# From OCR to Content Interpretation: Towards a Scalable Workflow for Arabic Literature in the Digital Humanities

Maura Tarquini

Università degli Studi di Sassari
(mtarquini@uniss.it)

Elisa Gugliotta

Università degli Studi di Sassari, Université Grenoble Alpes
(egugliotta@uniss.it)

**Abstract**

This article explores how Arabic literary texts can function as epistemic devices in Digital Humanities when computational methods are aligned with qualitative interpretation. After situating Arabic DH within global digitization efforts and pedagogical frameworks, we present an OCR-to-analysis workflow tailored to Arabic, comparing Tesseract and Qari-OCR. We also discuss the potential and limitations of applying automatic emotion recognition to literary texts, underscoring the need for a qualitative reading of al-Aswānī's *Awrāq ʿIṣām ʿAbd al-ʿĀṭī*, which highlights anger, hatred, and sadness as dominant emotions shaping a narrative of alienation and critique. The study demonstrates how digitization, analysis, and interpretation converge to enrich research and pedagogy within our project.

**Parole chiave**

Digital Humanities; Arabic Literature; OCR

**DOI**

https://doi.org/10.58015/2036-2293/806

## 1. Introduction

In recent decades, Digital Humanities (DH) have evolved from the simple application of computational techniques into an autonomous theoretical domain that redefines the nature of humanistic knowledge[1]. Central to this shift is Moretti's notion of *distant reading*[2], which extends literary analysis beyond individual texts to systemic patterns and statistical distributions[3]. Meanwhile, large-scale digitization initiatives such as Project Gutenberg (1971), Internet Archive (1996), Perseus Digital Library, HathiTrust, and Europeana[4] have made vast text corpora accessible, democratizing knowledge and opening new analytical perspectives[5]. In the Western context, these developments, supported by strong institutional investment[6] and a mature technological ecosystem, have established a genuine computational paradigm encompassing entire cultural systems[7].

Arabic literature, however, occupies a different position. The scarcity of structured digital corpora and the intrinsic challenges of Arabic computational processing have long confined DH practices to a conservative role. Only recently have projects such as Alwaraq, the Qatar Digital Library, OpenITI, and the Shamela Library begun to bridge the Arabist field with global DH tools, adapting timelines and methods to linguistic specificities. Yet, the consonantal script, morphology, and orthographic ambiguity continue to complicate Arabic NLP. The absence of robust digital infrastructures further constrains the development of innovative didactic practices and the integration of computational methodologies into Arabic language and literary studies.

The potential of DH applied to Arabic literary texts is significant: literature, as an identity, political, and emotional space on a macroscopic scale[8] reveals dynamics often imperceptible through traditional reading. We argue here that the study of the Arabic language cannot be separated from its literature; relying solely on journalistic or media texts offers an incomplete education, unable to foster the broader social transformations that language learning entails. The teaching of literature in foreign language pedagogy has evolved from the grammar-translation method[9], centered on classical texts and rule-based decoding, to the communicative turn initiated by Chomsky[10] and formalized by Hymes[11], who defined language as an instrument of authentic interaction. Later, Krashen[12] highlighted the role of socio-affective factors in creating a motivating environment for linguistic acquisition. Within this framework, literary texts serve as authentic linguistic and cultural resources, engaging learners emotionally and intellectually, and fostering reflection and intercultural empathy[13].

---

[1] David Berry, *Understanding digital humanities*, Basingstoke, Palgrave Macmillan, 2012.
[2] Franco Moretti, *Distant reading*, London, Verso, 2013.
[3] Franco Moretti, *Graphs, maps, trees: Abstract models for a literary history*, London, Verso, 2007.
[4] Europeana was launched in 2008 as an initiative of the European Union to digitize and make cultural and literary heritage accessible.
[5] Ted Underwood, *Distant horizons: Digital evidence and literary change*, Chicago, University of Chicago Press, 2019.
[6] See footnote 4.
[7] Ted Underwood, *Distant horizons: Digital evidence and literary change*, Chicago, University of Chicago Press, 2019.
[8] Ivi, pp. 155-160.
[9] Emanuele Borello, *La traduzione nella storia della glottodidattica*, «I saperi del tradurre. Analogie, affinità, confronti», a cura di Claudio Montella, Giovanni Marchesi, (pp. 147–172), Milano, FrancoAngeli, 2007.
[10] Noam Chomsky, 1957, *Syntactic Structures*, The Hague, Mouton, 1957.
[11] Dell Hymes, *On communicative competence*, «Sociolinguistics: Selected readings», John Baptiste Pride, Janet Holmes, (pp. 269–293). Harmondsworth, Penguin, 1972.
[12] Stephen Krashen, *Principles and Practice in Second Language Acquisition*, Oxford, Pergamon Press, 1982.
[13] Louise Rosenblatt, *The Reader, the Text, the Poem: The Transactional Theory of the Literary Work*, Carbondale (IL), Southern Illinois University Press, 1978.

Literary texts possess a holistic character that transcends mere aesthetic enjoyment, offering linguistic input that stimulates imagination and fosters forms of interpretive collaboration[14]. Literature is, in fact, an "ally to language"[15], as it engages learners on both affective and cognitive levels, promoting processes of intercultural empathy that facilitate the understanding of the Other and, consequently, linguistic acquisition. Ultimately, literature stimulates critical thinking, creativity, and intercultural awareness, enabling students to experience language as a social and emotional practice, enhancing motivation and the ability to reflect on communicative processes[16].

The main objective of our project[17] is to select excerpts from works of Arabic literature in order to develop digital didactic materials. We begin by presenting a state-of-the-art overview of Arabic literary studies within the DH context (§2.1). We then describe recent advances in the digital tools we have employed, including Arabic Optical Character Recognition (OCR, §2.2) and Automatic Emotion Recognition in Arabic written texts (AER, §2.3). As a case study, we digitized ʿAlāʾ al-Aswānī's short story *Awrāq ʿIṣām ʿAbd al-ʿĀṭī* (*The Notebooks of ʿIṣām ʿAbd al-ʿĀṭī*) from *Nirān Ṣadīqa* (*Friendly Fires*)[18] using OCR, within a broader corpus-building workflow (§3). Digitalization serves humanistic inquiry by combining quantitative analysis (§4.1) with qualitative interpretation of emotional and narrative dimensions (§4.2), outlining a model where DH tools enrich both critical research and pedagogy in modern Arabic language and literature (§5).

## 2. State of Art

### *2.1 Digital Humanities and Arabic Literature: Theoretical Paradigms and Applications*

The theoretical debate on DH reached a turning point with Underwood's metaphor of *distant horizons*[19], which invites us to consider the literary as an integrated system. The digital shift enables a move from isolated texts to macro-ecosystems, integrating *close* and *distant reading* within a hybrid approach where quantitative data support new critical interpretations[20]. Along this line, Eve[21] reframes computational methods as complementary to traditional hermeneutics, emphasizing their ethical and interpretive potential[22]. A paradigmatic example of this integration is the project *The Computational Study of Culture: Cultural Analytics for Modern Arab and Muslim Studies* (2018)[23] which, through a corpus of 2.3 billion words, reconstructed intellectual networks and discursive patterns in Arabic modernity, showing how literary and cultural texts have contributed to processes of identity formation[24]. These results, framed within Fiormonte, Numerico, and Tomasi's notion of DH as "the translation of humanistic questions into computational form"[25], show that texts are not mere data but complex cultural objects

[14] Mary Lou McCloskey, Lynn Stack, *Voices in literature: Integrated language and literature for ESOL*, Boston, Heinle & Heinle, 1993, p. vii.
[15] Cristopher John Brumfit, Ronald Alan, Carter, *Literature and language teaching*. Oxford, Oxford University Press, 1986, p. 1.
[16] Nadezhda Daskalovska, Violeta Dimova, Why should literature be used in the language classroom?, *Procedia – Social and Behavioral Sciences, 46*, 2012, 1182–1186, p. 1182. <https://doi.org/10.1016/j.sbspro.2012.05.271> (Accessed: Nov. 10, 2025)
[17] Project-dedicated GitHub repository name anonymized for review purposes.
[18] ʿAlāʾ al-Aswānī, *Nirān Ṣadīqa*, al-Qāhira: Dār al-Šurūq, 2007.
[19] Ted Underwood, *Distant horizons: Digital evidence and literary change*, Chicago, University of Chicago Press, 2019.
[20] Ivi, pp. 1-5.
[21] Martin Paul Eve, *The Digital Humanities and Literary Studies*, Oxford, Oxford University Press 2022.
[22] Ivi, pp. 3-6.
[23] The research team composed of Eid Mohamed, Umar Ryad, Emad Mohamed, Talaat Mohamed, Raheem Sarwar, Mai Zaki, and Michael Frishkopf has produced publications and results up to 2024.
[24] Eid Mohamed, *The potential and limits of Arabic digital humanities*, «Journal of Cultural Analytics», 9(3), 2024, pp. 1-11 <https://doi.org/10.22148/001c.116818> (Accessed: Aug. 21, 2025).

whose digital representation generates new critical and ethical perspectives[26], that acquire special significance when dealing with the fluid and context-bound nature of Arabic literary categorizations.

With the French conquest of Algeria in 1830 and the subsequent phase of colonial expansion, Arabic entered European university curricula[27]. This early stage, shaped by Orientalist paradigms, remained confined to textual decoding and structural analysis. By contrast, the twenty-first century has witnessed the emergence of a renewed approach grounded in communicative competence, reflecting the diversity of real-world language use. Such interactional dynamics extend beyond grammar and lexis, encompassing behavioral codes, social values, and systems of meaning continuously negotiated within Arab culture[28] since "society shapes language just as language shapes society"[29]. In recent decades, the integration of literature within communicative approach has grown significantly, with increasing emphasis on the use of authentic materials that expose learners to the sociocultural dynamics of the Arab world[30]. The emotional charge of literary narratives facilitates motivation, linking linguistic awareness with empathy and personal engagement[31]. Embodying "beautiful ideas described in beautiful words"[32], Arabic literature engages learners both aesthetically and intellectually.

DH studies applied to the study of Arabic language open new perspectives for didactic experimentation, where digital corpora and literary sensitivity jointly contribute to the understanding of language as lived practice. Corpora of naturally-occurring texts provide samples of genuine language, since they are produced by speakers and writers with real communicative goals[33]. In this sense, literature emerges as an epistemic bridge, integrating linguistic and cultural skills while promoting critical reflection and engagement. Our project aims to integrate Arabic literature into an accessible digital ecosystem, interpreting digital tools as devices that enhance both textual research practices and modes of didactic transmission.

This project follows a clear methodological path, viewing markup and text classification as interpretive acts that honor Arabic literary tradition. It adopts a hybrid approach combining distant and close reading: large-scale corpus analysis uncovers linguistic, stylistic, and thematic patterns, while textual annotation enables close, philologically grounded readings valuable for both research and pedagogy. In this sense, the project embodies Eve's hybrid and pragmatic vision of DH, where quantitative analysis informs qualitative interpretation, and close reading is enriched by the broader digital context. For Arabic literature in the digital age, this approach connects micro-textual analysis with long-term historical, cultural, and social dynamics, maintaining both interpretive depth and critical scope.

---

[25] Domenico Fiormonte, Teresa Numerico, Francesca Tomasi, *The digital humanist: A critical inquiry*. New York, Punctum Books, 2015, p.12.

[26] Ivi, pp. 73-83.

[27] Giuliano Lancioni, *Insegnamento dell'arabo e certificazione: una panoramica*, «*Didattica dell'arabo e certificazione linguistica*», a cura di Giuliano Lancioni, Cristina Solimando, (pp. 11–30), Roma, Roma TrE-Press, 2018.

[28] Fawaz Omari, "The Culture of Language as a Means or an End: A Comparative Look at Teaching Arabic for Speakers of Other Languages", *Dirasat: Human and Social Sciences*, vol. 39, no. 2, Deanship of Scientific Research, University of Jordan, 2012.

[29] Ivi, p. 395.

[30] Kassem M. Wahba, "Arabic Language Use and the Educated Language User", in *Handbook for Arabic Language Teaching Professionals in the 21st Century*, a cura di Kassem M. Wahba, Zeinab A. Taha e Liz England. New York–London, Routledge, 2010, p. 140.

[31] María Pilar Agustín Llach, "Teaching Language through Literature: *The Waste Land* in the ESL Classroom", *Odisea*, no. 8, 2007, p. 10.

[32] Rafik Toʾima, Mohammad Mannaʾ, Teaching Arabic and Religion between Science and Art. Cairo, Dar al-Fikr al-ʿArabi, 2001, p. 19.

[33] Laura Gavioli, Guy Aston, *Enriching reality: Language corpora in language pedagogy*. «*ELT Journal*», *2001, 55*(3), pp. 238–246. <https://doi.org/10.1093/elt/55.3.238> (Accessed: Aug. 15 2022), p. 240.

Within this methodological and theoretical framework, our project applies its hybrid model to a concrete case study: the digitization and analysis of ʿAlāʾ al-Aswānī's short story *Awrāq ʿIṣām ʿAbd al-ʿĀṭī*. The text was chosen especially for its sociopolitical resonance, which allows the integration of digital and hermeneutic approaches in exploring contemporary Egyptian subjectivity. ʿAlāʾ al-Aswānī's global prestige illustrates Underwood's *long arc of prestige*, where literary recognition endures across transnational networks of readership and critique[34]. The narrative centers on ʿIṣām's introspective monologues, which explore the disillusionment and moral decay of Egyptian society during the transition from ʿAbd al-Nāṣir to Sadāt. Caught between private and public crises, ʿIṣām embodies the individual's condition amid political upheaval, revealing the tension between historical memory and subjective experience. The story transforms collective faith in state authority (central to Nasserist discourse) into a sense of betrayal and impotence, turning negative emotions into critical awareness, while positive emotions are confined to the depiction of imaginary or hypothetical realities.

### *2.2 Available Tools for Arabic OCR*

Research on Optical Character Recognition (OCR) for Arabic has made significant progress, yet it remains constrained by ligatures, diacritics, and right-to-left directionality. Recent studies show that average accuracy rarely achieves optimal levels in the presence of complex layouts or vocalization marks[35], while in terms of linguistic post-correction they demonstrate that only a drastic reduction of noise ensures tangible benefits in retrieval tasks[36]. Tesseract[37] remains the most accessible and widely used system, and is therefore often adopted as a baseline in DH. However, its limitations for Arabic are well known[38]. Alongside Tesseract, the DH community has experimented with more specialized tools such as Kraken[39], which allows model training on domain-specific corpora. Nevertheless, a true paradigm shift has been introduced by Multimodal Large Language Models (MLLMs). Benchmarks such as KITAB-Bench[40] show that MLLMs can reduce the Character Error Rate (CER) by up to 60% compared with traditional OCR. Yet, the generalist design of these models does not fully meet the high-fidelity requirements of Arabic OCR, particularly regarding the rendering of diacritics and typographic variants. Against this backdrop, among the most recently released models specifically targeting Arabic is Qari-OCR[41], which takes into account not only linguistic features such as diacritics, but also typographic variation across fonts.

---

[34] Ted Underwood, *Distant horizons: Digital evidence and literary change*, Chicago, University of Chicago Press, 2019, pp. 68-100.

[35] Mahmoud S. Kasem, Mohamed Mahmoud, Hyun-Soo Kang, *Advancements and Challenges in Arabic Optical Character Recognition: A Comprehensive Survey*, «ArXiv», 2023, <https://doi.org/10.48550/arXiv.2312.11812> (Accessed: Aug. 24, 2025); Safiullah Faizullah et al., *A Survey of OCR in Arabic Language: Applications, Techniques, and Challenges*, «Applied Sciemces», 2023, 13, 7. <https://doi.org/10.3390/app13074584> (Accessed: Aug. 24, 2025).

[36] Walid Magdy, Kareem Darwish, *Arabic OCR error correction using character segment correction, language modeling, and shallow morphology*, in *Proceedings of EMNLP2006*, Sydney, July, Eds. Jurafsky, Gaussier, ACL, 2006; Walid Magdy, Kareem Darwish, *Effect of OCR error correction on Arabic retrieval*, «Information Retrieval», 11, 5, 2008, pp. 405-425.

[37] Ray Smith, *An overview of the Tesseract OCR engine*. In *ICDAR2007* [2007], vol. II, Curitiba, IEEE, 2007, pp. 629-633.

[38] See footnote 1.

[39] Benjamin Kiessling, *Kraken - A Universal Text Recognizer for the Humanities*, DH2019, Jul 2019, Utrecht. <https://hal.science/hal-04936936/> (Accessed: Aug. 24, 2025).

[40] Ahmed Heakl et al., *KITAB-Bench: A Comprehensive Multi-Domain Benchmark for Arabic OCR and Document Understanding*. In *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, ACL, pp. 22006–22024.

[41] Ahmed Wasfy et al., *QARI-OCR: High-Fidelity Arabic Text Recognition through Multimodal Large Language Model Adaptation*, «ArXiv», 2025, <https://doi.org/10.48550/arXiv.2506.02295> (Accessed: Aug. 24, 2025).

For these reasons, we have decided to test Qari and compare its performance with the one of Tesseract in the digitization process of *Awrāq ʿIṣām ʿAbd al-ʿĀṭī*.

*2.3 Automatic Emotion Recognition in Arabic Written Data*

Among the main challenges in Automatic Emotion Recognition (AER) for Arabic texts is the lack of large, finely annotated datasets suitable for the fine-tuning of pretrained language models[42]. One notable example is the *Multilingual GoEmotions Classifier*[43], a fine-tuning of the multilingual BERT base model[44] on the *Multilingual GoEmotions* dataset[45], which is a multilingual extension of *GoEmotions*[46], trained to classify text into 27 distinct emotion categories plus a neutral class. To our knowledge, this remains the only available model that includes Arabic and provides a more fine-grained classification than Ekman-inspired systems[47]. Within the latter category, the *bert-base-arabic-emotion-analysis-v2* model[48], a fine-tuning of the pretrained Arabic BERT base model[49] on the *Emotone_ar* dataset[50], is taken in this study as a reference model. Another available model, the *bert-base-arabic-finetuned-emotion*[51], classifies Arabic texts into a set of labels largely consistent with Ekman's framework.
Given the limited availability of such models, qualitative analyses of model outputs remain essential to assess the reliability and cultural adequacy of automatic emotion classification in Arabic.

**3. Data**

Our corpus consists of 19,684 words (1,934 sentences), the majority of which correspond to the lexicon of MSA, while only about 5% of the vocabulary can be considered specific to Egyptian Arabic (see Table 1).

| | N. of words | Percentage |
|---|---|---|
| MSA words | 18,698 | 94.99% |

[42] Yimei Xu et al., *Improving Arabic Multi-Label Emotion Classification using Stacked Embeddings and Hybrid Loss Function*, pre-print of IEEE Access, 2025, <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=11037678>.
[43] The model is available on Hugging Face, at the following link: <https://huggingface.co/AnasAlokla/multilingual_go_emotions> (Accessed: Nov. 7, 2025).
[44] Jacob Devlin et al., *Bert: Pre-training of deep bidirectional transformers for language understanding*, in *Proceedings of NAACL-HLT 2019*, Minneapolis, June, Ed. ACL, 2019, pp. 4171-4186.
[45] The dataset is available on Hugging Face, at the following link: <https://huggingface.co/datasets/AnasAlokla/multilingual_go_emotions> (Accessed: Nov. 7, 2025).
[46] Dorottya Demszky, et al., *GoEmotions: A Dataset of Fine-Grained Emotions*, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. July, Ed. ACL, 2020, pp. 4040–4054.
[47] Paul Ekman, *An argument for basic emotions*, «*Cognition and Emotion*, *6*(3–4)», pp. 169–200, 1992, <https://doi.org/10.1080/02699939208411068>, (Accessed: Aug. 21, 2025).
[48] The model is available at: <https://huggingface.co/alpcansoydas/bert-base-arabic-emotion-analysis-v2> (Accessed: Nov. 7, 2025).
[49] Ali Safaya, Abdullatif Moutasem, Yuret Deniz, *KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media*, in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Barcelona, ACL, pp. 2054-2059.
[50] Amr Al-Khatib, Samhaa El-Beltagy, *Emotional tone detection in arabic tweets.* In *International Conference on Computational Linguistics and Intelligent Text Processing,* Springer International Publishing, 2017, pp. 105-114.
[51] Hatem Noaman, *Improved Emotion Detection Framework for Arabic Text using Transformer Models,* «Advanced Engineering Technology and Application» 12, 2, 2023, pp. 1-11. The model is available at: <https://huggingface.co/hatemnoaman/bert-base-arabic-finetuned-emotion> (Accessed: Nov. 7, 2025).

| | | |
|---|---|---|
| EGY words | 986 | 5.01% |
| Total | 19,684 | 100% |

*Table 1: Corpus numbers*

Regarding Egyptian Arabic, while it is often classified among the so-called "dialects", generally described in the literature as low-resource languages[52], it nevertheless stands out as a variety that is relatively rich in data and resources: Egypt has historically been the center of both cinematic and literary production in colloquial Arabic[53].

Tesseract was adopted as a baseline, and Qari as a representative of new multimodal architectures. Their performance is evaluated through standard metrics: for Character and Word Error Rates (respectively, CER and WER).

### *3.1 Token-Based Language Identification*

To gauge OCR performance on colloquial input, tokens were classified as MSA or EGY. A first test with the CAMeL-tools[54] diatopic identifier, forced into binary (MSA vs EGY). However, this approach yielded distorted results: the text appeared to be 85% Egyptian and only 15% MSA, which is implausible and reflects the bias introduced by forcing a binary choice on the model, which is in fact trained to distinguish city varieties (e.g., CAI for Cairene, ALX for Alexandrian, ASW for Aswan, etc.). In the top-k probabilities, both MSA and the EGY classes often appear as marginal, while other varieties prevail, as reported in Table 2.

| Token | Binarized Output | Prediction scores |
|---|---|---|
| هذه | MSA | **MSA**:0.46, MUS:0.41, SAN:0.09, RIY:0.01, KHA:0, JED:0, DOH:0, BAG:0, BAS:0, ALG:0, SAL:0, DAM:0, AMM:0, ALE:0, JER:0, CAI:0, BEI:0, **ASW**:0, **ALX**:0 |
| المجموعة | EGY | SAL:0.20, JER:0.18, JED:0.13, AMM:0.11, MUS:0.10, RIY:0.10, DAM:0.06, DOH:0.05, BAS:0, ALE:0, KHA:0, BEI:0, BAG:0, **CAI**:0, ALG:0, SAN:0, **ASW**:0, **ALX**:0, **MSA**:0 |
| لها | EGY | RIY:0.32, DOH:0.22, JED:0.20, SAL:0.06, MUS:0, DAM:0, SAN:0, AMM:0, JER:0, BAS:0, ALE:0, KHA:0, BAG:0, **CAI**:0, BEI:0, **ASW**:0, ALG:0, **ALX**:0, **MSA**:0 |
| حكاية | EGY | SAL:0.04, RIY:0.01, JER:0.01, DAM:0, JED:0, DOH:0, AMM:0, ALG:0, MUS:0, ALE:0, BAS:0, BEI:0, KHA:0, BAG:0, **CAI**:0, SAN:0, **ASW**:0, **ALX**:0, **MSA**:0 |
| غريبة | EGY | DAM:0.23, RIY:0.17, SAL:0.14, DOH:0.11, JER:0.07, JED:0.06, BAS:0.04, ALE:0.04, MUS:0.03, AMM:0.02, BEI:0.02, KHA:0, BAG:0, SAN:0, **MSA**:0, **CAI**:0, ALG:0, **ASW**:0, **ALX**:0 |
| فقد | EGY | RIY:0.59, MUS:0.26, SAL:0.03, JED:0.03, KHA:0.02, DOH:0.01, BAG:0, JER:0, BAS:0, AMM:0, DAM:0, SAN:0, ALE:0, **MSA**:0, **CAI**:0, ALG:0, BEI:0, **ASW**:0, **ALX**:0 |
| فرغت | EGY | SAL:0.22, JED:0.22, JER:0.14, RIY:0.11, DAM:0.07, DOH:0.06, AMM:0.04, MUS:0.03, ALE:0.01, KHA:0.01, BEI:0, BAS:0, BAG:0, **CAI**:0, SAN:0, **ASW**:0, ALG:0, **ALX**:0, **MSA**:0 |
| من | EGY | RIY:0.22, SAL:0.16, DAM:0.14, AMM:0.09, JED:0.09, MUS:0.07, DOH:0.05, JER:0.05, KHA:0.02, ALE:0.01, BAS:0.01, BEI:0.01, BAG:0, SAN:0, **CAI**:0, **ASW**:0, ALG:0, **ALX**:0, **MSA**:0 |
| كتابتها | EGY | SAL:0.25, RIY:0.14, JER:0.13, AMM:0.12, JED:0.09, DAM:0.09, DOH:0.04, ALE:0.02, MUS:0.01, KHA:0.01, **CAI**:0.01, BEI:0, BAS:0, **ASW**:0, SAN:0, BAG:0, **ALX**:0, ALG:0, **MSA**:0 |
| عام | EGY | JED:0.27, DOH:0.14, RIY:0.14, DAM:0.09, SAL:0.07, JER:0.06, AMM:0.04, MUS:0.03, BEI:0.01, ALE:0.01, KHA:0, SAN:0, **CAI**:0, BAS:0, BAG:0, **ASW**:0, **ALX**:0, ALG:0, **MSA**:0 |

*Table 2: Camel-tools dialect identification task*

[52] Nizar Y. Habash, *Introduction to Arabic natural language processing*, San Rafael (California, USA), Morgan & Claypool Publishers, 2010; Omar Zaidan, Chris Callison-Burch, *Arabic dialect identification*. «Computational Linguistics», 40, 1, 2014, pp. 171-202.

[53] Niloofar Haeri, *Sacred language, ordinary people: Dilemmas of culture and politics in Egypt*. Berlin, Springer, 2003; Shafik, Viola. *Egyptian Cinema: Hollywood on the Nile*, «Oxford Islamic Studies Online», 2022, <http://www.oxfordislamicstudies.com/article/opr/t343/e0209> (Accessed: Aug. 24, 2025).

[54] Ossama Obeid et al., *CAMeL tools: An open source python toolkit for Arabic natural language processing*. In *Proceedings of LREC2020*, Marseille, ELRA, pp. 7022-7032.

These results should not be seen as a failure of the model: it is designed for sentence- rather than token-level recognition and performs better in phrasal contexts (see Ex.1[55], where all three sentences were classified as EGY, though the second would be MSA; in any case, no other colloquial varieties were involved). Our goal, however, was not to validate a dialect classifier, but to provide an operational resource for assessing whether OCR difficulties stem from Egyptian Arabic lexicon rather than MSA.

*Example 1:*

-مش كده يا عبده! كده . كتير
صوت أمى الخافت الملح يطن فى أذنى وأنا أجتاز : الردهة
-دى عملة يا عصام.. تقطع الواب؟! شفت أبوك كان فرحان به قد إيه! تقوم ! تقطعه؟

For this reason, rather than persisting with the binarization of predictions, we opted to draw directly on the MSA morphological database distributed with CAMeL-tools, treating it as a reference lexicon. The procedure adopted was as follows: each token in *Awrāq ʿIṣām ʿAbd al-ʿĀṭī* was searched within the MSA lexicon; if present, the token was classified as MSA, otherwise as EGY. Although simplified, this approach proved more consistent with our objectives: to assess the extent to which an OCR system encounters difficulties with non-standard forms and, more broadly, to estimate the proportion of non-MSA (and thus more problematic) lexicon within the narrative. The final results of the classification are reported in Table 1.

### *3.2 Error Analysis in the OCR-Based Digitization Process*

The workflow comprised three steps: page acquisition, optical recognition with both models, and manual correction of Tesseract's output to build a comparative gold standard. This allowed systematic documentation of recurrent errors linked to the graphic specificities of Arabic (cursive script, contextual forms, ligatures, diacritics, typographic variants (e.g., the syntagm "the love", *al–ḥubb*, in two different fonts الحبّ vs. الحبّ), and RTL issues with numbers and punctuation). For comparison, model outputs and the gold standard were aligned at sentence and character level using *difflib*[56] and anchoring tokens; in cases of omissions, a special <GAP> marker preserved alignment and ensured reliable comparison despite partial misrecognition.

Errors were extracted at the character level and classified into linguistically motivated categories:

1. Diacritic (*tanwīn*, *sukūn*, *šadda*, etc.);
2. graphemic (ligatures, groups of letters distinguished only by dots, graphemes overlapped for stylistic or typographic purposes);
3. punctuation (confusion between comma, period, hamza, semicolon);
4. numerals (Arabic vs Western digits);
5. insertion;
6. deletion;
7. others.

[55] Our translation: -Isn't that so, ʿAbduh! That's too much. / My mother's faint, insistent voice hums in my ears as I cross the hallway: / -That's money, ʿIṣām… you'd tear up the receipt?! Did you see how happy your father was with it! And now you'd go and tear it up?!

[56] Standard Python library module: <https://docs.python.org/3/library/difflib.html> (Accessed: Nov. 14, 2025).

*3.2.1 Tesseract Model Results*

Regarding Tesseract's performance on our corpus, character recognition accuracy reached 96–97% (CER 3.5%), while word-level accuracy stood at 86% (WER 14%). The most frequent error type was deletions (519 occurrences), i.e., omissions of characters, most likely due to segmentation errors or unrecognized diacritics. This was followed by punctuation errors, with the following subclasses and frequencies:

- comma (،) misrecognized as hamza (ء) (83 occ.);
- comma (،) substitued with angled quotation marks («») (184 occ.);
- dashes introducing direct speech (-) misinterpreted (93 occ.);
- confusions of numerals, e.g., "٩" into "4", or "٨" into "4" (129 occ.).

The latter reflects persistent difficulties in recognizing the forms of Arabic-Indic digits.

Tesseract errors predominantly occur at the graphemic (artistic ligatures, typographic allomorphs), and segmentation levels (omitted dialogue dashes, substituted punctuation, irregular spacing). Graphemic distortions may still allow sentence segmentation, whereas segmentation errors compromise tokenization and parsing. In practice, Tesseract often simplifies the text by dropping diacritics, reducing ligatures, and normalizing less frequent punctuation into more common forms (e.g., "،" into "." or "«" and "»"). One might therefore describe this as an implicit strategy of text normalization in the presence of noisy elements.

Table 3 presents CER and WER scores for the errors generated by the model, broken down by tokens classified as MSA and EGY.

| Language | CER | WER | Char err. | Chars ref. | Word err. | Words ref. |
|---|---|---|---|---|---|---|
| MSA | 0.03 | 0.08 | 2,838 | 73,417 | 1,660 | 18,698 |
| EGY | 0.04 | 0.11 | 208 | 4,749 | 113 | 986 |

*Table 3: Tesseract metrics based on language classification*

Table 3 shows that although most of the text is MSA (73,000 characters, 18,698 words vs. 4,749 characters, 986 words for EGY), error rates are higher for Egyptian tokens: CER 4.38% vs. 3.87% for MSA, and WER 11.46% vs. 8.88%. In absolute terms, this corresponds to 113 erroneous words out of 986 in EGY, against 1,660 out of 18,698 in MSA, meaning that, despite their smaller share, Egyptian tokens contribute proportionally more errors.

*3.2.2 Qari-OCR Model Results*

Qari was adapted from a general-purpose multimodal architecture (Qwen2-VL-2B-Instruct[57]) to Arabic OCR through fine-tuning on synthetic corpora. This enables more robust handling of vocalized script, typographic styles, and graphic variants often missed by systems like Tesseract, originally designed for broader contexts. Trained on clean, well-segmented artificial images, however, Qari risks domain bias and struggles with complex real-world texts such as our corpus, marked by dense punctuation and Arabic-Indic numerals. Table 4 seems to confirm these limitations.

| Language | CER | WER | Char ed. | Chars ref. | Word err. | Words ref. |
|---|---|---|---|---|---|---|
| MSA | 0.07 | 0.18 | 5,733 | 73,417 | 3,384 | 18,698 |
| EGY | 0.11 | 0.30 | 553 | 4,749 | 299 | 986 |

*Table 4: Qari-OCR metrics based on language classification*

[57] Peng Wang et al., *Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution*, «ArXiv», 2024, <https://doi.org/10.48550/arXiv.2409.12191> (Accessed: Aug. 24, 2025).

Although Qari attains fair accuracy in MSA (CER 7.8%, WER 18.1%), its error rates nearly double those of Tesseract, with an even stronger impact on Egyptian Arabic (CER 11.6%, WER 30.3%), where less standardized morphology heightens fragility. Yet these metrics penalize Qari more than qualitative inspection suggests: it outperforms Tesseract in punctuation and numeral recognition, despite frequent overcorrections. For instance, in Egyptian print, especially in al-Aswani's style, the grapheme *yāʾ* often appears without dots (as an *alif maqṣūra*). Qari systematically normalizes this grapheme into a dotted *yāʾ* (e.g., **فى** often rendered as **في** ), a choice that negatively affects the statistics but produces texts more compatible with modern NLP resources. Similarly, dialogue dashes are often detected but sometimes transcribed with a different mark (e.g., "–" as "-"), which alignment tools count as substitution errors, even though they constitute acceptable recognition. The same applies to the *fatḥa tanwīn* ً(), which Qari tends to reproduce, sometimes in the wrong position (e.g., **وأخيرًا** as **وأخيرا**): a formal error, but still a step forward compared with Tesseract, which consistently ignored it. Qari also handles punctuation more richly: it may add or duplicate marks, but still captures turn-taking and sentence segmentation, which Tesseract often missed. Thus, while sensitive to typography and domain bias, Qari offers qualitative gains in punctuation, numeral recognition, and normalization. Its evaluation should therefore consider not only CER/WER scores, but above all the downstream impact on linguistic and computational analysis of Arabic texts.

## 4. Preliminary Analysis of *Awrāq ʿIṣām ʿAbd al-ʿĀṭī* by ʿAlāʾ al-Aswānī

### *4.1 The Quantitative Analysis of Emotions in the text*

For the automatic analysis of emotions expressed in Arabic texts, three models were employed: *multilingual_go_emotions_V1.2* (MGoE2), *bert-base-arabic-emotion-analysis-v2* (BAE2), and *bert-base-arabic-finetuned-emotion* (BAFE)[58].

MGoE2 assigns each sentence a probability distribution across the 28 categories, producing a vector of continuous scores. For each textual unit (sentences from the OCR-processed and manually validated text), the three models identified the dominant emotion. To enable comparison with other models, the GoEmotions categories were mapped onto Ekman's six basic emotions (Table 5).

| Ekman Set | GoEmotions Labels |
|---|---|
| Anger | anger, annoyance, disapproval |
| Disgust | disgust |
| Fear | fear, nervousness |
| Joy | joy, amusement, excitement, gratitude, pride, relief, admiration, love, optimism, desire, approval, caring |
| Sadness | sadness, disappointment, grief, remorse |
| Surprise | surprise, realization, curiosity, confusion, embarrassment |
| Neutral | neutral |

*Table 5: Mapping of MGoE2 classes into Ekman classification set*

This reduction introduces an asymmetry in the emotion distribution, as a larger number of subcategories fall under *Joy* compared to, for example, *Disgust*, which remains singly represented. This choice may inflate the apparent frequency of *Joy*, but reflects the study's aim to ensure a shared interpretive framework across different models rather

[58] The analysis was conducted using the text-classification pipeline of the *Transformers* library (cf. Thomas Wolf et al., *Transformers: State-of-the-Art Natural Language Processing*, in *Proceedings of EMNLP: System Demonstrations*, 2020, Punta Cana, October, Ed. ACL, 2020, pp. 38-45).

than a fine-grained classification of emotional nuances. BAFE also differs from Ekman's taxonomy, as it does not include *Disgust* among its categories but introduces *Sympathy* and *Love* as additional positive emotions. For comparability, *Sympathy* and *Love* were mapped to *Joy*, given their valence. *Neutral* is available only in the MGoE2 taxonomy, while the other models classify exclusively non-neutral emotional content. In Table 6 we present the comparable results of the three aligned models applied to our data.

| Labels | MGoE2 (%) | BAE2 (%) | BAFE (%) |
|---|---|---|---|
| *Anger* | **8.95** | 3.36 | 4.76 |
| *Disgust* | 1.24 | **24.56** | 0 |
| *Fear* | 1.91 | 7.91 | 5.53 |
| *Joy* | **15.41** | **17.84** | **30.77** |
| *Sadness* | **6.36** | **32.94** | **22.75** |
| *Surprise* | **8.74** | 13.39 | **36.19** |
| *Neutral* | 57.39 | 0 | 0 |
| TOTAL | 100 | 100 | 100 |

*Table 6: Ekman classes frequencies following the three AER models*

The resulting distribution shows a prevalence of emotions such as *Joy* and *Sadness*, while other emotions like *Anger* and *Fear* are less represented. The emotion *Disgust* appears to be particularly frequent in BAE2, which also seems to use the label *Surprise* less often. This pattern is consistent with the predominantly descriptive or reflective nature of the analyzed texts[59]. The Fleiss' Kappa coefficient, computed as a measure of inter-model agreement (0.0340; without *Neutral*: 0.2208, e.g., barely fair), indicates a low level of convergence among the three classification models.

Taken together, these results confirm that, while the models can be useful for large-scale AER, their performance on our data is not fully reliable. This is evident from both the quantitative agreement scores and qualitative observations, for instance, all three models classified the following passage as *Surprise*, whereas in context it clearly conveys *Anger*, or even *Disgust*: *"They are not superior in any of that... So what, then, distinguishes the Egyptians?! Where are their virtues? I challenge anyone to name me a single Egyptian virtue!"*[60]. Such misclassifications likely stem from punctuation bias, where marks influence the models' emotional inference, leading them to interpret emphasis as *Surprise* rather than negative emotions. A limitation of this phase concerns the lack of manual segmentation of the corpus, which required annotation to be performed on individual text lines. However, to identify emotionally homogeneous sections, sequences of consecutive labels were observed and subsequently selected for didactic purposes. In this way, the emotional coherence of paragraphs was used as a criterion to explore how different emotions are expressed in Arabic.

However, these results highlight the need for a qualitative analysis of the data, aimed at better understanding the divergences among models and manually validating the most plausible emotional interpretations.

### *4.2 The Qualitative Analysis of Emotions in the text as a Key to Interpretation*

Following Underwood's hybrid approach and the Eve's conception of DH (§2.1), Al-Aswānī's short story becomes a literary arena where the emotional and social categories of an era marked by disillusionment and identity fractures are redefined. The work func-

[59] It is worth noting that the model is multi-label, meaning it can assign multiple emotions to a single sentence, and selecting only the dominant label simplifies the text's affective complexity. Moreover, its extension to Arabic, though effective, remains constrained by the original GoEmotions dataset (based on English Reddit content), with potential cultural and semantic biases.

[60] Our translation.

tions as a cultural device that reveals how emotions evolve into political categories, intertwining the private dimension of feeling with the public sphere of history.

The basic emotions theory proposed by Paul Ekman[61] offers a useful framework for literary analysis in DH, due to the lack of AER tools for Arabic texts (§2.3). Recent scholarship shows that a mere quantification of lexical occurrences risks reducing the complexity of emotional experience to numerical data[62]. For instance, some passages expressing *Anger* or *Sadness* were classified as *Joy* or *Neutral* emotions. This misalignment between labels and emotional nuance calls for a qualitative reading capable of recovering the symbolic dimension of sentiment.

1. Discursive strategies reveal how emotions function as instruments of critique. Through the alternation between confession and generalization, the narrator shifts from intensely personal statements, such as:

*All those who knew my worth and resented my success; I hated them all* — جميع كل من عرفوا قدري وأحنقهم تفوقي كل الذين كرهتهم

to universal formulations, like:

*The Egyptian knows nothing but obedience* — المصري لا يعرف إلا الطاعة

In this oscillation, anger ceases to be a private sentiment and becomes a collective diagnosis. Irony and negation further sustain this tension:

*If I were not Egyptian, I would wish to be Egyptian* — لو لم أكن مصريًا لوددت أن أكون مصريًا

The latter statement hides self-disgust beneath a façade of patriotic irony, revealing the psychological camouflage through which the narrator voices dissent while preserving the illusion of conformity. Likewise, assertions such as:

*Truth no longer exists except in books* — الحقيقة لم يعد لها وجود إلا في الكتب

epitomize the replacement of action with abstraction: thought becomes the last refuge of agency, and the intellectual's impotence is sublimated into moral authority.

2. Metaphors further anchor emotions within symbolic imagery that fuses the personal and the collective.

*The Egyptian is nothing but a servant* — المصري مجرد خادم

The metaphor of the "servant" condenses the protagonist's humiliation and articulates a discourse of historical dependency. Anger thus becomes a form of cultural reflection on subalternity and the failure of national emancipation.

*I am this puppet and the big hand is the hand of fate* — أنا هذه الدمية، واليد الكبيرة هي يد القدر

Similarly, the "puppet" metaphor dramatizes the loss of agency and the internalization of helplessness: the subject is no longer an actor of history but its object, suspended between determinism and despair. The same tone of defeat reverberates in the following passages, where collective rhetoric collapses into personal guilt:

*We repeat empty slogans day and night about our great Egyptian people, yet they drip with sorrow and helplessness* — رنانة جوفاء نرددها ليل نهار عن شعبنا المصري العظيم والمحزن

*At that moment, sorrow overcame me and I blamed myself* — عندئذ ينتابني الحزن وألوم نفسي

Finally, the "paternal" metaphor crystallizes the allegory of national failure:

*My father was a good man, but he was useless* — كان أبي رجلاً طيبًا، لكنه لم يكن يصلح لشيء

whose defeated idealism mirrors the moral and political exhaustion of an entire generation.

3. Narrative choices shape the emotional progression of the story, turning the succession of affective states into a cognitive trajectory, from anger to guilt, from the social

[61] See footnote 47.

[62] Bo Pang, Lillian Lee, *Opinion mining and sentiment analysis*, «Foundations and Trends in Information Retrieval» 2(1–2), 2008, pp. 1–135.

to the intimate. Internal focalization confines the reader within ʿIṣām's consciousness, creating a claustrophobic space where the voice becomes both subject and prison. The fragmentation of temporality, oscillating between memory and disillusionment, mirrors the fractured psyche of the protagonist and the instability of a society in transition. In this disarticulated temporal frame, sadness manifests as the impossibility of coherence or redemption. Structurally, the narrative follows a symmetrical and cathartic movement: it opens with public anger:

*I hate Egyptians and I hate Egypt* أكره المصريين وأكره مصر

and closes with private sorrow:

*At that moment, sorrow overcame me and I blamed myself* عندئذ ينتابني الحزن وألوم نفسي

The emotional arc thus humanizes rage, transforming denunciation into awareness and resentment into moral introspection.

The distribution of emotions across the text reveals how different *thematic nuclei* intertwines throughout the narrative, acquiring varying intensity depending on context. From an analytical standpoint, five overarching macro-themes emerge:

4. National identity and the relationship with Egypt;
5. social and interpersonal relations;
6. political corruption and social decay;
7. the cultural sphere and individual aspirations;
8. everyday life and personal memory.

The sentences cited above condense an emotional progression from anger to self-blame. The narrative voice is dominated by negative sentiment, reflecting alienation and disillusionment. From a hermeneutic perspective, *Awrāq ʿIṣām ʿAbd al-ʿĀṭī* functions as a cultural and affective device in which private sentiment becomes a lens for understanding historical transformation and collective identity. Pedagogically, this emotional configuration offers valuable material for language teaching: the story's emotionally charged discourse provides authentic input for exploring affective vocabulary and figurative language, while fostering empathy, intercultural reflection, and critical awareness. Within a digital didactic platform, such qualitative engagement complements computational analysis, enabling learners to experience Arabic as a living and affectively mediated language that connects linguistic knowledge with historical, social, and affective awareness

## 5. Conclusions

In this article, we presented our research project on the digitization of Arabic literary texts for didactic purposes, grounded in the conviction that language cannot be separated from its literature. The project integrates computational methods with humanistic interpretation to examine how Arabic fiction can function simultaneously as an object of scholarly inquiry and as a pedagogical resource. The study proceeded in two complementary stages: a quantitative phase, involving OCR processing and AER, and a qualitative phase based on close reading and interpretive analysis.

This dual approach revealed *joy* (which appears only in passages where the protagonist retreats into fantasies and mental wanderings) alongside *anger* and *sadness* as dominant emotional poles, transforming individual alienation into collective critique. However, it should also be taken into account that models trained on social media data may not be efficient in capturing the density and ambiguity of literary discourse. The automatic emotion labels assigned by the models did not always correspond to the actual affective content of the text, highlighting the necessity of qualitative inquiry to recover the symbolic and epistemic dimension of emotion.

From a glottodidactic perspective, the emotional spectrum emerging from the text can be effectively integrated into a communicative and affective syllabus. The selected excerpts provide authentic material for the exploration of emotional vocabulary, idiomatic structures, and pragmatic acts such as expressing anger, disappointment, or self-reflection. During the motivational phase, learners engage with the emotional and cultural context of the story; the exploration phase focuses on lexical and semantic analy-

sis; and the production phase encourages students to reformulate, narrate, or compare their own emotional experiences in Arabic.

Starting from a literary text like this, it becomes possible to design didactic units inspired by communicative principles that lower the affective filter and enhance emotional participation, creativity, and spontaneous linguistic output. Such sequences promote empathy, intercultural awareness, and critical reflection, aligning with functional approaches to language pedagogy. Ultimately, the integration of quantitative and qualitative methodologies illustrates how DH can bridge computational analysis and didactic innovation, positioning Arabic literature as a living field of linguistic, emotional, and civic formation.