

## **Dialects, Emojis, and Engagement: A Corpus-Based Study of Arabic Linguistic Features in TikTok Comments**

**Menik Mahmudah<sup>1</sup>; Mohammad Ahsanuddin<sup>1</sup>; Muhamad Solehudin<sup>2</sup>; Nur Hanifansyah<sup>2</sup>; Aufa Hanin Salsabil<sup>3</sup>; Aida Zavirah Fayruza<sup>1</sup>**

<sup>1</sup>Universitas Negeri Malang, Indonesia; <sup>2</sup>Universitas Islam Internasional Darullughah Wadda'wah, Indonesia; <sup>3</sup>International Islamic University Malaysia, Malaysia.

### **Abstract:**

This study employed a mixed-methods corpus-based discourse analysis to examine the linguistic features of Arabic used in TikTok comments. The data consisted of 2,160 Arabic comments collected from 36 TikTok videos published between September 2025 and January 2026. After data cleaning, the final corpus contained 28,742 tokens and 6,914-word types. Quantitative corpus procedures were conducted using AntConc 4.3.1 and Sketch Engine to generate frequency lists, normalized frequencies, concordance lines, and lexical clusters. The quantitative results were subsequently integrated with qualitative discourse-pragmatic interpretation informed by Computer-Mediated Communication, pragmatics, and discourse analysis. The findings show that Arabic in TikTok comments is predominantly informal, interactive, and multimodal. The dominant features include dialectal expressions, religious language, orthographic elongation, repetition, emoji usage, Arabic-English code-switching, and engagement-driven language. Emoji and multimodal elements were the most frequent feature (38.7%), followed by dialectal expressions (28.3%) and engagement-driven language (19.5%). These results indicate that communication on TikTok resembles spoken language that is expressive and socially oriented rather than formal written Arabic. Theoretically, this study extends Arabic corpus linguistics by focusing on an underexplored digital platform. Practically, the findings provide useful implications for Arabic digital language pedagogy and the development of NLP systems that can better handle non-standard Arabic varieties. The study also calls for expanding Arabic NLP frameworks to accommodate multimodal and non-standard digital registers.

**Keywords:** Arabic Corpus Linguistics; Tiktok Comments; Digital Discourse; Dialectal Arabic; Code-Switching; Multimodality

Submitted : April 27<sup>th</sup> 2026  
Revised : August 5<sup>th</sup> 2026  
Accepted : August 14<sup>th</sup> 2026

**Correspondence Author:**  
Nur Hanifansyah  
[nurhanifansyah@uiidalwa.ac.id](mailto:nurhanifansyah@uiidalwa.ac.id)



**Open Access Journal**

Availabel at: <https://jurnalfaktarbiyah.iainkediri.ac.id/index.php/asalibuna>

## **Dialek, Emoji, dan Keterlibatan: Studi Berbasis Korpus mengenai Fitur Linguistik Bahasa Arab dalam Komentar TikTok**

**Menik Mahmudah<sup>1</sup>; Mohammad Ahsanuddin<sup>1</sup>; Muhamad Solehudin<sup>2</sup>; Nur Hanifansyah<sup>2</sup>; Aufa Hanin Salsabil<sup>3</sup>; Aida Zavirah Fayruza<sup>1</sup>**

<sup>1</sup>Universitas Negeri Malang, Indonesia; <sup>2</sup>Universitas Islam Internasional Darullughah Wadda'wah, Indonesia; <sup>3</sup>International Islamic University Malaysia, Malaysia.

### **Abstrak:**

Penelitian ini menggunakan analisis wacana berbasis korpus dengan desain metode campuran untuk mengkaji karakteristik linguistik bahasa Arab dalam komentar TikTok. Data penelitian terdiri atas 2.160 komentar berbahasa Arab yang dikumpulkan dari 36 video TikTok yang dipublikasikan antara September 2025 dan Januari 2026. Setelah proses pembersihan data, korpus akhir berisi 28.742 token dan 6.914 tipe kata. Prosedur kuantitatif berbasis korpus dilakukan menggunakan AntConc 4.3.1 dan Sketch Engine untuk menghasilkan daftar frekuensi, frekuensi ternormalisasi, baris konkordansi, dan klaster leksikal. Hasil kuantitatif tersebut kemudian diintegrasikan dengan interpretasi kualitatif wacana-pragmatik berdasarkan teori Computer-Mediated Communication, pragmatik, dan analisis wacana. Hasil penelitian menunjukkan bahwa bahasa Arab dalam komentar TikTok bersifat informal, interaktif, dan multimodal. Fitur dominan meliputi ekspresi dialektal, ungkapan religius, pemanjangan ortografis, repetisi, penggunaan emoji, alih kode Arab-Inggris, serta bahasa berbasis keterlibatan (engagement). Emoji dan elemen multimodal muncul paling tinggi (38,7%), diikuti ekspresi dialektal (28,3%) dan bahasa berbasis keterlibatan (19,5%). Temuan ini menunjukkan bahwa komunikasi di TikTok lebih menyerupai bahasa lisan yang ekspresif dan berorientasi relasi sosial daripada bahasa tulis formal. Secara teoretis, penelitian ini memperluas kajian linguistik korpus bahasa Arab pada platform digital yang masih kurang dieksplorasi. Secara praktis, hasil penelitian ini bermanfaat bagi pengembangan pedagogi bahasa Arab digital dan sistem NLP bahasa Arab yang adaptif terhadap ragam non-standar. Penelitian ini juga menekankan perlunya pengembangan dan perluasan kerangka kerja NLP bahasa Arab agar mampu mengakomodasi ragam bahasa digital yang bersifat multimodal serta non-standar.

**Kata Kunci:** Linguistik Korpus Bahasa Arab; Komentar TikTok; Wacana Digital; Bahasa Arab Dialek; Alih Kode; Multimodal.

Submitted : April 27<sup>th</sup> 2026  
Revised : August 5<sup>th</sup> 2026  
Accepted : August 14<sup>th</sup> 2026

**Correspondence Author:**  
Nur Hanifansyah  
[nurhanifansyah@uiidalwa.ac.id](mailto:nurhanifansyah@uiidalwa.ac.id)



**Open Access Journal**

Availabel at: <https://jurnalfaktarbiyah.iainkediri.ac.id/index.php/asalibuna>



## INTRODUCTION

In recent years, the rapid development of digital communication technologies has significantly influenced the way languages are used, including Arabic (Arisandi & Habib, 2025; Baharun et al., 2026; Sapawi & Yusoff, 2025; Seyidov & Çitil, 2024). Social media platforms have become central spaces for interaction, where users produce large amounts of informal, spontaneous, and user-generated content (Hamdan, 2021; Nabilah et al., 2025). Among these platforms, TikTok has emerged as one of the most popular and rapidly growing platforms, characterized by short-form content and highly interactive comment sections. In 2023, TikTok was named the leading video-based social media app, with 1.5 billion active monthly users (Kemp, 2025; Nguyen & Diederich, 2023). In this environment, Arabic is no longer used in its traditional formal form, but rather in dynamic, hybrid, and often non-standard ways (djeghri zarouta, 2025; Kasmaluddin, 2026; Muh sabilar rosyad et al., 2023).

One of the major issues in contemporary Arabic linguistics is the shift from Modern Standard Arabic (MSA) to colloquial varieties (‘āmmiyya) in digital communication (Alanazi, 2024; Daoudi, 2011; La Rosa, 2022; Leblebici & Rostom, 2025). Users frequently employ dialects, mix languages, use non-standard spelling, and incorporate emojis and repetitive forms to express meaning (Hanadi et al., 2024; Tarihoran et al., 2022). This phenomenon raises important questions about how Arabic is evolving in digital spaces and how these changes reflect broader sociolinguistic transformations. Studies have confirmed that social media accelerates this linguistic shift, with younger Arab users increasingly preferring informal, dialect-based communication over MSA (Alkadi et al., 2026; AlMotairi & Hadwan, 2024). Despite the increasing presence of Arabic in online platforms, there is still limited understanding of the linguistic characteristics that define Arabic usage in emerging digital environments such as TikTok.

Previous studies in Arabic corpus linguistics have primarily focused on formal texts or specific platforms such as Twitter. These studies often examine aspects such as sentiment analysis, dialect identification, or computational modelling (Alkadi et al., 2026; AlMotairi & Hadwan, 2024; Al-Shenaifi et al., 2024). However, they tend to overlook the complex and multimodal nature of communication in newer platforms. Mohamed and Zaghoulani (2024) examined expressions of depression among Arab Twitter users and found that emotional expression is strongly associated with religious language, pronouns, and dialectal variation. Similarly, Al-Khawaldeh et al. (2024) investigated irony in Jordanian Arabic on social media, demonstrating that users employ various linguistic and rhetorical devices such as sarcasm, emoji, and code-switching.

From a computational perspective, Sadeghi et al. (2024) developed a large-scale Arabic Twitter corpus for age prediction using linguistic and metadata features, highlighting the potential of social media data for language analysis. Doughan et al. (2025) introduced ArabSis, a corpus designed for multi-dimensional sentiment analysis in Arabic, emphasizing the importance of capturing emotional nuances in digital texts. In addition, Sarbazi-Azad et al. (2021) proposed ExaAEC, a multi-label emotion classification corpus of Arabic tweets, demonstrating the feasibility of modeling complex emotional expressions using annotated datasets and deep learning techniques.

Despite these contributions, several limitations remain. First, most previous studies focus either on specific emotional aspects (e.g., depression, irony) or



computational tasks such as classification and prediction. Second, existing research is largely limited to Twitter data, with little attention to newer platforms such as TikTok. Third, there is a lack of studies that comprehensively analyze linguistic features, including interactional and multimodal elements, in Arabic digital discourse. In particular, TikTok remains underexplored in linguistic research, especially from a corpus-based perspective. This indicates a clear research gap in understanding how Arabic is used in TikTok comment sections, where communication is fast, informal, and highly interactive.

This study advances the field by employing a mixed-methods corpus-based discourse analysis of Arabic language use in TikTok comments. It integrates descriptive quantitative corpus procedures with qualitative discourse-pragmatic interpretation to examine a broad range of linguistic features, including dialectal expressions, orthographic variation, repetition, emoji usage, code-switching, and engagement-driven language. By moving beyond emotion detection and computational classification, the study provides a more comprehensive account of how linguistic forms function within TikTok's highly interactive and platform-specific communicative environment.

Therefore, this study aims to analyze the linguistic features of Arabic used in TikTok comments using a corpus-based approach. The research seeks to answer the following questions: (1) What are the dominant linguistic features of Arabic used in TikTok comments? and (2) How do these features reflect the nature of digital interaction and communication on TikTok?

This study is limited to textual data collected from TikTok comment sections and focuses only on written elements such as comments and emoji. It does not include audiovisual content or spoken language. In addition, the analysis is based on a selected corpus and does not aim to generalize all forms of Arabic used in social media. Despite these limitations, the study provides a focused examination of authentic language use in a specific digital context.

The significance of this study lies in its contribution to the fields of corpus linguistics, sociolinguistics, and digital media studies. By examining real-world data from TikTok, this research offers insights into contemporary Arabic language practices, particularly the use of dialects, informal expressions, and digital-specific linguistic features. The findings are expected to contribute to the development of Arabic language research in digital contexts and support future studies in Natural Language Processing (NLP), language education, and media analysis.

This research contributes to the field in several ways. Theoretically, it extends Arabic corpus linguistics to underexplored digital platforms, particularly TikTok, and emphasizes the importance of non-standard and dynamic language forms. Methodologically, it integrates corpus-based analysis with qualitative interpretation to capture real-world language use. Practically, the findings provide insights for Arabic language pedagogy and the development of more adaptive Arabic NLP systems that reflect contemporary digital communication.

In conclusion, this study positions itself as an effort to bridge the gap between traditional Arabic linguistic studies and the rapidly evolving nature of digital communication. By focusing on TikTok as a data source, it provides a novel perspective on how Arabic is used, transformed, and negotiated in modern online interaction.



## **METHOD**

This study employed a mixed-methods corpus-based discourse analysis to investigate the linguistic features of Arabic used in TikTok comments. The quantitative component consisted of descriptive corpus procedures, including token and type calculation, raw and normalized frequency analysis, category distribution, concordance analysis, and lexical cluster identification. The qualitative component involved discourse-pragmatic interpretation of recurrent linguistic patterns, particularly their functions in expressing affect, identity, politeness, stance, interactional alignment, and engagement. The two components were integrated during the interpretation stage, in which quantitative frequency patterns were examined alongside their contextual occurrences in concordance lines and representative extracts. This integrated design enabled the study to identify dominant linguistic tendencies while also explaining how these features function within the communicative and platform-specific environment of TikTok.

The corpus consisted of 2,160 Arabic TikTok comments collected from 36 TikTok videos posted between 1 September 2025 and 31 January 2026. After data cleaning, the final dataset contained 28,742 tokens and 6,914 unique word types. The comments were selected using purposive sampling based on specific criteria, namely that the comments were primarily written in Arabic script, reflected natural user interaction rather than automated or bot-generated text, and contained at least one identifiable linguistic feature such as dialectal expressions, orthographic elongation, emoji usage, code-switching, or interactional formulas (Zappavigna & Logi, 2024). Comments were excluded if they were duplicated, unreadable due to script corruption, fully non-Arabic, or consisted only of repeated symbols without lexical content. The videos were selected from public TikTok accounts that focus on entertainment content, particularly comedy, prank, and live-stream interactions. A total of 36 videos were chosen using purposive sampling based on three criteria: (1) high comment activity, (2) presence of Arabic-language comments, and (3) relevance to informal and interactive discourse. To minimize sampling bias, the videos were collected from different accounts and across multiple posting dates within the selected time frame, ensuring variation in content types and user interaction patterns.

The data were collected using a documentation method in which the researcher manually captured comments from selected TikTok videos and transcribed them into a structured text corpus. All data were stored in UTF-8 format to ensure compatibility with corpus analysis tools. To maintain ethical considerations, usernames were anonymized and replaced with a neutral label, while direct mentions were generalized. Non-linguistic elements such as URLs were removed unless they contributed to the meaning of the text. Emoji were retained as analytical units due to their significant role in conveying meaning in digital communication.

The corpus was processed using AntConc 4.3.1 to generate word-frequency lists, concordance lines, and lexical clusters. Before calculating token and type frequencies, the corpus underwent a minimal preprocessing procedure designed to improve consistency without removing linguistically relevant variation. Usernames, URLs, duplicated comments, and unreadable non-linguistic strings were removed, while multiple spaces and inconsistent punctuation were standardized. Arabic diacritics and the elongation mark (taṭwīl/kashīda, -) were removed because they did not contribute to the lexical analysis. However, dialectal spellings, non-standard





orthographic forms, repeated letters, and expressive elongation were retained because they constituted important analytical features of TikTok discourse.

Tokenization was based on whitespace and punctuation boundaries. Each orthographic word separated by a space was counted as one token. Attached Arabic clitics, including conjunctions and prepositions such as **و**, **ف**, **ب**, **ك**, and **ل**, as well as the definite article **ال**, were not separated from their lexical bases. Consequently, forms such as **والله** and **بالبيت** were treated as single tokens rather than morphologically segmented units. Orthographic variants involving hamza, alif, alif maqṣūra, and final yā' were also retained in their original user-generated forms to preserve naturally occurring digital variation. The reported type count therefore represents distinct orthographic word forms rather than lemmatized or morphologically segmented lexical items.

Two versions of the corpus were maintained during the analysis. The first was a raw corpus that preserved emoji, letter repetition, dialectal spelling, and orthographic elongation for qualitative and multimodal coding. The second was a preprocessed corpus used to calculate lexical token and type frequencies. Emoji were retained as separate analytical units in the raw corpus but were excluded from the lexical type count. The Word List tool was used to calculate token frequencies and identify dominant lexical items, while the Concordance tool enabled the contextual analysis of recurrent expressions. In addition, the Clusters/N-Grams function was used to identify repeated and rhythmic patterns commonly found in TikTok comments. The corpus was also uploaded to Sketch Engine to verify tokenization and obtain normalized frequencies per 1,000 words. The same preprocessing and tokenization rules were applied consistently across the corpus to ensure comparability among the different data segments.

A coding scheme was developed to classify linguistic features into several categories, including dialectal expressions, religious expressions, interactional formulas, orthographic elongation, repetition, emoji and multimodal elements, code-switching, engagement-driven language, low-information comments, identity expressions, and fragmented structures. Each comment could be assigned to more than one category due to the overlapping nature of linguistic features in digital discourse. The coding process was conducted systematically to ensure consistency across the dataset.

To ensure reliability, 20% of the corpus (432 comments) was independently coded by a second researcher using the same coding framework. Inter-rater agreement was measured using Cohen's Kappa, resulting in a score of 0.82, which indicates substantial agreement. Any discrepancies in coding were discussed and resolved through consensus, leading to refinement of category definitions. In addition, analytical validity was strengthened through triangulation by combining quantitative frequency analysis with qualitative interpretation based on concordance data.

Low-information or minimal comments were defined as comments containing limited propositional content, such as isolated words, numbers, emoji, short formulaic responses, or brief participation signals. This category was distinguished from repetition or rhythmic language, which involved the deliberate recurrence of lexical items or phrases for emphasis, rhythm, or attention. The term "spam" was avoided because the study did not independently verify whether the comments were automated or intentionally deceptive.



This study adopts an integrated analytical framework that combines corpus linguistics with theories from computer-mediated communication (CMC), pragmatics, and discourse analysis. Frequency analysis was first used to identify dominant linguistic patterns, followed by interpretation of these patterns in relation to communicative functions such as speech acts, politeness strategies, stance expression, and interactional alignment (Cristofaro, 2023). This framework allows the study to move beyond descriptive categorization and provide deeper insights into how language is shaped by the affordances and interactional norms of TikTok as a digital platform.

Accordingly, the quantitative and qualitative components were not treated as separate analyses but were integrated to explain both the distribution and communicative functions of the identified linguistic features.

The study was reviewed by the Universitas Islam Internasional Darullughah Wadda'wah and was determined to be exempt from full ethical approval because it analyzed publicly accessible TikTok comments without direct interaction with users or the collection of sensitive personal information. The data collection procedure followed the ethical guidelines of the institution and the applicable TikTok platform policies at the time of data collection. Usernames, profile information, direct mentions, and other potentially identifying details were removed or anonymized. No attempt was made to identify, contact, or profile individual users, and representative comments were reproduced only in anonymized form for linguistic analysis

## RESULTS AND DISCUSSION

### Dominant Linguistic Features in TikTok Comments

The analysis of the collected corpus reveals that Arabic language use in TikTok comments is highly informal, dynamic, and interaction-driven. Several dominant linguistic features were identified, including the use of dialectal expressions, religious language, orthographic variation, repetition, emoji, code-switching, and engagement-oriented language.

First, the data show a strong prevalence of colloquial Arabic (‘āmmiyya) over Modern Standard Arabic (MSA). The representative examples presented in Extract 1 illustrate the use of regional and spoken-like expressions in everyday digital interaction. These forms indicate that TikTok functions as a communicative space in which users favor immediacy, familiarity, and conversational naturalness over formal written conventions.

**Table 1.** Dialectal Expressions in Arabic TikTok Comments

| Original Arabic Text | Latin Transliteration | English Translation |
|----------------------|-----------------------|---------------------|
| ما عندي              | mā ‘andī              | I do not have [it]. |
| هلا والله            | halā wallāh           | Hello! / Welcome!   |
| حدا                  | ḥadā                  | Someone / anyone.   |

**Analytical note:** These examples represent colloquial forms associated with spoken Arabic. Their use reflects informality, regional identity, interpersonal familiarity, and spontaneous digital interaction. The original user-generated spelling has been retained to preserve the authenticity of the corpus.



Second, religious expressions also occur frequently in the corpus. As illustrated in Extract 2, users employ blessings, supplications, and invocations in their interactions. These expressions do not merely communicate religious meanings but also function as culturally embedded pragmatic resources for expressing goodwill, solidarity, emotional support, and interpersonal closeness.

**Table 2.** Religious Expressions in Arabic TikTok Comments

| Original Arabic Text | Latin Transliteration | English Translation        |
|----------------------|-----------------------|----------------------------|
| امين يارب            | āmīn yā rabb          | Amen, O Lord.              |
| الله يبارك فيك       | Allāh yubārik fīk     | May God bless you.         |
| ربي يحفظ             | rabbī yaḥfaẓ          | May my Lord protect [you]. |

**Analytical note:** These expressions function as blessings, supplications, and positive-politeness resources. Within TikTok interaction, they enable users to communicate goodwill, maintain social harmony, and establish emotional alignment with other participants.

**Note:** The Arabic examples are reproduced using their original user-generated orthography. Latin transliterations follow a simplified scholarly system, while the English translations are context-sensitive. Information added to clarify omitted or implicit meanings is placed in square brackets.

Another important feature is orthographic variation, particularly in the form of elongation and non-standard spelling. Words such as “كسبوروو” and “شيروووو” demonstrate how users extend letters to emphasize meaning or express emotional intensity. This type of variation reflects the flexibility of written Arabic in digital contexts.

The data also reveal extensive use of repetition and rhythmic expressions, such as “تعالون تعالون تعالون” and “بل بل بل”. These forms function to attract attention, reinforce meaning, and create a sense of interaction. In some cases, the language resembles spoken or performative expressions rather than traditional written text.

In addition, emoji and multimodal elements play a significant role in communication. Users frequently combine text with emoji such as ❤️, 😂, and 😊 to express emotions and attitudes. This indicates that meaning in TikTok comments is constructed not only through words but also through visual symbols.

The presence of code-switching and lexical borrowing is also observed, as seen in expressions such as “good live” and “تربلي”. This reflects the influence of global digital culture and the integration of English into Arabic online communication.

Furthermore, the data show the existence of engagement-driven language, including promotional and directive expressions such as “تعالون”, “شير”, and “حط” “الرابط في البايو”. These expressions indicate that language is often used strategically to increase interaction and visibility on the platform.

Finally, Low-information or minimal comments, including isolated words, numbers, symbols, emoji, and short formulaic responses, accounted for 14.7% of the corpus. Although these comments contain limited propositional information, they may function as participation signals that mark users’ presence, support visibility, and sustain interaction within comment threads. This category differs from





repetition or rhythmic language, which involves the recurrence of lexical items or phrases for emphasis, emotional expression, or attention management.

### Qualitative Observations of User Interaction

From a qualitative perspective, TikTok comments reflect a highly interactive and socially oriented communication environment. Users tend to engage in brief, spontaneous exchanges that resemble spoken conversation rather than structured written discourse.

Many comments express personal emotions and identity, as seen in phrases like “ياحبيبي” and “قد ايش انا”, which indicate closeness and self-expression. The frequent use of direct address and mention (e.g., @username) further emphasizes the interactive nature of the platform.

Additionally, users often produce fragmented and incomplete sentences, suggesting that clarity and grammatical accuracy are less important than immediacy and engagement. In this context, communication is driven more by social interaction than by linguistic correctness.

Overall, the qualitative findings suggest that TikTok comments function as a form of digital conversation, where language is used flexibly to express emotion, build relationships, and sustain interaction.

### Summary of Linguistic Features in TikTok Comments

To provide a clearer overview of the findings, the identified linguistic features are summarized in Table 1. The table highlights the main categories, examples, and their communicative functions within TikTok comments.

**Table 3.** Linguistic Features of Arabic in TikTok Comments

| No | Linguistic Feature                 | Example (Arabic)                    | Function / Interpretation                                                 |
|----|------------------------------------|-------------------------------------|---------------------------------------------------------------------------|
| 1  | Dialectal Expressions              | ما عندي، هلا والله، حدا             | Reflect informal spoken language (‘ammiyya)                               |
| 2  | Religious Expressions              | امين يارب، الله يبارك فيك، ربي يحفظ | Express cultural and religious identity                                   |
| 3  | Interactive Formulas               | السلام عليكم، عليكم السلام          | Maintain social interaction and politeness                                |
| 4  | Orthographic Elongation            | كسبوووو، شيروووو                    | Show emphasis and emotional intensity                                     |
| 5  | Repetition / Rhythmic Language     | تعالون تعالون، بل بل بل             | Attract attention and create engagement                                   |
| 6  | Emoji & Multimodality              | ❤️😂😂                                | Reinforce emotions and meanings                                           |
| 7  | Code-Switching                     | good live، تربل                     | Reflect global digital influence                                          |
| 8  | Engagement Language                | شير، تعالون، حظ الرابط في البايو    | Encourage interaction and participation                                   |
| 9  | Low-Information / Minimal Comments | تكبيس، نعم، ١، 😂                    | Signal participation, maintain visibility, or sustain interaction despite |



|    |                       |                     |                                                |
|----|-----------------------|---------------------|------------------------------------------------|
|    |                       |                     | limited propositional content                  |
| 10 | Identity Expression   | ياحبيبي، قد ايش انا | Express personal identity and emotion          |
| 11 | Fragmented Structures | في شي ضروري اكتب    | Reflect spontaneous and informal communication |

The final corpus contained 2,160 comments, 28,742 tokens, and 6,914 unique types. The average comment length was 13.3 tokens, indicating a short, interaction-driven style consistent with TikTok's rapid-response environment. The lexical profile showed a high concentration of emotionally marked and interactional expressions, including blessings, directives, vocatives, and repeated forms. Emoji occurred in 38.7% of all comments, confirming the multimodal nature of TikTok discourse.

### Frequency Distribution of Linguistic Features

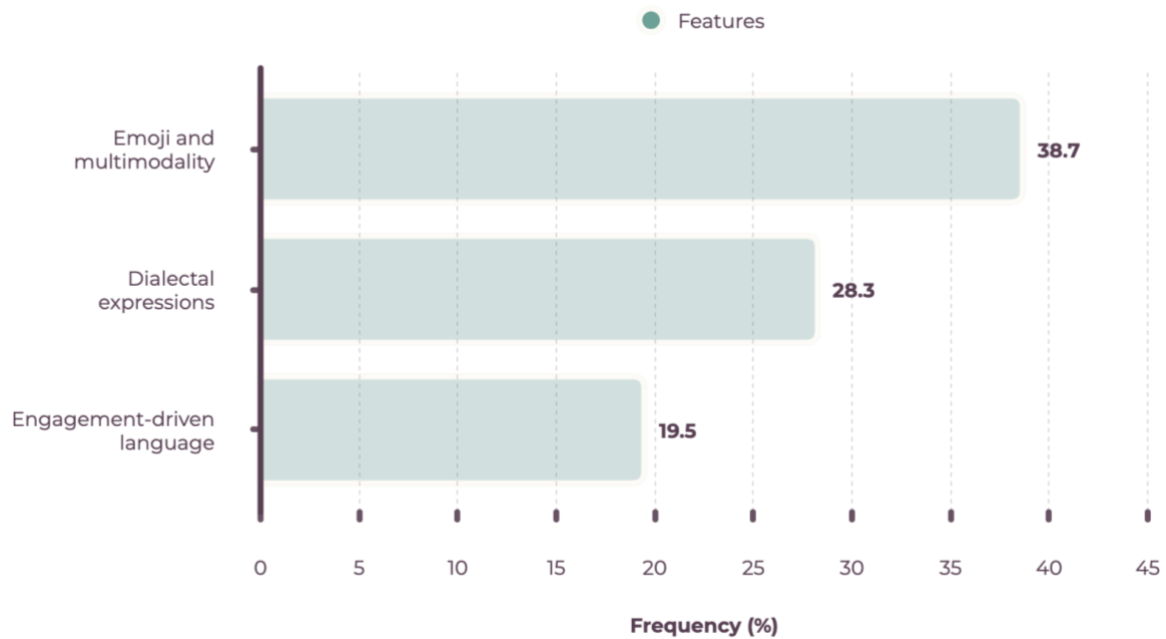
**Table 4.** Frequency of Linguistic Features in Arabic TikTok Comments (N = 2,160)

| No | Linguistic Feature               | Raw Count | % of Comments |
|----|----------------------------------|-----------|---------------|
| 1  | Dialectal expressions            | 612       | 28.3%         |
| 2  | Religious expressions            | 331       | 15.3%         |
| 3  | Interactional formulas           | 298       | 13.8%         |
| 4  | Orthographic elongation          | 274       | 12.7%         |
| 5  | Repetition / rhythmic language   | 247       | 11.4%         |
| 6  | Emoji & multimodality            | 836       | 38.7%         |
| 7  | Code-switching (Arabic-English)  | 146       | 6.8%          |
| 8  | Engagement-driven language       | 421       | 19.5%         |
| 9  | Low-information/minimal comments | 318       | 14.7%         |
| 10 | Identity & affective expressions | 284       | 13.1%         |
| 11 | Fragmented structures            | 259       | 12.0%         |

**Note:** A single comment may contain multiple features; therefore, percentages do not total 100%.

The most frequent feature was emoji and multimodality (38.7%), followed by dialectal expressions (28.3%) and engagement-driven language (19.5%). This pattern indicates that TikTok comments prioritize affective signaling and interaction management over formal linguistic accuracy.

**Figure 1.** Features Frequency in TikTok Comments



### High-Frequency Lexical Items

To identify lexical tendencies, a word frequency list was generated using AntConc. Table 3 shows the most frequent content-bearing items, excluding function words and particles.

**Table 5.** Top Lexical Items in the Corpus

| Rank | Token  | Frequency | Normalized Freq (per 1,000 words) |
|------|--------|-----------|-----------------------------------|
| 1    | الله   | 182       | 6.33                              |
| 2    | يارب   | 126       | 4.38                              |
| 3    | شير    | 109       | 3.79                              |
| 4    | تعالون | 102       | 3.55                              |
| 5    | حبيبي  | 96        | 3.34                              |
| 6    | أمين   | 91        | 3.17                              |
| 7    | رابط   | 84        | 2.92                              |
| 8    | لايك   | 77        | 2.68                              |
| 9    | فولو   | 73        | 2.54                              |
| 10   | والله  | 68        | 2.37                              |

The prominence of religious forms (الله, يارب, أمين), engagement commands (شير, رابط, لايك, فولو), and affective address terms (حبيبي) confirms that TikTok comments are shaped by both social bonding and algorithm-oriented participation practices.

### Dialectal Expressions and Informality

Dialectal forms such as “هلا والله”, “ما عندي”, and “حدا” occurred in 28.3% of comments, indicating a clear preference for colloquial Arabic (‘āmmiyya) over Modern Standard Arabic (MSA). From a CMC perspective, this preference can be



explained by the platform's immediacy and informality, where users prioritize speed, spontaneity, and relational closeness over grammatical standardization. In discourse terms, dialect indexes in-group identity and authenticity, signaling that the speaker belongs to a shared social or regional community.

### Religious Expressions as Pragmatic Resources

Religious expressions such as "الله يبارك فيك", "امين يارب", and "ربي يحفظ" appeared in 15.3% of comments. Rather than functioning as purely theological content, these expressions often operated as pragmatic markers of solidarity and positive politeness.

Concordance analysis of the node word "الله" revealed that it predominantly appears in collocational patterns with blessing formulas (e.g., "والله", "الله يبارك", "العظيم") and supplication structures (e.g., "يارب الله"). The left-span collocates within a window of L2 were dominated by exclamative particles ("و", "يا"), while right-span collocates included benefactive verbs ("يرزق", "يبارك", "يحفظ"), confirming the pragmatic function of religious language as a solidarity marker rather than doctrinal expression.

For example, "الله يبارك فيك" was frequently used in response to praise or requests, functioning as a relational response that maintains social harmony. In pragmatic terms, such expressions can be interpreted as blessings and face-supportive acts, reinforcing emotional warmth and interpersonal alignment.

### Orthographic Elongation and Emphasis

Orthographic elongation (e.g., "كسبوووو", "شيروووو") occurred in 12.7% of comments. This feature visually extends vowels or consonants to encode emphasis, urgency, or emotional intensity. In CMC, elongation compensates for the absence of vocal stress and intonation. Instead of speaking louder, users stretch letters to produce a similar effect in writing. Concordance analysis showed that elongated forms were especially common in directives ("شيروووو") and affective expressions ("حلوووو"), indicating their role in persuasion and emotional display.

### Repetition and Rhythmic Language as Attention Strategies

Repetition (e.g., "تعالون تعالون تعالون", "بل بل بل") appeared in 11.4% of comments. These repeated sequences are not random noise but function as attention-seeking discourse moves. From a discourse perspective, repetition increases visibility and rhythm, making comments more noticeable in fast-moving comment threads. This aligns with TikTok's interaction economy, where comments compete for attention and engagement. In many cases, repeated forms also create a performative style that resembles spoken chanting or cheering.

### Emoji and Multimodality

Emoji occurred in 38.7% of comments, often alongside short text. The most common emoji included ❤️, 😂, 😊, and 🔥. Their frequent use indicates that meaning is constructed not only through words but also through visual symbols. In CMC theory, emoji serve as paralinguistic substitutes by signaling tone, emotion, and stance that would otherwise be conveyed through voice and gesture in face-to-face interaction. In this corpus, emoji often intensified positive affect, softened directives, or marked irony.



### Code-Switching and Lexical Borrowing

Code-switching (Arabic–English) was found in 6.8% of comments, including forms such as “good live”, “follow”, and “link”. Borrowed words were often adapted into Arabic phonology and orthography (e.g., “لايك”, “فولو”). This pattern reflects the influence of global digital culture, where English serves as a shared lexical resource for platform-specific functions. In discourse terms, code-switching indexes digital fluency and aligns users with transnational online norms.

### Engagement-Driven Language and Platform Logic

Engagement-driven language (e.g., “تعالون”, “حط الرابط في البايو”) appeared in 19.5% of comments. These expressions function as directives, encouraging actions such as sharing, liking, or following. Pragmatically, these comments are less about content and more about participation. They are strategic utterances shaped by platform incentives such as visibility, follower growth, and algorithmic amplification. This supports the claim that social media language is not only communicative but also performative and goal-oriented.

### Low-Information Comments as Participation Signals

Low-information comments (e.g., repeated words, isolated symbols, short formulaic responses) accounted for 14.7% of the corpus. Although these comments carry minimal propositional content, they play a meaningful role in interaction. From a discourse perspective, such comments function as participation signals that mark presence, support visibility, and sustain thread activity. Therefore, they should not be dismissed as noise but understood as part of the communicative ecology of TikTok.

Taken together, the results indicate that Arabic TikTok comments are shaped by three interconnected forces: dialectal informality, multimodal expression, and platform-driven interaction. Compared with earlier Arabic social media studies that focused mainly on Twitter, the present findings show that TikTok discourse is more performative and engagement-oriented, with frequent use of emoji, repetition, and directive language. Compared with patterns reported in previous Twitter-based studies (Sadeghi et al., 2024), the present findings suggest that TikTok discourse may display a particularly strong orientation toward performative and engagement-related practices. These patterns suggest that digital Arabic is increasingly influenced by the affordances of short-form video platforms, where visibility, emotional signaling, and social alignment are central to communication.

This study has several limitations. First, the corpus was manually collected and limited to a specific period (September 2025–January 2026), which may not fully represent all Arabic TikTok discourse. Second, the analysis focused only on textual comments and emoji, excluding audiovisual elements such as voice, gesture, and video context. Third, while inter-rater reliability was established, some categories (e.g., low-information comments vs. spam) remain interpretively flexible. Future studies should use larger corpora, multimodal analysis, and cross-platform comparisons to strengthen generalizability.

Theoretically, this study extends Arabic corpus linguistics by demonstrating that non-standard and dynamic forms, such as elongation, emoji, and repeated directives, are central to meaning-making in digital environments. It also supports CMC and discourse-pragmatic frameworks that treat online language as





interactional, multimodal, and platform-sensitive. Practically, the findings can inform Arabic language pedagogy by helping educators introduce learners to contemporary digital usage, including dialectal variation and online politeness strategies. These findings also carry pedagogical implications. Culturally embedded Arabic environments can encourage spontaneous language use, communicative confidence, and learner engagement, while structured video-based instruction can provide exposure to naturally paced spoken Arabic and contextual vocabulary (Lazuardi et al., 2026; Mukhasibi & Wahyudi, 2026). Accordingly, carefully curated and anonymized TikTok comments may be used as supplementary materials to introduce learners to dialectal variation, digital pragmatics, and contemporary patterns of Arabic interaction. The results also provide useful insights for Arabic NLP development, particularly for sentiment analysis, dialect identification, and moderation systems that must handle non-standard orthography and multimodal signals.

The findings of this study confirm and extend previous research on Arabic digital discourse. First, the dominance of dialectal expressions (28.3%) supports earlier studies showing that Arabic users increasingly prefer colloquial varieties over Modern Standard Arabic (MSA) in informal online communication. This aligns with Alkadi et al. (2026), who reported a shift from formality to familiarity in written Arabic on social media, and with Hanadi et al. (2024), who found that digital environments encourage non-standard phonological and orthographic representations. In the present study, forms such as “*ما عندي*” and “*هلا والله*” indicate that TikTok users privilege immediacy and social closeness, which is consistent with computer-mediated communication (CMC) research emphasizing speed and interpersonal orientation in online interaction.

Second, the high frequency of religious expressions (15.3%) confirms findings by Mohamed and Zaghouani (2024), who observed that Arabic social media users frequently employ religious language to express emotion and solidarity. However, this study extends their findings by showing that on TikTok, expressions such as “*الله يبارك فيك*” and “*امين يارب*” are not only emotional markers but also function as pragmatic tools of politeness and relational alignment. In this sense, religious language in TikTok comments is less about doctrinal content and more about maintaining social harmony and positive face in fast-paced digital interaction.

Third, the prominence of orthographic elongation (12.7%) and repetition (11.4%) demonstrates how Arabic users adapt written language to compensate for the absence of vocal prosody, intonation, volume, and gesture in text-based interaction. From a Computer-Mediated Communication (CMC) perspective, users can develop and communicate interpersonal meanings through textual and graphic cues despite the reduced availability of nonverbal signals in digitally mediated interaction (Walther, 1992). Repeated letters in forms such as “*شيروووو*” and “*حلوووو*” visually reproduce prolonged pronunciation, while repeated expressions such as “*تعالون تعالون تعالون*” create rhythm and strengthen the illocutionary force of the message. These practices demonstrate that digitally mediated Arabic is not simply an abbreviated form of writing but a visually expressive register shaped by the communicative affordances and constraints of the platform.



The findings can also be interpreted through Brown and Levinson's (1987) politeness framework. Directives such as "شير" and "تعالون" may constitute potential face-threatening acts because they impose an expected action on other users. However, orthographic elongation, repetition, affectionate forms of address, religious expressions, and emoji may modify the interpersonal force of these directives. In some contexts, elongation and repetition intensify the request, but when combined with positive affective markers, they can also create a playful, familiar, and solidarity-oriented tone. Such forms may therefore function as positive-politeness strategies by emphasizing shared enthusiasm, group membership, familiarity, and relational closeness.

Emoji and other visual markers further support the interpersonal interpretation of these expressions. Previous research has demonstrated that emoticons can intensify verbal messages, communicate affect, and perform functions comparable to nonverbal behavior in face-to-face interaction (Derks et al., 2008). In the present corpus, emoji accompanying elongated or repeated expressions may strengthen enthusiasm, soften directives, mark humorous intent, or signal interpersonal alignment. Rather than merely transmitting propositional information, these textual and multimodal resources allow users to negotiate social relationships within TikTok's fast-paced interactional environment.

While previous Arabic social media studies have often focused on sentiment classification or irony detection, the present findings indicate that elongation and repetition also perform interactional and politeness-related functions. They enable users to compensate for reduced nonverbal cues, manage the intensity of speech acts, express affective involvement, and attract attention in crowded comment sections. Nevertheless, these interpretations remain context-dependent because the same elongated or repeated form may intensify a directive, express enthusiasm, create humor, or reinforce solidarity depending on its lexical content, accompanying emoji, and position within the interaction.

Fourth, the very high occurrence of emoji and multimodality (38.7%) reinforces previous claims that digital Arabic communication is increasingly multimodal. Al-Khawaldeh et al. (2024) identified emoji as a key rhetorical device in Jordanian Arabic social media, and the present study confirms this pattern in TikTok comments. However, this study adds a platform-specific insight: emoji on TikTok are not merely emotional supplements but often serve as paralinguistic substitutes that intensify directives, soften requests, and mark stance in short-form, rapid-response interaction.

Fifth, the presence of code-switching (6.8%) and lexical borrowing reflects the influence of global digital culture, in line with Tarihoran et al. (2022), who reported increased code-mixing among younger social media users. In this corpus, forms such as "لايك", "فولو", and "good live" indicate that English lexical items are integrated into Arabic digital discourse, especially for platform-related actions. This supports the argument that code-switching in social media is not random but functionally tied to digital participation and transnational online identity.



Finally, the findings on engagement-driven language (19.5%) and low-information comments (14.7%) extend prior corpus studies that typically focus on semantically rich or sentiment-laden text. Unlike traditional corpus approaches that may treat repetitive or minimal comments as noise, this study demonstrates that such forms play an important role in sustaining visibility and interaction on TikTok. This confirms a key insight from discourse-pragmatic studies: in social media environments, linguistic value is not determined only by propositional meaning but also by participation function. In other words, comments such as “تسبب” or repeated tokens are part of the communicative ecology of the platform, helping users remain visible and socially present.

Overall, this study contributes to the existing literature by showing that Arabic TikTok discourse is shaped by a combination of dialectal informality, multimodal expression, and platform-driven pragmatics. Compared with earlier work focused largely on Twitter, the present findings highlight that TikTok communication is more interactional, performative, and algorithm-aware. This suggests that future Arabic corpus linguistics should move beyond standard text analysis and incorporate platform affordances, multimodal features, and participation-oriented language practices into its analytical frameworks.



## **CONCLUSION**

This study has examined the linguistic features of Arabic used in TikTok comments through a mixed-methods corpus-based discourse analysis. The findings reveal that Arabic in TikTok is predominantly informal, interactive, and multimodal, characterized by the dominance of dialectal expressions, the frequent use of religious phrases, orthographic elongation, repetition, emoji, and engagement-driven language. These features indicate that TikTok communication reflects spoken-like, emotionally expressive, and socially oriented discourse rather than formal written Arabic.

Despite these contributions, this study has several limitations, including its reliance on a relatively small and manually collected dataset and its focus solely on textual elements without considering audiovisual aspects of TikTok communication. Future research is therefore encouraged to expand the corpus size, incorporate multimodal analysis, and explore comparative studies across different social media platforms. Practically, the findings of this study can inform the development of more adaptive Arabic NLP systems, improve digital language teaching strategies, and enhance understanding of online communication practices.

In conclusion, this study shows that Arabic in TikTok comments is predominantly informal, interactive, and multimodal. The findings highlight the importance of including non-standard orthography, emoji, and engagement-driven expressions in Arabic corpus analysis. Future research should expand the dataset, include multimodal features, and compare linguistic patterns across platforms to improve the generalizability of findings.



## REFERENCES

- Alanazi, M. S. (2024). The use of Modern Standard Arabic and colloquial Arabic in translation tasks: A new perspective. *Cogent Arts & Humanities*, 11(1), 2366572. <https://doi.org/10.1080/23311983.2024.2366572>
- Alkadi, H., Almoaily, M., & Alsuhailbani, Y. (2026). From Formality to Familiarity: A Content Analysis of Linguistic Shifts in Written Arabic on YouTube. *International Journal of Applied Linguistics*, 36(1), 904–915. <https://doi.org/10.1111/ijal.12825>
- Al-khawaldeh, N., Mashaqba, B., Al-Khawaldeh, S., Al Huneety, A., & AlQbailat, N. (2024). Irony in Jordanian Arabic on Social Media: A Corpus-Based Analysis of Linguistic and Rhetorical Devices. *International Journal of Society, Culture and Language*, 12(2). <https://doi.org/10.22034/ijscsl.2024.2021255.3363>
- AlMotairi, R. K., & Hadwan, M. (2024). Sentiment Analysis Methods for Arabic Content on Social Media: A Systematic Review. *Ingénierie Des Systèmes d'Information*, 29(1), 389–396. <https://doi.org/10.18280/isi.290138>
- Al-Shenaifi, N., Azmi, A. M., & Hosny, M. (2024). Advancing AI-Driven Linguistic Analysis: Developing and Annotating Comprehensive Arabic Dialect Corpora for Gulf Countries and Saudi Arabia. *Mathematics*, 12(19), 3120. <https://doi.org/10.3390/math12193120>
- Arisandi, Y., & Habib, Moh. T. (2025). Optimizing YouTube for Interactive Arabic Learning in Pesantren: Effective Content Creation Strategies. *International Journal of Arabic Language Teaching*, 7(02), 239–254. <https://doi.org/10.32332/ijalt.v7i02.10363>
- Baharun, S., Hanifansyah, N., Habib, Moh. T., Arifin, A., Syaifullah, M., & Faqihah, N. A. (2026). YouTube Content Production for Arabic Language Learning: Cross-National Study. *Online Learning In Educational Research (OLER)*, 6(1), 13–27. <https://doi.org/10.58524/oler.v6i1.982>
- Brown, P., & Levinson, S. C. (1987). *Politeness: Some Universals in Language Usage* (Vol. 4). Cambridge University Press. <https://doi.org/10.1017/CBO9780511813085>
- Cristofaro, M. D. (2023). *Corpus Approaches to Language in Social Media* (1st edn). Routledge. <https://doi.org/10.4324/9781003225218>
- Daoudi, A. (2011). Globalization, Computer-mediated Communications and the Rise of e-Arabic. *Middle East Journal of Culture and Communication*, 4(2), 146–163. <https://doi.org/10.1163/187398611X571328>
- Derks, D., Bos, A. E. R., & von Grumbkow, J. (2008). Emoticons and Online Message Interpretation. *Social Science Computer Review*, 26(3), 379–388. <https://doi.org/10.1177/0894439307311611>
- djeghri zarouta, nassira. (2025). The Use of Arabic language on social media: Between Diffusion and Hybridization -Descriptive and Analytical Study. 666–635, (2)9, *المفكر*.
- Doughan, Z., Itani, S., & Itani, S. (2025). ArabSis: Arabic Corpus Sentiment Analysis. *IEEE Access*, 13, 81083–81095. <https://doi.org/10.1109/ACCESS.2025.3567755>
- Hamdan, H. J. (2021). The Use of Arabic Neologisms in Social Media Applications. *International Journal of Arabic-English Studies*, 21(1). <https://doi.org/10.33806/ijaes2000.21.1.3>
- Hanadi, A. A., Loureiro-Rodríguez, V., & Fidan Acar, E. (2024). Makkan Arabic in the digital age: A sociolinguistic analysis of the representation of fricative, stop,





- and sibilant variation in WhatsApp text messages. *Journal of Arabic Sociolinguistics*, 2(2), 135–154. <https://doi.org/10.3366/arabic.2024.0029>
- Kasmaluddin. (2026). Duolingo-Based Arabic Vocabulary Acquisition in Islamic Educational Contexts. *Al-Muhawwaroh: Jurnal Pendidikan Bahasa Arab*, 2(1), 1–11. <https://doi.org/10.38073/almuhawwaroh.v2i1.4358>
- Kemp, S. (2025). *TikTok Users, Stats, Data, Trends, and More*. DataReportal – Global Digital Insights. <https://datareportal.com/essential-tiktok-stats>
- La Rosa, C. (2022). Linguistic Varieties and Social Radios in the Tunisian Sahel The Case of News Broadcasting: The Case of News Broadcasting. *Annali di Ca' Foscari. Serie orientale*, (1), JournalArticle\_8297. <https://doi.org/10.30687/AnnOr/2385-3042/2022/01/004>
- Lazuardi, A. M., Baharun, S., Solehudin, M., & Husein, A. F. (2026). Enhancing Arabic Listening Comprehension Skills through YouTube-Based Learning: A Community Service Program at Ribath Imam Al-Ghazali. *Community, Innovation, Education, and Empowerment*, 1(1), 1–8.
- Leblebici, D., & Rostom, M. (2025). “Alexa learned Arabic”: A translanguaging and multimodal perspective on language and media ideologies. *Discourse, Context & Media*, 66, 100909. <https://doi.org/10.1016/j.dcm.2025.100909>
- Mohamed, A., & Zaghouani, W. (2024). Expression of Depression Among Arab Twitter Users Using Arabic Corpus Analysis. *Procedia Computer Science*, 244, 76–85. <https://doi.org/10.1016/j.procs.2024.10.180>
- Muh sabilar rosyad, Muhammad Afifuddin, & Muhammad Afthon Ulin Nuha. (2023). HYBRID TRANSLATION: TINJAUAN MEDIA PENERJEMAHAN INDONESIA-ARAB BERBASIS OFFLINE DAN ONLINE PERSPEKTIF MAHASISWA PENDIDIKAN BAHASA ARAB. *JALIE: Journal of Applied Linguistics and Islamic Education*, 7(02), 225–248.
- Mukhasibi, A., & Wahyudi, I. (2026). Immersive Arabic Islamic Environment and Student Engagement: Evaluating the Linguistic Impact of World Arabic Language. *Farasyah: Journal of Linguistics and Language Education*, 1(1), 1–12.
- Nabilah, N., Mufidah, H. N., Nafisah, S., Baihaqi, M., Ulin Nuha, A. N., & Sa'adi, Z. R. (2025). The Influence of Social Media on Arabic Language Change in Online Communication. *Arabi: Journal of Arabic Studies*, 10(1), 14–25. <https://doi.org/10.24865/ajas.v10i1.753>
- Nguyen, H., & Diederich, M. (2023). Facilitating knowledge construction in informal learning: A study of TikTok scientific, educational videos. *Computers & Education*, 205, 104896. <https://doi.org/10.1016/j.compedu.2023.104896>
- Sadeghi, R., Akbari, A., & Jaziryan, M. M. (2024). ExaAUAC: Arabic Twitter user age prediction corpus based on language and metadata features. *Discover Artificial Intelligence*, 4(1), 48. <https://doi.org/10.1007/s44163-024-00145-0>
- Sapawi, M. S. M., & Yusoff, N. M. R. N. (2025). Integrating technology into the Arabic language curriculum: A systematic review of trends, strategies and cultural dimensions. *Social Sciences & Humanities Open*, 12, 101974. <https://doi.org/10.1016/j.ssaho.2025.101974>
- Sarbazi-Azad, S., Akbari, A., & Khazeni, M. (2021). ExaAEC: A New Multi-label Emotion Classification Corpus in Arabic Tweets. *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)*, 465–470. <https://doi.org/10.1109/ICCKE54056.2021.9721493>



- Seyidov, R., & Çitil, A. (2024). The Impacts of Contemporary Educational Technologies on Learning Arabic. *EVOLUTIONARY STUDIES IN IMAGINATIVE CULTURE*, 81–92. <https://doi.org/10.70082/esiculture.vi.2131>
- Tarihoran, N., Fachriyah, E., Tressyalina, & Sumirat, I. R. (2022). The Impact of Social Media on the Use of Code Mixing by Generation Z. *International Journal of Interactive Mobile Technologies (ijIM)*, 16(07), 54–69. <https://doi.org/10.3991/ijim.v16i07.27659>
- Walther, J. B. (1992). Interpersonal Effects in Computer-Mediated Interaction: A Relational Perspective. *Communication Research*, 19(1), 52–90. <https://doi.org/10.1177/009365092019001003>
- Zappavigna, M., & Logi, L. (2024). *Emoji and Social Media Paralanguage* (1st edn). Cambridge University Press. <https://doi.org/10.1017/9781009179829>