Thesis Titan at CheckThat! 2023: Language-Specific Fine-tuning of mDeBERTaV3 for Subjectivity Detection⋆

Notebook for the CheckThat! Lab Task 2 at CLEF 2023

Folkert Atze Leistra[1,†], Tommaso Caselli[1,†]

[1] *Rijksuniversiteit Groningen, Oude Kijk in't Jatstraat 16, 9712 EK Groningen, The Netherlands*

Abstract

The detection of subjectivity in natural language plays a crucial role in various applications, such as sentiment analysis, fake news detection, and fact-checking systems. However, effectively and accurately detecting subjectivity across different languages presents substantial challenges due to linguistic variations and cultural nuances. This paper describes the system we developed for 2023 CheckThat! Lab Task 2 on subjectivity detection using a multilingual model, `mDeBERTaV3-base`. In particular, we use a common multilingual dataset to fine-tune multiple `mDeBERTaV3-base` models using language specific development data to specialize the systems towards a target language and reduce the impact of the class imbalance in the training data. In this way, we managed to rank first in German, Italian and Turkish, second in Arabic and over a Multilingual dataset, and third in English.

Keywords

Sentence semantic, Subjectivity detection, Multilinguality, mDeBERTaV3

1. Introduction

The widespread use of social media has resulted in an unprecedented production of unstructured data (e.g., textual messages, images, multimodal content, among other forms), which is now accessible to a wide range of individuals from diverse backgrounds. The lack of strict forms of control on what is published makes it urgent to be able to reliably distinguish *opinions* from *facts*. This ability plays an essential role in differentiating between subjective from objective content. Subjectivity detection has strong ties with opinion mining and is often seen as a subtask of sentiment analysis [1, 2]. The development and deployment of subjectivity detection systems would also be beneficial for other tasks such as argument mining [3], fake news detection [4], and (automated) fact-checking [5].

The CLEF 2023 CheckThat! Lab [6] offers a valuable platform to address the challenges associated with subjectivity detection. In particular, Task 2 of this edition, "Subjectivity in News Articles" provides an extensive multilingual setting (Arabic, Dutch, English, German, Italian, and Turkish) for the identification of the subjectivity status of sentences extracted from news

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

f.a.leistra@student.rug.nl (F. A. Leistra); t.caselli@rug.nl (T. Caselli)
https://www.researchgate.net/profile/Folkert-Leistra (F. A. Leistra); https://www.rug.nl/staff/t.caselli (T. Caselli)
0009-0004-4981-5961 (F. A. Leistra); 0000-0003-2936-0256 (T. Caselli)

CEUR Workshop Proceedings (CEUR-WS.org)

articles [7]. The increasing polarization of the public debate is impacting the news production cycle and the way news articles are written [8], thus making it even more important to be able to discriminate between the account of events, i.e., what has happened in the world, and their opinions and interpretations.

The task is framed as a binary classification, whose goal is to assign the subjectivity status (*SUBJ* for subjective, and *OBJ* for objective) to a given sentence. Sentences that express the personal perspective of the author are considered subjective, regardless of the truthfulness of the statement [9, 6]. If the sentence presents an objective view of the covered topic, it is considered objective. The subjectivity status is assigned to sentences in isolation. This makes the task more challenging as systems cannot access representations of the full text, or portions of it, and use them to enhance the knowledge of the context of occurrence of each sentence.

After an initial round of experiments focusing on the use of monolingual pre-trained language models, we shifted to a multi-lingual one, namely `mDeBERTaV3-base` [10]. By leveraging the capabilities of a single pre-trained model capable of handling multiple languages, our approach exemplified its effectiveness. To promote reproducibility, we have made all fine-tuned models and the code used to obtain them publicly available at our GitHub repository. [1]

The remainder of this contribution is structured as follows: Section 2 provides an overview of the data that we used to fine-tune `mDeBERTaV3-base`. Section 3 illustrates our approach and the results on the development set that we used to finalize our models. Results and leaderboard ranking for each language are presented in Section 4. In Section 5 we discuss different approaches that we tried, but did not give the expected results. Lastly, Section 6 concludes this paper and outlines directions for future work.

2. Data

Table 1 provides an overview of the label distributions for all the languages composing Task 2. Each language is accompanied by its own training, development, and test distribution. Notably, the data sizes and label distributions vary quite largely across the languages. In general, there is a skewed distribution towards the OBJ class, although with varying degrees: Arabic, English, and Italian present the largest data imbalance in favor of the OBJ class, while Turkish has the smallest difference. The imbalance is in large part due to the selected domain, i.e., news articles, where the writing style tend to be flat and adhere to the facts that are being reported. Exceptions to this general pattern occur, with differences due to cultural backgrounds and traditions in media reporting across different countries as well as to the current status of the public debate [11, 12, 13]. The data annotation process involved multiple teams responsible for annotating each language.

The Multilingual data was conceived by the task organizers as an independent language. This means that the training, development, and split distributions have been constructed by mixing data from all other languages regardless of their original splits, resulting in the presence of validation instances of some individual languages in its training set. This has prevented us from directly use the dataset to train a single model and deploy it on all the other languages. To obviate to this issue we developed our multilingual training dataset, *Adapted Multilingual* (detailed in

[1] https://github.com/folkertleistra/mDeBERTaV3-subjectivity

Table 1
The distribution for the training, validation and test data per language. The *Adapted Multilingual* has been created by sampling data for all the languages (Arabic, Dutch, English, German, Italian and Turkish).

Language	Train		Validation		Test	
	OBJ	SUBJ	OBJ	SUBJ	OBJ	SUBJ
Arabic	905	280	227	70	363	82
Dutch	489	311	107	93	263	237
English	532	298	106	113	116	127
German	492	308	123	77	194	97
Italian	1231	382	167	60	323	117
Turkish	422	378	100	100	111	129
Multilingual	4,371	2,257	300	300	300	300
Adapted Multilingual	2,957	1,957	-	-	-	-

Table 2
Sentences labeled as OBJ in the *Adapted Multilingual* dataset, the coverage per language (percentages in parentheses), and the coverage with respect to the original monolingual training data.

Language	OBJ	% Original
Arabic	654 (22.11%)	72.26%
Dutch	364 (12.30%)	74.43%
English	383 (12.95%)	71.99%
German	353 (11.93%)	71.74%
Italian	898 (30.36%)	72.94%
Turkish	305 (10.31%)	72.27%

last row in Table 1). The *Adapted Multilingual* dataset has been obtained by combining all the training data from each language, excluding the Multilingual dataset from the task organisers. As a result, we obtained a large unbalanced dataset consisting of 6,028 sentences, with 4,071 labeled as *OBJ* and 1,957 as *SUBJ*. To achieve a more balanced dataset, we have retained all the subjective sentences and randomly selected 2,957 objective ones. The *Adapted Multilingual* dataset has been used to fine-tune `mDeBERTaV3-base`. Given that we have randomly sampled the *OBJ* class, we have checked the distribution of each language and to which proportion of the original training data they correspond to. Table 2 summarizes the language distribution for the *Adapted Multilingual* training. Although, roughly speaking, for each language we obtained $\approx 72\%$ of their original training data, the distribution in the *Adapted Multilingual* mirrors the unbalanced distribution of each language as detailed in Table 1.

3. Approach

The adoption of a multilingual approach was mainly guided by the relatively small amount of training data for all languages, excluding Italian and Arabic. However, we wanted to optimize the results *per language*. This means that while concatenating the training materials will help the model to learn from multiple and more varied examples, we want to avoid the model "forgetting" the specific target language. To this end, we used the language-specific development data during the fine-tuning process.

Model The mDeBERTaV3-base model [10] is an improved multilingual version of the original DeBERTa [14]. In this model, the original Masked Language Modeling (MLM) pre-training objective is replaced with Replaced Token Detection (RTD), which is more sample-efficient. The newly introduced gradient-disentangled embedding sharing method has improved the training efficiency resulting in a better pre-trained model with respect to the original version. The model structure has a hidden size of 768, 12 layers and 12 attention heads. mDeBERTaV3-base has been trained on the 2.5T CC100 multi-lingual dataset, with 250k tokens of SentencePiece vocabulary, the same as mT5 [15], and for 500k steps. The model has obtained new state-of-the art results on the XNLI benchmark [16].

Grid Search To optimize the results on each language, we performed random grid searches, consisting of 100 iterations for each model via the software framework provided by Weights and Biases [17]. Table 3 provides a summary of the hyperparameter choices that we considered, including number of fine-tuning epochs, batch size, learning rate, warm-up steps and weight decay. In each grid search experiment we used a maximum tokenization length of 40, obtained as the maximum number of tokens from the Multilingual split. We only experimented with AdamW as optimizer as this optimizer can yield better training loss and the models generalize better in comparison to models trained with Adam [18].

Table 3
The (hyper)parameters that could be chosen during each of our random grid searches.

Parameter	Values
Epochs	2, 3, 4, 5, 6, 7, 8
Batch Size	16, 32, 64
Learning Rate	2e-5, 3e-5, 4e-5, 5e-5, 6e-5
Warmup Steps	100, 200, 300, 400, 500
Weight Decay	0.0, 0.1, 0.2, 0.3, 0.4, 0.5

Table 4
(Hyper)parameters that were using for mDeBERTaV3 per language.

language	Batch Size	Max Epochs	Best epoch	LR	Warmup Steps	Weight Decay
Arabic	16	4	2	5e-5	500	0.0
Dutch	64	6	1	4e-5	100	0.2
English	64	3	1	6e-5	200	0.0
German	16	5	2	4e-5	100	0.2
Italian	32	2	2	5e-5	300	0.0
Turkish	64	2	2	6e-5	300	0.1
Multilingual	64	8	3	3e-5	500	0.3

Table 5
Development data results for mDeBERTaV3, trained on *Adapted Multilingual* dataset and validated per language. We report macro F1, Precision and Recall, as well as the F1 for the SUBJ class.

language	F1	Precision	Recall	SUBJ F1
Arabic	0.8308	0.8937	0.7961	0.7288
Dutch	0.7033	0.7235	0.7137	0.7256
English	0.8262	0.8265	0.8260	0.8333
German	0.8342	0.8394	0.8303	0.7919
Italian	0.8068	0.7999	0.8151	0.7200
Turkish	0.9100	0.9107	0.9100	0.9118
Multilingual	0.8516	0.8523	0.8517	0.8548

Fine-tuning To fine-tune our `mDeBERTaV3-base` models, we have conducted grid searches of 25 iterations for each language using the *Adapted Multilingual* dataset. On the other hand, we run independent grid searches and fine-tuning on the orginal Multilingual dataset provided by the organizers. The final model that we chose for each language was based on the macro average F1 score that was obtained on the validation set of that language. Moreover, we mainly looked at the macro F1 scores for the epoch in which the validation loss was the lowest, while ensuring that overfitting did not occur on the training set. Table 4 shows the final (hyper)parameters that were used for each fine-tuned `mDeBERTaV3-base` model per language. In general, our models obtained the lowest validation loss with very few training epochs, with a maximum of three on the original Multiligual data. Furthermore, a smaller learning rate with a small amount of weight decay resulted in the most impactful setting. Table 5 visualizes the final scores of our models for each language on the development data. We achieved the highest macro F1 score for Turkish (0.9100) and the lowest on Dutch (0.7033). The performances for the other language are relatively similar between 0.80 and 0.83. For the Multilingual data, we obtained the second best score in absolute terms (macro F1 0.8516). On the other hand, the F1 scores for the SUBJ class present more variation, with relatively high scores for English, Turkish, and Multilingual (in line with the macro-F1), while the scores are lower for all the other languages, excluding Dutch. On close inspection, it appears that the distribution of the classes in the development set plays a major role in the behavior of the fine-tuned models. In particular, we observe that for those languages where the class distribution tends to be largely skewed towards the majority class, i.e., OBJ, systems underperform on the SUBJ class. This is the case for Arabic, German, and Italian. For all other languages, the class distribution is either perfectly balanced (e.g., Turkish and Multilingual) or slightly unbalanced (e.g., English and Dutch). As a matter of fact the delta between the macro F1 scores and the F1 score for the SUBJ class is positive, i.e., in favor of the SUBJ class. Assuming that the test data distributions will not differ largely from the development ones, we can expect a similar behaviour of the fine-tune models.

4. Results and Discussion

In Table 6, we report the overview of the results on the official test data, including the ranking. Our models obtained the top results in each language, ranking first for Dutch, German, Italian and Turkish, second for Arabic and Multilingual, and third for English. In general, at test time, we observe the same behaviour in terms of differences between the macro F1 score and the SUBJ F1 that were present in the development data. This clearly indicates that, besides the class imbalance of the training, the language specific development data play a key role in the fine-tuning process both in specializing the model for a specific language and for balancing their performances on the two classes. By observing the leaderboard, the differences across the top ranking systems on each language vary a lot, ranging between 0.1 point for Arabic, English, and the Multilingual dataset, up to 0.8 for German. Without having access to the code and the description of the other participants, we can only speculate that these differences could be due to the optimisation of the various models and the way they have been trained.

Table 6
Test results for our fine-tuned models. Scores are directly taken from the official CheckThat! Lab 2023 leaderboard and correspond to macro F1 and SUBJ F1. Rank indicates our position on the leaderboard for each language.

language	F1	SUBJ F1	Rank
Arabic	0.78	0.64	#2
Dutch	0.81	0.80	#1
English	0.77	0.79	#3
German	0.82	0.77	#1
Italian	0.76	0.65	#1
Turkish	0.90	0.91	#1
Multilingual	0.81	0.81	#2

A further remarkable result concerns Turkish. For this language we obtained a very high macro F1 score (0.90), only 0.3 points higher than the second best system. While representing roughly the same amount of training data as English, Dutch, and German (i.e., 14% of the*Adapted Multilingual*), previous work [19, 20] has highlighted how Turkish presents a battery of linguistic devices (evidentiality markers, prepositions, modality suffixes, modal adjectives and adverbs, among others) that mark in an overt manner the subjectivity status of a sentence, apparently making its identification easier when compared to other languages. Although to a lesser extent, a similar behavior (especially for connectives) can be observed for Dutch, an aspect that can help to explain the very good results on the SUBJ class in this language as well.

By plotting the predictions across multiple language specific contingency matrices, we observe that, for all languages except Dutch and Arabic, the fine-tuned models tend to over-predict the *OBJ* label, following the data distribution of the training data. On the other hand, for Turkish, the model performs really well only misclassifying 12 times *OBJ* as *SUBJ* and 12 times *SUBJ* as *OBJ*.

5. What did not work

In this section, we discuss two other approaches that we tried but did not give the expected results. We focused in modelling by using different algorithms and paradigms rather than attempting to extend the training materials.

Different Models We have experimented with fine-tuning various models, both multilingual as well as monolingual models for the Dutch language. The multilingual models that we experimented with were the cased and uncased version of `mBERT` [21]. Regarding the monolingual models, our experimentation involved `BERTje` [22] and `RobBERT` [23]. Although all of these models demonstrated reasonable performance on our validation data, we decided against them. Regarding the multilingual models, `mDeBERTaV3-base` consistently outperformed `mBERT` across all validation data. The monolingual Dutch models achieved similar performance as `mDeBERTaV3-base`. However, the main advantage of `mDeBERTaV3-base` here is the ability of training on larger quantities of data resulting in more room for improvements.

BiLSTM To explore a different architecture, we extracted the embedding representations from the last four layers of `mDeBERTaV3-base` without fine-tuning any parameters. These contextual embeddings were then concatenated and utilized as input for a BiLSTM model, on top of which we performed random grid searches. The results did not exhibit any improvements compared to fine-tuning.

6. Conclusion and Future Work

With this contribution, we focused on fine-tuning a `mDeBERTaV3-base` model that could be used with relative ease across multiple languages for the detection of subjectivity in newspaper sentences. We have shown that fine-tuning `mDeBERTaV3-base` on an adapted multilingual dataset using language specific development sets to maximise the language specific information is a powerful approach. resulting in good results. Notably, we obtained the first place in four languages and never ranked lower than third. Furthermore, our results indicates that language specific development sets data can also help to address class imbalance challenges.

Future work will explore different strategies to create the multilingual training data. In our approach, we have randomly down-sampled the OBJ class and kept all the data for the SUBJ class. We believe this approach can be revised and possibly lead to enhanced performance by working both on the training and the development distributions. For instance, an alternative could be to down-sample only languages which present a very skewed distribution between the classes (e.g., Arabic and Italian) and leave the rest as is. As a complementary step, we could develop balanced development data for each language as they play a key role during fine-tuning.

References

[1] I. Chaturvedi, E. Cambria, R. E. Welsch, F. Herrera, Distinguishing between facts and opinions for sentiment analysis: Survey and challenges, Information Fusion 44 (2018) 65–77.

URL: https://www.sciencedirect.com/science/article/pii/S1566253517303901. doi:https://doi.org/10.1016/j.inffus.2017.12.006.

[2] B. Liu, Sentiment analysis: Mining opinions, sentiments, and emotions, Cambridge university press, 2020.

[3] J. Lawrence, C. Reed, Argument mining: A survey, Computational Linguistics 45 (2020) 765–818.

[4] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, ACM SIGKDD explorations newsletter 19 (2017) 22–36.

[5] P. Atanasova, P. Nakov, L. Màrquez, A. Barrón-Cedeño, G. Karadzhov, T. Mihaylova, M. Mohtarami, J. Glass, Automatic fact-checking using context and discourse information, Journal of Data and Information Quality (JDIQ) 11 (2019) 1–27.

[6] A. Barrón-Cedeño, F. Alam, T. Caselli, G. Da San Martino, T. Elsayed, A. Galassi, F. Haouari, F. Ruggeri, J. M. Struss, R. N. Nandi, G. S. Cheema, D. Azizov, P. Nakov, The clef-2023 checkthat! lab: Checkworthiness, subjectivity, political bias, factuality, and authority, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2023, pp. 506–517.

[7] 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. Zaghouani, 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.

[8] A. E. Wilson, V. A. Parker, M. Feinberg, Polarization in the contemporary political and media landscape, Current Opinion in Behavioral Sciences 34 (2020) 223–228. URL: https://www.sciencedirect.com/science/article/pii/S2352154620301078. doi:https://doi.org/10.1016/j.cobeha.2020.07.005, political Ideologies.

[9] F. Antici, L. Bolognini, M. A. Inajetovic, B. Ivasiuk, A. Galassi, F. Ruggeri, Subjectivita: An italian corpus for subjectivity detection in newspapers, in: CLEF, volume 12880 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 40–52.

[10] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021. arXiv:2111.09543.

[11] R. Coward, Speaking personally: The rise of subjective and confessional journalism, Bloomsbury Publishing, 2013.

[12] E. Jakaza, M. Visser, ‘subjectivity’in newspaper reports on ‘controversial’and ‘emotional’debates: An appraisal and controversy analysis, Language Matters 47 (2016) 3–21.

[13] P. Chong, Valuing subjectivity in journalism: Bias, emotions, and self-interest as tools in arts reporting, Journalism 20 (2019) 427–443.

[14] P. He, X. Liu, J. Gao, W. Chen, DeBERTa: Decoding-enhanced bert with disentangled attention, arXiv preprint arXiv:2006.03654 (2020).

[15] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, arXiv preprint arXiv:2010.11934 (2020).

[16] A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, V. Stoyanov, Xnli: Evaluating cross-lingual sentence representations, 2018. arXiv:1809.05053.

[17] L. Biewald, Experiment tracking with weights and biases, 2020. URL: https://www.wandb.com/, software available from wandb.com.

[18] I. Loshchilov, F. Hutter, Fixing weight decay regularization in adam, CoRR abs/1711.05101 (2017). URL: http://arxiv.org/abs/1711.05101. arXiv:1711.05101.

[19] M. SARGIN, Reflection of subjectivity and objectivity in turkish newspaper reportage., Electronic Turkish Studies 9 (2014).

[20] D. Çokal, D. Zeyrek, T. J. Sanders, Subjectivity and objectivity in turkish causal connectives? results from a first corpus study on çünkü and için, Discourse Meaning: The View from Turkish 341 (2020) 223.

[21] 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.

[22] W. de Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. v. Noord, M. Nissim, BERTje: A Dutch BERT Model, arXiv:1912.09582, 2019. URL: http://arxiv.org/abs/1912.09582.

[23] P. Delobelle, T. Winters, B. Berendt, RobBERT: a Dutch RoBERTa-based Language Model, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 3255–3265. URL: https://www.aclweb.org/anthology/2020.findings-emnlp.292. doi:10.18653/v1/2020.findings-emnlp.292.

[24] 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.