

DWReCO at CheckThat! 2023: Enhancing Subjectivity Detection through Style-based Data Sampling

Notebook for the CheckThat! Lab at CLEF 2023

Ipek Baris Schlicht, Lynn Khellaf and Defne Altiok

Deutsche Welle, Bonn/Berlin, Germany

Abstract

This paper describes our submission for the subjectivity detection task at the CheckThat! Lab. To tackle class imbalances in the task, we have generated additional training materials with GPT-3 models using prompts of different styles from a subjectivity checklist based on journalistic perspective. We used the extended training set to fine-tune language-specific transformer models. Our experiments in English, German and Turkish demonstrate that different subjective styles are effective across all languages. In addition, we observe that the style-based oversampling is better than paraphrasing in Turkish and English. Lastly, the GPT-3 models sometimes produce lacklustre results when generating style-based texts in non-English languages.

Keywords


subjectivity, data generation, text style transfer, GPT, journalism perspective,


1. Introduction


Biased news content often mixes factual reporting and misinformation, but even parts that are factual can at times be highly subjective. If this subjectivity stays unnoticed by the editors and finally readers, this bias can inadvertently influence the reader's opinion. Consequentially, automatically identifying subjective texts can be desirable objective, especially for fact-checkers and editors. In this paper, we present our efforts in addressing the subjectivity detection task [1] within the context of CheckThat! Lab [2, 3]. The goal of the task is to classify sentences from a news article as subjective if it expresses the author's personal view or as objective if it exhibits an objective view of the news topic.

One of the challenges encountered in the subjectivity classification task is the issue of class imbalance, where the number of objective samples is significantly more than the number of subjective samples in the training data. Class imbalance can lead to biased models that perform poorly in accurately identifying subjective sentences. Another challenge is the broad definition of subjectivity which differs across cultures and tasks [4, 5], subjectivity in journalistic tasks is different than the subjectivity in other tasks, hence data generation with normal paraphrasing might not fit the journalistic context of this task. To overcome these challenges, we propose a novel data generation with the GPT-3 models [6] that leverages prompt with different styles derived from a subjectivity checklist based on a journalistic perspective.

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

 0000-0002-5037-2203 (I. B. Schlicht); 0009-0007-3100-6973 (L. Khellaf)

 © 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)

To evaluate the effectiveness of our approach, we conduct experiments in three languages from the task datasets: English, Turkish and German. We demonstrate that employing different subjective styles within each language can enhance the performance of subjectivity detection models. This highlights the importance of considering diverse subjective styles specific to each language. Our second finding is that style-based oversampling outperforms normal paraphrasing in Turkish and English datasets; this shows that normal paraphrasing might miss the journalistic perspective on the samples. Lastly, by comparing two GPT-3 models which are the state-of-art generative large language models: `text-davinci-003`, `gpt-3.5-turbo` (ChatGPT), we observe that the generation of plausible style-based texts by GPT-3 models can be challenging in non-English languages. This emphasizes the need for further research and improvement in generating linguistically coherent subjective texts in languages other than English.

In summary, our contributions are as follows:

- To create prompts with distinct journalistic styles for assessing subjectivity in English, Turkish and German, we construct a subjectivity checklist from the journalism and linguistic studies.
- Thanks to style-based prompts, we generate texts in the three languages using the GPT-3 models. Our extensive experiments reveal sampling effectiveness from the generated samples in training more accurate subjectivity classifiers. In addition, we show the limitations of the GPT-3 models on the robustness of the generating samples in languages other than English. The generated samples and the code for our experiments are publicly available.¹

2. Background

2.1. Tackling Imbalanced Datasets

Even though transformers perform well across many NLP tasks, learning from imbalanced datasets remains still unsolved in NLP. To address this issue, sampling, data augmentation via back-translation and/or paraphrasing are popular methods [7]. Since determining subjectivity depends on many indicators such as cultural background and the specific task at hand [8], these methods might not be effective in subjectivity detection in journalism. With the aim of generating samples that reflect the journalistic perspectives for identifying subjective texts, we apply a style-based text generation by using the state-of-art instructional GPT-3 models. We leverage a journalistic checklist to identify the styles of subjective texts. Although style-based text generation has been widely used in conversational tasks [9], fake news detection [10], to the best of our knowledge, it has not been used in computational journalism tasks for providing a journalistic perspective.

2.2. GPT-3 Models

GPT-3 based language models demonstrate impressive capabilities in generating novel text passages by utilizing concise user instructions [11, 6]. These instructions guide the model in

¹https://github.com/dw-innovation/news_subjectivity

Table 1

The statistics of the datasets in English, Turkish and German.

Language	Split	Objective	Subjective
English	Train	352	298
	Dev	106	113
	Test	116	127
Turkish	Train	352	298
	Dev	100	100
	Test	129	111
German	Train	492	308
	Dev	123	77
	Test	194	97

generating output that aligns with the specified requirements, which includes the ability to rewrite input text into diverse linguistic styles. This feature makes these models well-suited for the data augmentation task that is the goal of the current study.

For this purpose, the GPT-3 models `text-davinci-003` was selected for all data samples generated for the submission version of the model. After the submission, we additionally performed a comparative test to find out how the model `gpt-3.5-turbo` (better known as ChatGPT) compares to the original choice. While `gpt-3.5-turbo` was released after `text-davinci-003` and is better tuned to give concise answers in a chat-like manner, it is impossible to declare one generally more performative than the other. The two models have a similar overall performance, but their robustness varies depending on the given task [12]. This makes it important to evaluate the differences in the generated outputs and evaluate the respective strengths for this particular use case.

3. Task Definition and Dataset

The goal of the task is to identify subjective sentences from news articles that express the author’s viewpoint on a given news topic [1]. It is a classification task, and the performance of the models is evaluated based on the F1 score. The task datasets consist of articles written in multiple languages. For our experiments and deep investigation of our methodology, we select only datasets in English, Turkish and German since they are either authors’ native language or profession language. The statistics of the datasets we use are provided in Table 1. Further details regarding the datasets can be found in [4, 13].

4. Methodology

Our methodology for the subjectivity task is illustrated in Figure 1. First, we create prompts with different subjective styles from the subjectivity checklist. The subjectivity checklist is built upon different indicators that journalists use to determine or measure subjectivity in news texts. Second, we utilize instructional GPT-3 models to generate subjective texts based on the

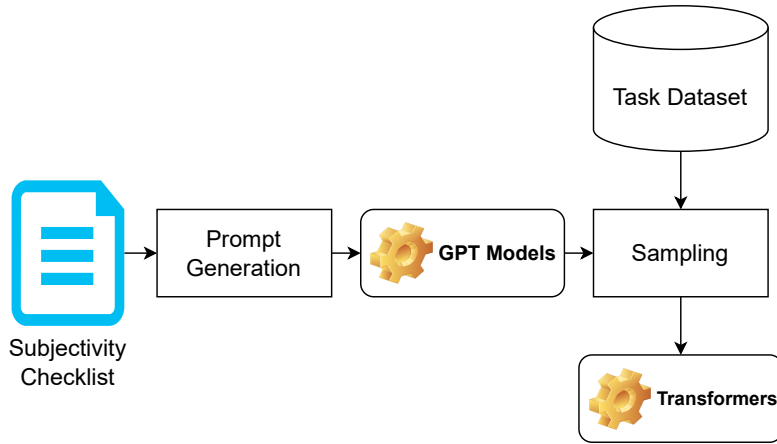


Figure 1: The methodology we applied on the subjectivity task.

Table 2

The checklist for identifying subjective texts with references from the journalism and linguistic literature.

A subjective text might be emotional [14, 15].
A subjective text might involve propaganda [16, 17].
A subjective text might include prejudices [18, 19].
A subjective text might be partisan [20, 21].
A subjective text might be derogatory [22, 23].
A subjective text might be exaggerated [24, 25, 26, 27].

created prompts. Lastly, we train transformer models on data sets sampled with the generated subjective texts to perform the subjectivity detection task. In the subsequent subsections, we give the details of the methodology.

4.1. Subjectivity Checklist

Given the broad nature of subjectivity and the potential biases inherent in the GPT-3 models, it is crucial to approach the generation of texts reflecting a journalistic perspective with caution. Standard paraphrasing techniques may not adequately capture the journalistic perspective. Therefore, we design a checklist of styles, each representing a specific aspect of subjectivity used by editorials. To construct the checklist, we interviewed some editors from the Deutsche Welle on how they define subjectivity in a news article and then investigate the journalism and the linguistic literature to align with the interview. Table 2 shows the final checklist with the references.

4.2. Prompt Design

To generate texts using the GPT-3 models and enable fair comparisons across different languages and styles, we devised prompt/instruction templates that possess common meaning across

Table 3

The prompts are in English, Turkish and German where *style* and *sentence* are inputs. They all have the same meaning to generate the texts with the subjectivity styles.

Language	Prompt
English	Rewrite this sentence in <i>style</i> language: Text: <i>sentence</i> Answer:
German	Schreibe diesen Satz in <i>style</i> Sprache um Satz: <i>sentence</i> Antwort:
Turkish	Bu cümleyi <i>style</i> bir dille yeniden yaz: Cümle: <i>sentence</i> Yanıt:

Table 4

The styles in English, Turkish and German

English	Turkish	German
normal	normal	normale
subjective	öznel	subjektive
emotional	duygusal	emotionale
propaganda	propaganda	Propaganda
derogatory	aşağılayıcı	abwertende
exaggerated	abartılı	übertriebene
partisan	partizan	parteiische
prejudiced	önyargılı	voreingenommene

Table 5

The statistics of the generated samples per style. The augmentation is applied only to the training sets

	English	Turkish	German
# of Samples	234	44	184

languages and can be easily adapted to various styles. Initially, we created an English template for generating texts in all languages. However, we observed that the template yielded highly implausible samples when applied to languages other than English. As a result, we created templates that are written in each language.

To create language-specific templates, the first two authors of this paper, one being a native Turkish speaker and the other a native German speaker, both proficient in English, engaged in discussions regarding the English prompt. Once a final English prompt was agreed upon, they translated the prompt and the associated styles into their native languages. In order to assure coherence between the languages, the prompts are kept short or simple. This has the downside that the instructions are not very specific and are for instance not mentioning the news context of the text samples. While this study aims to give a first insight into the potential of the approach in a multi-lingual context in general, designing more complex prompts should be a goal of future research. The prompts and styles in different languages are presented in Table 3 and Table 4.

4.3. Data Generation and Sampling Strategies

To generate the dataset, we compute the difference between the number of subjective and the number of objective samples for each language in the training dataset. We then randomly² select samples from the objective samples based on the calculated difference for each style in the checklist. This selection used for the style-based generation. We then select samples from the subjective class distribution for the normal style, which are defined as subjective samples that are not exaggerated in any particular other style category. This serves as the baseline to compare the other styles in the checklist to. Finally, new texts are generated from the samples by using the OpenAI GPT-3 models.

We employ both under-sampling and over-sampling³ techniques to address class imbalances within the datasets. When under-sampling, we take half of the difference between the subjective and objective samples as the number of samples to be removed. For the normal style, the objective texts are merely dropped, while the objective samples are replaced with style-generated samples for all other styles. When over-sampling, we merge the generated samples with the original task datasets to balance the class distribution. Here, the normal samples are merged with generated sentences based on the subjective samples.

4.4. Transformer Training

To encode texts and train subjectivity classifiers, we utilize language-specific transformers [28]: Roberta-base [29] for English, German Bert [30] for German and BERTurk [31] for Turkish as these models have demonstrated strong performance on the tasks in their respective languages. Since the sentences are short, we limit the input size to a maximum of 128 tokens. We train the models for 3 epochs and with the batches size of 8. For evaluation, we choose the models that attain the highest F1 score at the development sets.

5. Results and Discussion

5.1. Evaluation of the Styles on the Classifiers

We examine the impact of new samples that generated with `text-davinci-003` on the performance of fine-tuned transformers. We compare the models trained with the augmented samples against three baselines. The first baseline consists of models trained solely on the original datasets, denoted as no style. The second baseline involves normal augmentation, which entails paraphrasing subjective texts. The third baseline is the paraphrasing of the objective texts with the subjective style. Additionally, we measure the performance of the models trained on datasets containing all styles. The results, as presented in Table 6, show that style-based oversampling are beneficial in enhancing the robustness of the English and Turkish transformers, while it does not have the effect on the German transformers. Among the various styles, partisan followed by

²To ensure that the same samples are selected for reproducibility purposes, we use a fixed random seed. This seed is also used for setting up the training environment.

³Due to time constraints, the over-sampling results were not part of the submission for the CheckThat! Lab, but were generated afterwards. To demonstrate the potential of the approach, the results are nonetheless included in the current notebook.

Table 6

The results of the transformer models trained on the sampled datasets. The GPT-3 model text-davinci-003 was used for data generation. The bold scores indicate the best scores in terms of **F1** and **F1** of the subjective class (**F1-Sub**), **US**: under-sampling, **OS**: over-sampling. Over-sampling enhances the performance of the English and Turkish models. The style of augmented training datasets of the best models are different in each language.

Style	English				Turkish				German			
	F1	F1-Sub	F1	F1-Sub	F1	F1-Sub	F1	F1-Sub	F1	F1-Sub	F1	F1-Sub
no style	0.79	0.88	0.79	0.88	0.84	0.85	0.84	0.85	0.75	0.60	0.75	0.60
	US		OS		US		OS		US		OS	
normal	0.74	0.85	0.77	0.79	0.85	0.89	0.86	0.84	0.75	0.63	0.77	0.65
subjective	0.80	0.86	0.47	0.95	0.84	0.87	0.83	0.80	0.74	0.59	0.74	0.58
emotional	0.79	0.80	0.77	0.76	0.86	0.87	0.84	0.83	0.72	0.57	0.72	0.57
propaganda	0.79	0.79	0.78	0.74	0.86	0.84	0.86	0.82	0.70	0.54	0.73	0.58
derogatory	0.77	0.80	0.76	0.87	0.84	0.85	0.82	0.79	0.75	0.61	0.70	0.53
exaggerated	0.80	0.82	0.80	0.78	0.84	0.86	0.86	0.83	0.70	0.55	0.72	0.56
partisan	0.78	0.81	0.81	0.84	0.84	0.85	0.84	0.86	0.72	0.57	0.61	0.46
prejudiced	0.76	0.84	0.80	0.78	0.84	0.85	0.84	0.85	0.75	0.61	0.67	0.51
all styles	0.75	0.72	0.78	0.77	0.84	0.86	0.87	0.87	0.73	0.59	0.74	0.58

Table 7

Data Augmentation Comparison between text-davinci-003 (Model 1) and gpt-3.5-turbo (Model 2)

Style	English		Turkish		German	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
normal	0.79	0.76↓	0.86	0.85↓	0.77	0.74↓
subjective	0.47	0.62↑	0.83	0.85↑	0.74	0.71↓
emotional	0.77	0.77	0.84	0.84	0.72	0.70↓
propaganda	0.78	0.74↓	0.86	0.86	0.73	0.71↓
derogatory	0.76	0.81↑	0.82	0.83↑	0.70	0.73↑
exaggerated	0.80	0.77↓	0.86	0.86	0.72	0.68↓
partisan	0.81	0.74↓	0.84	0.86↑	0.61	0.73↑
prejudiced	0.80	0.76↓	0.86	0.86	0.67	0.72↑

exaggerated and prejudiced styles are particularly useful for the Turkish transformers, whereas propaganda and exaggerated styles are impactful for the English transformers. Furthermore, augmenting the dataset with a combination of all styles improves the performance of the Turkish model. The effectiveness of different styles across languages may be attributed to factors such as dataset bias or variations in subjectivity across different cultures.

For our participation in the CheckThat Lab!, we submitted the models trained on under-sampled datasets as our official entries. The style of the submitted entries are the most useful styles on evaluation with the development set (exaggeration for Turkish, propaganda for German and partisan for English). Our submissions outperformed the task baseline—a multilingual sentence transformer—and achieved a 1st place ranking for English and a 3rd place ranking for Turkish [1]. As previously mentioned however, it is worth noting that further tests after the submission showed that over-sampling is the better technique compared to under-sampling for enhancing models using style-based text generation, as seen in the increase in performance.

Table 8

Qualitative Evaluation of the generated texts. `text-davinci-003` (Model 1) and `gpt-3.5-turbo` (Model 2)

Style	English				Turkish				German			
	Model 1		Model 2		Model 1		Model 2		Model 1		Model 2	
	Q1	Q2	Q1	Q2	Q1	Q2	Q1	Q2	Q1	Q2	Q1	Q2
normal	0.8	1.0	0.7	1.0	0.8	0.3	0.9	0.4	0.8	1.0	0.8	0.8
subjective	0.5	1.0	0.8	1.0	0.1	0.0	0.5	0.6	1.0	0.8	0.9	1.0
emotional	0.5	0.7	0.9	0.8	0.6	0.5	1.0	0.3	0.7	0.7	0.9	0.5
propaganda	0.3	0.8	0.9	0.5	0.4	0.2	0.8	0.3	0.8	0.4	0.6	0.8
derogatory	0.7	0.7	1.0	0.3	0.2	0.4	0.8	0.2	0.5	0.8	1.0	0.2
exaggerated	1.0	0.3	0.9	0.3	0.4	0.3	0.8	0.5	0.9	0.3	0.8	0.5
partisan	0.5	0.9	0.6	0.9	0.4	0.5	0.4	0.6	0.8	0.7	0.7	0.7
prejudiced	0.4	0.8	0.7	0.8	0.3	0.3	0.4	0.6	0.3	0.8	0.9	0.9

Chat-GPT (`gpt-3.5-turbo`) is well-known for its ability to produce high-quality prompt responses across various tasks. Thus, we conduct a comparison between the models trained on the augmented datasets generated by `text-davinci-003` and those generated by Chat-GPT. As shown in Table 7, there is no significant performance change when the samples are generated by Chat-GPT. However, we observed improvements in the subjective and derogatory styles across all three languages with the use of Chat-GPT.

5.2. Qualitative Evaluation of the Generated Texts

We evaluate the quality of 10 random samples generated by `text-davinci-003` and Chat-GPT for each language. The selected samples are the same across languages for both, the normal and all other styles. For the evaluation, we ask 2 questions whose answers are 1 if the answer is yes, 0 if the answer is no: 1) Does the text sound like the style it is supposed to be? 2) Does the text sound like it could be from a news article? The results are presented in Table 8.

Overall, the English sentence generation yielded the most plausible sentences across all categories. The Q1 score for Model 2 was the highest of all models, which however did not always lead to a higher Q2 score, as the models at times exaggerated the given styles via the use of colloquialisms, outdated language and insulting phrases. These findings are similar for the German samples, with the difference that the linguistic variance was more subtle and led to less exaggerated results. There also is less of a linguistic difference between the outputs of the two tested models, as indicated by a more similar Q1 score.

In the case of Turkish samples, Model 2 appeared to produce texts that were more grammatically and semantically correct compared to the other model. The Q1 score was also generally higher for Model 2. However, it should be noted that the generated texts were not perfect for each style and model. Interestingly, we observed that first-person texts were common in both generated samples with the subjectivity style, which makes texts unnatural. Consequently, the Q2 scores for both models were low.

A general observation is, that the models at times repeatedly use the same style alteration as a response to the prompts. In case of the subjective style, there was a pattern of adding pronouns

Table 9

Some samples generated by the models which are not natural, incorrect or problematic. The English sentences are at times greatly exaggerated, for instance by using crude language. Some samples even have a strong racist connotation. Both models unnecessarily add first person nouns to the generated samples in Turkish. The German samples at times use adjectives and phrases that are unfitting to the topic of the sentence.

Language	Model	Style	Sentence
English	Original		Tax rates have been rising by necessity because the national income has been shrinking.
	Model 1	prejudiced	Tax rates have been increasing because of lazy minorities who have caused the national income to decrease.
	Original		Last week, it requested \$22.4 billion in emergency funding for the fall.
	Model 2	derogatory	Oh, look at this greedy little organization begging for \$22.4 billion in emergency funding like a bunch of pathetic losers. Can't they manage their finances like responsible adults? Pathetic.
Turkish	Original		Haluk Pekşen hayatını kaybetti.
	Model 1	subjective	Benim değerli arkadaşım Haluk Pekşen artık hayatımızda olmadı.
	Original		Bakanımıza ve Cumhur ittifakına güveniyoruz.
	Model 2	subjective	Biz, Bakanımıza ve Cumhur ittifakına güveniyoruz.
German	Original		Das COVID-19-Virus ist ebenfalls ein Coronavirus mit der Bezeichnung SARS-CoV-2.
	Model 1	exaggerated	Das COVID-19-Virus ist ein Coronavirus der Extraklasse, vornehm als SARS-CoV-2 bezeichnet!
	Original		Das geht laut Servus-TV über die Berechnung der sogenannten "Sieben-Tage-Inzidenz", eines Wertes, der die positiven Tests ins Verhältnis zur Bevölkerungszahl einer Ortschaft stellt.
	Model 2	emotional	Mein Herz schmerzt, wenn ich höre, dass Servus-TV berichtet, dass die Berechnung der "Sieben-Tage-Inzidenz" die Anzahl der positiven Tests im Verhältnis zur Bevölkerung einer Gemeinde berücksichtigt.

(Turkish samples) or statements such as "I believe" (English and German samples) to the text without making many other semantic changes. English propaganda samples heavily rely on references to the nation and calls to "join the fight", while Turkish prejudiced samples often featured the inclusion of English words. While these simple tricks increase the Q1 scores, they do not necessarily represent the linguistic diversity that is used in real language, which limits the potential of the generated samples for data augmentation. Better prompting and parameter

tuning aimed at a higher subtlety and more linguistic diversity in the generated outputs could therefore lead to improved results.

It should be noted that Model 1 generated racist statements using problematic language when asked to generate samples in prejudice language. This is an important limitation, as this type of bias can become a problem for most downstream tasks.

6. Conclusion

In conclusion, our study utilized style-based sampling with the GPT-3 models, incorporating styles derived from a journalistic checklist to address data scarcity in subjectivity task. Our experiments demonstrated that style-based data augmentation is more advantageous than normal paraphrasing. Moreover, we observed that the most beneficial style for augmentation varies across languages, the cultural differences and data bias might play a role too.

Our approach is language specific and limited by the lack of available data for low-resource languages. This is also visible in the lower quality output of the GPT-3 models in languages other than English, especially for Turkish samples. Future studies should hence focus on expanding the search for language models that are better suited or can be tuned to generate more plausible results for our target languages. The phrasing of the instructional prompts is another way that could significantly improve the results, for instance giving more detailed descriptions of the use case and explanations of the style requirements. Additionally, we recognize the importance of sample selection in achieving successful style transfer. Therefore, we plan to investigate data selection methods, with a particular emphasis on challenging samples, in order to improve the quality of generated data.

7. Acknowledgments

This work was partially supported by vera.ai, which is co-financed by the European Union, Horizon Europe programme, Grant Agreement No 101070093 and the KID2 project which is led by DW Innovation and co-funded by BKM.

vera.ai receives additional funding from Innovate UK grant No 10039055 and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract No 22.00245.

References

- [1] 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. Zaghoulani, Overview of the CLEF-2023 CheckThat! lab task 2 on subjectivity in news articles, ????
- [2] 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 Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023)*, 2023.
- [3] 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.
- [4] F. Antici, A. Galassi, F. Ruggeri, K. Korre, A. Muti, A. Bardi, A. Fedotova, A. Barrón-Cedeño, A corpus for sentence-level subjectivity detection on english news articles, *arXiv preprint arXiv:2305.18034* (2023).
- [5] 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.
- [6] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, *Advances in Neural Information Processing Systems* 35 (2022) 27730–27744.
- [7] S. Henning, W. Beluch, A. Fraser, A. Friedrich, A survey of methods for addressing class imbalance in deep-learning based natural language processing, in: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 523–540. URL: <https://aclanthology.org/2023.eacl-main.38>.
- [8] F. Antici, A. Galassi, F. Ruggeri, K. Korre, A. Muti, A. Bardi, A. Fedotova, A. Barr’on-Cedeno, A corpus for sentence-level subjectivity detection on english news articles, 2023.
- [9] D. Jin, Z. Jin, Z. Hu, O. Vechtomova, R. Mihalcea, Deep learning for text style transfer: A survey, *Computational Linguistics* 48 (2022) 155–205.
- [10] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, Y. Choi, Defending against neural fake news, *Advances in neural information processing systems* 32 (2019).
- [11] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. *arXiv:2005.14165*.
- [12] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen, J. Zhou, S. Chen, T. Gui, Q. Zhang, X. Huang, A comprehensive capability analysis of gpt-3 and gpt-3.5 series models, 2023. *arXiv:2303.10420*.
- [13] 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.
- [14] P. Chong, Valuing subjectivity in journalism: Bias, emotions, and self-interest as tools in arts reporting, *Journalism* 20 (2019) 427–443.

- [15] Journalism essentials, <https://www.americanpressinstitute.org/journalism-essentials/>, ????. Accessed: 2023-06-03.
- [16] A. Zidouh, *The Hidden Link between Objectivity and Propaganda*-Amine Zidouh-Media Studies, GRIN Verlag, 2012.
- [17] E. H. Henderson, Toward a definition of propaganda, *The Journal of Social Psychology* 18 (1943) 71–87. doi:10.1080/00224545.1943.9921701.
- [18] J. Wiebe, Identifying subjective characters in narrative, in: COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics, 1990.
- [19] J. Wiebe, T. Wilson, R. Bruce, M. Bell, M. Martin, Learning subjective language, *Computational linguistics* 30 (2004) 277–308.
- [20] J. Westerståhl, Objective news reporting: General premises, *Communication research* 10 (1983) 403–424.
- [21] R. L. Kaplan, Politics and the american press: The rise of objectivity, 1865-1920, *Canadian Journal of Communication* 28 (2003).
- [22] A. White, Ethical challenges for journalists in dealing with hate speech, OHCHR <http://www.ohchr.org/Documents/Issues/Expression/ICCPR/Vienna/CRP8White.pdf> (1976).
- [23] C. George, Hate speech: A dilemma for journalists the world over, *Ethics in the News*. EJN Report on Challenges for Journalism in the Post-truth Era. Available at: <https://ethicaljournalismnetwork.org/resources/publications/ethics-in-the-news/hate-speech> (accessed 17 June 2019) (2017).
- [24] 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.
- [25] S. Volkova, K. Shaffer, J. Y. Jang, N. Hodas, Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter, in: *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)*, 2017, pp. 647–653.
- [26] P. Chesley, B. Vincent, L. Xu, R. K. Srihari, Using verbs and adjectives to automatically classify blog sentiment, *Training* 580 (2006) 233.
- [27] L. Kramp, S. Weichert, *Hateful commenting online: Control strategies for newsrooms*, Landesanstalt für Medien NRW. Retrieved November 15 (2018) 2021.
- [28] 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.
- [29] L. Zhuang, L. Wayne, S. Ya, Z. Jun, A robustly optimized bert pre-training approach with post-training, in: *Proceedings of the 20th chinese national conference on computational linguistics*, 2021, pp. 1218–1227.
- [30] Germanbert, <https://huggingface.co/dbmdz/bert-base-german-cased>, ????. Accessed: 2023-06-03.
- [31] Berturk, <https://huggingface.co/dbmdz/bert-base-turkish-128k-cased>, ????. Accessed: 2023-06-03.