ALTRALANG Journal
e-ISSN: 2710-8619 p-ISSN: 2710-7922
Volume 8 Issue 1 / June 2026 pp. 199-217

Beyond the Mainstream of AI: Empowering NLP Models to Decipher Non-Canonical Morphological Processes in CMC Discourse at Algeria

Ali Berrabah[1*]
[1]*University of Tiaret Ibn Khaldoun, Algeria*
ali.berrabah@univ-tiaret.dz
Hamida Doulate Serouri[2]
[2]*University of Tiaret Ibn Khaldoun, Algeria*
serourihamida.doulate@univ-tiaret.dz
Mustafa Hamidi[3]
[3]*University of Chlef Hassiba Benbouali, Algeria*
m.hamidi@univ-chlef.dz

Received: 13/03/2025, **Accepted:** 12/01/2026, **Published:** 05/06/2026

ABSTRACT: *The proliferation of Computer-Mediated Communication (CMC) and the advent of social networking sites since the late 1980s and beyond have ignited a captivating realm where technology, communication, and culture intertwine, offering a fertile ground for Natural Language Processing (NLP) to delve into linguistic data. Due to the linguistic mosaic and/or the complex linguistic melting pot in the Algerian Speech Repertoire, new informal-written styles, non-standardised syntactic patterns, and non-normative morphological processes have appeared in CMC discourse. The paper explores the challenges posed by current AI-driven NLP models to decode the non-canonical morphological processes in CMC discourse in Algeria. The ultimate objectives are to assess the in/effectiveness of current AI-driven NLP models in analysing such linguistic practices and mainly to suggest potential improvements or viable solutions to overcome any obstacles or limitations encountered during the analysis. In a mixed-method research paradigm, 6 interviews were conducted with MA2 linguistic-dissertations supervisors together with a questionnaire directed to 24 MA2 students of linguistics in the Department of English at Ibn Khaldoun University of Tiaret, was employed to gather comprehensive insights from both informants and respondents regarding their experiences, perceptions, and suggestions related to the analysis of the non-canonical morphological processes in CMC in Algeria via current AI-driven NLP models. Besides, the study incorporated the Content Analysis Approach (CAA) to analyse such processes in the examples provided by the informants. Our findings demonstrate sparsity in current AI-driven NLP models, which still create suffering for researchers who have dared to deconstruct those non-canonical morphological processes. Therefore, it is recommended that researchers in the linguistic sphere collaborate with those in the computing sciences fieldwork to create non-canonical-text-based software packages to effectively analyse the non-standardized morphological processes in CMC discourse in the Algerian context.*

KEYWORDS: Artificial Intelligence (AI), Computer-Mediated Communication Discourse (CMCD), the Algerian Speech Repertoire, Natural Language Processing (NLP), Non-Canonical Morphological Processes

* Corresponding author
ALTRALANG Journal / © 2026 The Author(s). Published by the University of Oran 2 Mohamed Ben Ahmed, Algeria.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

1. Introduction

Communication strategies have undergone significant transformations since the late of 20th century and the early beginning of the 21st century all thanks to the incredible evolution of Computer-Mediated Communication (CMC) and Information Communication Technologies (ICTs). These advancements have provided people with a plethora of communication options, including emails, text messaging, instant messaging, and social networking sites. These platforms have revolutionised how people interact, allowing them to connect with others across the globe, both synchronously and asynchronously, and at minimal cost or even for free. The continuous progress in CMC, especially in Mobile Mediated Communication (MMC) and social networking sites, has not only introduced new social contact practices but has also reshaped the hierarchy of communicative strategies, with totally innovative linguistic processes playing a crucial role in this transformation.

While wireless communication in Algeria experienced a relatively slower pace compared to neighbouring countries and the global landscape, the prevalence of mobile phone services during the early years of the twenty-first century has brought into play a profound shift in the realm of interpersonal communication. This new way of everyday communication has introduced unique and inventive linguistic patterns, such as informal writing styles, word borrowings, and code-switching or mixing from both genetically related and unrelated language varieties. The content of these messages, on the one hand, enables individuals, particularly the youth, to access information and engage in discussions about their personal interests or private experiences without the need for face-to-face interaction. However, on the other hand, unconventional linguistic practices have also become a fertile ground for research, captivating the attention of a wide range of linguists seeking to gain a deeper understanding of the substantial socio-psychological factors and motivations that lie behind such communicative behaviour through unravelling their structural properties.

Despite the progress made in AI and NLP, the complexity and nuance of non-canonical language patterns, herein morphological processes, in CMC discourse in Algeria pose significant challenges, highlighting the limitations of current AI-driven NLP models in accurately decoding and comprehending these linguistic practices.

Given that our interest lies in comprehending the underlying reasons, acquiring an understanding of why, and identifying the limitations/constraints of current AI-driven NLP models in deconstructing non-standard language/non-canonical varieties, mainly the formative linguistic features and/or morphological processes in CMC discourse in the Algerian context, a range of intertwined research questions have arisen. Some of these questions include:

1. What specific linguistic features or phenomena within the Algerian speech repertoire pose challenges for AI-driven NLP models in accurately processing and analysing text-based mediated communication?
2. How far do the limitations of current AI-driven NLP models impact the ability to analyse and understand the linguistic mosaic/melting-pot, hybridity and non-canonical morphological processes that exist in CMC discourse in Algeria?
3. What are the implications of AI-driven NLP models' inadequacy in analysing the linguistic mosaic of Algeria for natural language understanding, sentiment analysis, or other language-related tasks?

To answer the aforementioned research questions and accomplish the goals of this enquiry, the subsequent research hypotheses have been formulated. The hypotheses serve the dual purpose of addressing the constraints and complexities associated with existing AI-powered NLP models in examining CMC discourse in Algeria.

1. The current AI-driven NLP models face challenges in processing and understanding informal and non-standard morphological processes in CMC discourse in Algeria leading to inaccuracies in language analysis tasks.

2. The limitations of current AI-driven NLP models result in reduced accuracy and performance when analysing Algerian text-based mediated communication containing hybrid formative language forms (non-canonical morphological processes) alongside borrowings from multiple coexisting language varieties.
3. The inability of current AI-driven NLP models to accurately analyse the linguistic mosaic of Algeria, herein morphological processes, can have implications for various language-related applications, including sentiment analysis, contextual end cuts, machine translation, and information retrieval, potentially leading to biases, inaccuracies, and poor performance in those tasks.

The primary focus of this study revolves around the attainment of the following objectives:

1. To identify the specific linguistic features or phenomena within the Algerian speech repertoire that pose challenges for AI-driven NLP models in accurately processing and analysing text-based mediated communication.
2. To examine the extent to which the limitations of current AI-driven NLP models hinder the analysis and comprehension of the linguistic mosaic, hybridity, and non-canonical morphological processes present in CMC discourse in Algeria.
3. To explore the implications of the inadequacy of AI-driven NLP models in analysing the linguistic mosaic of Algeria for tasks such as natural language understanding, sentiment analysis, and other language-related applications.

2. Literature Review

In this section, it is crucial to provide a concise overview of important concepts related to the study. It is of a dual purpose: firstly, it facilitates a seamless transition from theory to practical application, and secondly, it helps clarify key points for the reader by referencing influential works on CMC and the development of AI-driven NLP models. Additionally, we will explore the emergence of novel linguistic features in CMC discourse in Algeria.

2.1 Unveiling the Digital Realm: CMC Discourse as a Linguistic Endeavour

The advancement of technology has given rise to CMC, which has replaced traditional face-to-face interactions with digital means of communication. Scholars have started to examine the use of language in CMC since the early 1980s, employing an interdisciplinary approach known as Computer-Mediated Discourse Analysis (CMDA), initially adopted by Herring (2004). This approach aims to compare and contrast CMC interactions with face-to-face conversations, focusing on the unique communicative style observed in digital means.

With the widespread use of ICTs, the definition of CMC has expanded to include any digitally mediated communication. People now engage in various forms of CMC, such as email, instant messaging, text messaging, and online audio or video chatting. The concept of CMC has evolved to encompass different modes and forms of discourse, including synchronous, which refers to real-time online conversations with active turn-taking systems, exemplified by teleconferencing, Zoom Meetings, and live voice and video calls on platforms like Facebook, Messenger, Viber, etc, and asynchronous communication that involves independent offline interactions with delayed feedback, including email, written messages on social media platforms, voice memos, and recorded videos.

Crystal's influential work in the early 2000s revolutionized linguistic research, particularly in analysing synchronous and asynchronous modes of digital communication. CMC can be categorized as inclusive, allowing unknown viewers to react to posted messages on social networking sites, or exclusive, involving private and intimate exchanges such as emails or text messages. However, there has been a greater emphasis on studying verbal interactions in discussion boards and chat channels, while written content on websites and weblogs has received less attention. The scarcity of data and limited accessibility to mediated communication platforms in the early stages may have contributed to this imbalance.

2.2 NLP: A Multidisciplinary Approach to Computational Language Analysis

NLP is an important branch of AI that involves the computational modelling of human language. It focuses on studying the interactions between humans and computers using natural language. NLP encompasses various technologies and processes that analyse the meaning of words, phrases, sentences, and documents, as well as syntactic and semantic structures. Additionally, it develops applications such as machine translation, question-answering systems, information retrieval, dialogue systems, text generation, and recommendation systems. This section explores the multidisciplinary nature of NLP, its relationship with computational linguistics, and its connections to other fields of study.

NLP can be defined as the computational modelling of human language, aiming to understand and process natural language interactions. It involves the analysis of words, phrases, sentences, and documents, as well as syntactic and semantic processing. Key technologies in NLP include word breaking, syntactic parsers, and semantic parsing. These technologies form the foundation for various NLP applications, such as machine translation, question-answering systems, and text generation.

NLP is often contrasted with computational linguistics, with NLP considered to be more applied. While both fields deal with modelling human languages, computational linguistics often rejects statistical techniques, which are seen as essential in NLP. NLP draws on linguistic theory to varying degrees, and its multidisciplinary nature allows it to integrate concepts from linguistics, cognitive science, psychology, philosophy, and mathematics.

In recent years, alternative terms like ‘Language Technology’ or ‘Language Engineering’ have gained popularity alongside NLP. These terms reflect the broader scope of NLP, which extends beyond linguistic analysis to encompass the development of practical language-based applications. The choice of terminology may vary depending on the context and specific focus of the research or application.

The term ‘language’ is often used to refer to standard or canonical language varieties, distinct from speech or hybrid forms of language commonly used in CMC. CMC encompasses various digital platforms and communication channels, where non-standard or non-canonical language varieties are prevalent. NLP plays a crucial role in analysing and understanding language usage in CMC contexts, considering the unique characteristics and linguistic features of these interactions.

NLP is inherently multidisciplinary, drawing from various fields of study. Linguistics provides insights into language structures and theories that inform NLP research. Cognitive science offers an understanding of human language processing and comprehension, while psychology contributes to the study of language behaviour and user interactions. Philosophy offers philosophical foundations and ethical considerations, and mathematics, particularly logic, underpins formal language theory and certain computational techniques.

Within computer science, NLP intersects with several subfields. Formal language theory provides a theoretical framework for language modelling and analysis. Compiler techniques are relevant for developing efficient NLP algorithms and systems. Theorem proving contributes to logical reasoning and inference in NLP applications. Machine learning plays a significant role in training models and improving NLP performance. Human-computer interaction focuses on designing intuitive and user-friendly NLP interfaces.

NLP is a multidisciplinary field that encompasses the computational modelling of human language. It involves analysing the meaning of words, syntactic and semantic structures, and developing applications for various language-based tasks. NLP is closely related to linguistics, cognitive science, psychology, philosophy, and mathematics, drawing from these disciplines to enhance language understanding and processing. Furthermore, NLP has connections to computer science, specifically in areas such as formal language theory, compiler techniques, theorem proving, machine learning, and human-computer interaction. The multidisciplinary nature of NLP highlights its significance in advancing language technology and facilitating human-computer communication in the digital age.

2.3 Algeria's Linguistic Mosaic: History, Diversity, and Digital Creativity

The linguistic landscape in Algeria is a fascinating mosaic, shaped by centuries of cultural exchange and conquest. The diverse linguistic repertoire found in Algeria today is a result of prolonged acculturation, characterized by direct and indirect contact between genetically related and unrelated language varieties brought by successive waves of conquests and invasions throughout its extensive historical epochs. Language contact (aka Contact Linguistics, coined by Weinreich in "Languages in Contact: Findings and Problems", 1953) situations have played a pivotal role in shaping Algeria's linguistic richness and complexity, giving rise to phenomena such as diglossia, language switching/mixing, lexical borrowing, individual bilingualism, and plurilingualism (Berrabah, 2014).

One cannot underestimate the influence of language contact on the linguistic dynamics of Algeria. The encounter between different languages and cultures has led to the emergence of a unique linguistic hybridity within the Algerian Speech Repertoire. Additionally, the prevalence of informal formative linguistic features has become a prominent characteristic, especially in CMC and social networking sites, predominantly among the youth. It is crucial to recognize the historical and sociolinguistic factors that have shaped the linguistic landscape of Algeria. The country's linguistic diversity is evidence of its complex history, marked by various waves of conquest and cultural exchange (Berrabah & Benabed, 2021).

The impact of language contact on Algeria's linguistic repertoire cannot be understated, as it has resulted in a rich tapestry of language varieties and linguistic phenomena, best exemplified by Algerian Dialectal Arabic (ADA), the predominant variety in the speech of the Algerians, encompasses a fusion of mainly Arabic and French, with the adoption or adaptation of many words or expressions from other languages such as Spanish, Turkish, English, etc. In essence, ADA holds significant prominence in the speech repertoire, and accounts for the biggest slice of the cake of approximately "80% of the whole population" (Boucherit, 2002, p. 56), which makes it par excellence the Algerians' mother tongue.

CMC discourse in Algeria showcases an overabundance of informal linguistic phenomena and morphological processes, including *'Unconventional Acronyms/Initialisms'*, *'Clippings (Abbreviations)* ', *'Phonetic Approximations'* that involve representing words using letters or combinations of letters that approximate their pronunciation, *'Leetspeakings or number homophones'*, which involve manipulating numbers based on their visual resemblance to letters, *'Phonographs or letters homophones'* that represent the use of numbers instead of letters to reflect phonetic or phonological similarities, *'Misspellings and/or Typos'* (un/intentional omitting or adding letters or syllables in words, are widely utilized as lexical shortenings or shortcuts techniques in CMC texting), *Eye dialect'*, which is a deliberate misspelling textual device used in standard languages or the representation of regional dialects, *'Respellings'* that involves intentionally altering the spelling of a word due to the absence of corresponding letters in different languages, *'Neologisms, and Coinage'* are just a few examples of the linguistic creativity observed in the interactions within these digital spaces. These linguistic innovations are often employed by the younger generation and contribute to the ever-evolving nature of Algerian language use in digital communication (Berrabah, 2023).

3. Research Methodology, Participants and Data Collection Procedures

To address the requirements of validating prior findings, examining the research enquiries mentioned earlier, accomplishing the research objectives, and testing the formulated hypotheses, the current study employed a meticulously designed mixed methods research framework that seamlessly amalgamated quantitative and qualitative methodologies. The methodology encompassed two distinct data collection instruments, each serving a specific purpose in the investigation.

The first component involved conducting semi-structured interviews with six (06) esteemed supervisors specializing in linguistics, who were actively engaged in researching and analysing linguistic phenomena. These interviews were conducted within the Department of English at Ibn Khaldoun University of Tiaret,

Algeria. The interview consisted of five (5) carefully crafted items, tailored to elicit comprehensive insights from the supervisors regarding their experiences, perceptions, and suggestions about the analysis of non-canonical morphological processes in CMC within the Algerian context, utilizing usual NLP techniques.

The second component comprised a quasi-structured questionnaire, administered to twenty-four (24) Master's students specialized in linguistics within the aforementioned department. This questionnaire consisted of seven (7) items, including a combination of three open-ended questions and four closed-ended questions. The closed-ended questions encompassed various measurement levels, such as dichotomous, nominal, ranking order of preference, and filter or contingency questions. The objective of this questionnaire was to gather valuable insights from the linguistic students regarding their perceptions, experiences, and recommendations on the analysis of non-canonical linguistic features in CMC discourse in Algeria, utilizing NLP techniques.

These two research tools, the interview and the questionnaire, were employed in tandem to ensure a comprehensive understanding of the research topic, by capturing the perspectives of both knowledgeable informants (the supervisors) and actively engaged respondents (the students). The integration of both qualitative and quantitative data allowed for a multifaceted analysis of the research objectives, enabling a robust exploration of the research enquiries and hypotheses.

In addition to the aforementioned mixed methods approach, the study incorporated another research approach known as the Content Analysis Approach (CAA). This approach was utilized to scrutinize the non-canonical/non-standardised morphological processes in the examples provided by the informants. The objective was twofold: firstly, to corroborate the reliability of the results obtained from the questionnaire and interview, and secondly, to enhance the credibility of the research findings by employing a critical comparative discourse analysis of a significant corpus of linguistic practices in CMC discourse within the Algerian context. This additional step aimed to affirm the validity and accuracy of the study's conclusions, further reinforcing the robustness of the research outcomes.

By employing a meticulously designed mixed methods research framework, consisting of semi-structured interviews, a quasi-structured questionnaire, and a Content Analysis Approach, this study aimed to comprehensively examine the analysis of non-canonical/non-standardised morphological processes in CMC discourse in Algeria through the lens of NLP. The integration of multiple methodologies and data collection instruments ensured a rigorous and holistic exploration of the research topic, enabling the attainment of reliable and conclusive findings.

4. Data Analysis and Discussion

In this section, our primary focus is to ensure the reliability of our research by conducting a meticulous data analysis. our interpretations are firmly grounded in concrete scientific and empirical evidence to facilitate the reader's ability to observe, comprehend, and evaluate the limitations of current AI-driven NLP models in capturing the morphological processes within CMC discourse in Algeria. To accomplish this, we conducted separate analyses of the developed research tools to highlight the inherent limitations of current AI-driven NLP models to examine the morphological processes in CMC discourse in Algeria.

4.1 Questionnaire Analysis

Through the analysis of the questionnaire, we hope to contribute to a deeper understanding and assessment of the limitations of current AI-driven NLP models in capturing the morphological processes in CMC discourse in Algeria. Thus, the responses gathered from all questionnaire items are statistically presented in Tables 1-5, accompanied by rigorous and thorough interpretations of the findings.

Q Q	**1/How often do you use NLP-based tools/ applications for analysing CMC discourse in Algeria?**		**2/How satisfied are you with the accuracy of NLP models in capturing the informal written data in CMC discourse in Algeria?**	
Answers	Never	5 (20.83%)	Very dissatisfied	7 (29.17%)
	Rarely	2 (8.33%)	Dissatisfied	6 (25%)
	Often	4 (16.67%)	Neutral	4 (16.67%)
	Sometimes	6 (29.17%)	Satisfied	2 (8.33%)
	Always	7 (25%)	Very satisfied	5 (20.83%)
Total		**24 (100 %)**	**Total**	**24 (100 %)**

Table 1. *Frequency and Satisfaction of the Current AI-Driven NLP-Based Tool Usage for Analysing CMC Discourse in Algeria*

Based on the responses from a sample of 24 participants of the first question in Table 1, the frequency of using NLP-based tools or applications for analysing CMC discourse in Algeria is 20% (the Arithmetic Mean of the percentages). This indicates a relatively low level of adoption and utilization within the sample, which highlights varying degrees of engagement with AI-driven NLP models as follows: it was found that 20.83% of the respondents '*never*' use such tools or applications, indicating a lack of utilization in this context. Additionally, 8.33% of the participants reported '*rarely*' employing NLP-based tools or applications, suggesting infrequent usage. On the other hand, 16.67% of the respondents indicated that they '*often*' use these tools, implying a higher level of engagement in analysing Algerian CMC discourse. Another 29.17% reported using NLP-based tools or applications '*sometimes*', suggesting moderate but not consistent usage. Furthermore, 25% of the participants stated that they '*always*' utilize NLP-based tools or applications for this purpose, indicating a consistent and regular application.

The relatively low average percentage of respondents using NLP-based tools or applications for analysing CMC discourse in Algeria may be attributed to various factors. Limited awareness among the respondents about the existence and benefits of such tools in the Algerian context could contribute to the lower utilization. Additionally, resource constraints, including limited access to suitable hardware and a lack of expertise in NLP technologies, may hinder widespread adoption. The scarcity of Algerian-specific training data and language models, as well as cultural and linguistic variations in the discourse, could pose challenges for NLP-based analysis. Furthermore, alternative approaches or traditional research methods might be preferred or more commonly used by the respondents. Further investigation is necessary to gain a comprehensive understanding of the reasons behind the limited usage of NLP-based tools in analysing Algerian CMC discourse.

The satisfaction level with the accuracy of NLP models in capturing the informal written data in CMC discourse in Algeria varies amongst our participants as revealed in the percentages of the second question in Table 1. The analysis reveals that approximately 29.17% of the respondents expressed being '*very dissatisfied'*, while 25% reported being '*dissatisfied'*. In contrast, 16.67% of the participants indicated a '*neutral'* stance, suggesting a lack of strong satisfaction or dissatisfaction. A smaller segment of 8.33% stated being *satisfied*, while 20.83% expressed being '*very satisfied*' with the accuracy of NLP models. The calculated Arithmetic Mean (AM) of the percentages is 20%, indicating an overall moderate level of satisfaction among the respondents. These findings suggest that while a significant portion of the sample expresses dissatisfaction, there is also a notable proportion that holds positive views regarding the accuracy of NLP models in capturing CMC discourse in Algeria. Individual experiences, expectations, and contextual factors are likely to influence these varied perspectives.

The moderate level of satisfaction expressed by the respondents regarding the accuracy of NLP models in capturing informal written data in Algerian CMC discourse can be ascribed to several factors. Firstly, the presence of many formative linguistic features poses a challenge as NLP models trained on standard or formal language variants may struggle to interpret colloquial expressions and non-standard grammar and irregularities commonly found in informal language. Secondly, the limited availability of Algerian-specific

training data can impede the models' performance and accuracy in capturing the nuances of the CMC discourse in Algeria. Additionally, the cultural context embedded in the discourse, such as implicit cultural knowledge and context-specific expressions, may be difficult for NLP models to understand accurately.

Q Q	**3/ Have you encountered instances where NLP models fail to accurately analyse CMC discourse in Algeria? Please! Can you provide examples if possible?**			**4/ How well do you think the current NLP models understand and interpret the cultural and linguistic nuances specific to CMC discourse in Algeria?**	
Answers	**Yes** 21 (87.5%)	Illogical word order	1 (4.76%)	NLP models are not well-trained to handle unconventional abbreviations, acronyms, lexical shortcuts, words written with the letters of another language, morphological coinage, and innovative terms	13 (54.17%)
		Missing elements in sentences	2 (9.52%)		
		Lack of punctuation	3 (14.29%)		
		Unconventional acronyms or abbreviations	3 (14.29%)		
		Word combinations	2 (9.52%)	NLP models are not accustomed to analysing formative grammatical features with unusual punctuation	11 (45.83%)
		Single-letter representations	1 (4.76%)		
		Orthographic mistakes	4 (19.05%)		
		Transliterations	5 (23.81%)		
		Total	**21** (87.5%)		
	No	**Total**	**3** (12.5%%)	**Total**	**24** (100 %)

Table 2. *Instances of the Current AI-Driven NLP Models Failures and Assessment of Understanding Cultural/Linguistic Nuances in CMC Discourse in Algeria*

Based on the responses of the third question in Table 2., 87.5% of the respondents reported encountering instances where NLP models failed to accurately analyse CMC discourse in Algeria. This indicates a significant majority of the participants have experienced difficulties with NLP models in this regard. The examples provided by the respondents, such as: *illogical word order* 1 (4.76%), *missing elements in sentences* 2 (9.52%), *lack of punctuation* 3 (14.29%), *unconventional acronyms or abbreviations* 3 (14.29%), *word combinations* 2 (9.52%), *single-letter representations* 1 (4.76%), *orthographic mistakes* 4 (19.05%), and *transliterations* 5 (23.81%) highlight specific instances where the models may struggle to accurately analyse the language variations in CMC discourse in Algeria.

The exceptionally high frequency of instances reported by our respondents, where NLP models struggle to accurately analyse CMC discourse in Algeria, can be attributed to the linguistic intricacies present within the Algerian speech repertoire. This complexity arises from the coexistence and utilization of multiple language varieties within the same sentence or structure, particularly when the discourse involves the use of unconventional transliterations.

The respondents' answers concerning the fourth question in Table.2 highlighted several key points that current NLP models have difficulty understanding and interpreting the cultural and linguistic nuances specific to CMC discourse in Algeria. First, 13 (54.17%) respondents pointed out that these models are not well-trained to handle unconventional abbreviations, acronyms, lexical shortcuts, words written with the letters of another language, morphological coinage, and innovative terms. Second, 11 (45.83%) respondents stated that the models are not accustomed to analysing formative grammatical features with unusual punctuation. These limitations demonstrate the necessity for further training and development of NLP models to effectively capture the unique cultural and linguistic aspects of Algerian CMC discourse.

Q Q	**5/ Do you believe that the current NLP models adequately capture the language choices and code-switching patterns observed in CMC discourse in Algeria? Why or why not?**		
Answers	**Yes** 22 (91.67%)	The linguistic mosaic in Algeria is specific and socio-culturally bound.	7 (31.82%)
		There is a huge regional variation that's why a glossary can never be collected.	4 (18.18%)
		Language variation in Algeria is difficult even for the Algerians to understand not to deconstruct.	5 (22.73%)
		Those NLP models are not well-trained to interpret non-standardised linguistic practices.	6 (27.27%)
			22 (91.67%)
	No		**2** (8.33%)

Table 3. *The Current AI-Driven NLP Models Performance, Limitations, and Challenges in Analysing Informal Written Data in CMC Discourse in Algeria*

Based on the responses obtained from the fifth question in Table.3, only a small percentage 2 (8.33%) of the respondents believe that the current NLP models adequately capture the language choices and code-switching patterns observed in CMC discourse in Algeria. On the other hand, the majority of respondents 22 (91.67%) expressed their belief that the current NLP models do not adequately capture these linguistic aspects.

Among the 22 respondents (91.67%) who provided similar answers and expressed their disagreement, their explanations shed light on the challenges faced by NLP models in capturing the nuances of Algerian CMC discourse. Firstly, 7 respondents (31.82%) emphasized that '*the linguistic mosaic in Algeria is specific and socio-culturally bound'*. This suggests that the diverse linguistic landscape of Algeria, with its various dialects and regional variations, poses a significant challenge for NLP models to accurately capture and interpret the language choices and code-switching patterns prevalent in CMC discourse in Algeria. Additionally, 4 respondents (18.18%) highlighted '*the immense regional variation within Algeria, which makes it difficult to collect a comprehensive glossary*'. This regional variation further complicates the task for NLP models, as the linguistic practices and code-switching patterns can vary significantly across different regions of the country. Furthermore, 5 respondents (22.73%) pointed out that *language variation in Algeria is not only complex for outsiders or foreigners but even for Algerians themselves*. This suggests that the intricacies of language choices and code-switching patterns in CMC discourse in Algeria may be challenging for both NLP models and human speakers alike, making it difficult for the models to accurately capture and analyse these linguistic dynamics. Lastly, 6 respondents (27.27%) highlighted *'the lack of adequate training of NLP models to interpret the non-standardized linguistic practices observed in CMC discourse in Algeria'*. This limitation implies that the models' training data may not encompass the diverse linguistic variations, code-switching patterns, and informal language use present in CMC discourse in Algeria, leading to a suboptimal performance in accurately capturing these language dynamics.

In a nutshell, based on the 22 respondents (91.67%) who provided similar answers expressing their disagreement, it is evident that the current NLP models are perceived to inadequately capture the language choices and code-switching patterns observed in CMC discourse in Algeria. The linguistic mosaic, regional variation, language complexity, and insufficient training of the models are cited as significant challenges that hinder their effectiveness in accurately understanding and interpreting the nuances of CMC discourse in Algeria.

Q Q	**6 / In your opinion, what are the major limitations or challenges faced by NLP models in analysing the informal written data in CMC discourse in Algeria?**	
Answers	Neither the students nor their teachers have received adequate training	6 (25%)
	The lack of training in using software packages to analyse language varieties.	4 (16.67%)
	Dialectal variations within Algeria, which hinder the collection of a unified glossary of words	3 (12.5%)
	The phenomenon of language shifts and the mixing of different language varieties within the same sentence or word	4 (16.67%)
	The lack of sufficient linguistic data and research on the linguistic practices and word formation processes in in CMC discourse in Algeria	7 (29.17%)
Total	**24 (100 %)**	

Table 4. *Major Limitations and Challenges Faced by NLP Models in Analysing Informal Written Data in CMC Discourse in Algeria*

Several perceptions and opinions were shared by the respondents regarding the major limitations or challenges faced by NLP models in analysing informal written data in CMC discourse in Algeria raised in the sixth question which is represented in Table 4. Firstly, **6** respondents (25%) highlighted that *neither the students nor their teachers have received adequate training* on NLP models. In the same vein, four respondents (16.67%) stressed that *the lack of training in using software packages to analyse language varieties*, resulting in a limited ability to utilize NLP models effectively. This indicates that a significant proportion of the respondents perceive the lack of training sessions on NLP models as a challenge, which hampers the effective utilization and understanding of NLP models for analysing informal written data in CMC discourse.

Another notable challenge mentioned by the respondents relates that to the specific linguistic characteristics observed in Algerian discourse that is intricate and specific to their culture. Three respondents (12.5%) pointed out the *dialectal variations within Algeria, which hinder the collection of a unified glossary of words*. This suggests that the linguistic melting-pot in Algerian discourse poses a challenge for NLP models due to the complexity of dialectal variation, which further complicate the analysis, making it difficult for linguists to compile a comprehensive glossary or unify language variations.

Furthermore, four respondents (16.67%) emphasized *the phenomenon of language shifts and the mixing of different language varieties within the same sentence or word*. This highlights the intricacies of language mixing in CMC discourse in Algeria, which can pose challenges for NLP models in accurately analysing and interpreting the data.

Finally, seven respondents (29.17%) underscored the *lack of sufficient linguistic data and research on the linguistic practices and word formation processes in in CMC discourse in Algeria*. They worried that this limitation leads to insufficient training data for NLP models, preventing them from adequately capturing the cultural nuances, references, local contexts, and specific expressions unique to Algerian culture and society.

Q Q	**7 / What improvements or features would you like to see in NLP models to better capture the linguistic mosaic and informal language variations in CMC discourse in Algeria?**	
Answers	Enhanced training on informal language	8 (33.33%)
	Integration of Algerian dialects	2 (8.33%)
	Improved morphological analysis	4 (16.67%)
	Contextual understanding	3 (12.5%)
	Handling language mixing	5 (20.83%)
	Collaboration with linguistic experts	2 (8.33%)
Total	**24 (100 %)**	

Table 5. *Perspectives on Enhancing NLP Models for Capturing the Linguistic Mosaic and Informal Language Variations in CMC Discourse in Algeria*

Respondents provided valuable insights and suggestions on improvements or features they would like to see in NLP models to better capture the linguistic mosaic and informal language variations in CMC discourse in Algeria. One important suggestion is '*to enhance the training of NLP models to specifically address informal language variations*' found in CMC discourse in Algeria. This would involve training the models to understand dialectal variations, code-switching/mixing patterns, and the unique linguistic formative features commonly used in informal written communication. The suggestion received support from 8 respondents (33.33%), highlighting the significance of this improvement.

Additionally, 2 respondents (8.33%) pointed out *'the need to integrate Algerian dialects into NLP models to ensure they accurately capture the linguistic diversity present in Algeria'*. This suggestion, mentioned by recognizes the importance of considering dialectal variations in the development of NLP models for analysing CMC discourse in Algeria.

Others also highlighted *'the importance of improving morphological analysis to handle unconventional abbreviations, lexical shortcuts, and morphological coinage specific to CMC discourse in Algeria*'. This suggestion received support from 4 respondents (16.67%), underlining the need for NLP models to effectively analyse the unique morphological aspects of the informal written data.

Another significant suggestion is *'to enhance the contextual understanding of NLP models'*. Respondents stressed the need for models to have a deeper understanding of the local context and cultural references to accurately interpret and analyse the informal language variations in CMC discourse in Algeria. This suggestion was mentioned by 3 respondents (12.5%), highlighting the importance of contextual awareness in NLP models.

Furthermore, 5 respondents (20.83%) emphasized '*the importance of developing NLP models that can effectively handle language mixing'*, where different languages or language varieties are used within a single sentence or word. This would require improved language identification and code-switching/mixing detection capabilities. This suggestion recognised the complexity of language mixing in Algerian CMC discourse.

In addition to technical improvements, 2 respondents (8.33%) also suggested *'closer collaboration between NLP researchers and linguistic experts*' who possess a deep understanding of CMC discourse in Algeria. By involving linguistic experts in the development and evaluation of NLP models, the accuracy and effectiveness of capturing the linguistic mosaic and informal language variations can be improved.

4.2 Interview Analysis

This section focuses on examining the interview (Qs 1-5) in terms of how effective NLP models are in capturing the morphological processes in non-canonical language varieties in CMC Discourse. The goal is to assess the efficiency of these models and provide recommendations for enhancing their performance in this area. Additionally, throughout Content Analysis Approach (CAA), we endeavoured to deconstruct the informants' examples (cf., Table 6) and to validate the claims made and connect them to potential challenges in NLP's analysis of non-canonical language varieties.

1. Can you name or mention any of the NLP models that you have heard about or used?

Our informants cited several notable NLP models and resources, which all have been developed to address various linguistic tasks. One such model is BERT (Bidirectional Encoder Representations from Transformers), which has gained recognition for its strong performance on various NLP tasks. Another prominent model is GPT (Generative Pre-trained Transformer), including sequential versions like GPT-2, GPT-3 and GPT-4, which have demonstrated impressive language generation capabilities. Transformer-

XL is a variant that focuses on handling long-range dependencies in text, while RoBERTa (A Robustly Optimized BERT Pretraining Approach) improves upon BERT's (Bidirectional Encoder Representations from Transformers) pre-training techniques. ALBERT (A Lite BERT) addresses the memory and computational constraints of BERT. T5 (Text-to-Text Transfer Transformer) is a model trained on diverse NLP tasks. Moving beyond models, VADER (Valence Aware Dictionary and sEntiment Reasoner) is a rule-based sentiment analysis tool often utilized for analysing sentiment in social media text. Lexico3, on the other hand, is a lexical analysis resource offering insights on word meanings, usage, and relationships. These tools, such as VADER and Lexico3, are valuable resources in sentiment analysis and lexical analysis, respectively, contributing to the advancement of NLP research and applications.

2. Can you describe your experiences and challenges when using NLP models to analyse the informal written data in CMC discourse in Algeria?

Based on the provided answers to the second question from the informants, it is evident that there are several challenges when using NLP models to analyse informal written data in CMC discourse in Algeria. Let's analyse and discuss each of the challenges mentioned: To begin with, *'the failure to count the number of occurrences of linguistic phenomena or processes presents a significant challenge*'. NLP models struggle to accurately identify and quantify specific linguistic features or patterns in the informal written data, hindering detailed linguistic analysis and meaningful insights.

Furthermore, *'the failure to conceptualize word formation processes and identify linguistic practices is another challenge*'. NLP models may not effectively capture the morphological and lexical variations or the structural features that exist in CMC discourse in Algeria. The creative word formations and unique linguistic practices found in informal language pose difficulties for accurate understanding and analysis by the models.

In addition, *'the inability to limit sentence endings and identify their types*' poses challenges for NLP models. Effective segmentation and classification of sentences in CMC discourse in Algeria become difficult, impeding accurate analysis and interpretation.

Moreover, *'the issue with limiting the frequency of two or three-letter words reveals a limitation in the models' handling of specific linguistic constraints*'. The models may not accurately filter words based on length requirements, resulting in an excessive number of irrelevant words included in the analysis.

Another challenge is *'the failure to decipher code-switching and mixing syntactic/structural features*'. NLP models struggle to handle the complex language phenomena that exist in CMC discourse in Algeria. Code-switching, where speakers switch between languages or dialects makes it challenging for models to interpret the syntactic structures and linguistic features associated with each language variety.

Additionally, '*the misunderstanding of transliterated words, expressions, and sentences'* indicates difficulties faced by NLP models in accurately interpreting and analysing text with transliterations or phonetic representations from other languages or dialects. Pronunciation variations or inconsistent transliteration conventions contribute to this challenge.

Furthermore, *'the unfamiliarity with culturally-bound words reveals a lack of cultural knowledge*' within the NLP models. Algerian CMC discourse often includes words or expressions specific to the local culture, which may not be recognized or understood by models trained on more general language data.

Lastly, the challenge posed by *'words containing symbols, signs, numbers, transposed letters, combined words, or missing/added letters*' highlights the difficulty NLP models face in analysing text with non-standard spelling, typographical errors, or unconventional word formations. These variations, commonly found in informal written data, can confound the models' ability to accurately analyse and interpret the text.

3. Based on your expertise, what specific word formation processes or phenomena within CMC discourse in Algeria do you believe pose the greatest challenges for NLP models? Why? Could you provide specific examples?

In light of the informants' answers to the question about the specific linguistic features or phenomena within CMC discourse in Algeria that pose the greatest challenges for NLP models, several key insights emerge. The informants identified several challenges: transliteration and related issues, leetspeak and phonographs, unconventional abbreviations and acronyms, and lexical shortcuts. Transliteration presents difficulties for NLP models due to inconsistencies in spelling, diacritics, and pronunciation when Arabic words are written using Latin characters. Leetspeak and phonographs, which involve replacing letters with alphanumeric characters, create deviations from standard linguistic conventions, making it challenging for models to interpret these non-standard representations. Unconventional abbreviations and acronyms commonly used in CMC discourse introduce challenges due to their context-dependent and non-standard nature, requiring models to possess contextual knowledge to accurately interpret them. Additionally, lexical shortcuts, which involve truncating words or phrases for efficiency, further complicate language understanding for NLP models. These linguistic challenges have arisen mainly from the non-standard language usage, inconsistencies in spelling and pronunciation, and context-dependency inherent in CMC discourse in Algeria.

In the analysis of examples presented in Table 6, as a heart of this research, a multitude of unforeseen non-canonical morphological processes in CMC discourse in Algeria came to light. These processes pose challenges for current AI-driven NLP models. One issue highlighted by the informants is transliteration, which involves writing words using a different alphabet. In the case of the non-canonical language variety ADA, there are various linguistic issues related to transliteration.

First and foremost, we encounter the perplexing issue of *'multiple script'*. This refers to the unconventional alteration of spelling for the same word with the same meaning, resulting in seemingly distinct lexemes. For instance, the term 'عيد' /ʕi:d/, meaning 'celebration' or 'festival,' has been transcribed as 'Aid, Eid, eid, Id, 3id,' and so on.

Another issue is an intriguing phenomenon known as *'misspellings and/or typos*'. These intentional or unintentional omissions or additions of letters or syllables serve as lexical shortcuts in CMC texting. Take, for example, 'khe***m***dma' /χe***m***dma/, which stands in for 'khedma' /χedma/, meaning 'work' or 'job.'

Furthermore, we stumble upon the fascinating practice of *'combining words* into a single entity'. This involves attaching two or more words, and defying conventional linguistic boundaries. A striking example is 'Aidkoummoubarak,' instead of 'Aidkoum moubarak' /ʕi:dkəm mubɑ:rɑk/, a popular greeting conveying blessings and well wishes during Muslims' religious holidays or festive occasions. Equally intriguing is the occurrence of *'separated words*', where morphemes are detached from their roots as if they were standalone entities. Consider 'laal aaml jay,' which should be 'laalaamljay' /lel ʕɑ:m eldʒɑ:j/, denoting 'for the coming/next year.'

Moreover, we come across the curious use of *'Eye Dialect'*, intentional misspellings that reflect standard languages or regional dialects. A prime example is 'yet***k***abel' /jət***k***a'bel/), representing 'yet**q**abel' /jət**q**a'bel/, meaning 'may accept'. Additionally, we uncover the fascinating process of *'respellings'* due to the absence of corresponding letters across different languages. This intentional alteration of spelling manifests in forms such as 'Ɜaye***t***lak' /ʕaj***t***lək/, contrasting with 'Ɜaye**ŧ**lek' /ʕaj**t**ʕlək/, which means 'I called you.'

Moving on, we encounter *'phonetic approximations*', where words are represented using letters or combinations that closely approximate their pronunciation. An illustrative example is 'wlh,' standing for 'والله' /wə ælˈɑːh/, conveying the expression 'I swear Allah'.

processes	Examples
Multiple Script of the same word	**1)**The phrase ‘إن شاء الله’ /ɪn ʃɑːʔ ælˈɑːh/ (If Allah wills) has been unconventionally written like: inchallah, Inchallah, inchalh, anchallah, challah, inshalah, nchlh, Nchallah, inchaallah, in Sha2 Allah**,** etc **2)**The term ‘عيد’ /ʕiːd/ (‘celebration’ or ‘festival’, which is commonly used to refer to Muslims’ religious holidays or festive occasions) has been transcribed like: Aid, Eid, eid, Id, 3id, etc
Misspellings and/or Typos	**1)** el**h**eir/əl**h**iːr/ *instead of* el**kh**eir=الخير /əlχiːr/(The goodwill, benevolence, goodwill) **2)** khe**m**dma /χe**m**dma/*in place of* khedma=خدمة/χedma/ (work/job) **3)** masot**l**ch /mɑː sɒˈte**l**ʃ/ *rather than* masote**k**ch= ما صتكش /mɑː sɒˈte**k**ʃ/ (I did not find you) **4)** Sa**h**a /sæ'**h**æ/ *in place of* Sa**ħ**a /sæˈ**ħ**æ/ (Ok, health, etc)
Combined Words	**1)** manegderch=ما نقدرش/mɑː negderʃ/ (I cannot), **2)** wmataz3afch =ماتزعفش /wəmɑː tezʕefʃ / (and do not be angry), **3)**Aidkoummoubarak= عيدكم مبارك/ʕiːdkəm mubɑːrɑk/ (a common greeting used to wish someone a blessed or happy holiday, particularly during the Islamic festival of Eid), **4)ṭ**wltwana=طولت و انا /ṭˤaˈwəlt wɑnɑː/ (I waited a lot and I),**5)** mknwlw= ما كان والو/mɑː kɑːn wɑluː/ (there is nothing) , **6)** matansach= ما تنساش/mɑː tensɑːʃ/ (do not forget), **7)** manzidechnahdar= ما نزيدش نهدر /mɑː nziːdtʃ nehdar/ (I will not talk anymore), **8)**wma7kmtkch=و ما حكمتش /wə mɑːħkemtʃ/ (and I could not reach)
Separated Words or Morphemes	**1)** lel jami3a=لل جامعة /lel dʒɑːmiʕə/ (to the university), **2)** bi khir=ب خير /bχiːr/(fine/well), **3)** bel ongli= بال اونقلي /belɒngliː/ (in English), **4)** laal **a**aml jay=لل عام الجاي /lel ʕɑːm eldʒɑːj/(for the coming/next year), **5)** l ga**d**ia= ال قضية/ lgadˤijə/ (the issue/problem)
Eye Dialect and Respellings	**1)**yahaf***dh***ak /jaha***d***ək/ *as opposed to* yahaf**đ**ak/jahafðˤək/=يحفظك (May protect you),**2)** 3aye**ṭ**lak /ʕaj**ṭ**ləkˤ/ *contrasted with* 3aye**t**lek /ʕaj**t**ˤləkˤ/=عيطتلك(I called you), **3)** 3a***th***ama/ʕa***θ***ama/ *rather than* 3a**đ**ama= عظم/ʕaðˤama/(multiply reward), **4)** Rama***d***an /ræmə***d***æn/ *in preference to* Rama***ḍ***an=رمضان /ræməḍˤæn/(the 9th month of the Muslim year, during which time Muslims fast), **5)** ***g***adan/***g***adan/ *in place of* **gh**adan=غدا /**ɣ**adan/ /(tomorrow)**, 6)** rou***h*** /ru***h***/ *instead of* rou**ħ**/ru**ħ**/=روح (go), **7)** ***r***adwa /***r***adwa/ *rather than* **gh**adwa=غدوة /**ɣ**adwa/(tomorrow), **8)** Be3a***t***eli /bəʕa***t***li/ *in lieu of* Be3a**th**eli=بعثلي /bəʕa**θ**li/ (send me), **9)** ma***h***ta***j***ek/ma***h***ta͡dʒak/ *more willingly than* ma**ħ**ta**j**ek =محتاجك /maħta**d͡ʒ**ak/ (I need you), **10)**yet***k***abel/jət**k**a'bel/ *representing* yet**q**abel/jət**q**a'bel/ يتقبل(may accept), **11)** ***k***ali /***k***aːli/ *rather than* **kh**ali= خالي /χaːli/(my uncle)
Phonetic Approximation	**1)**Slm=سلام /səˈlɑːm/(Peace be upon you), **2)**w=و /wəˈ/ (and), **3)**f=في /fiː/ (at/in), **4)**mnkm=منكم/minkum/(from you), **5)**yslmk= يسلمك /jəselmək/(May Allah bless you), **6)** l =لل /liː(l)/(the/to), **7)** ahl=اهل /æhl/ (family), **8)**wlh=والله /wə ælˈɑːh/ (I swear Allah), **9)** mn=من /min/ (from/of), **10)**b=ب /biː/ (with/by/in/at/or),
Leetspeakings and Phonographs	**1)**Defa**3**eli=دفعلي /dəfaːʕəli/ (submit to me), **2)**mataz**3**afch=ماتزعفش /maˈtəzʕafʃ/ (don’t be angry), **3)**sa**7**bi=صاحبي /sæ**ħ**bi/(my friend), **4)9**riba=قريبة/**g**riba/(closer or nearer), **5)**n**9**ar**3**lk=نقارعلك/n**g**a:r**ʕ**ək/(I will wait for you), **6)6**bib=طبيب/**tˤ**biːb/(doctor), **7)** Insha**2** Allah=إن شاء الله /ɪn ʃɑːʔ ælˈɑːh/ (If Allah wills),
Words With Unusual Affixations	**1)**liafichawah=افيشاوه لي/liːaːfiʃaːwah/(that they displayed), **2)**Bipili=بيبيلي/bipiːli/(give me a beep), **3)**fleexily=فليكسيلي/flɛːksili/(Top up my balance/recharge my mobile phone), **4)** تليفونك /tiːlifunək/(your mobile phone), **5)**الشوفو/əˈʃʊfu/(the bath-heater), **6)**السطاسيون /əs.tˤæsijun/ (the bus station or Fuelling station), **7)** بلاپيس/blæpiːs/(with the mobile phone chip), **8)**الكونيكسيو /əlkɔniːksju/ (connection), **9)** المساج/əlmiːsaːʒ/(the message **10)**الپرامات /əlpæraːmaːtər/(the settings), **11)**السمانة /əlsæmaːna/(the week), **12)**الصالة /əlˈsæla/(reception room/hall), **13)** سكانيهم/skæniːhum/(scan them), **14)**اللوطو /əˈluːtˤuː/(the car), **15)** الميكانيسيان/əlmeːkaniːsijaːn/(the mechanic) **16)**يصوني /jəswuni/(ringing), **17)**الكريدي /əlkriːdi/(credit)
Non-conventional/ Inherent Polysemy	Saha /sæˈhæ/ *instead of* Saħa /sæˈħæ/ (Ok, health/strength, thanks, hi/hello, recovery, understood, shut up, bye, etc.)
Fusion-scripting	نيvو/niːvo / (level) ,لاوار / lagaːr / (station), الvار /elvaːr / (VAR: Video Assistant Referee), غzة/ɣaza/ (Gaza) ،فيلم /film/ (film) يديو/"vidiəʊ/ (video) vمدن /ədmin/ (admin) ،اع g /gaːʕ/ (all)
Unconventional Acronyms	**1) Cc** (Fr **C**omment **ç**a va? i.e., How are you doing?) *vs.* **Cc** (Eng **c**ubic **c**entimetre or **C**arbon **c**opy), **2) Cv** (Fr **c**a **v**a?) *vs.* **Cv** (**c**urriculum **v**itae)
Unvowelized / Non-vowelized Words in Arabic.	**1) اهله** *might be read either* **أَهْلَه** /ʔahˈlahu/ (Show us the crescent of Ramadan, or show it to us**)** *or* **أَهْلُهُ** /ʔaˈhlahu/(his family) **2) مهمة** *might be inferred either* **مُهِمة** /muhɪˈma/ (important/significant) *or* **مَهمّة** /mahaˈma/ (task/mission) **3) قبل** *might be guessed either* **قُبِلَ** /qabila/ (accept/agree) *or* **قَبْلَ** /qabəl/ (before) **4) يكبر** *might be predicted either* **يَكْبُر**/yakbɔr/ (getting older or ageing) *or* **يُكَبِّرُ** /jukaˈbiru/ (to magnify or declare the greatness of Allah)

Table 6. *Examples of Morphological Processes Posing Challenges for NLP Models in CMC Discourse*

Attractively, in Table 6 also, we come by *'leetspeakings or number homophones*', where numbers are manipulated based on their visual resemblance to letters. For instance, 'sa**7**bi' represents 'صاحبي' /sæ***ħ***bi/), meaning 'my friend'. Similarly, we chance upon *'Phonographs or letters homophones*', wherein numbers replace letters to reflect phonetic or phonological similarities. Consider 'Defa3eli,' representing 'دفعلي' /dəfaːʕəli/, meaning 'submit to me.'

Furthermore, we bump into *'transliterated words with unusual affixations*'. This process involves combining roots and affixes from foreign or genetically unrelated varieties within a single word. A notable example is 'سكانيهم' /skæniː***hum***/, which means 'scan them.'

Then, '*fusion-scripting*' refers to a linguistic phenomenon observed in CMC discourse. It involves a unique morphological process known as fusion or coalescence, wherein one or more letters are intentionally or unintentionally substituted by letters from a genetically unrelated or divergent language variety. These substituted letters are then merged within the same word. This fusion occurs due to the differences in phonetic segments or sounds between the primary language variety (ADA) and other coexisting foreign varieties. Speakers feel compelled to incorporate a letter from the foreign variety into the primary one because the latter lacks an equivalent letter. For example, the word "ال**v**ار"/elvaːr/ represents the fusion-scripting of the acronym "VAR" (Video Assistant Referee).

Another remarkable phenomenon is *'non-conventional polysemy*,' that Ortega-Andrés & Vicente (2019) defined as "senses have to be generated online via pragmatic mechanisms. It is also habitual to refer to the kind of polysemy …discussed by Pustejovsky (1995) as 'inherent polysemy',” (p.2). Accordingly, *'non-conventional/ inherent polysemy'* occurs where the same transliterated word conveys different meanings depending on the context. For instance, 'Sa***ħ***a' /sæ'***ħ***æ/ instead of 'Saħa' /sæˈħæ/ can mean 'Ok,' 'health/strength,' 'thanks,' 'hi/hello,' 'recovery,' 'understood,' 'shut up,' 'bye,' and much more.

Lastly, we come across *'unconventional acronyms/initialisms*', wherein uncommon acronyms are used. An intriguing example is '**C**c', standing for the French phrase '**C**omment **ç**a va?' (How are you doing?). These unconventional acronyms and initialisms add an additional layer of complexity to the CMC discourse.

In addition to the morphological processes related to transliteration, another linguistic phenomenon emerges, presenting a distinct challenge for current AI-driven NLP models. This phenomenon is context-dependent and revolves around the process of vowelisation in Arabic. Vowelisation in Arabic involves the use of symbols placed above or below letters to aid speakers or readers in easily identifying the word's category, pronunciation, and meaning. However, in CMC discourse, a contrasting morphological process occurs, termed *'unvowelized' or 'non-vowelized words in Arabic'*. This process involves the exclusion of those symbols, leading to potential misunderstandings or miscomprehension of the words and the overall message. For instance, consider the word 'مهمة' /muˈhma/, which may be inferred in isolation either as 'مُهِمة' /muhɪˈma/) meaning 'important' or 'significant,' or as 'مَهمّة' /mahaˈma/ meaning 'task' or 'mission.' The absence of vowel symbols in CMC discourse makes it challenging for AI-driven NLP models to accurately decipher the intended meaning, leading to potential confusion. The exclusion of vowelisation in CMC discourse highlights the importance of understanding the broader context and relying on contextual cues for accurate interpretation. This phenomenon underscores the need for AI-driven NLP models to incorporate contextual analysis and disambiguation techniques to overcome the challenges posed by non-vowelized words in Arabic.

Overall, addressing these non-canonical morphological processes in CMC discourse is crucial for enhancing the accuracy and effectiveness of AI-driven NLP models in understanding and generating text within Algerian contexts. By unravelling the intricacies of transliteration and its associated phenomena and recognizing the significance of vowelisation and its absence in CMC discourse, linguists can develop more robust language processing models that account for context-dependent linguistic phenomena. This, in turn, enhances the accuracy and effectiveness of AI-driven NLP systems in understanding and generating text-based software packages within Arabic contexts, fostering clearer and more meaningful communication in the digital realm.

4. What do you think are the key reasons behind the models' inadequacy in capturing the linguistic nuances and informal language variations present in CMC discourse in Algeria?

The inadequacy of current AI-driven NLP models in capturing the linguistic nuances and informal language variations present in CMC discourse in Algeria can be attributed to several key reasons. Firstly, these models were primarily formulated to handle conventional linguistic processes and standardized linguistic patterns, and they have been trained on existing standardized linguistic structures. As a result, they may struggle to accurately capture the dynamic and ever-changing nature of CMC discourse in Algeria, which involves frequent code-switching and borrowing between language varieties. Secondly, there is a lack of empirical research and analysis on the content of language choice in non-canonical varieties of languages, with many linguistic endeavours being more theoretical or relying solely on perceptions and attitudes about language use. The scarcity of authentic research conducted on real-life regional variations and non-canonical Algerian dialects hampers the development of comprehensive linguistic resources and datasets necessary for training the models effectively. Furthermore, the objectives of teaching linguistics or conducting research in linguistics in Algerian contexts may be vague, resulting in limited efforts to collect glossaries or linguistic resources that encompass the non-canonical Algerian dialects. These factors collectively contribute to the models' inadequacy in capturing the linguistic nuances and informal language variations in CMC discourse in Algeria.

5. From your perspective, what steps should be taken to improve the accuracy and effectiveness of NLP models in analysing the informal written data in CMC discourse in Algeria?

According to our informants, to improve the accuracy and effectiveness of NLP models in analysing informal written data in CMC discourse in Algeria, several key steps can be taken. Firstly, collaborative team linguistic works should be conducted to create a comprehensive glossary database that captures the linguistic practices and word formation processes specific to CMC discourse in Algeria. This collaborative effort would involve researchers, linguists, and experts working together to document the diverse linguistic patterns and variations found in informal written data, providing NLP models with a solid resource for accurate analysis. Secondly, organizing training sessions for teachers, students, and researchers on using NLP software packages is essential. These sessions would equip individuals with the necessary skills and knowledge to effectively utilize NLP tools and analyse the informal written linguistic patterns in CMC discourse. By empowering educators and researchers, they can contribute to advancing the accuracy and effectiveness of NLP models in this domain. Additionally, other possibilities include building domain-specific language models trained on Algerian CMC discourse, incorporating sentiment analysis techniques to capture emotional elements, involving native speakers as annotators for high-quality labelled datasets, and continuously evaluating and improving the models based on user feedback. These steps and possibilities collectively strive to enhance the accuracy and effectiveness of NLP models in analysing informal written data in CMC discourse in Algeria, leading to more comprehensive and context-aware language understanding.

5. Results and Discussion

The analysis of informant and respondent data reveals significant challenges faced by current AI-driven NLP models in accurately capturing the linguistic nuances present in CMC discourse in Algeria. The inadequacies stem from the models' inability to handle non-standard language variations, context-dependent linguistic phenomena, and the unique socio-cultural elements embedded within this informal written data. Overcoming these limitations is crucial for developing more robust and context-aware NLP systems tailored to the Algerian context.

The core results highlight several key issues that hinder the performance of existing NLP models. Firstly, the models struggle to accurately identify and quantify specific linguistic features and patterns found in informal Algerian CMC data, such as creative word formations, code-switching, and unconventional abbreviations. Secondly, the models face difficulties in handling non-standard morphological processes like transliterations, leetspeak, and phonetic approximations, which deviate from conventional linguistic norms. Additionally, the lack of cultural knowledge within the models impedes their understanding of culturally-bound words and expressions common in Algerian CMC discourse.

A significant challenge arises from the non-canonical morphological processes prevalent in Algerian CMC discourse. Phenomena like multiple script variations, misspellings or typos, eye dialect and respellings, combined words, separated morphemes, words with unusual affixations, non-conventional/ inherent polysemy, fusion-scripting and unconventional acronyms pose obstacles for accurate text analysis. Furthermore, the models encounter issues with transliterated words exhibiting unusual affixations, fusion-scripting (letter substitutions from different language varieties), and non-conventional polysemy (context-dependent multiple meanings). The absence of vowelisation in Arabic words within CMC discourse also introduces ambiguity, requiring contextual cues for accurate interpretation.

The respondents and informants attribute these challenges to several key factors. Firstly, NLP models are primarily trained on standardized linguistic structures, rendering them ill-equipped to handle the dynamic and ever-changing nature of the non-canonical CMC discourse in Algeria. Secondly, there is a lack of empirical research and authentic linguistic resources capturing the regional variations and non-canonical Algerian dialects, hindering the development of comprehensive training datasets. Additionally, the objectives of linguistic research in Algerian contexts may be vague, resulting in limited efforts to collect glossaries or linguistic resources encompassing these non-canonical varieties.

To address these limitations, collaborative efforts among researchers, linguists, and experts are crucial for creating comprehensive glossary databases that capture the linguistic practices and word formation processes specific to Algerian CMC discourse. Furthermore, organizing training sessions on NLP software packages, building domain-specific language models trained on CMC data in Algeria, incorporating sentiment analysis techniques, involving native speakers as annotators, and continuously evaluating and improving the models based on user feedback are essential steps. By implementing these measures, NLP models can better capture the linguistic mosaic and informal language variations in CMC discourse in Algeria, leading to more accurate and context-aware language understanding and analysis.

6. Conclusion

In conclusion, the study reveals that current AI-driven NLP models face significant challenges in accurately capturing the intricate linguistic complexities present within CMC discourse in Algeria. The core impediments stem from the models' inability to handle the non-standard language variations, context-dependent linguistic phenomena, and socio-cultural nuances inherent in this informal written data. Specific issues include difficulties identifying creative word formations, code-switching, transliterations, and culturally-bound expressions (this confirms our first presumed hypothesis). Furthermore, non-canonical morphological processes like multiple script variations, misspellings or typos, eye dialect and respellings, combined words, separated morphemes, words with unusual affixations, non-conventional/ inherent polysemy, fusion-scripting, unconventional acronyms and un-vowelisation pose substantial obstacles (this approves our second assumed hypothesis). These shortcomings can be attributed to the models' training on standardized linguistic structures, the lack of comprehensive research and resources capturing Algerian dialectal variations, and vague objectives guiding linguistic studies in this context (this supports our third pre-set hypothesis). To overcome these limitations, concerted efforts involving linguists, researchers, and experts are essential for creating glossary databases, organizing NLP training sessions, developing domain-specific language models, incorporating sentiment analysis, collaborating with native annotators, and

continuously refining the models' text-based software packages on user feedback. Further research is also warranted on structural and textual analysis of CMC discourse in Algeria to uncover additional linguistic patterns that can inform more robust NLP model development. By addressing these challenges through tailored approaches, NLP models can achieve enhanced accuracy and contextual understanding, ultimately enabling more robust analysis of the rich linguistic mosaic present in Algerian CMC discourse.

References

- Aronoff, M. (1976). *Word formation in generative grammar*. Cambridge: MIT Press.
- Bauer, L (1983). *English word formation*. In R. Hogg & D. Denison (Eds.), *A history of the English language* (pp. 352-395). Cambridge University Press.
- Bauer, L. (2001). *Morphological production and English derivation: A corpus study in linguistics*. London: Cambridge University Press.
- Berrabah, A. & Benabed, A. (2021). Contact-induced phonological change of the phoneme /s/ in the speech of EFL learners and teachers in Algeria: a case study. International journal of linguistics, literature and translation (IJLLT) IV-2, pp. 136-143.
- Berrabah, A. & Benabed, A. (2021). The Discourse of text messaging as a locus of contact-induced linguistic change in Algeria: the case of Relizane speech community, *Faslo El-Khitab' by the Research Laboratory of 'Argumentative Discourse, its Origins, References and Perspectives in Algeria'*, Ibn Khaldoun University of Tiaret, Algeria, V10, N4, Pp 245-258
- Berrabah, A. (2014). The Ideological beliefs towards code-switching/mixing practices and their effects on communication efficiency via 'SMS language' in Relizane: A critical discourse analysis. Unpublished thesis of Magister. University of Oran
- Berrabah, A. (2023). A sociolinguistic perspective to the study of SMS discourse features in Algeria (Unpublished doctoral dissertation). Djillali Liabes University of Sidi Bel Abbes, Algeria.
- Bouamrane, A. (1988). *Arabic-French code-switching in Algeria*, University of Oran.
- Boucherit, A. (2002). Algérie: de l'Arabe à l'Arabisation, *Language Contact and Language Conflict in Arabic. Variations on a sociolinguistic theme*, Routledge Curzon, pp. 54- 69
- Crystal, D. (2001). *Language and internet*. Cambridge, Cambridge University Press.
- Crystal, D. (2008). Texting. *ELT Journal: English Language Teachers Journal*, *January*: 77-83.
- Crystal, D. (2008). *Txtng: the gr8 db8*. Oxford, Oxford University Press.
- Deumert, A. & S. O. Masinyana. (2008). "Mobile language choices _ The use of English and isiXhosa in text messages (SMS)". *English World-Wide* 29. 117_147.
- Halliday, M. A. K. 2003. "Written Language , Standard Language , Global Language." *World Englishes* 22 (4): 405–18.
- Harrat, S., Meftouh, K., Abbas, M., & Smaïli, K. (2014). Building resources for Algerian Arabic dialects. In *15th Annual Conference of the International Communication Association Interspeech, Sep 2014, Singapour, Singapore*. Retrieved from https://hal.inria.fr/hal-01066989
- Harrat, S., Meftouh, K., Abbas, M., Hidouci, W. K., & Smaïli, K. (2016). An Algerian dialect: Study and resources. *International Journal of Advanced Computer Science and Applications (IJACSA), 7(3), 384-396*. https://doi.org/10.14569/IJACSA.2016.070353
- Herring, S. C. (2004). "Slouching toward the ordinary: Current trends in computer-mediated communication." *New Media & Society* 6 (1): 26–36.
- Herring, S. C. (2007). A faceted classification scheme for computer-mediated discourse. Language @Internet, 4. URL: http://www.languageatinternet.de/articles/2007/761/.
- Herring, S. C. (2008). "Foreword." *In Handbook of Research on Computer Mediated Communication*, edited by Sigrid Kelsey and Kirk St.Amant, xxxv–xxxvi. Hershey . New York: Information Science Reference.

- Ibrahim, R. (2013). Reading in Arabic: New Evidence for the Role of Vowel Signs. *The Edmond J. Safra Brain Research Centre for the Study of Learning Disabilities and Department of Learning Disabilities.*
- Kelsey, S. and Amant, K. 2008. "Preface." In *Handbook of Research on Computer Mediated Communication*, edited by Kelsey, S. and Amant, K., xxxvii–xxxviii. Hershey. New York: Information Science Reference.
- Ortega-Andrés, M., & Vicente, A. (2019). Polysemy and co-predication. Glossa: A Journal of General Linguistics, 4(1), 1-23. https://doi.org/10.5334/gjgl.564
- Plank, B. (2016). What to do about non-standard (or non-canonical) language in NLP [PDF]. Retrieved from http://arxiv.org/pdf/1608.07836.pdf.
- Saadane, H., & Habash, N. (2015). A conventional orthography for Algerian Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing* (pp. 69-79). Beijing, China: Association for Computational Linguistics.
- Shawkat, N. & Parveen, H. (2017). *Content analysis*. Retrieved from https://www.researchgate.net/publication**.**
- Stafford, L., & Hillyer, J. D. (2012). Information and communication technologies in personal relationships. *Review of Communication, 12(4),* pp. 290-312.