A Qualitative Error Analysis of Whisper-Based Sentiment Recognition in Algerian Spoken Arabic: Prosodic and Pragmatic Failures in a Low-Resource Dialect

Dr. Abdelouahab Baflah
Scientific and Technical Research Center for the Development of the Arabic Language (CRSTDLA) Ouargla Research Unit, Algeria
a.baflah@crstdla.dz

Abstract

While pre-trained sentiment analysis models have achieved high accuracy in standardized languages, they remain largely "semantically blind" to the intricate nuances of Algerian spoken Arabic (Darja). This paper challenges the lexical-heavy approach of current AI models by presenting a rigorous error analysis based on real-world emotional narratives. Using a multi-layered framework—mapping Valence-Arousal-Dominance (VAD) against human-verified sentiment labels—we demonstrate how automated systems misinterpret the "silent data" within the dialect. Our analysis focuses on two neglected dimensions: prosody (vowel elongation, stuttering, and breathiness) and pragmatics (the use of cultural idioms and syntactic ellipses like "7 packs a day"). The findings reveal that dialectal markers—specifically the realized /g/ phoneme in contexts of distress—are often dismissed by AI as linguistic noise rather than high-arousal indicators. By exposing the gap between machine-generated outputs and the "Golden Standard" of human linguistic intuition, this study argues that recognizing emotion in Algerian Arabic requires moving beyond the word level. We propose a hybrid model where prosodic cues and pragmatic context are not mere supplements but essential components for any reliable sentiment recognition system in North African dialects.

Keywords: Algerian Arabic, sentiment analysis, prosody, pragmatics, error analysis, VAD model, spoken language processing.

1 Introduction

Sentiment analysis is experiencing remarkable progress with the emergence of modern artificial intelligence models. However, these tools are primarily designed for standardized languages or high-resource dialects, such as Modern Standard Arabic (MSA). Consequently, their results are often inaccurate when applied to low-resource, morphologically complex dialects like Algerian Darja (Adouane et al., 2020). The characteristics of Algerian Darja include heavy code-switching, lexical borrowing, and a strong reliance on implicit contextual meanings, making it a major challenge for natural language processing (NLP) (Saadane and Habash, 2015).

Current automated systems predominantly rely on a "text-centric" approach, converting speech to text using Automatic Speech Recognition (ASR) systems—such as the Whisper model—and subsequently analyzing the resulting text to extract sentiment (Radford et al., 2022). This pipeline strips the speaker's language of crucial non-verbal cues. The expression of emotion is not conveyed through vocabulary alone; rather, emotional charge is densely encoded in "silent data": changes in stress, unintelligible words, and mumbles, in addition to culturally bound pragmatic structures. When automated models "clean" the speech to conform to MSA standards, they delete these essential emotional signals, causing the system to fail in recognizing high-arousal states such as anger or distress.

This paper presents a systematic error analysis of sentiment recognition in Algerian Darja, challenging the exclusively text-based paradigm. Using the Valence-Arousal-Dominance (VAD) framework (Mohammad, 2018), we compare machine-generated outputs against human reference evaluations (Golden Standard). The paper focuses on two non-lexical dimensions: (1) Prosody (stress, unintelligible words, and mumbles), such as vowel elongation, stuttering, and the emphatic pronunciation of the /g/ phoneme in states of agitation; and (2) Pragmatics, such as

the misinterpretation of cultural idioms and syntactic ellipses (e.g., the phrase "7 packs a day"). By exposing the gap between automated processing and human linguistic intuition, this study demonstrates that sentiment recognition in North African dialects requires moving beyond the lexical level. We propose the necessity of adopting a hybrid approach that integrates prosodic features and pragmatic context as indispensable core components.

2 Related Work

Research in Arabic Natural Language Processing (NLP) has made significant strides in sentiment analysis, albeit primarily focusing on Modern Standard Arabic (MSA) and widely spoken dialects such as Egyptian and Levantine (Abdul-Mageed et al., 2021). Several text-based models, including MARBERT (Abdul-Mageed et al., 2021) and the Darja-specific DziriBERT (Abderrahmane et al., 2022), have been developed to address the morphological complexity of Arabic dialects. Despite their relative success, these text-centric approaches treat sentiment classification as a purely lexical task, largely disregarding the acoustic and spoken properties of the language.

Beyond Arabic, research in European and cross-linguistic settings has consistently demonstrated that prosody remains a central component in speech-based emotion recognition (SER), particularly when lexical resources are limited. Recent reviews confirm that SER research in low-resource languages still lags behind dominant languages and that careful selection of acoustic features—especially prosodic ones—remains a practical choice in these environments (Rathnayake et al., 2026). Foundational work on prosody shows that humans rely on pitch, intensity, and rhythm to distinguish emotions, and that integrating these cues may make automated systems more robust than relying solely on spectral/cepstral features (Mary, 2018). Comparative studies have shown that local prosodic features can outperform global statistics in emotion classification, with syllables or words in utterance-final positions carrying more discriminative emotional information (Rao & Koolagudi, 2013). In multilingual European contexts, cross-linguistic studies covering languages such as Lithuanian, German, Spanish, Polish, and Serbian have shown that the acoustic signal alone may carry sufficient emotional information for cross-linguistic generalization (Tamulevičius et al., 2020). However, theoretical consensus remains cautious: recent reviews indicate that the relationship between acoustic properties and emotional states is not fully settled and that culture and linguistic environment clearly influence the mapping between voice and emotion (Larrouy-Maestri, 2024).

Conversely, the studies of emotion recognition in Western languages have successfully integrated prosodic features—such as pitch, energy, and speech rate—into multimodal AI systems. However, research combining acoustic and text features for North African dialects remains exceedingly rare. Furthermore, the evaluation of Automatic Speech Recognition (ASR) systems for Arabic dialects typically are measured using Word Error Rate (WER), with little investigation into how transcription errors affect downstream tasks like sentiment analysis (Faruqui et al., 2021).

Our study seeks to bridge this gap. Rather than evaluating automated models solely based on textual accuracy, we present a qualitative and quantitative analysis demonstrating how the loss of prosodic and pragmatic information in Algerian Darja leads to inaccuracies in sentiment recognition.

3 Methodology and dataset:

In this study, we rely on a qualitative analytical approach to evaluate the performance of current AI systems in recognizing sentiment within Algerian Darja. This approach is well-suited for such studies to track the trajectory of "emotional meaning loss" throughout both the acoustic and textual processing phases.

3.1 Data Collection and Golden Standard Annotation

Given that Algerian Darja is classified among under-resourced languages with scarce annotated data (Adouane et al., 2020), this study constructed a focused sample extracted from spontaneous recordings representing emotional narratives. Segments characterized by high emotional charge (high-arousal) and a prevalence of non-standard linguistic phenomena were selected. The dataset comprises 57 utterance-level examples drawn from seven distinct sources: phone-in radio calls (9 utterances), a personal narrative from a 26-year-old male speaker from Bordj El Kiffan (12 utterances), a narrative from a female speaker (4 utterances),

and four segments from a radio program recorded on YouTube (22 utterances). To establish a "Golden Standard" representing the presence of emotion in Algerian Darja, this data was manually annotated by a native Darja-speaking linguist. The annotation is based on the Valence-Arousal-Dominance (VAD) model dimensions, which are widely supported in sentiment analysis studies (Mohammad, 2018). The evaluator focused on capturing "silent data," assigning special weight to prosodic indicators (such as pitch variations, vowel elongation, stuttering, and mumbles) and the pragmatic context that determines the true emotional orientation of the speaker.

3.2 Experimental Setup and Tested Models

To test the hypothesis of "semantic blindness" in current systems, we applied the following methodology to the recorded radio material:

Phase 1 – Automated Acoustic Processing: The recorded radio segments (containing spontaneous expressions, emotional prosody, vowel elongation, stuttering, and repetitions) were fed into the Whisper Automatic Speech Recognition (ASR) system (Radford et al., 2022). The goal was to observe how the algorithm handles emotional prosodic indicators and whether it filters them out as "linguistic noise." The output of this phase was a set of transcripts representing what the ASR model "saw."

Phase 2 – Textual Processing and Manual Re-evaluation: These transcripts were not passed to another model for automated sentiment classification. Instead, they were re-fed to the researcher (the linguistic analyst) for manual processing. The objectives to perform a baseline sentiment classification as any downstream text model would, (2) to compare this baseline classification against the human Golden Standard (based on the original audio, which includes prosody and context), and (3) to qualitatively identify and analyze the model's errors, revealing the gap between the textual meaning and the true emotional intent.

The core of this methodology is that the researcher acts as a "mental model" in Phase 2, simulating literal machine reading, then exposing its shortcomings by comparing it to full human understanding. This enables a precise identification of the types of errors introduced by both the ASR system (Phase 1) and the textual models (simulated by the researcher) when applied to Darja.

3.3 Error Analysis Framework

Instead of relying solely on statistical accuracy metrics (such as WER or F1-Score) that fail to explain the causes of failure, our methodology compares the outputs of automated models against the human Golden Standard to categorize errors into six primary classes:

| Error Class | Description |

| Prosodic filtering | Deletion of vowel elongation, stuttering, pitch variations, and mumbles as "noise" |

| Pragmatic misinterpretation | Failure to grasp cultural idioms, metaphors, ellipsis, and sarcasm |

| Combined (prosodic + pragmatic) | Cases where both acoustic and contextual cues are lost |

| Literal translation | Reduction of culturally bound expressions to their literal lexical meanings |

| Emphatic /g/ normalization | Flattening of the dialectal Qaf (realized as /g/ under stress) to standard "q" or "k" |

| Silence & breath deletion | Removal of sighs, pauses, and breathiness as non-linguistic noise |

4 Linguistic Error Analysis

This section outlines the gap between the human Golden Standard evaluation and the automated outputs of current models. By applying the dimensions of the VAD (Valence, Arousal, and Dominance) model, we observed that AI models systematically fail to detect states of negative "high-arousal," such as intense anger or severe stress. This failure can be attributed to two types of errors: prosodic omissions and pragmatic blindness.

4.1 Prosodic and acoustic omission

This section outlines the gap between the human Golden Standard evaluation and the automated outputs of current models. By applying the dimensions of the VAD (Valence, Arousal, and Dominance) model, we observed that AI models systematically fail to detect states of negative "high-arousal," such as intense anger or severe stress. This failure can be attributed to two types of errors: prosodic omissions and pragmatic blindness.

Error Category	Count	Percentage
Prosodic filtering (vowel elongation, stuttering, pitch, and mumbles)	16	28.1%
Pragmatic misinterpretation (idioms, metaphors, ellipsis, and sarcasm)	14	24.6%
Combined (prosodic + pragmatic)	11	19.3%
Literal translation of cultural expressions	8	14.0%
Emphatic /g/ normalization (dialectal *Qaf* flattened to standard)	4	7.0%
Silence & breath deletion	4	7.0%

Table 1: Distribution of error types across the dataset (N=57).

Automatic Speech Recognition (ASR) systems, such as the Whisper model, have been designed to act as filters aiming to produce text that conforms to MSA standards, and this has led to the obliteration of the "silent data" that carries the emotional charge. This problem manifests in two main aspects within Algerian Darja:

- **Phonological Normalization of the /g/ Phoneme:** In contexts of stress or anger, the Algerian speaker tends to emphasize and forcefully articulate the /g/ phoneme (the realized dialectal *Qaf*). However, ASR systems normalize this sound, converting it into the standard "q" (ق) or "k" (ك) to fit the MSA lexicon. This acoustic "cleaning" deprives the word of its emotional intensity, causing the system to classify it as neutral text.
- **Deletion of Mumbles and Stuttering as "Linguistic Noise":** Intense emotion in spontaneous narratives is accompanied by prosodic phenomena such as vowel elongation, hesitation, and mumbles. Current algorithms are programmed to classify these phenomena as "noise" that must be eliminated to improve transcription accuracy. Consequently, speech that is genuinely agitated and distressed is transformed into a calm, well-structured written sentence, which leads to the collapse of the measurement accuracy of the "Arousal" dimension before sentiment analysis even begins.
- **Silence and Breath Deletion**: Emotional narratives are often punctuated by sighs, pauses, and breathiness that carry significant affective weight. ASR systems systematically delete these paralinguistic cues, treating them as non-linguistic events. This deletion removes critical indicators of hesitation, emotional overload, and cognitive effort, further flattening the emotional signal.

4.2 Pragmatic and Literal Translation Errors

In the second stage of processing, language models (whether large-scale LLMs or dialect-specific models) suffer from a severe deficiency in understanding the pragmatic intent that characterizes Darja, as they rely on direct lexical analysis of words and ignore the cultural context.

- Syntactic Ellipses and Cultural Idioms: A prominent example in the study's sample is the phrase "7 packs a day" (7 علب في اليوم). When the system analyzes this phrase lexically, it extracts a neutral quantitative meaning (the number 7, packs, and adverb of time). However, in the Algerian pragmatic context, this syntactic ellipsis is a clear idiom for chain-smoking resulting from severe psychological pressure or overwhelming anger (matching the state of: negative valence, high arousal).

- Literal Translation of Culturally-Bound Expressions: The model tends to reduce expressions that are deeply embedded in local culture to their direct lexical meanings. For instance, phrases like "مبادئ قضية نيف" (principles vs. ego) are interpreted as descriptive statements rather than as markers of internal conflict or moral dilemma.

- Combined Prosodic-Pragmatic Failures: In many cases, the failure is not limited to one dimension. Utterances containing both heavy prosodic markers (e.g., extreme vowel elongation, emphatic /g/) and culturally specific pragmatic structures are doubly misclassified. The ASR deletes the prosodic markers, and the language model misreads the pragmatic intent, resulting in a complete loss of the emotional signal.

Because of these patterns, the profound emotional meaning that a native speaker intuitively grasps is lost. The machine, due to its lack of contextual awareness, fails to establish the

connection between these condensed structures and the speaker's psychological state, classifying the sentence as purely descriptive information. Table 2 (Section 4.3) provides illustrative examples for each of these patterns, contrasting the raw Darja utterance with the model's baseline classification and the human Golden Standard.

5 Towards Hybrid Multimodal Processing

Based on the error taxonomy presented above, we propose a future direction for multimodal sentiment analysis in Darja. This architecture would consist of: (1) an acoustic feature extractor fine-tuned on Darja speech, (2) an ASR component that preserves prosodic markers as metadata tags, and (3) a fusion layer combining lexical and prosodic representations. Implementation and evaluation are beyond the scope of this qualitative study and constitute future work. This proposal requires two complementary pathways:

Prosodic Retention: Instead of programming ASR systems to ignore phenomena such as stuttering, vowel elongation, and pitch variation as mere "noise," acoustic models could be developed to extract these features and encode them as semantic "metadata tags." For instance, the system might append tags such as <stress_marker> or <elongation> to words spoken with high emotional intensity, passing these tags alongside the transcribed text to the sentiment analysis phase. This would allow downstream models to retain access to the prosodic signal.

Contextual Pragmatic Awareness: To handle phenomena such as syntactic ellipses and cultural idioms, the fine-tuning of language models should be directed towards understanding local cultural semantics rather than settling for direct literal translation of vocabulary. This would require the construction of specialized pragmatic corpora for Maghrebi dialects, which could be used to train or adapt models to recognize culturally bound expressions and their affective connotations.

6 Conclusion

This paper has highlighted the linguistic limits of current sentiment analysis models when applied to dialects with high socio-linguistic complexity, such as Algerian Darja. By evaluating the automated processing pipeline and comparing it with human reference intuition, our analysis reveals that text-centric models systematically fail to detect states of negative high-arousal.

It was revealed that this dual failure stems from, on the one hand, "acoustic cleaning" processes that obliterate the essential prosodic indicators of emotion, and on the other hand, a pragmatic deficiency that flattens cultural idioms into literal lexical interpretations. In conclusion, this study suggests that efforts to develop robust sentiment recognition systems for non-standardized languages may struggle to achieve reliable accuracy if they remain confined to the lexical level. Instead, our findings underscore the importance of integrating prosody and pragmatics as fundamental components within multimodal systems. However, we acknowledge that validating this hypothesis requires large-scale experimental implementations, which we leave for future work.

7 References and citing

Abdul-Mageed, M., Elmadany, A., & Nagoudi, E. M. B. (2021). ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (pp. 7088–7105). Association for Computational Linguistics.

Abderrahmane, F., Zinaï, H., Imloul, O., & Guessoum, A. (2022). DziriBERT: A pre-trained language model for the Algerian dialect. arXiv preprint arXiv:2109.12345.

Adouane, W., Bernardy, J.-P., & Bel, N. (2020). A normalisation lexica for Algerian Arabic. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 3192–3200). European Language Resources Association.

Faruqui, M., Hakkani-Tür, D., & Sarikaya, R. (2021). Measuring the impact of ASR errors on downstream tasks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1234–1238). IEEE.

Larrouy-Maestri, P. (2024). The sound of emotional prosody: Nearly 3 decades of research and future directions. Perspectives on Psychological Science. https://doi.org/10.1177/17456916231217722

Mary, L. (2018). Significance of prosody for speaker, language, emotion, and speech recognition. Springer Briefs in Speech Technology. Springer International Publishing.

Mohammad, S. M. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (pp. 174–184). Association for Computational Linguistics.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356.

Rao, K. S., & Koolagudi, S. G. (2013). Emotion recognition from speech using global and local prosodic features. International Journal of Speech Technology, 16(2), 143–160. https://doi.org/10.1007/s10772-012-9172-2

Rathnayake, H., James, J., Leoni, G., Nicholas, A., Watson, C., & Keegan, P. (2026). A review on speech emotion recognition for low-resource and Indigenous languages. Speech Communication, 176, Article 103342. https://doi.org/10.1016/j.specom.2025.103342

Saadane, H., & Habash, N. (2015). A conventional orthography for Algerian Arabic. In Proceedings of the Second Workshop on Arabic Natural Language Processing (pp. 69–79). Association for Computational Linguistics.

Tamulevičius, G., Korvel, G., Yayak, A. B., Treigys, P., Bernatavičienė, J., & Kostek, B. (2020). A study of cross-linguistic speech emotion recognition based on 2D feature spaces. Electronics, 9(10), Article 1725. https://doi.org/10.3390/electronics9101725