By incorporating these state-of-the-art techniques into our computational models, we aim to advance the accuracy and robustness of subjectivity classification systems.

3. Data overview

This task is provided in seven languages: Arabic, Dutch, English, German, Italian, and Turkish [2]. In our work, we adopt a multilingual method by merging all available datasets into one and then training a multilingual model. All details in Table 2.

Upon examining Table 2, it becomes apparent that class imbalance is present, which poses a particular concern. To address this issue, we have implemented a range of techniques as detailed subsequently.

Upsampling is a method used to tackle imbalanced data by increasing the number of samples in the minority class. It can improve performance for underrepresented classes but has drawbacks like overfitting due to duplicated or similar instances, leading to limited generalization on unseen data.

Downsampling is a method used to handle imbalanced data by decreasing the number of instances in the majority class. It aims to achieve balance by randomly removing samples from the majority class. However, downsampling has limitations, including the loss of valuable

information and smaller dataset size, which may not be ideal for training complex models.

4. Proposed approach

In our work, we have adopted a multilingual approach to data aggregation, consolidating diverse datasets from multiple languages into a comprehensive dataset and training a multilingual model. By merging datasets from different languages, we harness the inherent linguistic variations and semantic richness present across languages. This approach enhances the model’s ability to capture a wide range of linguistic patterns and effectively handle cross-lingual tasks, facilitating better generalization across different language domains. By including multiple languages in the training data, the model becomes more adaptable, robust, and capable of handling diverse textual inputs.

However, certain languages often suffer from limited available resources, known as low-resource languages, leading to suboptimal results when training models individually due to data scarcity and lack of linguistic resources. To address this challenge, our multilingual data aggregation strategy proves advantageous. By merging datasets across languages, the model can learn from the patterns and structures present in resource-rich languages and transfer that knowledge to low-resource languages. This cross-lingual transfer mitigates data scarcity issues and enhances the model’s performance on low-resource languages, resulting in more accurate and reliable outcomes.

Our approach involves not only merging diverse datasets from various languages into a single comprehensive dataset but also fine-tuning pre-trained language models and employing sampling techniques to address the class imbalance. By fine-tuning pre-trained language models, we leverage their existing linguistic understanding and adapt them to the specific subjectivity detection task. Additionally, we employ sampling techniques to ensure a balanced representation of subjective and objective instances. We used both Upsampling and Downsampling in a comparative study to assess which technique was more suitable for the desired task. This comprehensive methodology combining multilingual data aggregation, fine-tuning, and class imbalance handling results in a robust and accurate subjectivity detection model capable of effectively classifying subjective and objective content in different languages, contributing to a deeper understanding of textual sentiment and perspective. In this perspective, we used two multilingual models:

BERT-Multilingual [3] is a pretrained language model that can be finetuned on various downstream natural language processing tasks, including named entity recognition, sentiment analysis, and question answering, across multiple languages. It is trained on a large corpus of monolingual text from 104 languages, including low-resource languages, making it a valuable tool for cross-lingual transfer learning. The model uses a transformer architecture and employs a masked language modeling objective during pretraining. BERT-Multilingual has achieved state-of-the-art results on many benchmark NLP tasks, making it a widely used and highly influential model in the NLP community.

XLM-RoBERTa [4] is a cross-lingual language model and is an extension of RoBERTa. XLM-RoBERTa is pre-trained on monolingual and multilingual datasets, including 100 languages, using masked language modeling (MLM) and translation language modeling (TLM) objectives.

Table 3

Model performance comparison for multilingual subjectivity detection

Model	Accuracy	F1	Precision	Recall	F1-Macro	MacroP	MacroR
BM1	0.7817	0.7698	0.8141	0.7300	0.7811	0.7847	0.7817
BM2	0.8167	0.8161	0.8188	0.8133	0.8167	0.8167	0.8167
BM3	0.7700	0.7738	0.7613	0.7867	0.7699	0.7703	0.7700
XR1	0.7550	0.7361	0.7977	0.6833	0.7537	0.7603	0.7550
XR2	0.7800	0.7815	0.7763	0.7867	0.7800	0.7800	0.7800
XR3	0.7500	0.7525	0.7451	0.7600	0.7500	0.7501	0.7500

The model employs a larger batch size and more data augmentation techniques, such as noise and token shuffling, to improve performance. XLM-RoBERTa outperforms previous state-of-the-art models on several cross-lingual benchmarks, such as XNLI and the MLQA multilingual question-answering dataset.

5. Results and discussion

In this study, we evaluated six different model configurations for classifying subjective and objective claims in a multilingual dataset containing Arabic, Dutch, English, German, Italian, and Turkish languages. The dataset was imbalanced, with a larger number of objective sentences compared to subjective sentences. To tackle the class imbalance issue, we employed Oversampling and Downsampling techniques in four out of the six configurations. Table 3 shows the results obtained on a fraction of the Development dataset (50% of the Development dataset was used for hyperparameter optimization). We found that the following hyperparameters produced the best results: number of epochs= 4; learning rate= $1e-5$; batch size= 16.

The results presented in this study were obtained prior to the release of the Test dataset, with the objective of identifying the most appropriate configuration to be employed during the evaluation phase. This pre-release assessment aimed to discern the optimal setup for subsequent evaluations and ensure the reliability and validity of the findings. For the finetuned models with different sampling techniques, we use the following contractions:

- BM1: BERT-Multilingual;
- BM2: BERT-Multilingual + Oversampling;
- BM3: BERT-Multilingual + Downsampling;
- XR1: XLM-RoBERTa;
- XR2: XLM-RoBERTa + Oversampling;
- XR3: XLM-RoBERTa + Downsampling.

Upon analysis of Table 3, we have derived the following observations:

- BM2 performs best in terms of accuracy, F1-Score, F1-Macro, Precision Macro, and Recall Macro. This indicates that handling the class imbalance using Oversampling helps improve the model’s performance.

- BM3 yields better results than the baseline BM1 model, although it underperforms compared to the BM2 approach. This suggests that Oversampling is more effective at addressing the class imbalance in this case.
- XR models generally underperform when compared to BM models. It is possible that the XR architecture is less suited to this specific task or dataset or that additional hyperparameter tuning is required to improve its performance.
- Similar to BM, XR models also benefit from Oversampling, showing an increase in performance compared to the baseline XR1. However, the improvement is not as significant as that observed for BM1.
- During the evaluation cycle, we employed **BM2**, which yielded F1-Macro of **0.78**. Unfortunately, due to technical problems, we were unable to produce the required results within the evaluation cycle deadline. Consequently, despite achieving a commendable fourth-place result, we did not secure a position on the leaderboard. Notably, the first-place position was attained with a slightly higher F1-Macro of 0.82.

6. Conclusion

Our work focuses on a multilingual approach to data aggregation, merging diverse datasets from multiple languages to train a comprehensive multilingual model. By leveraging linguistic variations and semantic richness across languages, the model captures a wide range of patterns and excels in cross-lingual tasks, enhancing generalization. We employ sampling techniques and fine-tune pre-trained language models to address class imbalance. Two models are used BERT-Multilingual and XLM-RoBERTa. BERT-Multilingual achieves the highest performance, demonstrating significant improvements across multiple metrics. Oversampling effectively addresses class imbalance and further enhances its performance. Downsampling also yields better results but is outperformed by Oversampling. XLM-RoBERTa models generally underperform compared to BERT-Multilingual and may require additional hyperparameter tuning. While both models benefit from Oversampling, BERT-Multilingual achieves a commendable fourth-place position in the evaluation, securing high accuracy and F1-Macro. Unfortunately, technical issues prevented the submission of final results within the deadline, narrowly missing the first-place position achieved by a slightly higher F1-Macro. Overall, our approach yields a robust and accurate subjectivity detection model, facilitating sentiment analysis and enhancing textual understanding across languages.

References

- [1] A. Barrón-Cedeño, F. Alam, A. Galassi, G. Da San Martino, P. Nakov, , T. Elsayed, D. Azizov, T. Caselli, G. Cheema, F. Haouari, M. Hasanain, M. Kutlu, C. Li, F. Ruggeri, J. M. Struß, W. Zaghoulani, 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 In-*

- teraction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), 2023.
- [2] A. Galassi, F. Ruggeri, A. B.-C. no, 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. Turkmen, M. Wiegand, W. Zaghouni, Overview of the CLEF-2023 CheckThat! lab task 2 on subjectivity in news articles, in: Working Notes of CLEF 2023—Conference and Labs of the Evaluation Forum, CLEF '2023, Thessaloniki, Greece, 2023.
 - [3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2019).
 - [4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, et al., Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2020).
 - [5] B. Pang, L. Lee, A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, arXiv preprint cs/0409058 (2004).
 - [6] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, arXiv preprint arXiv:1404.2188 (2014).
 - [7] F. Antici, L. Bolognini, M. A. Inajetovic, B. Ivasiuk, A. Galassi, F. Ruggeri, Subjectivita: An italian corpus for subjectivity detection in newspapers, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12, Springer, 2021, pp. 40–52.
 - [8] J. Wiebe, E. Riloff, Creating subjective and objective sentence classifiers from unannotated texts, in: Computational Linguistics and Intelligent Text Processing: 6th International Conference, CICLing 2005, Mexico City, Mexico, February 13-19, 2005. Proceedings 6, Springer, 2005, pp. 486–497.
 - [9] N. Das, S. Sagnika, A subjectivity detection-based approach to sentiment analysis, in: Machine Learning and Information Processing: Proceedings of ICMLIP 2019, Springer, 2020, pp. 149–160.
 - [10] 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, in: Proceedings of Text2Story—Sixth Workshop on Narrative Extraction From Texts, held in conjunction with the 45th European Conference on Information Retrieval (ECIR 2023), volume 3370, CEUR-WS. org, 2023, pp. 103–111.