

ArImpliSent: A Large-Scale Dataset for Arabic Implicit Sentiment Analysis

Ghozlene Chenni, Sadik Bessou, Hanane Boutouta, Nour elhouda Chenni

Department of Computer Science, Faculty of Sciences, Setif 1 University - Ferhat Abbas

{ghozlene.chenni,bessou.s,hanane.boutouta}@univ-setif.dz

nour3che@gmail.com

Abstract

Despite significant advances in sentiment analysis, existing resources primarily focus on explicit sentiment, where sentiment is directly expressed through affective words. As a result, implicit sentiment, conveyed through context, events, or figurative language, remains under-explored, particularly in morphologically rich and low-resource languages such as Arabic. To address this gap, we introduce ArImpliSent, a new large-scale Arabic dataset specifically designed for implicit sentiment analysis. The dataset comprises approximately 22K manually verified instances and was constructed by aggregating several widely used Arabic sentiment and emotion datasets and applying rigorous lexicon-based filtering strategies. We conducted a rigorous manual verification phase to further ensure dataset quality. Baseline evaluations, ranging from traditional machine learning methods to state-of-the-art pretrained language models, confirm that implicit sentiment analysis remains a highly challenging task for current NLP systems. These findings emphasize the necessity of more advanced models and resources capable of capturing the subtle contextual cues underlying implicit sentiment in Arabic text. ArImpliSent will be available for research purposes.¹

1 Introduction

Sentiment analysis (SA) has become one of the most widely studied tasks in Natural Language Processing (NLP), driven by the rapid growth of user-generated content on social media platforms. Significant progress has been achieved for Arabic SA through the development of large annotated datasets and powerful pre-trained language models such as AraBERT, MARBERT, CAMELBER, and QARiB. As a result, state-of-the-art systems now achieve strong performance on several Arabic sentiment benchmarks.

Despite these advances, all existing Arabic sentiment resources focus on *explicit sentiment*, where polarity is directly expressed through sentiment-bearing words such as “excellent”, “wonderful”, or “terrible”. In real-world communication, however, sentiment is often conveyed implicitly through events, situations, metaphors, sarcasm, or commonsense knowledge rather than explicit sentiment vocabulary. For example, the statement *لقد تاجروا بنا يا صديقي* “They traded us (like commodities), my friend.” expresses a negative attitude despite containing no overtly negative words. Correctly interpreting such cases requires contextual reasoning beyond lexical sentiment cues.

This distinction presents a major challenge for current SA systems. Previous studies have shown that neural models frequently rely on lexical shortcuts and sentiment-word memorization, achieving high performance on explicit benchmarks while struggling when sentiment must be inferred from context. Consequently, progress on Arabic SA may be overestimated by existing datasets, which predominantly contain explicit expressions of polarity.

To address this gap, we introduce **ArImpliSent**. To the best of our knowledge, it is the first large-scale dataset for implicit Arabic SA. ArImpliSent is constructed by aggregating publicly available Arabic sentiment and emotion datasets and applying a strict multi-lexicon filtering strategy that removes instances containing explicit sentiment or emotion vocabulary. The resulting corpus is further refined through manual verification to improve annotation reliability and dataset quality.

ArImpliSent contains 22,034 manually verified instances spanning Modern Standard Arabic and multiple regional dialects. The dataset provides a challenging benchmark that requires models to infer sentiment from contextual cues rather than direct lexical indicators. To establish benchmark performance, we evaluate a diverse set of traditional

¹<https://github.com/chen-ghozla/ArImpliSent>

machine learning methods and state-of-the-art Arabic pre-trained language models. Experimental results reveal a substantial performance gap, highlighting the difficulty of implicit sentiment understanding and the limitations of current approaches.

Our contributions are summarized as follows:

- We introduce **ArImpIiSent**, the first large-scale benchmark specifically designed for implicit Arabic SA.
- We propose a rigorous lexicon-based methodology for extracting instances with implicit sentiment through multi-level exclusion filtering.
- We provide extensive benchmark experiments using traditional machine learning models and state-of-the-art Arabic Transformer based models.
- We demonstrate the limitations of current NLP models in implicit SA and highlight the need for more advanced approaches for Arabic text.

2 Related Work

2.1 Arabic Sentiment Analysis

Arabic SA has received considerable attention over the past decade. The rapid growth of Arabic social media has fueled the development of numerous SA datasets, with Twitter emerging as the dominant source due to its abundance of opinion-rich content and linguistic diversity.

Early datasets, including ArTwitter (Abdulla et al., 2013) and AJGT (Alomari et al., 2017), each contain around 2,000 manually annotated tweets with binary sentiment labels, covering MSA and Jordanian Arabic. Subsequent efforts focused on increasing corpus size and expanding linguistic coverage. ASTD (Nabil et al., 2015) provides approximately 10,000 tweets annotated into four sentiment classes (positive, negative, mixed, and neutral), but they are highly imbalanced toward the neutral class. ArSAS (Elmadany et al., 2018) contains about 21,000 manually annotated tweets covering MSA and multiple Arabic dialects, while AT-SAD (Abu Kwaik et al., 2020) comprises 36,000 tweets annotated using a combination of distant supervision, self-training, and manual validation. The ASAD dataset (Alharbi et al., 2020) further expanded the scale to approximately 95,000 manually annotated tweets, making it one of the largest

Arabic Twitter sentiment datasets with broad dialectal coverage. In addition, SemEval-2017 Task 4 (Arabic) (Rosenthal et al., 2017) introduced a standardized benchmark for Arabic Twitter SA. While these datasets have driven progress in Arabic NLP, they do not isolate implicit expressions.

2.2 Datasets for Implicit SA

Several benchmark datasets have been developed for implicit SA in high-resource languages such as English and Chinese. SMP2019-ECISA (smp, 2019) has emerged as the standard benchmark for Chinese implicit sentiment analysis, containing sentences without explicit opinion words. In English, different resources have been developed to address various aspects of implicit SA. CLIPeval (Russo et al., 2015), derived from SemEval-2015 Task 9, focuses on identifying the implicit polarity (positive, negative, or neutral) associated with an event instance, while EveSA (Zhou et al., 2021) is an event-centered Implicit SA benchmark that provides annotated event triplets, event types, and sentence-level sentiment polarity labels. These two datasets primarily emphasize event-centric sentiment inference. More recent benchmarks broaden the scope of Implicit SA by evaluating models on implicit sentiment expressed through contextual cues, figurative language, humor, irony, and commonsense knowledge. These include SemEval-2021 Task 7 (Meaney et al., 2021), which contains 6,179 implicit and 3,821 non-implicit sentiment texts; the Reddit ISA Dataset (Weller and Seppi, 2019), comprising 15,909 texts, of which 2,025 exhibit implicit sentiment; and Pun-of-the-Day (Yang et al., 2015), containing 4,827 texts, of which 2,424 are implicit sentiment instances. Beyond English and Chinese, ViISA (Luu and Thin, 2025) extends ISA research to Vietnamese, providing a sentence-level dataset annotated with only positive and negative labels. Table 1 summarizes these public benchmark datasets for implicit SA.

For Arabic, the need for richer sentiment resources has been widely recognized (Abu Kwaik et al., 2020); however, existing datasets are primarily designed for explicit sentiment analysis, leaving implicitly expressed opinions largely unexplored. Although Arabic has recently witnessed initial efforts toward implicit affect understanding, with datasets proposed for Arabic Implicit Emotion Recognition (Boutouta et al., 2025a,b), these resources focus on identifying implicit emotional

categories rather than sentiment polarity. This suggests a critical gap in the resources necessary to advance Arabic implicit SA. Our work is motivated by creating a gold standard for benchmarking Arabic implicit SA, providing a foundation for developing future research in this field.

3 The ArImpliSent Dataset

3.1 Data Source

We constructed our dataset by leveraging seven existing manually annotated Arabic sentiment datasets, all collected from Twitter. These include ASAD (Alharbi et al., 2020), ASTD (Nabil et al., 2015), ArSAS (Elmadany et al., 2018), ArTwitter (Abdulla et al., 2013), ATSAD (Abu Kwaik et al., 2020), SemEval-2017 (Rosenthal et al., 2017), and AJGT (Alomari et al., 2017).

In addition, to address severe class imbalance observed during the initial aggregation, particularly the scarcity of positive instances, we incorporate the Arabic Emotion Twitter Dataset (AETD) (Al-Khatib and El-Beltagy, 2017).

Collectively, these datasets cover Modern Standard Arabic as well as at least four regional dialects, including Egyptian, Jordanian, Gulf, and mixed dialects. Table 2 summarizes their size, label taxonomy, dialectal coverage, and annotation quality.

3.2 Data Preprocessing

This step is an initial step that aims to reduce the inconsistencies and normalise the data into a coherent form so that it can be handled easily. In our implementation, we used the following steps:

- **Cleaning:** removing unknown characters, diacritics, punctuation, URLs, emojis, etc.
- **Elongation removal:** removing the repeated letters which might appear specially in social media data.
- **Instance Deletion:** Tweets that contain less than 3 words, are nonsensical, or rely entirely on external links to convey sentiment were permanently removed.
- **Letter normalisation:** letters which appeared in different forms were transformed into a single form. We replace $\hat{\text{ا}}$ with ا , $\hat{\text{و}}$ with و

and ي with ى .

3.3 Dataset filtering

To extract only tweets in which sentiment is conveyed implicitly rather than through explicit affective words, we applied a strict instance-level filtering pipeline.

Lexicons are traditionally utilized as features to aid classification. In contrast, our work uniquely applies these lexicons as a *negative filter* to detect and exclude tweets containing explicit sentiment-bearing terms, yielding a purely implicit corpus.

We compiled a comprehensive Explicit Vocabulary List, denoted as V_{exp} , by aggregating three resources:

1. **NRC Emotion and Sentiment Lexicon:** (Mohammad et al., 2016) Adapted for Arabic.
2. **NileULex:** (El-Beltagy, 2016) An Arabic sentiment lexicon containing both MSA and Egyptian dialect terms.
3. **Moataz saad Emotion Lexicon:** (Saad, 2015) We mapped discrete emotion terms to polarities (e.g., words associated with Anger, Sadness, Fear, and Disgust to Negative; Joy and Surprise to Positive).

Let T represent a tweet and W_T represent the set of tokens in T . If the intersection of the tweet’s tokens and the explicit vocabulary is not empty:

$$W_T \cap V_{exp} \neq \emptyset \quad (1)$$

the entire tweet T is discarded.

The surviving subset consists exclusively of naturally occurring, syntactically intact tweets that contain zero explicit sentiment or emotion words but retain their original sentiment labels.

3.4 Annotation

Constructing our corpus from existing manually annotated Arabic sentiment datasets eliminated the need for exhaustive re-annotation, leveraging high-quality human annotations to form a solid foundation for implicit sentiment identification and verification.

To obtain a unified sentiment representation, all source datasets were standardized into three sentiment categories: Positive, Negative, and Neutral. For the AETD dataset, a targeted emotion-to-polarity mapping was applied, whereby instances

Dataset	Year	Language	Domain	Size	Sentiment Classes
SMP2019-ECISA (smp, 2019)	2019	Chinese	Social Media	~23K	Positive, Neutral, Negative
CLIPeVal (Russo et al., 2015)	2015	English	Event-centric Text	1,651	Positive, Neutral, Negative
Pun-of-the-Day (Yang et al., 2015)	2015	English	Puns	4,827	Implicit vs. Non-implicit
Reddit ISA Dataset (Weller and Seppi, 2019)	2019	English	Reddit Discussions	15,909	implicit vs. Non-implicit
EveSA (Zhou et al., 2021)	2021	English	Event-centered Twitter	~7K	Positive, Neutral, Negative
SemEval-2021 Task 7 (Meaney et al., 2021)	2021	English	Social Media	10K	Implicit vs. Non-implicit
ViISA (Luu and Thin, 2025)	2025	Vietnamese	Social Media	302	Positive, Negative

Table 1: Public Benchmark Datasets for Implicit SA

originally labeled as Joy, Love, and Sympathy were mapped to the Positive sentiment class. This mapping contributed to a more balanced distribution across the three sentiment categories.

3.5 Annotator Agreement and Manual Verification

To ensure the quality and reliability of the dataset, a manual verification phase was conducted to confirm both the correctness of the inherited sentiment labels and the implicit nature of each instance. Three annotators (Native Arabic speakers) independently reviewed a randomly selected subset of 500 instances, assessing each instance along two dimensions:

1. The correctness of the sentiment label (Positive, Negative, or Neutral)
2. Whether the instance naturally conveyed implicit sentiment, i.e., without relying on explicit affective words.

Inter-annotator agreement was measured using Fleiss’ Kappa (κ) (Fleiss, 1971), yielding $\kappa = 0.61$ for sentiment label correctness, indicating substantial agreement, and $\kappa = 0.91$ for implicitness verification, indicating near-perfect agreement. These results confirm that while sentiment polarity judgments involve a degree of subjectivity, the distinction between implicit and explicit sentiment was consistently and reliably identified by annotators.

Following this assessment, a verification approach was extended to the entire initial dataset, with annotators resolving disagreements through discussion until consensus was reached. During this process, instances found to contain explicit sentiment were removed, accounting for 10% of the initial data. This filtering process resulted in a final dataset of 22,034 manually verified implicit sentiment instances.

4 Dataset Analysis

Table 3 shows retention rates ranging from 7.33% (AJGT) to 20.15% (ATSAD gold), with an overall retention of 15.13% across 145,715 tweets, yielding 22,034 implicit instances. The AETD figure (3,563) reflects only the Joy/Love/Sympathy subset of its full 10,065 tweet corpus, filtered through the same pipeline as the other datasets.

Table 4 shows the final label distribution: 20.41% Negative, 59.28% Neutral, and 20.31% Positive. The Neutral class remains dominant because the original source datasets are inherently imbalanced, with neutral instances substantially outnumbering positive and negative ones. Consequently, the resulting corpus naturally preserves this class distribution despite our emotion-mapping augmentation.

Table 5 illustrates examples of implicit sentiment along with their corresponding triggers from our

Dataset	Size	Classes	Coverage	Collection	Annotation	Raters	Quality
ArTwitter (Abdulla et al., 2013)	2K	Pos/ Neg	MSA + Jordanian	Manual Collection	Manual	3	–
ASTD (Nabil et al., 2015)	10K	Pos/ Neg/ Mixed/ Obj	Egyptian Dialect	Twitter API Crawling	Manual	3	–
SemEval 2017 Task 4 (Rosen-thal et al., 2017)	10K	Pos/ Neg/ Neu	Multiple Dialects	Twitter API Crawling	CrowdFlower	>5	–
AJGT (Alomari et al., 2017)	1.8K	Pos/ Neg	MSA + Jordanian	Manual Collection	Manual	3	–
ArSAS (Elmadany et al., 2018)	21K	Pos/ Neg/ Neu/ Mixed	Multiple Dialects	Twitter Crawling	Manual	500	Fleiss’ $\kappa = 0.79$
ASAD (Alharbi et al., 2020)	95K	Pos/ Neg/ Neu	MSA + Multiple Dialects	Twitter API + Manual Selection	Manual	3	Fleiss’ $\kappa = 0.56$
ATSAD (Abu Kwaik et al., 2020)	8K	Pos/ Neg	Multiple Dialects	Twitter API Crawling	Manual validation of weakly supervised corpus	3	Gold standard validated subset

Table 2: Comparison of Arabic sentiment analysis datasets. All datasets were collected from Twitter.

dataset. Positive examples rely on metaphor (e.g., comparing a sibling to a source of light) with no explicit positive words. Negative examples use sarcastic simile and implied grievance (e.g., the verb *traded*) without explicit negative adjectives. Neutral examples range from purely transactional to mildly complaint-adjacent, showing that the Neutral class is more heterogeneous than a single category.

5 Baseline Experiments

5.1 Models

We evaluate two non-neural baselines (Bessou and Chenni, 2021) alongside nine pre-trained Arabic and multilingual Transformer models, all fine-tuned for three-class sequence classification.

- **TF-IDF + LinearSVC.** A bag-of-n-grams baseline (unigrams and bigrams, 50K features, sublinear term-frequency scaling) paired with a linear SVM.
- **TF-IDF + Multinomial Naive Bayes.** A second non-neural baseline using the same

feature space as the SVM above, but with a probabilistic generative classifier, allowing us to assess whether the choice of classifier — margin based versus probabilistic — affects performance independently of the feature representation.

- **AraBERT v2** (Antoun et al., 2020) is the first large-scale Arabic BERT model, pre-trained primarily on MSA news and Wikipedia text.
- **MARBERT and MARBERT v2** (Abdul-Mageed et al., 2021) are trained on a billion-tweet multi-dialect Arabic corpus, making them better suited to social-media register than MSA-only models.
- **CAMeL-BERT** (Inoue et al., 2021) is trained on a mix of MSA, dialectal, and Classical Arabic.
- **XLM-RoBERTa** is a multilingual Transformer included as a non-Arabic-specific reference point.
- **QARiB** (Abdelali et al., 2021) is trained on

Source Dataset	Original Size	Discarded (Explicit)	Remaining (Implicit)	Retention Rate
ASAD	95,000	80,369	14,631	15.40%
ArSAS	19,897	17,543	2,354	11.83%
ASTD	10,000	8,307	1,693	16.93%
SemEval-2017	9,455	7,665	1,790	18.93%
ATSAD gold	4,000	3,194	806	20.15%
ArTwitter	2,000	1,711	289	14.45%
AJGT	1,800	1,668	132	7.33%
AETD (Joy/Love/Sympathy)	3,563	3,224	339	9.51%
TOTAL	145,715	123,681	22,034	15.12 %

Table 3: Detailed Corpus Distribution: Explicit Collection vs. Implicit Annotation

Class	Count	Percentage
Negative	4,497	20.41%
Neutral	13,061	59.28%
Positive	4,476	20.31%
Total	22,034	100 %

Table 4: Label distribution of the resulting ArImpliSent Dataset.

Arabic tweets and web text with an emphasis on dialect diversity.

- **AraRoBERTa-SA** (AlYami and Al-Zaidy, 2022) is a RoBERTa style Arabic sentiment model previously benchmarked on ASTD and SS2030.
- **AraBERT v2 twitterB / twitterL** are the base and large variants of AraBERTv2 further pre-trained on Twitter data, included to test whether Twitter-domain pre-training narrows the explicit-to-implicit gap.

5.2 Experimental setup

We balance ArImpliSent by exploiting the fact that the Negative and Positive classes are already naturally similar in size in the full corpus (4,428 and 4,408 instances, respectively; Table 4); we downsample the majority Neutral class to match ($\approx 4,500$), yielding a balanced pool of unique, non-duplicated instances. This balanced corpus is split into 70% train, 10% development, and 20% test, stratified by label. All Transformer models are fine-tuned for up to 5 epochs with early stopping on development Macro-F1 (patience of 2 epochs), a maximum sequence length of 128 tokens, batch size of 32, learning rate of 2×10^{-5} with linear warm-up over 10% of training steps, and weight decay of 0.01. Each model is fine-tuned with three random

seeds (42, 123, 456); we report mean and standard deviation of per-class and Macro-F1 scores on the held-out test set. The TF-IDF + LinearSVC baseline uses unigrams and bigrams with sublinear term-frequency scaling and 50,000 maximum features. All experiments were conducted on a workstation running Linux Mint with an NVIDIA GeForce RTX 5060 Ti GPU (16 GB VRAM).

5.3 Results

Table 6 presents the baseline performance on the balanced ArImpliSent test set, reporting per-class F1 and Macro-F1 scores averaged over three random seeds. Among the evaluated models, MARBERT v2 achieves the highest Macro-F1 of 66.6, followed closely by AraBERT v2 Twitter L (66.4), AraBERT v2 Twitter B (66.3), and MARBERT (66.2). These results indicate that models pre-trained on Arabic social media data generally outperform both the non-neural baselines and models trained primarily on broader Arabic or multilingual corpora.

Overall, even the best performing model (MARBERT v2, Macro-F1 = 66.6) remains well below typical explicit sentiment benchmarks reported for the same model families on existing Arabic datasets (commonly 80+ Macro-F1 on ASTD or ArSAS), confirming that implicit Arabic sentiment constitutes a substantially harder challenge for current state-of-the-art models than its explicit counterpart.

5.4 Error Analysis

To better understand the challenges posed by implicit sentiment, we examine representative errors made by our top-performing model, MARBERT v2, on the ArImpliSent test set. Table 7 presents misclassified instances categorized according to their primary implicit sentiment trigger. The errors reveal several recurring challenges, including situation-dependent sentiment, commonsense

Example	Class	Trigger
ليست اختا لي فقط انها مصدر الضوء في عمري "She is not just my sister; she is the source of light in my life."	Positive	Metaphor
ضع امنياتك في سجده "Place your wishes in a prostration (prayer)."	Positive	Commonsense
صوتها زي ايفكت الفضائي الي في سناب شات "Her voice is like the space effect on Snapchat."	Negative	Metaphor
لقد تاجروا بنا يا صديقي "They traded us like commodities, my friend."	Negative	Metaphor
حاولت الحصول علي رقم السجل الزراعي من خلال بوابه حصاد لكن لم يظهر لي اي شي "I tried to obtain the agricultural registry number through the Hasad portal, but nothing appeared."	Neutral	Event
سجلت الوالدة كيف اعرف متى تستحق السكن وهي من مستفيدي الضمان الاجتماعي "My mother is registered; how do I know when she becomes eligible for housing, given that she is a social welfare beneficiary?"	Neutral	Event

Table 5: Representative examples from ArImpliSent illustrating different implicit sentiment triggers.

Model	Negative F1	Neutral F1	Positive F1	Macro-F1
<i>Non-neural baseline</i>				
TF-IDF + LinearSVC	55.4	50.7	55.4	53.9
TF-IDF + MultinomialNB	58.6	52.0	58.2	56.2
<i>Pre-trained language models (mean $\pm$ std, 3 runs)</i>				
AraBERT v2	56.1 $\pm$ 9.8	52.7 $\pm$ 5.3	60.3 $\pm$ 6.5	56.3 $\pm$ 7.2
MARBERT	68.1 $\pm$ 1.2	60.5 $\pm$ 0.4	70.0 $\pm$ 0.9	66.2 $\pm$ 0.1
MARBERT v2	69.8 $\pm$ 0.5	61.5 $\pm$ 0.7	68.5 $\pm$ 0.2	66.6$\pm$0.2
CAMeL-BERT	68.3 $\pm$ 1.0	60.3 $\pm$ 0.3	68.6 $\pm$ 1.1	65.7 $\pm$ 0.8
XLNet-RoBERTa	57.3 $\pm$ 1.0	56.6 $\pm$ 1.1	60.2 $\pm$ 1.2	58.0 $\pm$ 0.5
QARIB	68.1 $\pm$ 1.3	60.2 $\pm$ 0.8	68.1 $\pm$ 0.6	65.4 $\pm$ 0.5
AraRoBERTa-SA	59.7 $\pm$ 1.2	53.5 $\pm$ 0.9	59.7 $\pm$ 0.5	57.6 $\pm$ 0.3
AraBERT v2 twitter _B	68.4 $\pm$ 1.5	60.9 $\pm$ 0.6	69.7 $\pm$ 0.4	66.3 $\pm$ 0.7
AraBERT v2 twitter _L	68.8 $\pm$ 1.1	60.5 $\pm$ 0.3	69.9 $\pm$ 1.0	66.4 $\pm$ 0.5

Table 6: Baseline results on ArImpliSent test set. We report per-class F1 and Macro-F1 for Transformer models averaged over three random seeds (mean $\pm$ std). Best Macro-F1 is in **bold**.

reasoning, sarcasm, and metaphorical language.

6 Conclusion

In this paper, we introduced ArImpliSent, a novel, large-scale dataset for implicit Arabic SA. By employing a rigorous methodology of instance-level lexicon filtering, emotion mapping for class balancing, and manual refinement, we established a high quality benchmark comprising purely natural, implicit sentences. Our baseline experiments reveal that state-of-the-art Arabic language models struggle significantly with implicit reasoning, highlighting the necessity of this dataset for future NLP research. In future work, we plan to explore commonsense augmented modeling approaches to improve performance on this benchmark and to

extend ArImpliSent’s dialectal coverage further.

Limitations

A primary limitation of automated lexicon filtering is lexicon coverage. While we aggregated three distinct lexicons to cover MSA and dialects, the highly morphologically rich nature of Arabic means that some rare explicit morphological variations might have bypassed the filter. Furthermore, interpreting implicit sentiment can sometimes be highly subjective; despite our manual verification, annotators bias regarding what constitutes "implicit sentiment" may be present.

Ethics Statement

ArImpliSent is constructed from previously pub-

Example (Arabic / English)	Gold	Predicted	Primary Trigger
في لقطات من حياتي نفسي احولها لفديو كليب واحتفظ فيها فأسطوانه "I wish I could turn some moments from my life into a music video and keep them on a record."	Positive	Neutral	Situation
هو باين كده ان انا الوحيد الذي خدت الاجازة كلها ع السرير "It seems I'm the only one who spent the entire vacation in bed."	Negative	Positive	Situation
ماشاءله عالاخلاق الحلوه الي بسببها بخسر كل الناس "What wonderful manners—they are the reason I am losing everyone."	Negative	Positive	Sarcasm
كن كالغيمة أثقل ما فيه المطر "Be like the cloud—the heaviest thing it carries is rain."	Positive	Neutral	Metaphor

Table 7: Misclassified Tweets by MARBERTv2 on ArImpIiSent test set, categorized by their primary implicit sentiment trigger.

lished Arabic Twitter sentiment and emotion datasets, including ASAD, ASTD, ArSAS, ArTwitter, ATSAD, SemEval-2017, AJGT, and AETD. Most source datasets are publicly available for research use under their respective authors' terms, while ASAD was obtained directly from its authors for research purposes. We do not collect any new data directly from social media platforms, and no personally identifiable information beyond what was present in the source corpora is introduced during our filtering or annotation process. Manual verification and inter-annotator agreement assessment were conducted by native Arabic-speaking annotators who were informed of the research purpose and compensated fairly for their time.

References

2019. Smp2019-ecisa: Evaluation of chinese implicit sentiment analysis. <https://www.biendata.xyz/competition/smp2019/>. Shared task dataset of the 8th National Conference on Social Media Processing (SMP 2019).
- Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations. *arXiv preprint arXiv:2102.10684*.
- Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. **ARBERT & MARBERT: Deep bidirectional transformers for Arabic**. In *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)*, pages 7088–7105, Online. Association for Computational Linguistics.
- Nawaf A Abdulla, Nizar A Ahmed, Mohammed A Shehab, and Mahmoud Al-Ayyoub. 2013. Arabic sentiment analysis: Lexicon-based and corpus-based. In *2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT)*, pages 1–6. IEEE.
- Kathrein Abu Kwaik, Stergios Chatzikyriakidis, Simon Dobnik, Motaz Saad, and Richard Johansson. 2020. **An Arabic tweets sentiment analysis dataset (AT-SAD) using distant supervision and self training**. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 1–8, Marseille, France. European Language Resource Association.
- Amr Al-Khatib and Samhaa R El-Beltagy. 2017. Emotional tone detection in arabic tweets. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 105–114. Springer.
- Basma Alharbi, Hind Alamro, Manal Alshehri, Zuhair Khayyat, Manal Kalkatawi, Inji Ibrahim Jaber, and Xiangliang Zhang. 2020. Asad: A twitter-based benchmark arabic sentiment analysis dataset. *arXiv preprint arXiv:2011.00578*.
- Khaled Mohammad Alomari, Hatem M ElSherif, and Khaled Shaalan. 2017. Arabic tweets sentimental analysis using machine learning. In *International conference on industrial, engineering and other applications of applied intelligent systems*, pages 602–610. Springer.
- Reem AlYami and Rabah Al-Zaidy. 2022. **Weakly and semi-supervised learning for Arabic text classification using monodialectal language models**. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 260–272, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. **AraBERT: Transformer-based model for Arabic language understanding**. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resources Association.

- Sadik Bessou and Ghazlane Chenni. 2021. Efficient measuring of readability to improve documents accessibility for arabic language learners. *arXiv preprint arXiv:2109.08648*.
- Hanane Boutouta, Abdelaziz Lakhfif, Ferial Senator, and Chahrazed Mediani. 2025a. [Enhancement of implicit emotion recognition in arabic text: Annotated dataset and baseline models](#). *IEEE Access*, 13:165096–165116.
- Hanane Boutouta, Abdelaziz Lakhfif, Ferial Senator, and Chahrazed Mediani. 2025b. [A transformer-based hybrid model for implicit emotion recognition in arabic text](#). *Engineering, Technology and Applied Science Research*, 15(3):23834–23839.
- Samhaa R. El-Beltagy. 2016. [NileULex: A phrase and word level sentiment lexicon for Egyptian and Modern Standard Arabic](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 2900–2905, Portorož, Slovenia. European Language Resources Association (ELRA).
- AbdelRahim A Elmadany, Hamdy Mubarak, and Walid Magdy. 2018. An arabic speech-act and sentiment corpus of tweets. In *The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools*. European Language Resources Association (ELRA).
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.
- Huy Gia Luu and Dang Van Thin. 2025. [Exploring the power of large language models for Vietnamese implicit sentiment analysis](#). In *Proceedings of the 18th International Natural Language Generation Conference*, pages 202–214, Hanoi, Vietnam. Association for Computational Linguistics.
- J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. [SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense](#). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 105–119, Online. Association for Computational Linguistics.
- Saif Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. [Sentiment lexicons for Arabic social media](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 33–37, Portorož, Slovenia. European Language Resources Association (ELRA).
- Mahmoud Nabil, Mohamed Aly, and Amir Atiya. 2015. [ASTD: Arabic sentiment tweets dataset](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519, Lisbon, Portugal. Association for Computational Linguistics.
- Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. [SemEval-2017 task 4: Sentiment analysis in Twitter](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.
- Irene Russo, Tommaso Caselli, and Carlo Strapparava. 2015. [SemEval-2015 task 9: CLIPeval implicit polarity of events](#). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 443–450, Denver, Colorado. Association for Computational Linguistics.
- Motaz Saad. 2015. *Mining documents and sentiments in cross-lingual context*. Ph.D. thesis, Université de Lorraine.
- Orion Weller and Kevin Seppi. 2019. [Humor detection: A transformer gets the last laugh](#). 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)*, pages 3621–3625, Hong Kong, China. Association for Computational Linguistics.
- Diya Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. [Humor recognition and humor anchor extraction](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376, Lisbon, Portugal. Association for Computational Linguistics.
- Deyu Zhou, Jianan Wang, Linhai Zhang, and Yulan He. 2021. [Implicit sentiment analysis with event-centered text representation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6884–6893, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.