

ThatiAR: Subjectivity Detection in Arabic News Sentences

Reem Suwaileh^{1*}, Maram Hasanain², Fatema Hubail³, Wajdi Zaghouani⁴, Firoj Alam^{2*}

¹Hamad bin Khalifa University, Qatar,

²Qatar Computing Research Institute, Qatar

³Free University of Berlin, Germany,

⁴Northwestern University in Qatar, Qatar

{rsuwaileh, fialam, mhasanain}@hbku.edu.qa, fatemahubail93@gmail.com, wajdi.zaghouani@northwestern.edu

Abstract

In this study, we present the first large dataset, *ThatiAR*, for subjectivity detection in Arabic, consisting of ~3.6K manually annotated sentences, and GPT-4o based explanations. In addition, we include instructions (both in English and Arabic) to facilitate LLM based fine-tuning. We provide an in-depth analysis of the dataset, annotation process, and extensive benchmark results, including PLMs and LLMs. Our analysis of the annotation process highlights that annotators were strongly influenced by their political, cultural, and religious backgrounds, especially at the beginning of the annotation process. The experimental results suggest that LLMs with in-context learning provide better performance. We release the dataset and resources to the community.

Code — <http://llmebench.qcri.org>

Datasets — <https://huggingface.co/datasets/QCRI/ThatiAR>

Extended version — <https://arxiv.org/pdf/2406.05559>

1 Introduction

Detecting subjectivity¹ in news sentences is crucial for several reasons (Mohammad 2016; Antici et al. 2024a). It helps identifying media bias by distinguishing between objective reporting and subjective content, thereby enhancing the credibility of news sources (Stru et al. 2024). This differentiation is also vital in combating misinformation and fake news by flagging opinion-based content for further verification (Zhou and Zafarani 2020). In Figure 1, we present an example of a subjective sentence that can be misleading and cause fear among citizens. The highlighted part of the text in the example is subjective. It is important to note that subjectivity alone does not imply misleading intent; subjective expressions can serve legitimate communicative or emotional purposes in journalistic writing.

With the reliance on social media as platforms of expression, users often resort to informality, dialects, and a combi-

*These authors contributed equally.

Copyright   2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹Subjectivity, according to Abo et al. (2019), “refers to aspects of language used to express feelings, opinions, evaluations, and speculations and, as such, it incorporates sentiment”.

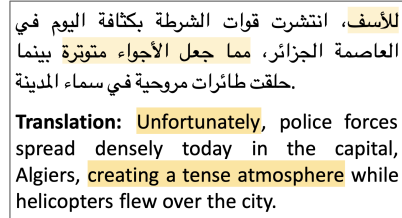


Figure 1: An example of a subjective sentence that can be misleading and cause fear.

nation of languages. When seeking news reports and statements, readers turn to news outlets or social media platforms for knowledge and assessments of current events (Shearer and Mitchell 2021). While readers may consider news from reliable outlets as objective sources of information, research shows that news reports are often partisan, subjective, and reflective of the news agency’s standpoint (Shokri et al. 2024). Therefore, analyzing subjectivity provides insights into public sentiment and the social impact of news. It can empower readers to make informed decisions and encourage critical thinking by highlighting subjective reports.

While there has been research efforts to develop methods and systems to automatically identify such content, the majority of studies focus on English or other high-resourced languages. However, the field is growing to incorporate low-resourced languages, including Arabic, Bulgarian, Turkish and Urdu (Abdul-Mageed 2015; Galassi et al. 2023; Stru et al. 2024).

Research on subjectivity in Arabic content (Abdul-Mageed and Diab 2012; Mourad and Darwish 2013; Abdul-Mageed and Diab 2014) addresses the complexities of language usage, primarily focusing on Modern Standard Arabic (MSA) and regional dialects, however, the resulting resources are not publicly available. Therefore, in this study, we focused on Arabic, with a special emphasis on news content. Given the lack of resources for developing AI-based systems in Arabic, we introduce *ThatiAR*,² a reasonably large and well-balanced data set consisting of manually annotated news sentences. While annotating *ThatiAR*, we ad-

²Translated in Arabic as ذاتي (“Thati”) meaning “subjective” in English.

dressed three research questions: (*Q1*) What are the emerging characteristics of news reports with regards to subjectivity? (*Q2*) How do annotators of diverse backgrounds approach news reports? and, (*Q3*) Should current subjectivity annotation strategy and guidelines be further developed to account for socially complex, and cultural-specific content?

We conducted extensive experiments to create a benchmark using different Pre-trained Language Models (PLMs) and Large Language Models (LLMs) that can serve as a foundation for future research. Given that current LLMs consistently push the boundaries of NLP and achieve state-of-the-art performance in tasks such as machine translation, summarization, sentiment analysis, and more complex applications like legal document analysis and creative writing (Bang et al. 2023; Ahuja et al. 2023; Hendy et al. 2023; Khondaker et al. 2023; Abdelali et al. 2024), therefore, we used GPT-4o to generate explanations for why a sentence is labeled as subjective or objective. Additionally, we developed instructions for each data point, resulting in a comprehensive instruction-following dataset. We are expanding upon an existing body of research on subjectivity with new, recent data and benchmarks. The points contribute to the distinctiveness of our research:

- We developed *ThatiAR*, a dataset consisting of $\sim 3.6K$ manually annotated news sentences. This is largest dataset compared to any other subjectivity dataset released so far. Three times larger than CheckThat! 2023 (Stru et al. 2024).
- *ThatiAR* is benchmarked for the subjectivity task using a diverse set of SOTA NLP models, including PLMs and LLMs.
- *ThatiAR* is the first Arabic dataset to include instructions and explanations for annotations, addressing the current demands of advancements in the NLP field, particularly with LLMs.

2 Related Work

Subjectivity detection has been studied across multiple contexts, with research primarily focused on sentiment analysis in English (Refaee and Rieser 2014; Savinova and Moscoso Del Prado 2023). The focus was broadened to include multilingual approaches (Mihalcea, Banea, and Wiebe 2007; Banea et al. 2008), and other contexts such as bias detection (Aleksandrova, Lareau, and M enard 2019; Hube and Fetahu 2019) and fact-checking (Vieira et al. 2020).

The subjectivity detection datasets were constructed using different syntactic and semantic techniques such as domain-specific heuristics (Das and Sagnika 2020)), statistical models (Pang and Lee 2004), and manual annotation (Abdul-Mageed and Diab 2011). Syntactic approaches are limited as they depend on domain- and language-specific knowledge, while semantic methods, especially those grounded on annotation guidelines, have recently gained favor (Chaturvedi et al. 2018). However, challenges such as interpretative bias, annotation ambiguity, and edge cases made the dataset creation more complex (Geva, Goldberg, and Berant 2019). To address these issues, a prescriptive data annotation methodology R ottger et al. (2022), particularly in the context of fact

verification (Antici et al. 2024b), is a promising direction. This methodology has been used in the Subjectivity Task in the CheckThat! Lab in both 2023 (Galassi et al. 2023) and 2024 (Stru et al. 2024) editions.

The datasets contain various levels of granularity, including sentence (Rustamov, Mustafayev, and Clements 2013), segment (Benamara et al. 2011), and document-level (Antici et al. 2021a) analysis. While the majority of available corpora are in English, efforts have been made to extend subjectivity detection to other languages, including Arabic (Banea, Mihalcea, and Wiebe 2010; Abdul-Mageed, Diab, and Korayem 2011; Abdul-Mageed, Diab, and K ubler 2014), German (Banea, Mihalcea, and Wiebe 2010), French (Benamara et al. 2011; Banea, Mihalcea, and Wiebe 2010), Italian (Antici et al. 2021a), Romanian (Banea, Mihalcea, and Wiebe 2010), and Spanish (Banea, Mihalcea, and Wiebe 2010). Moreover, subjectivity datasets created in the 2023 and 2024 editions of the CLEF CheckThat! lab cover seven languages including Arabic, Bulgarian, Dutch, English, German, Italian, and Turkish. However, most of the multilingual efforts often rely on machine translation and ontological methods, which can introduce noise and affect the quality of annotations. Additionally, the Arabic resources are either inaccessible to researchers or limited in size.

As for systems, recent efforts include deep learning techniques, such as RNNs and LSTMs, were also developed and improved the performance due to their ability to capture the contextual nuances within textual data (Irsoy and Cardie 2014). The PLMs with their transferred learning abilities made notable improvement especially when fine-tuned. Systems participated in *Subjectivity Task* in CheckThat! 2024 were developed using transformer-based models such as BERT-base-uncased and XLM-RoBERTa-base being the most frequent choice in both monolingual and multilingual setups (Stru et al. 2024). The multilingual setup was developed with automatic translation system.³ Nevertheless, there have been attempts to use LLMs such as Google’s Gemini⁴, LLaMA-3-8b, and Mistral-7B-Instruct-v0.2, classical neural networks such as CNN, LSTM, and BiLSTM, classical machine learning techniques such as SVMs and random forests, ensemble classifiers, and data augmentation (Stru et al. 2024). Focusing on the Arabic language, none of the participating teams achieved a macro-F1 score above 0.50, demonstrating its challenges.

3 Dataset

In this section, we discuss the effort carried out to construct *ThatiAR* dataset. In Figure 2, we provide a complete workflow of our data collection (Section 3.1), manual annotation (Section 3.2) and analysis (Section 3.3).

3.1 Data Collection

To prepare a set of sentences for subjectivity annotation, we went through three phases discussed in the below sections.

³<https://www.deepl.com/en/translator>

⁴<https://gemini.google.com>

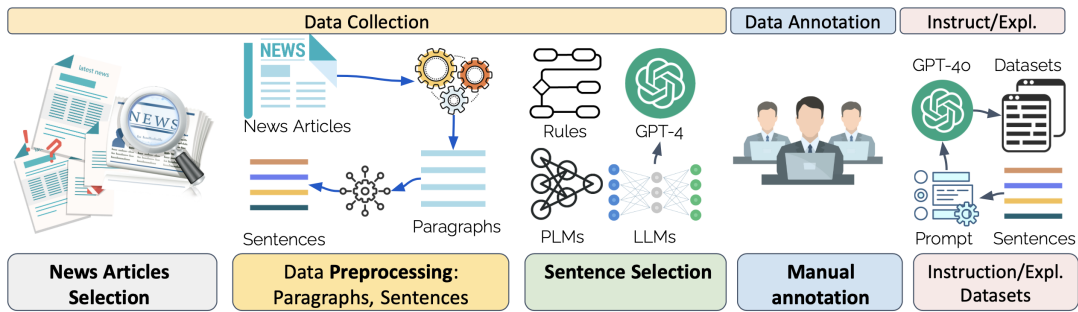


Figure 2: The pipeline of the data collection, annotation, and instruction/explanation datasets development process.

News Article Selection We selected the AraFacts dataset (Sheikh Ali et al. 2021), which contains claims verified by Arabic fact-checking websites. Each claim is associated with web pages that either propagate or refute the claim. In total, we collected 1,159 new articles from AraFacts. To address the issue of skewed distribution of sentence types in news articles, which tend to favor objective sentences, a graduate student manually searched for opinionated articles published by various Arabic news outlets (e.g., Sky News Arabia, Alarabiya). This effort resulted in selecting 221 new articles. Our pool of articles includes content from over 500 news outlets, covering a wide range of categories such as politics, social issues, arts and culture, and health, among others.

Preprocessing We parsed the web pages using three different scrapers, favoring the longest output. The tools used were Goose3,⁵ Newspaper3k,⁶ and Trafilatura.⁷ After extracting the text content, we segmented the body of text into paragraphs and sentences, resulting in a total of 15,947 sentences. The parsing and segmentation involves rule based approaches to filter and remove noisy html tags.

Sentence Selection We applied two sampling strategies. The first strategy was rule-based, considering only sentences with a length between 10 and 45 words to select not overly long self-contained sentences. The second strategy utilized four transformer-based models and GPT-4 to label the subjectivity of the sentences. The goal of this strategy was to select sentences with at least one “subjective” vote, thereby oversampling potentially subjective sentences for annotation. For this purpose, we trained five models: ARABERTv2 (Antoun, Baly, and Hajj 2020), ARBERTv2 (Abdul-Mageed, Elmadany, and Nagoudi 2021), MARBERTv2 (Abdul-Mageed, Elmadany, and Nagoudi 2021), GIGABERTv4 (Lan et al. 2020), and GPT-4 (Achiam et al. 2023). These models were fine-tuned on the entire Arabic subjectivity datasets from the CheckThat! 2023 lab (Galassi et al. 2023). As a result of length-based filtering and model-based labeling, we selected 4,524 sentences for manual annotation.

⁵<https://goose3.readthedocs.io>

⁶<https://newspaper.readthedocs.io/>

⁷<https://trafilatura.readthedocs.io>

3.2 Data Annotation

To annotate *ThatiAR*, we opt to employ human annotators on Amazon Mechanical Turk (mTurk) crowdsourcing platform. Given that we used the mTurk platform, the demographic information of the annotators is not known to us. We adopted the annotation guidelines from a previous study which have been widely used for different languages (Antici et al. 2021b) and tailored them for Arabic language with examples. We discuss the annotation guidelines with additional examples in Appendix C (see Supplemental material). In a nutshell, we define *subjective* sentences as expressions of the writer’s feelings, literary tastes, or personal interpretations of topics and events. Sentences containing sarcasm, support, or offensive language are also considered subjective. In contrast, *objective* sentences present facts, events, and topics supported by verifiable data. They may also include widely recognized expressions or phrases that were not originally authored by the writer.

To ensure the clarity and coherence of the guidelines, and the mTurk annotation configuration, we ran multiple pilot studies that consumed around 850 sentences from our pool.

To ensure the quality of annotations, we sampled a set of 115 annotated sentences from CheckThat! 2023 lab (Galassi et al. 2023). We use these sentences in two ways: (i) 10 questions for pre-qualification test that an annotator has to pass before being eligible to start the actual HITs, and (ii) 105 questions for ongoing-qualification that an annotators has to maintain an acceptable accuracy throughout the annotation process. For both we requested the worker accuracy above 60%.

We finally set up the design of the annotation interface and configurations as follows. We ran 245 HITs, each containing no more than 15 sentences and 5 quiz questions. We initially collected 3 annotations per sentence and dynamically requested up to 2 more annotations when the majority agreement of 66.6% was not met, to guarantee the reliability of annotations. We compensated annotators \$0.60 per HIT, costing around \$550 for the entire dataset. As a result, we obtained 3,661 sentences with 66.6% agreement, of which 1,579 were subjective and 2,082 were objective sentences. The sentences that did not pass the agreement score were removed from the final dataset. Table 1 shows the statistics of *ThatiAR*. In Table 2, we present a few annotated sentences from the *ThatiAR* dataset along with their English transla-

tions.

| Set | SUBJ | OBJ | All |
|------------|---------------|---------------|--------------|
| Train | 1,055 (66.8%) | 1,391 (66.8%) | 2,446 |
| Dev | 201 (12.7%) | 266 (12.8%) | 467 |
| Test | 323 (20.5%) | 425 (20.4%) | 748 |
| All | 1,579 | 2,082 | 3,661 |

Table 1: Statistics of *ThatiAR* dataset

3.3 Data Analysis

Annotation Agreement. To evaluate the reliability of human annotations, we computed the Inter-Annotator Agreement (IAA) using an agreement coefficient⁸ that averages the observed agreement across all annotators and sentences. We found the agreement to be approximately 0.54, indicating a reasonable level of agreement for the subjectivity annotation.

We further computed the Cohen’s Kappa (κ) coefficient between each of the first three annotators and the consolidated label (determined by majority voting) (Alam et al. 2021). The κ value of 0.54 indicates moderate agreement (See the Table 9, in Appendix C.3 as a part of supplemental material).

Deep Analysis. Prior to posting the data on the mTurk for annotation, we conducted pilot studies to refine our annotation guidelines. We then discussed with the annotators to deeply analyze their annotations and behavior. We focused on the key aspects that impact the understanding of Arabic news reporting and the quality of annotations. In the following, we discuss our analysis by discussing the examples reported Table 6 (in Appendix).

Bias in reporting and annotating. News reports often contain phrases and terms that can be interpreted in multiple ways. For example, in the Sentence #1 the phrase “الإقليم المضطرب” (“volatile region”) is a preliminary site of disagreement. The region may be described as volatile because it merits the description of Oxford dictionary definition: “liable to change rapidly and unpredictably, especially for the worse.” However, the perception of volatility could also be influenced by partisan news reporting that portrays China as oppressive and democracy as liberating. This raises the question of whether the term “volatile” is accurate or if it carries political, historical, or cultural biases of the journalists and news agencies. On the other hand, annotators with similar potential biases are likely to consider this news sentence objective, while those with differing biases may view it as subjective.

Subjectivity in reporting and annotating. To understand the sources of disagreement between annotators, we examined several instances that exhibits some aspects contributing to their subjectivity. For instance, sentence #2 ref-

⁸https://www.nltk.org/_modules/nltk/metrics/agreement.html#AnnotationTask.avg_Ao

erences “الاحتلال” (“the occupation”), which readers commonly understand to mean “Israel.” This term is politically loaded and functions as a critique of the Israeli occupation, placing blame on Israel as an occupying power and alluding to other historical occupations. Additionally, the phrase اختراق الحدود (“border breaching”) followed by “الأراضي الفلسطينية” (“Palestinian lands”) also carries significant political weight. If this report were from a news agency that supports Israeli claims to nationhood, it might not use “Palestinian lands” or refer to Israel as “the occupation.” The term “border breaching” implies unlawful activity, indicating subjectivity in the portrayal of events. The subjectivity in this sentence may not intended as a negative or politically motivated claim but rather emerges from religious and cultural contexts that are more easily understood by regional annotators.

Composite reporting. Multiple news sentences often report different matters within the same text segment. For example, sentence #3 combines three distinct headlines into one statement, each containing both subjective and objective descriptions. This discrepancy can lead to disagreement among annotators, as each annotator may focus on different parts of the sentence or interpret the main focus differently.

Perspectives of annotators. The perspective of annotators and their standpoint is a crucial element in their judgments. For instance, an annotator, who identifies herself as a feminist, had accounted for terms in sentence #4 “قام بالتآمر” (“conspired”) and “وهو عالم” (“knowing [well]”), and the usage of the term “فإن” (“if”) and “فسيتم التخلص” (“get rid of”), therefore judge it to stand collectively as a subjective sentence. The term “conspire” implies criminal or unlawful activities; knowing, a subjective term, espouses knowledge as more prevalent in one person than another; and “get rid” implies the parents, although violent, are disposable further dehumanizing them through the lens of criminal activity. Had this statement appeared in English, the statement would appear immediately subjective. However, in the construction of the sentence in Arabic, the initial clause functions as a factual statement, further justified by the following clause, and that subjectivity may only be interpreted as appearing in the last line with the term “rid”. To this end, we answer our three questions in light of our examination and analysis.

To address *Q1*, we confirm that news statements either reflect an accurate description of the entities and events being reported (objective view) or convey the reporter’s personal judgments and predictions about the impact of the news (subjective view). The subjective view is typically driven by political, historical, and cultural biases and subjectivities of the reporter or the news agency. For future recommendations, we suggest hiring annotators who are knowledgeable about the different types of subjectivity that can influence news reporting, to ensure unbiased annotations.

To address *Q2*, we affirm that annotators’ political, historical, and cultural backgrounds significantly influence their understanding of the news articles and consequently their judgments. We recommend giving the annotators the option

| # | Label | Sentence | Translation |
|---|-------|---|--|
| 1 | SUBJ | وجدت بوحيرد نفسها وهي فتاة تبلغ من العمر ٢٢ عاما - بين يدي ضباط المستعمر الفرنسي ينهش لحمها بكل الطرق. | Bouhired found herself - a 22-year-old girl - in the hands of French colonial officers, a prey whose flesh was being devoured in every way. |
| 2 | SUBJ | ولكنني لم أجد الوقت الكافي للتعرف عليه عن كثب ولكن عندما مررت بأوقات عصيبة، أعطتني العقيدة الإسلامية القوة اللازمة لمواجهة. | But I did not find the time to get to know Islam closely, but when I went through difficult times, the Islamic faith gave me the strength necessary to face COVID. |
| 3 | OBJ | كما تدخل نترات الأمونيوم في صناعة المتفجرات خاصة في مجال التعدين والناجم. | Ammonium nitrate is also used in the manufacture of explosives, especially in the field of mining. |
| 4 | OBJ | اشتية: السعوديون أعادوا القضية الفلسطينية للطاولة عند الحديث عن التطبيع مع إسرائيل | Shtayyeh: The Saudis put the Palestinian issue back on the table when talking about normalization with Israel |

Table 2: Example sentences from *ThatiAR* dataset.

to abstain when they cannot judge sentences. This can be compiled in the annotation tool design by adding the label “Others” with the ability to provide justification, forming open-ended annotations that would be more valuable for analysis and validation.

To address **Q3**, we highlight four points:

- *Semantic curation for data*: We endorse the importance of carefully preparing data for annotation for ensuring accurate results. The processing pipeline, including the segmenter, must consider both syntactic and morphological aspects of the sentences. Furthermore, focusing on annotating self-contained and concise sentences will enhance the overall quality of the annotations.
- *Abstention with open-ended annotations*: We recommend allowing annotators to abstain when they cannot judge sentences. This is mainly because not all sentences must be subjective or objective, some are neutral or ambiguous. This can be implemented in the annotation tool by adding an “Others” label with the option to provide justification. Additionally, requesting the rationale behind annotations would enhance their value. Such open-ended annotations would be more valuable for analysis and validation.
- *Domain-specific training for annotation*: We emphasize the importance of specifying the data source in the annotation guidelines. For news reporting, annotators should be trained to distinguish between factual statements and text influenced by biases, as this fine distinction separates objective from subjective sentences.
- *Validation phase for annotation*: We highlight the importance of implementing a validation phase where annotators can meet and discuss their annotations to minimize discrepancies due to subjectivity by looking at different opinions.

4 Experimental Setup

In this section, we detail the evaluation setup used to benchmark *ThatiAR* and explore the subjectivity of Arabic news articles. We conducted experiments in a monolingual setup using the *ThatiAR* dataset, as well as multilingual experi-

ments with publicly available datasets. The goal of the multilingual experiments was to investigate whether incorporating additional datasets could enhance performance.

4.1 Data

We used stratified sampling to split the data into training, development, and test sets in a 70:10:20 ratio per class. Table 1 shows the data statistics.

Monolingual experiments: We used the training and development splits of *ThatiAR* for model fine-tuning, while the test split was used for evaluation.

Multilingual experiments: We used three setups for the training data: (i) AR: *ThatiAR* training set alone, (ii) ML: the entire multilingual datasets from CheckThat! 2023 (Galassi et al. 2023) and 2024 (Struβet al. 2024), and (iii) ALL: combining both *ThatiAR* training and the entire CheckThat! multilingual datasets. In all setups, we test on *ThatiAR* test set. We show the data statistics in Table 3.

| Lang. | Set | SUBJ | OBJ | All |
|-------|-------|---------------|---------------|-------|
| AR | Train | 1,055 (66.8%) | 1,391 (66.8%) | 2,446 |
| | Dev | 201 (12.7%) | 266 (12.8%) | 467 |
| | Test | 323 (20.5%) | 425 (20.4%) | 748 |
| ML | Train | 2,580 (79.1%) | 4,778 (86.0%) | 7,358 |
| | Dev | 357 (11.0%) | 353 (6.4%) | 710 |
| | Test | 323 (9.9%) | 425 (7.65%) | 748 |
| ALL | Train | 3,635 (80.5%) | 6,169 (85.5%) | 9,804 |
| | Dev | 558 (12.4%) | 619 (8.6%) | 1,177 |
| | Test | 323 (7.2%) | 425 (5.9%) | 748 |

Table 3: Statistics of multilingual training data.

4.2 Models

We have used three categories of models in our experiments, random, majority, classical models, PLMs, and LLMs.

Simple Models: To establish reasonably performing baselines, we used three simple models: RANDOM, which as-

signs labels randomly to sentences; MAJORITY, which assigns the most prevalent label in the dataset to all sentences; and SVC (Platt 1998), which is trained with the Support Vector Machine algorithm using standard preprocessing and TF-IDF representation. The regularization parameter of SVC is set to its default value of $C=1.0$.

Pre-trained Language Models (PLMs): We fine-tuned several PLMs to evaluate their performance on the subjectivity task using the transformer toolkit (Wolf et al. 2020).

Monolingual Experiments: We fine-tuned ARABERT version 2 (Antoun, Baly, and Hajj 2020) and QARIB (Abdelali et al. 2021), both of which are initially trained on Arabic datasets.

Multilingual Experiments: We fine-tuned multilingual BERT (mBERT) (Devlin et al. 2019) and XLM-RoBERTa base (XLM-R) (Conneau et al. 2020). All these models were fine-tuned using the training dataset of *ThatiAR* or the entire multilingual data from the Subjectivity Task 2 in CheckThat! Lab 2023 (Galassi et al. 2023) and 2024 (Strußel et al. 2024).

Large Language Models (LLMs): To align with recent advancements in NLP, we experimented with Jais-13B Arabic model (Sengupta et al. 2023), GPT-4 (version 0314) (Achiam et al. 2023), Gemini-1.5 (Anil et al. 2023), Mistral (Jiang et al. 2023), and Llama3-8b⁹ in zero-shot setup. We also run GPT-4 in few-shot setup. For reproducibility, we set the temperature to zero for all experiments and designed the prompts using concise instructions similar to those given to human annotators when creating *ThatiAR*. We used the LLMeBench framework to run the experiments (Dalvi et al. 2024).

The use and evaluation of LLMs involve prompting and post-processing of output to extract the expected label. For each GPT-4 experimental setup we explored multiple prompts guided by the same instruction and format as recommended in in OpenAI playground. After having an expected prompt, we run complete evaluation.

Zero-Shot. For the zero-shot experiments, we designed prompts by providing natural language instructions that describe the task and specify the expected label.

Few-Shots. For the few-shot example selection, we used the maximal marginal relevance-based (MMR) method to construct example sets that are both relevant and diverse (Carbonell and Goldstein 1998). The MMR method calculates the similarity between a test example and the example pool (e.g., training set) and selects m examples (shots). We applied MMR on top of embeddings generated by multilingual sentence-transformers (Reimers and Gurevych 2019). We conducted experiments with 3-shot and 5-shot examples.

4.3 Evaluation Measures

We evaluate all models’ predictions using classification metrics including weighted Precision, Recall, and F1-score for the “Subjective” class.

⁹<https://ai.meta.com/blog/meta-llama-3/>

5 Results and Discussion

5.1 Monolingual Results

Table 4 presents the benchmark results on the test set of all models trained on the *ThatiAR* training split. JAIS outperforms all other models in zero-shot setup. This highlights the importance of using models trained on Arabic data. GPT-4, in few-shot learning, markedly surpasses all other models across all measures except Recall. The optimal setup for GPT-4 is the 3-shot setup, showing a reasonable improvement compared to the 0-shot and 5-shot setups. Notably, in terms of *Recall*, JAIS, the only model trained on Arabic, outperforms GPT-4. This could indicate a weakness in GPT-4 in identifying all “Subjective” sentences, despite achieving the highest *Precision* scores by more frequently assigning the “Subjective” label to sentences. While the precision and recall scores differ slightly between the 3-shot and 5-shot settings of GPT-4, the weighted F1 score for the subjective class remains the same (0.800), likely due to rounding and marginal differences in the harmonic mean. This indicates that, despite varying trade-offs between precision and recall, both setups exhibit comparable overall effectiveness in detecting subjectivity.

| Model | Acc | P | R | F1 |
|---------------------------|--------------|--------------|--------------|--------------|
| RANDOM | 0.568 | 0.284 | 0.500 | 0.362 |
| MAJORITY | 0.500 | 0.499 | 0.499 | 0.497 |
| SVC | 0.540 | 0.517 | 0.515 | 0.509 |
| QARIB | 0.523 | 0.519 | 0.523 | 0.520 |
| ARABERT | 0.592 | 0.582 | 0.592 | 0.566 |
| mBERT | 0.563 | 0.549 | 0.563 | 0.546 |
| XLM-R | 0.568 | 0.323 | 0.568 | 0.412 |
| JAIS _{0-shot} | 0.610 | 0.605 | 0.610 | 0.578 |
| LLAMA3 _{0-shot} | 0.468 | 0.731 | 0.543 | 0.431 |
| GEMINI _{0-shot} | 0.520 | 0.456 | 0.557 | 0.501 |
| MISTRAL _{0-shot} | 0.539 | 0.167 | 0.238 | 0.415 |
| GPT-4 _{0-shot} | 0.768 | 0.517 | 0.507 | 0.529 |
| GPT-4 _{3-shot} | 0.795 | 0.647 | 0.544 | 0.800 |
| GPT-4 _{5-shot} | 0.785 | 0.636 | 0.528 | 0.800 |

Table 4: Results of mono-lingual models on *ThatiAR*.

5.2 Multilingual Results

Table 5 shows the benchmark results on the test set of all models trained on *ThatiAR* and multilingual data. The performance difference between mBERT and XLM-R models is generally marginal across each setup. Both models achieve their best performance when fine-tuned with only Arabic data (AR setup). mBERT shows superior performance in the ALL setup, whereas XLM-R excels in the ML setup, demonstrating its robustness in the absence of Arabic training data.

6 Annotations with Rationals

We utilized GPT-4 to validate and rationalize the human subjectivity annotations. Specifically, for each sentence in *ThatiAR*, we prompted GPT-4 with the sentence and its label,

| Setup | Model | Acc | P | R | F1 |
|-------|-------|--------------|--------------|--------------|--------------|
| AR | MBERT | 0.563 | 0.549 | 0.563 | 0.546 |
| | XLM-R | 0.568 | 0.323 | 0.568 | 0.412 |
| ML | MBERT | 0.525 | 0.498 | 0.525 | 0.495 |
| | XLM-R | 0.532 | 0.505 | 0.532 | 0.500 |
| ALL | MBERT | 0.554 | 0.535 | 0.554 | 0.528 |
| | XLM-R | 0.532 | 0.502 | 0.532 | 0.494 |

Table 5: Results of multilingual models on *ThatiAR*. Refer to Section 4.1 for training setup, “Setup” column.

and asked it, as an expert linguist, to “*Write a simple and short explanation*” for its given annotation. We generated explanation in both Arabic and English languages, which we will release along with *ThatiAR* for the community. Table 10 (in Appendix as a part of supplemental material) shows the prompt and example output in both languages. To verify the quality of the explanation we have randomly checked a sample of 20 rationals, out of which 65% is acceptable.

7 Instruction Dataset

To instruct-tune LLMs, it is essential to create an instruction-following dataset. For this purpose, we used GPT-4o to generate instructions for the development and test sets. To reduce the API cost of GPT-4o, the generated instructions from the development set were then used to assign instructions randomly to the samples in the training dataset. Let D_{dev} be the development set. We denote the set of instructions generated by GPT-4o for D_{dev} as $I = \{\text{GPT-4o}(x) \mid x \in D_{\text{dev}}\}$.

Let D_{train} be the training set. Instructions from I are assigned randomly to each sample in $D_{\text{train}} = \forall x \in D_{\text{train}}$, assign $I_{\text{rand}}(x) \in I$; where $I_{\text{rand}}(x)$ denotes an instruction randomly selected from I . This ensures that each training sample is paired with an instruction. Note that we kept the instruction from the test set independent.

To create instructions for the development and test sets, we aimed to generate diverse instructions. In Listing 1, we present the prompt used to create these instructions. For different samples, we asked GPT to create various types of instructions, such as (i) simple, (ii) straightforward, and (iii) detailed. We randomly selected one type from the three and used in the placeholder *random_ins_type*. The placeholder *sentence* represents the input sentence. Please see section D.2 (in Appendix as a part of supplemental material) for further details.

To evaluate the effectiveness of the instruction-following dataset, we fine-tuned Llama-3.1 8B with 4-bit quantization for one epoch. The fine-tuned model achieved an F1 score of 0.480 on the test set, which is loosely comparable to the LLAMA3 0-shot results with an F1 score of 0.431. Note that LLAMA3 0-shot was evaluated using a full-precision model, whereas we fine-tuned Llama-3.1 8B with 4-bit quantization. Due to limitations in computational resources, we were unable to perform a full-scale evaluation to compare the performance with other PLM models, which we aim to address in future studies.

8 Conclusion and Future Work

In this study, we introduce *ThatiAR*, a large Arabic subjectivity dataset consisting of manually annotated news sentences. We provide a detailed discussion of the data collection and annotation process. We further conducted extensive experiments with PLMs and LLMs to demonstrate the utility of *ThatiAR*. Additionally, we provide rationales and an instruction-following datasets, which can be used in LLM-based model development. Given the complexity of annotation, future research should include more annotators from diverse backgrounds to further enhance the subjectivity annotation process. Our study is preliminary in nature and serves as an initial step towards understanding news media in terms of subjectivity. However, this study presents important considerations for scholars specifically interested in subjectivity and for the field of NLP in general.

9 Limitations

Subjectivity annotation is a complex task, which has also been noticed in other languages. Even though we provided clear guideline in Arabic to make sure that native speakers fully understand the task, however, it still become a challenge for that. Many mturk annotators did not pass our qualified test. It might be because they are from diverse background, and culture, which might have effected the annotation process.

Acknowledgments

The contributions of F. Alam, M. Hasanain, R. Suwaileh and W. Zaghouani were funded by the NPRP grant 14C-0916-210015, which is provided by the Qatar National Research Fund part of Qatar Research Development and Innovation Council (QRDI).

References

- Abdelali, A.; Hassan, S.; Mubarak, H.; Darwish, K.; and Samih, Y. 2021. Pre-Training BERT on Arabic Tweets: Practical Considerations. *arXiv preprint arXiv:2102.10684*.
- Abdelali, A.; Mubarak, H.; Chowdhury, S.; Hasanain, M.; Mousi, B.; Boughorbel, S.; Abdaljalil, S.; El Kheir, Y.; Izham, D.; Dalvi, F.; Hawasly, M.; Nazar, N.; Elshahawy, Y.; Ali, A.; Durrani, N.; Milic-Frayling, N.; Hawasly, M.; Durrani, N.; and Alam, F. 2024. LARA-Bench: Benchmarking Arabic AI with Large Language Models. In Graham, Y.; and Purver, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 487–520. St. Julian’s, Malta: Association for Computational Linguistics.
- Abdul-Mageed, M. 2015. *Subjectivity and sentiment analysis of Arabic as a morphologically-rich language*. Ph.D. thesis, Indiana University.
- Abdul-Mageed, M.; and Diab, M. 2011. Subjectivity and Sentiment Annotation of Modern Standard Arabic Newswire. In Ide, N.; Meyers, A.; Pradhan, S.; and Tomanek, K., eds., *Proceedings of the 5th Linguistic Annotation Workshop*, 110–118. Portland, Oregon, USA: Association for Computational Linguistics.

- Abdul-Mageed, M.; Diab, M.; and Korayem, M. 2011. Subjectivity and Sentiment Analysis of Modern Standard Arabic. In Lin, D.; Matsumoto, Y.; and Mihalcea, R., eds., *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 587–591. Portland, Oregon, USA: Association for Computational Linguistics.
- Abdul-Mageed, M.; Diab, M.; and Kübler, S. 2014. SAMAR: Subjectivity and sentiment analysis for Arabic social media. *Computer Speech & Language*, 28(1): 20–37.
- Abdul-Mageed, M.; and Diab, M. T. 2012. AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis. In *LREC*, volume 515, 3907–3914.
- Abdul-Mageed, M.; and Diab, M. T. 2014. Sana: A large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis. In *LREC*, 1162–1169.
- Abdul-Mageed, M.; Elmadany, A.; and Nagoudi, E. M. B. 2021. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 7088–7105. Online: Association for Computational Linguistics.
- Abo, M. E. M.; Shah, N. A. K.; Balakrishnan, V.; Kamal, M.; Abdelaziz, A.; and Haruna, K. 2019. Ssa-sda: subjectivity and sentiment analysis of sudanese dialect Arabic. In *2019 International Conference on Computer and Information Sciences (ICIS)*, 1–5. IEEE.
- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Ahuja, K.; Diddee, H.; Hada, R.; Ochieng, M.; Ramesh, K.; Jain, P.; Nambi, A.; Ganu, T.; Segal, S.; Ahmed, M.; Bali, K.; and Sitaram, S. 2023. MEGA: Multilingual Evaluation of Generative AI. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 4232–4267. Singapore: Association for Computational Linguistics.
- Alam, F.; Shaar, S.; Dalvi, F.; Sajjad, H.; Nikolov, A.; Mubarak, H.; Da San Martino, G.; Abdelali, A.; Durrani, N.; Darwish, K.; Al-Homaid, A.; Zaghoulani, W.; Caselli, T.; Danoe, G.; Stolk, F.; Bruntink, B.; and Nakov, P. 2021. Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Findings of the Association for Computational Linguistics: EMNLP 2021*, 611–649. Punta Cana, Dominican Republic: Association for Computational Linguistics.
- Aleksandrova, D.; Lareau, F.; and Ménard, P.-A. 2019. Multilingual Sentence-Level Bias Detection in Wikipedia. In Mitkov, R.; and Angelova, G., eds., *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 42–51. Shoumen, Bulgaria: Incoma Ltd. ISBN 9789544520564.
- Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Antici, F.; Bolognini, L.; Inajetovic, M. A.; Ivasiuk, B.; Galassi, A.; and Ruggeri, F. 2021a. SubjectivITA: An Italian Corpus for Subjectivity Detection in Newspapers. In Candan, K. S.; Ionescu, B.; Goeuriot, L.; Larsen, B.; Müller, H.; Joly, A.; Maistro, M.; Piroi, F.; Faggioli, G.; and Ferro, N., eds., *Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2021*, volume 12880 of *LNCS*, 40–52. Springer.
- Antici, F.; Bolognini, L.; Inajetovic, M. A.; Ivasiuk, B.; Galassi, A.; and Ruggeri, F. 2021b. SubjectivITA: An Italian Corpus for Subjectivity Detection in Newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, 40–52. Cham: Springer International Publishing.
- Antici, F.; Ruggeri, F.; Galassi, A.; Korre, K.; Muti, A.; Bardi, A.; Fedotova, A.; and Barrón-Cedeño, A. 2024a. A Corpus for Sentence-Level Subjectivity Detection on English News Articles. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 273–285. Torino, Italia: ELRA and ICCL.
- Antici, F.; Ruggeri, F.; Galassi, A.; Korre, K.; Muti, A.; Bardi, A.; Fedotova, A.; and Barrón-Cedeño, A. 2024b. A Corpus for Sentence-Level Subjectivity Detection on English News Articles. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 273–285. Torino, Italia: ELRA and ICCL.
- Antoun, W.; Baly, F.; and Hajj, H. 2020. AraBERT: Transformer-based Model for Arabic Language Understanding. In Al-Khalifa, H.; Magdy, W.; Darwish, K.; Elsayed, T.; and Mubarak, H., eds., *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, 9–15. Marseille, France: European Language Resource Association.
- Banea, C.; Mihalcea, R.; and Wiebe, J. 2010. Multilingual Subjectivity: Are More Languages Better? In Huang, C.-R.; and Jurafsky, D., eds., *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 28–36. Beijing, China: Coling 2010 Organizing Committee.
- Banea, C.; Mihalcea, R.; Wiebe, J.; and Hassan, S. 2008. Multilingual Subjectivity Analysis Using Machine Translation. In Lapata, M.; and Ng, H. T., eds., *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 127–135. Stroudsburg, PA and USA: Association for Computational Linguistics.
- Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; V. Do, Q.; Xu,

- Y.; and Fung, P. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 675–718. Indonesia: Association for Computational Linguistics.
- Benamara, F.; Chardon, B.; Mathieu, Y.; and Popescu, V. 2011. Towards Context-Based Subjectivity Analysis. In Wang, H.; and Yarowsky, D., eds., *Proceedings of 5th International Joint Conference on Natural Language Processing*, 1180–1188. Chiang Mai, Thailand: Asian Federation of Natural Language Processing.
- Carbonell, J.; and Goldstein, J. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 335–336.
- Chaturvedi, I.; Cambria, E.; Welsch, R. E.; and Herrera, F. 2018. Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Inf. Fusion*, 44: 65–77.
- Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. Online: Association for Computational Linguistics.
- Dalvi, F.; Hasanain, M.; Boughorbel, S.; Mousi, B.; Abdaljalil, S.; Nazar, N.; Abdelali, A.; Chowdhury, S. A.; Mubarak, H.; and Ali, A. 2024. LLMeBench: A Flexible Framework for Accelerating LLMs Benchmarking. In Aletras, N.; and De Clercq, O., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 214–222. St. Julians, Malta: Association for Computational Linguistics.
- Das, N.; and Sagnika, S. 2020. A Subjectivity Detection-Based Approach to Sentiment Analysis. In Swain, D.; Pattnaik, P. K.; and Gupta, P. K., eds., *Machine Learning and Information Processing*, 149–160. Singapore: Springer Singapore. ISBN 978-981-15-1884-3.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.
- FORCE11. 2020. The FAIR Data principles. <https://force11.org/info/the-fair-data-principles/>.
- Galassi, A.; Ruggeri, F.; no, A. B.-C.; Alam, F.; Caselli, T.; Kutlu, M.; Struss, J. M.; Antici, F.; Hasanain, M.; Köhler, J.; Korre, K.; Leistra, F.; Muti, A.; Siegel, M.; Turkmen, M. D.; Wiegand, M.; and Zaghouani, W. 2023. 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.
- Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.
- Geva, M.; Goldberg, Y.; and Berant, J. 2019. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 1161–1166. Association for Computational Linguistics.
- Hendy, A.; Abdelrehim, M.; Sharaf, A.; Raunak, V.; Gabr, M.; Matsushita, H.; Kim, Y. J.; Afify, M.; and Awadalla, H. H. 2023. How good are GPT models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.
- Hube, C.; and Fetahu, B. 2019. Neural Based Statement Classification for Biased Language. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 195–203. New York, NY, USA: Association for Computing Machinery. ISBN 9781450359405.
- İrsoy, O.; and Cardie, C. 2014. Opinion Mining with Deep Recurrent Neural Networks. In Moschitti, A.; Pang, B.; and Daelemans, W., eds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 720–728. Doha, Qatar: Association for Computational Linguistics.
- Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. I.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- Khondaker, M. T. I.; Waheed, A.; Nagoudi, E. M. B.; and Abdul-Mageed, M. 2023. GPTAraEval: A Comprehensive Evaluation of ChatGPT on Arabic NLP. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 220–247. Singapore: Association for Computational Linguistics.
- Lan, W.; Chen, Y.; Xu, W.; and Ritter, A. 2020. GigaBERT: Zero-shot Transfer Learning from English to Arabic. In *Proceedings of The 2020 Conference on Empirical Methods on Natural Language Processing (EMNLP)*.
- Mihalcea, R.; Banea, C.; and Wiebe, J. 2007. Learning Multilingual Subjective Language via Cross-Lingual Projections. In Zaenen, A.; and van den Bosch, A., eds., *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*, 976–983. The Association for Computer Linguistics.

- Mohammad, S. 2016. A Practical Guide to Sentiment Annotation: Challenges and Solutions. In Balahur, A.; van der Goot, E.; Vossen, P.; and Montoyo, A., eds., *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 174–179. San Diego, California: Association for Computational Linguistics.
- Mourad, A.; and Darwish, K. 2013. Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 55–64.
- Pang, B.; and Lee, L. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 271–278. Barcelona, Spain.
- Platt, J. 1998. Sequential minimal optimization: A fast algorithm for training support vector machines.
- Refaee, E.; and Rieser, V. 2014. Subjectivity and sentiment analysis of arabic twitter feeds with limited resources. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme*, volume 16.
- Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992.
- Röttger, P.; Vidgen, B.; Hovy, D.; and Pierrehumbert, J. B. 2022. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In Carpuat, M.; de Marneffe, M.; and Ruíz, I. V. M., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, 175–190. Association for Computational Linguistics.
- Rustamov, S.; Mustafayev, E.; and Clements, M. 2013. Sentence-Level Subjectivity Detection Using Neuro-Fuzzy Models. In Balahur, A.; van der Goot, E.; and Montoyo, A., eds., *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 108–114. Atlanta, Georgia: Association for Computational Linguistics.
- Savinova, E.; and Moscoso Del Prado, F. 2023. Analyzing Subjectivity Using a Transformer-Based Regressor Trained on Naïve Speakers’ Judgements. In Barnes, J.; De Clercq, O.; and Klinger, R., eds., *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, 305–314. Toronto, Canada: Association for Computational Linguistics.
- Sengupta, N.; Sahu, S. K.; Jia, B.; Katipomu, S.; Li, H.; Koto, F.; Afzal, O. M.; Kamboj, S.; Pandit, O.; Pal, R.; et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.
- Shearer, E.; and Mitchell, A. 2021. News use across social media platforms in 2020.
- Sheikh Ali, Z.; Mansour, W.; Elsayed, T.; and Al-Ali, A. 2021. AraFacts: The First Large Arabic Dataset of Naturally Occurring Claims. In Habash, N.; Bouamor, H.; Hajj, H.; Magdy, W.; Zaghouni, W.; Bougares, F.; Tomeh, N.; Abu Farha, I.; and Touileb, S., eds., *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 231–236. Kyiv, Ukraine (Virtual): Association for Computational Linguistics.
- Shokri, M.; Sharma, V.; Filatova, E.; Jain, S.; and Levitan, S. 2024. Subjectivity Detection in English News using Large Language Models. In De Clercq, O.; Barriere, V.; Barnes, J.; Klinger, R.; Sedoc, J.; and Tafreshi, S., eds., *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, 215–226. Bangkok, Thailand: Association for Computational Linguistics.
- Struß, J. M.; Ruggeri, F.; Barrón-Cedeño, A.; Alam, F.; Dimitrov, D.; Galassi, A.; Pachov, G.; Koychev, I.; Nakov, P.; Siegel, M.; Wiegand, M.; Hasanain, M.; Suwaileh, R.; and Zaghouni, W. 2024. Overview of the CLEF-2024 Check-That! Lab Task 2 on Subjectivity in News Articles. In Faggioli, G.; Ferro, N.; Galuščáková, P.; and García Seco de Herrera, A., eds., *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*, CLEF 2024. Grenoble, France.
- Vieira, L. L.; Jeronimo, C. L. M.; Campelo, C. E. C.; and Marinho, L. B. 2020. Analysis of the Subjectivity Level in Fake News Fragments. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, 233–240. New York, NY, USA: Association for Computing Machinery. ISBN 9781450381963.
- Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In Liu, Q.; and Schlangen, D., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.
- Zhou, X.; and Zafarani, R. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5): 1–40.

Paper Checklist

- (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes**
- (b) Do your main claims in the abstract and introduction accurately reflect the paper’s contributions and scope? **Yes**

- (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? [Yes, as described in Section 3.](#)
 - (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? [No](#)
 - (e) Did you describe the limitations of your work? [Yes, please see Section 9.](#)
 - (f) Did you discuss any potential negative societal impacts of your work? [NA, we do not see negative societal impact of our work.](#)
 - (g) Did you discuss any potential misuse of your work? [Yes, please see Section 9.](#)
 - (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? [Yes, please see Section 9, and Appendix A, and B.](#)
 - (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes](#)
1. Additionally, if your study involves hypotheses testing...
- (a) Did you clearly state the assumptions underlying all theoretical results? [NA](#)
 - (b) Have you provided justifications for all theoretical results? [NA](#)
 - (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? [NA](#)
 - (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? [NA](#)
 - (e) Did you address potential biases or limitations in your theoretical framework? [NA](#)
 - (f) Have you related your theoretical results to the existing literature in social science? [NA](#)
 - (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? [NA](#)
2. Additionally, if you are including theoretical proofs...
- (a) Did you state the full set of assumptions of all theoretical results? [NA](#)
 - (b) Did you include complete proofs of all theoretical results? [NA](#)
3. Additionally, if you ran machine learning experiments...
- (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes, please see Section 4, and Appendix A, B, and C.](#)
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes, please see Section 4, and Appendix A, B, and C.](#)
- (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No](#)
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes, see Appendix B.](#)
 - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? [Yes, please refer to Section 3.3 and 4.](#)
 - (f) Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? [No](#)
4. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
- (a) If your work uses existing assets, did you cite the creators? [Yes, please refer to Section 4.1.](#)
 - (b) Did you mention the license of the assets? [Yes, please refer to Appendix A.](#)
 - (c) Did you include any new assets in the supplemental material or as a URL? [Yes, we release the dataset as described in Appendix A.](#)
 - (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [Yes, please refer to Section 3.1 and 4.1.](#)
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [NA, the dataset we release does not contain personally identifiable information or offensive content.](#)
 - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? [No](#)
 - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? [No](#)
5. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
- (a) Did you include the full text of instructions given to participants and screenshots? [Yes, please refer to Appendix C. Full guidelines will be released with the dataset.](#)
 - (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? [NA, our work does not carry risks to our annotators.](#)
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes, please refer to Section 3.2.](#)
 - (d) Did you discuss how data is stored, shared, and de-identified? [Yes, please refer to Appendix A.](#)

Ethics and Broader Impact

We collected news articles from a range of Arabic media outlets and selected sentences for annotation. While we aimed to include diverse topics and perspectives, we acknowledge the potential for bias in our data sampling. Annotations are inherently subjective and may reflect the sociocultural biases of the annotators. To mitigate this, we recruited annotators from different Arabic-speaking countries, with diverse educational and professional backgrounds. We also developed detailed annotation guidelines and conducted multiple rounds of training to promote consistency. However, biases and disagreements remain, which we analyze in the discussion section. In any of the data collection and annotation process we do not collect any personally identifiable information.

The models developed using *ThatiAR* have significant potential for positive impact by helping to detect subjective and potentially biased or misleading content in Arabic news. This can assist fact-checkers, journalists, and policymakers in combating misinformation and promoting media literacy. However, we also recognize the potential for misuse, such as in censorship or political manipulation. We encourage users to consider the ethical implications of their applications. Furthermore, while *ThatiAR* is a step towards greater representation of Arabic in NLP research, much work remains to fully capture the linguistic diversity of Arabic and its dialects. Our annotators and data sources skew towards Modern Standard Arabic, which may not reflect everyday language use. Future work should prioritize inclusivity and linguistic diversity. We are releasing *ThatiAR* dataset and resources publicly to encourage research on Arabic subjectivity analysis. However, we urge researchers to be transparent about the limitations and potential biases of the dataset and any resulting models. Appropriate documentation should be provided to help end users make informed decisions about model deployment.

Supplemental Material

A Data Release

The *ThatiAR* dataset¹⁰ is released under the Creative Commons Attribution 4.0 International License: <https://creativecommons.org/licenses/by/4.0/legalcode>. The dataset includes the following files:

- Subjectivity manual annotations divided into training, development, and test sets, in CSV format. Each news sentence is represented by an id, text, and label.
- Annotation guidelines provided to the crowd annotators in Arabic.
- Explanation and instruction annotations generated automatically by the GPT-4o model, in JSONL format, with the same splits as the manually annotated data. Each news sentence is represented by an id, text, label, explanation, and instruction.
- Example scripts for running experiments, including PLMs (AraBERT model) and LLMs (GLUE model).

¹⁰anonymouse.com

B Details of the experiments

For the experiments, we used SVM, PLMs, and LLMs (GPT-4). All these scientific artifacts are used according to their terms and conditions for research purposes. Below, we discuss the parameters we used. Furthermore, we have made all our scripts available to ensure reproducibility.

Models and Parameters:

- **AraBERT**: L=12, H=768, A=12; the total number of parameters is 371M, where L is the number of layers (i.e., Transformer blocks), H is the hidden size, and A is the number of self-attention heads;
- **BERT Multilingual** (bert-base-multilingual-uncased) (mBERT): L=12, H=768, A=12, number of parameters (172M);
- **XLM-RoBERTa** (xlm-roberta-base): L=24, H=1027, A=16; the total number of parameters is 355M.
- N-gram with SVM: TF-IDF transformation and used C=1.0 in SVM.

To fine-tune PLMs, we used the following hyper-parameters.

- Batch size: 8;
- Learning rate (Adam): $2e-5$;
- Number of epochs: 10;
- Max seq length: 256.

We ran the PLM-based fine-tuning experiments with different seed values and report the results of the best runs on the development set of *ThatiAR*. We run our experiments on a cluster consisting of GPUs such as P100, V100, V100-NVLINK, and T4.

C Annotation Guidelines

For the annotation we adopted and refined the annotation guidelines discussed in (Antici et al. 2021b). To begin the annotations, annotators of diverse backgrounds were provided with a specific use-cases for subjective and objective sentences that we present in Tables 6, 7 and 8, respectively. We release the annotation guidelines with the dataset.¹¹

C.1 Subjective Use Cases

We define *subjective* sentences as expressions of feelings, literary tastes, or personal interpretations of topics and events. Below are a few use cases of subjective sentence with examples in Table 7:

- Sentences expressing personal opinions about events and topics, or containing rhetorical questions, or containing probabilities and expectations and building conclusions on them, e.g., Sentence #1.
- Sentences containing sarcasm or humor, according to the writer’s expression, e.g., Sentence #2.
- Sentences encouraging, supporting, or approving an action, e.g., Sentence #3.
- Sentences containing offensive expressions such as racism, tactlessness, etc., e.g., Sentence #4.
- Sentences containing a rhetorical expression and depiction of people and situations, such as “exaggeration”, that a writer uses to express his or her personal opinion, e.g., Sentence #5.

¹¹anonymouse.com

| # | Sentence | Translation |
|---|---|--|
| 1 | دمرت السلطات الصينية آلاف المساجد في شينجيانغ، حسبما ذكر مركز أبحاث أسترالي الجمعة، في أحدث تقرير عن انتهاكات واسعة لحقوق الإنسان في الإقليم المضطرب. | Chinese authorities have destroyed thousands of mosques in Xinjiang, an Australian think tank said Friday, in the latest report on widespread human rights violations in the volatile region. |
| 2 | وأعلن الاحتلال في أوقات المسيرات خلال الأعوام السابقة عن رفع حالة الاستنفار العسكري على الحدود لأكثر من أسبوعين، خشية اختراق الحدود من دول الطوق باتجاه الأراضي الفلسطينية، والذين من ضمنهم مئات من المتضامنين الدوليين. | The occupation announced during marches throughout recent years in reducing military presence in borders for more than two weeks for fear of breaching the borders from the countries of the enclave towards the Palestinian territories, including hundreds of international solidarity activists. |
| 3 | هالة صدقي تلجأ إلى الديانة الإسلامية لتحقيق رغبتها وهذه التفاصيل - مشاهير عالمية كم تقاضى جونغكوك من فرقة BTS للغناء في إفتتاح مونديال قطر؟ - المشاهير العرب مفاجأة من العيار الثقيل!! | Hala Sudqi resorts to the Islamic religion to achieve her desires and these are the details – celebrities globally how much does Jongkok from BTS make for singing in the opening ceremony of FIFA in Qatar? – Arabic celebrities, a big surprise! |
| 4 | وتبين أيضاً أن زوجة الحياي سبق ان قامت بقطع المساعدة المالية عن ابنهما، وأن هذا الأخير قام بالتأمر على والديه عن طريق حشو المسدس بالرصاص، وهو عالم بما دأب عليه أبوه من عادة تهديد أمه بالقتل عن طريق ذلك المسدس الفارغ، فإن نفذ تهديده مرة واحدة فسيتخلص من أمه وأبيه بضرية، أو رصاصة واحدة. | It was also found that the perpetrator's wife had previously cut off financial assistance from their son, and the son had conspired against his parents by loading the gun knowing that his father usually threatened his mother with death using an empty gun. If the father executes his threat one more time, then the son will get rid of both his parents at once or should we say with one bullet. |

Table 6: Examples of news sentences.

| # | Sentence | Translation |
|---|--|---|
| 1 | والدليل اغتيال البروفيسور التونسي محمود عبد القادر البرزقي وكان قد فك الشيفرة الوراثية للفيروس وعزلها مخبريا وإعادة تشكيلها ليتحول الفيروس الى لقاح مضاد وذلك أغضب أمريكا فقامت باغتياله إنها باختصار جريمة بحق البشرية. | The evidence is the assassination of Tunisian professor Mahmoud Abdel Qader Al-Bazrti, who had decoded the genetic code of the virus, isolated it in a laboratory, and reconstructed it so that the virus could be transformed into an anti-vaccine. This angered America, so it assassinated him. It is, in short, a crime against humanity. |
| 2 | لنعود كل سنة مع أول قطرة غيث إلى نقطة الصفر نشكو لهم من انسداد البالوعات فيقولوا لنا عليكم بالسباحة!!!! | Let us return back every year with the first drop of rain to ground zero and complain to them about the clogged drains, and they tell us that you should go swimming!!!! |
| 3 | هذا يعني حماقة أو بيستحمق الآخرين، فنحن نعتقد من واجبنا شرعاً أن نقاوم هذا الاحتلال بكل ما أوتينا من قوة ونعاقبه بنفس الطرق التي هو يستخدمها ضدنا. | This means foolishness or fooling others. We believe it is our legal duty to resist this occupation with all our might and punish it with the same methods it uses against us. |
| 4 | كما سقط مبارك وسيسقط السفية السيسي، وخطيئة أي ديكتاتور أن ينظر إلى الشعب من خلال عصابته المتنفعة، وأن يصدق نفاقهم ويكذب علامات الغضب الثوري. | Mubarak also fell and the foolish Sisi will fall, and it is the sin of any dictator to look at the people through his beneficial gang, believe their hypocrisy and deny the signs of revolutionary anger. |
| 5 | لا تظلم أحداً، فالظلم نار لا تطفئ في قلب صاحبها، ولو مرت عليه الأعوام. | Do not oppress anyone, for injustice is a fire that will never be extinguished in the heart of its perpetrator, even if years have passed. |

Table 7: Use cases of “subjective” sentences.

C.2 Objective Use Cases

We define *objective* sentences as a presentation of facts, events, and topics based on factual data. Below are a few use cases of objective sentence with examples in Table 8:

- Sentences containing news (Sentence #1), facts (Sentence #2) and laws (Sentence #3) conveyed by the writer of the sentence.
- Sentences describing the writer’s feelings or emotions without expressing any opinions on any topic, e.g., Sentence #4.
- Sentences containing opinions, claims, feelings, or viewpoints attributed to a third party other than the writer, e.g., Sentence #5.
- Sentences conveying the writer’s comments without explicitly stating any personal conclusion, interpretation, or expression of a personal opinion, so that the discussion is left open, e.g., Sentence #6.
- Sentences stating conclusions reached by the writer of the sentence, without expressing his personal position or opinion, or they are justified by hypotheses that are not related to personal opinions, e.g., Sentence #7.
- Sentences referring to an individual by a well-known nickname that was not given by the writer, e.g., Sentence #8.
- Common expressions and examples or sayings, e.g., Sentence #8.

C.3 Annotation Agreement

In Table 9, we provide detail of the annotation agreement.

| Setup | C.Kappa |
|-------------------------|---------------|
| Annotator1 vs. Majority | 0.5464 |
| Annotator2 vs. Majority | 0.5512 |
| Annotator3 vs. Majority | 0.5173 |
| Average | 0.5383 |

Table 9: Inter-annotator agreement using Cohen Kappa (α) for *ThatiAR* dataset

D Challenges

D.1 Annotation Challenges

Annotating for subjectivity presents significant challenges, especially when conducted via crowdsourcing platforms. One major obstacle is the lack of shared cultural, linguistic, and experiential backgrounds among annotators. As mTurk does not disclose demographic information about annotators, we did not analyze the effect of educational, cultural, and regional backgrounds on annotation. Such disparities inevitably influence how annotators interpreted sentences and judged subjectivity, leading to disagreements.

Furthermore, subjective annotations are inherently influenced by individual biases, standpoints, and opinions, which are difficult to control in a crowdsourced setting. Achieving reliable annotations required iterative refinement of guidelines, pilot studies, qualifications tests, and ongoing quality

checks – underscoring the complexities of crowdsourcing subjective annotations across diverse annotator pools.

D.2 Prompting Challenges

The performance of the model is highly dependent on the prompting strategy. Designing optimal prompts for each task is challenging and requires multiple iterations. Depending on the prompt, the output varies across all instances of the dataset. For the subjectivity task in this study, we experimented with (i) zero-shot and few-shot methods for label generation, (ii) generating explanations, and (iii) generating the instruction dataset. In Table 10, we provide examples of prompts in Arabic and English for generating explanations, along with the provided sentences and their labels.

```
1 prompt = f"You are an expert in
   creating instruction datasets to
   train AI models. \
2 Here, our idea is to create an
   instruction dataset for a
   subjectivity detection task. \
3 The task is to determine whether a
   sentence is subjective or
   objective. \
4 Write a '{random_ins_type}'
   instruction for this '{sentence
   }'. Do not include the sentence
   in the instruction."
```

Listing 1: Prompt to create instructions.

| # | Sentence | Translation |
|---|---|--|
| 1 | وأضافت نفس المصادر أن هناك أحزاب رفضت هذا الاقتراح نظراً لأن بن عيسى غير ملم بالوضع الاقتصادي الحرج التي تمر به تونس. | The same sources added that there are parties that rejected this proposal because Ben Aissa is not familiar with the critical economic situation that Tunisia is going through. |
| 2 | وتمثل الفيروس خطراً بشكل خاص على كبار السن وعلى من يعانون من مشاكل صحية، ولدى إيطاليا واحدة من أكبر نسب كبار السن في العالم. | The virus represents a particular danger to the elderly and those suffering from health problems, and Italy has one of the largest proportions of elderly people in the world. |
| 3 | في حالة وجود منظمة غير حكومية مرخص لها بتصدير أو إعادة تصدير الخدمات لسوريا بموجب هذا القسم في التاريخ السابق لتاريخ سن هذا القانون، فإن هذا القسم ينطبق على هذه المنظمة في تاريخ سن القانون وبعده إلى الحد نفسه وبنفس الطريقة التي كان ينطبق بها هذا القسم على هذه المنظمة في التاريخ السابق لتاريخ سن هذا القانون. | If a non-governmental organization licensed to export or re-export services to Syria under this section existed on the date prior to the date of enactment of this Act, this section shall apply to such organization on and after the date of enactment of this Act to the same extent and in the same manner as this section applied to such The organization on the date prior to the date of enactment of this law. |
| 4 | وكلما عدت من التحقيق والمحاكم منهكة إلى زنزاتي أجدد عهدي لله بأنني رغم كل التضحيات سأسير حتى مماتي على نفس الدرب لأنني نذرت نفسي للمسجد الأقصى. | Whenever I return from the investigation and the courts exhausted to my cell, I renew my pledge to God that despite all the restrictions, I will walk the same path until my death because I have vowed myself to Al-Aqsa Mosque. |
| 5 | وتقول إنه في المجتمع العلمي، هناك تعاريف متضاربة حول موعد تحقيق مناعة القطيع. | She says that in the scientific community, there are conflicting definitions about when herd immunity will be achieved. |
| 6 | لقد كانت جميلة احما على مسمى، انطلقت بأحلام الزهور فتعلمت الخياطة والرقص وكانت تحلم بأن تكون مصممة أزياء، ولكن القدر قادها لتكون مصممة أجمل نضال في تاريخ البشرية. | She was beautiful by her name. She started out with dreams of flowers, learned sewing and dancing, and dreamed of being a fashion designer, but fate led her to become the designer of the most beautiful struggle in human history. |
| 7 | في حال تعرض الفيروس لدرجة حرارة ٦٢ أو ٧٢ سوف يُقتل لذا لا يعيش في المناطق الحارة. | If the virus is exposed to a temperature of 26 or 27, it will be killed, so it does not live in hot areas. |
| 8 | وجاء ليصب في مصلحة القوى اليمينية الرافضة لمفهوم الوحدة، مثل تيار لوبان في فرنسا، أو حزب الاستقلال البريطاني، غير أن رافضي الدستور الفرنسيين لم يكونوا جميعاً من التيار اليميني بل كانوا أيضاً من التيار اليساري، وسبب رفضهم يرجع إلى انتقادهم للسياسة الاقتصادية المقترحة التي لا تهتم بتأمين ضمانات اجتماعية كافية. | It came to serve the interest of the right-wing forces that reject the concept of unity, such as the Le Pen movement in France, or the British Independence Party. However, the French opponents of the constitution were not all from the right-wing movement, but rather they were also from the left-wing movement, and the reason for their rejection is due to their criticism of the proposed economic policy that does not care about Providing adequate social guarantees. |
| 9 | الأدعية المأثورة: «اللهم إني أعوذ بك من البرص والجنون والجذام وسيئ الأسقام وأعوذ بك من همزات الشياطين وأعوذ بك رب أن يحضرون وصل اللهم على سيدنا محمد وآل محمد» | Supplications: “O God, I seek refuge in You from leprosy, madness, leprosy, and bad diseases, and I seek refuge in You from the instigations of the devils, and I seek refuge in You, my Lord, from their presence, and may God bless our master Muhammad and his family.” |

Table 8: Use cases of “objective” sentences.

| | |
|--------------|--|
| Input Prompt | أنت خبير لغوي. تستطيع فهم التعبير الذاتي في جمل المقالات الإخبارية. اشرح لماذا الجملة التالية: «ومع أن أداء القوات المصرية كان نصرًا لا يمكن لمنصف أن يماري فيه، فإن نظام أنور السادات كان امتدادًا للنظام العربي الرسمي، الذي استولده الاستعمار الغربي، وجعله بمثابة غشاوة استقلال خادعة تعمي عيون وبصائر الشعوب العربية، وتطمس على قلوبها فتغفل عن واقعها المرير». تعتبر «ذاتية». اكتب شرح موجز وبسيط. |
| GPT Response | الجملة المذكورة تعتبر ذاتية لأنها تحتوي على تقييم شخصي لأداء القوات المصرية ونظام أنور السادات. الكاتب يعبر عن رأيه الشخصي بأن أداء القوات المصرية كان نصرًا لا يمكن الجدل فيه، ولكنه ينتقد نظام السادات بوصفه امتدادًا للنظام العربي الذي أنشأه الاستعمار الغربي. الكاتب يستخدم عبارات مثل «غشاوة استقلال خادع» و«تعمي عيون وبصائر الشعوب العربية» للتعبير عن وجهة نظره الشخصية حول النظام السياسي في مصر والعالم العربي. هذه العبارات تظهر تحيز الكاتب وتقييمه الشخصي للأحداث، مما يجعل الجملة ذاتية. |
| Input Prompt | You are an expert linguist. You can understand the subjectivity of sentences in news articles. Explain why the sentence is "Subjective": «ومع أن أداء القوات المصرية كان نصرًا لا يمكن لمنصف أن يماري فيه، فإن نظام أنور السادات كان امتدادًا للنظام العربي الرسمي، الذي استولده الاستعمار الغربي، وجعله بمثابة غشاوة استقلال خادعة تعمي عيون وبصائر الشعوب العربية، وتطمس على قلوبها فتغفل عن واقعها المرير». Write a simple and short explanation. |
| GPT Response | The sentence in question contains subjective language because it includes value judgments and opinions rather than objective facts. Phrases like "نصرًا لا يمكن لمنصف أن يماري فيه" (a victory that no fair person can dispute) and "غشاوة استقلال خادعة" (a deceptive veil of independence) reflect the author's personal perspective and interpretation of events. Additionally, the use of emotionally charged words such as "تعمي" (blind) and "تطمس" (obscure) further indicates a subjective viewpoint. The sentence presents the author's opinion on the performance of the Egyptian forces and the nature of Anwar Sadat's regime in relation to Western colonialism and its impact on Arab societies. |

Table 10: Prompt for generating explanations for human-annotated sentences