STUDIA PHÆNOMENOLOGICA

XXVI (2026) 201-209

Challenges of Using Deep Learning to Analyse Texts Written in the Algerian Arabic Dialect

Dr. Rekia Djermouni[1], Pr. Réda Baba Ahmed[2]

[1]University of Mustapha Stambouli – Mascara, Algeria

[2]University of Mustapha Stambouli – Mascara, Algeria

Email: [1]rekia.djarmouni@univ-mascara.dz; [2]r.babaahmed@univ-mascara.dz

Abstract

The study of Arabic dialects, particularly Algerian Arabic, and texts written in them poses significant challenges for automatic processing and the tasks required by researchers and users, especially given the widespread emergence of dialect-written posts on social media platforms. Various problems have consequently become apparent that hinder the identification of words and expressions in these dialects. These include ambiguity due to the lack of digitised corpora, issues related to orthographic writing systems and interference between dialectal writing and Modern Standard Arabic or foreign languages. Accordingly, we conducted a study of a sample of Facebook posts published by students of the Department of Arabic Language and Literature at the University of Mascara. The posts were written in the Algerian dialect (Darja), and we observed the difficulties this presented in recognising the words they contained. This has led researchers to rely on deep learning techniques to achieve a certain level of success.

Keywords: deep learning, text recognition, Algerian dialect, ambiguity, semantic ambiguity.

Submission: 15.08.2025. Acceptance: 12.02.2026. Publication: 29.05.2026

Introduction:

Artificial intelligence refers to the design of algorithms that, when executed, can perform tasks typically assigned to humans. One of the most prominent of these tasks is learning, whereby a machine acquires knowledge by analysing data and extracting useful information in order to improve its performance in a specific area. This is followed by a training phase, during which the same tasks are carried out using new data.

Inspired by the human nervous system, particularly the way neurons operate, researchers devised a methodology for simulating the machine's learning process. This methodology has contributed to the development of deep learning technology, which has proven highly effective in analysing natural languages.

Natural languages exhibit variation and dialectal diversity; therefore, the study of Arabic dialects and texts written in them is currently a significant challenge and important area of research. This is largely due to the widespread use of dialectal posts on social media networks, despite more conservative users' resistance to this style of writing. They urge users to write in Modern Standard Arabic. Consequently, difficulties arise in automatically processing and analysing such texts using artificial intelligence techniques, including deep learning.

In this study, we will analyse a sample of Facebook posts written by students of the Department of Arabic Language and Literature at the University of Mascara. These posts are written in Darja (an Algerian dialect) and published on the department's Facebook page. Our aim is to identify the most significant issues arising from this choice of language and to examine its impact on the automatic processing of these texts using deep learning. This will enable us to propose a set of linguistic and orthographic constraints to help reduce ambiguity in word writing and facilitate recognition of the words used in these posts.

1. The Concept of Artificial Intelligence

In order to understand artificial intelligence, it is first necessary to define human intelligence. This is associated with cognitive capabilities that are unique to human beings, such as the ability to adapt to life circumstances, draw on previous experiences and knowledge, think and analyse, plan, solve problems, apply sound reasoning and sense the needs of others. It also encompasses learning quickly and making effective use of what has been learnt.

In contrast, artificial intelligence involves the simulation of human intelligence in an attempt to understand its nature. This is achieved through computer programs designed to mimic human thought processes and behaviour in relation to the environment.

Artificial intelligence is a branch of computer science. The focus is on creating advanced, high-speed software with high competence that replicates the processes of perception and logical inference typically performed only by humans. Such systems can carry out many difficult and complex tasks that would take humans a long time to accomplish. They are also characterised by their ability to learn from mistakes.

This process relies on a knowledge base containing the concepts, theories and procedures provided by humans to support appropriate decision-making.

Characteristics of artificial intelligence

Artificial intelligence has many features that make it an effective investment across various fields.

- widespread implementation in devices, smartphones and machines to assist with planning and problem solving using logic;
- speech and sound recognition, as well as the ability to control and move objects.
- Input understanding and analysis, whereby AI-enabled devices can accurately process user inputs and provide efficient, user-specific outputs.
- Continuous learning, whereby the learning process becomes automatic and self-directed without the need for human monitoring or supervision.
- The ability to process vast amounts of information.
- Pattern recognition: AI can identify and analyse similar patterns in data more effectively than human cognition.
- Problem-solving for unfamiliar challenges by applying cognitive capabilities to find solutions. (Al-Sayyid, 2020, pp. 21–23).

2. Language and artificial intelligence

Language is one of the most important cognitive activities and mental capacities that are unique to human beings. Artificial intelligence seeks to imitate and replicate these processes automatically. In

fact, a specific field within artificial intelligence is dedicated to this purpose: Natural Language Processing (NLP). This field investigates how to enable machines to understand natural language at different levels through analysis and generation. The ultimate goal is to create an interactive environment that facilitates communication between humans and machines in both written and spoken forms. (Atiya, 2019, p. 14).

Moreover, these AI efforts pursue several goals, the most prominent of which is to standardise the models developed by researchers to describe and explain how language works and the various systems that analyse, generate, acquire and process language automatically. Another goal is to design computer programs that provide language-related services, either directly or indirectly, such as machine translation, speech and text recognition, and information retrieval. These tools save researchers and users time and effort by facilitating their tasks.

The Link Between Linguistics, Computer Science, and AI

Moreover, the relationship between linguistics and computer science—and artificial intelligence in particular—is not a recent development. This is clearly evident in generative grammar. Chomsky emphasised this connection when he considered the system of rules that constitute the structure of the transformational-generative model to be governed by computational theory. As Chomsky states: 'In short, language appears to be, at its core, a rich and complex computational system that is fully and precisely organised and rigorous in its basic operations' (Chomsky, 1993, p. 107). This means that the grammatical rules internalised by speakers are, to a large extent, comparable to the formal rules that computers follow when carrying out computations.

The convergence between linguistic and communication/media theories can be attributed to their shared scientific and technological objectives. The scientific aim is to contribute to an understanding of the structure of the human mind and how it works from a linguistic perspective. Their technological aims include building a computer that can emulate human linguistic abilities and applying it to a wide range of fields.

Additionally, computers are regarded as a means of standardising cognitive models. Within artificial intelligence, software developers — i.e. specialists who make machines perform tasks typically dependent on human reasoning — share research interests with linguists, cognitive psychologists and neuroscientists in understanding how the human brain functions, solves problems and processes knowledge from the environment. They also study how the brain plans, acts and makes decisions (Malmkjær, 1991, pp. 28–29). (Malmkjær, 1991, pp. 28–29).

Hence, This Principle in Cognitive Science and Its Role in Computational Studies

Accordingly, this principle was adopted in cognitive psychology through its general view of the mental architecture of information. It was likened to the mechanisms of computer processing, in which the computer receives information, encodes it, and processes it through a sequence of stages and levels—either in a serial or in a parallel manner—before storing it in central memory and using it to carry out tasks.

A computer comprises two main components: hardware (the physical device) and software (the programming system). The software can process information independently of the physical device, while the hardware can perform computational operations based on instructions issued by the software.

Generative grammar played a major role in advancing computational studies, drawing from computer science and communication theory. This occurred through research into the properties of formal languages and their suitability for constructing syntactic and semantic descriptions of natural languages, as well as through the development of computer programming languages that facilitate interaction between users and the system. (Yousfi, 2006, p. 13). Additionally, computational scholars have benefited from the findings of linguists to develop algorithms that can be integrated into software designed specifically to perform the automated processing of natural languages. This includes the provision of appropriate morphological and syntactic analysers for describing natural languages. (Ismaili, 2009, p. 104).

There is no doubt that Arabic is one of the most important languages and should be given particular attention by artificial intelligence researchers. This is because it is a global language characterised by many scientific, religious, economic and other features. It also has a rich linguistic heritage in terms of both its lexicon and grammar, and is characterised by systematic linguistic operations such as inflection, derivation and case endings. This facilitates the description of the language and consequently its integration into machine-based systems.

However, this process is not without obstacles and difficulties. Some of these are shared with other languages, while others are specific to Arabic due to its unique characteristics.

3. Applying Deep Learning Techniques to Automatic Language Processing

Since the early days of deep learning, neural networks have been used for automatic language processing to achieve the same objective of representing discrete linguistic units, such as words, characters or specific morphological and semantic categories. The aim is to replace these discrete units with a continuous numerical representation (usually in vector form) so that they can be placed in a space where notions of similarity can be determined, thereby enabling generalisation. The concept of word embedding first emerged in the context of semantic networks. Furthermore, neural networks are characterised by their ability to represent a sentence in a manner that transcends a mere 'bag of words'.

This capability relies on two main types of architecture, both of which feature mechanisms for representing sequences, namely how they decompose input and how they retain the computation of dependencies and interactions between the words that make up a sentence. Convolutional networks constitute the first type. Inspired by the concept of convolution in signal processing, these networks can be considered a generalisation of n-gram models. They use sliding windows of different sizes to extract local features. These features are then combined to produce a representation of the sentence as a whole.

The second architectural type uses recurrent networks and their more recent variants, such as LSTM networks, to model long-term memory. In such recurrent architectures, the network processes the sentence step by step — for example, from left to right and word by word — updating its internal memory at each step and gradually accumulating an overall representation of the segment. To improve modelling of long-range dependencies, recurrent networks may be bidirectional, processing the sentence from both directions. (Allauzen & Schütte, 2019, pp. 8–10).

Reasons for the success of deep learning

One reason for the success of deep learning algorithms is that, unlike many other machine learning algorithms, they do not rely on fixed, pre-defined features. Instead, they learn important features from the data during training, provided a large enough dataset is available. The availability of large volumes of data has been facilitated by advances in storage media and the vast amount of data flowing over networks. (Al-Aryan et al., 2019, p. 151).

Despite the wide variety of applications in the automatic processing of Arabic, such as speech recognition and synthesis, automated reading and writing of text, machine translation, information extraction and indexing systems, etc., those that rely on deep learning technology are still in their early stages.

This is because deep learning technology is relatively accessible and does not require extensive expertise in machine learning, unlike traditional approaches. Moreover, the results of this technology have demonstrated significant superiority over conventional techniques.

Applications of deep learning in natural language processing

Natural language processing is the field concerned with the interaction between computers and humans through the natural languages used in everyday life. Researchers have proposed a character-level linguistic model that assigns a value to each character sequence via a probabilistic distribution. This algorithm has been applied to several languages, including Arabic.

Applications of Deep Learning in Arabic Speech Recognition

Speech recognition involves converting spoken language into a machine-readable textual representation. Researchers have achieved excellent results using deep learning techniques involving recurrent neural networks and language models to recognise Arabic speech patterns. This approach improved accuracy by 15.7%.

Applications of deep learning in recognising Arabic written characters

Using deep learning techniques in optical character recognition (OCR) for Arabic texts is one of the most beneficial applications for the Arabic language, although further improvement is still required. This is because Arabic has certain characteristics that differ from other languages, such as writing from right to left, connected letters, high similarity between some letters and variation in the shape of the same letter depending on its position in a word. Therefore, specific algorithms need to be developed for Arabic that differ from those designed for other languages such as English or Chinese. Additionally, ongoing efforts are being made to develop technologies for recognising handwritten text and to achieve near-perfect accuracy.

4. Challenges of studying social media

Social media platforms present three main computational challenges: Scale (volume), speed, and variety. Approaches to automatic language processing face even greater difficulties in particular due to the distinctive nature of social media content, namely that it is short, noisy and strongly context-dependent.

New technical languages and methods have emerged to address these challenges, such as identifying and modelling users' linguistic varieties and translating content into different languages. Determining

the dialect (or linguistic variety) is essential, as it is the first component of any natural language application dealing with Arabic and its varieties, such as machine translation, information retrieval from social media, sentiment analysis and opinion mining (Sadat, Kazemi & Farzindar, 2014a, pp. 35–40). (Sadat, Kazemi & Farzindar, 2014a, pp. 35–40).

Difficulties in processing text:

Algerian bloggers and tweeters tend to write in colloquial Arabic (dialect). These texts exhibit issues that overlap with those encountered when analysing Standard Arabic texts, such as spelling problems related to diacritics (vowel marks) and differences in how Arabic letters are written. There are also problems specific to the dialect, which are as follows:

Scarcity of computerised blogs

Automated text processing relies on computer-readable corpora, which are texts whose components are tagged with various linguistic, grammatical and stylistic information. This kind of corpus facilitates the search for information and the inference of hidden beliefs and emotions in texts. However, the digital library currently lacks computational corpora, tools, resources and works for Arabic and its local dialects.

Lexical richness and multiple sources

Local dialects contain a large vocabulary, some of which is borrowed from other languages. They also have morphological features that differ from Standard Arabic, such as the attachment of pronouns and negation. Furthermore, they are continuously evolving as new words enter from other languages. There are also many varieties that differ from one region to another at several levels. Sometimes, the meaning of a word in a local dialect differs from its meaning in Standard Arabic (Guellil, Azouaou, Saâdane & Semmar, 2017, p. 43).

The problem of ambiguity

Ambiguity, or an unclear meaning, can make it difficult to identify a word, determine its grammatical category and understand its meaning. This is because bloggers often abbreviate words, which may be missing letters, or they may not follow spelling rules. The problem worsens when Standard Arabic is used as the reference point because additional ambiguity issues arise from the mismatch between pronunciation and spelling (semantic ambiguity/shared forms), as well as differences in meaning and possible interpretations.

Lack of a reference standard for writing the dialect

Other problems also arise because dialect spelling is not based on any linguistic reference system; it is simply an agreement among users.

Writing in Latin letters

Bloggers often use the Latin alphabet to write their posts, either because they find it easier or because it is available on all devices, unlike Arabic script. This makes it more difficult to recognise the content. In addition, code-switching and language mixing occur: users may combine the dialect with another

language, such as French, or mix Standard Arabic with the dialect. This makes it difficult to determine which words were written in the dialect.

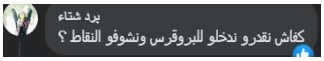
These issues make processing dialects more difficult in general, and they will not be resolved unless extensive linguistic resources are available and appropriate processing tools are developed (Guellil, Azouaou, Saâdane & Semmar, 2017, p. 98).

5. Sample analysis

The study sample includes some posts written by students of the Arabic Language and Literature department on the department's Facebook page (Arabic Language and Literature students, University of Mascara, Algeria, 2023). Through these posts, we will identify the difficulties that may arise when trying to recognize the words in those posts using deep learning techniques. After collecting the study sample, we can observe the following phenomena and comment on them:

1) Code-switching within the same sentence

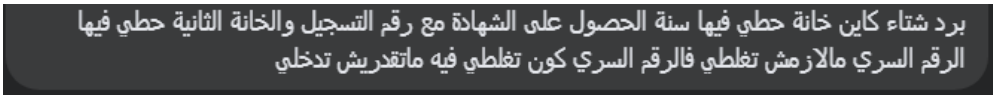
A noticeable feature of these posts is the combination of Standard Arabic and local dialect words within the same sentence. For example, when a student asks the page members:



This communication style, represented by bilingualism between Standard Arabic and the local dialect, will greatly complicate traditional processing, especially since it relies on built-in computational resources.

2) Mixing Arabic (standard or dialect) with foreign words

We also observe mixing between standard Arabic or dialect words and words from foreign languages, whether written using Arabic or Latin letters, within the same sentence, for example:

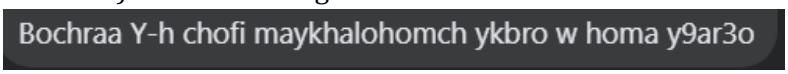


Comments written in Arabic or dialect using Latin letters are also common, for example:

This makes word recognition difficult because there is no agreed system for writing the dialect in academic circles, even if bloggers agree on one.

3) Widespread ambiguity

We also observe that ambiguity is very common in the words used in the posts, for example: ("criticism") and ("mascara") in the following:



Here, the poster does not mean the usual meaning of 'criticism'; rather, it refers to the course material on literary criticism that they have been assigned. Likewise, 'mascara' does not refer to a military camp or gathering of soldiers, but to a city in Algeria. This makes it hard to recognise the meanings of these two words, especially because the most likely interpretation is usually chosen first.

These difficulties in recognising the words used in students' posts present a challenge to traditional automated processing. However, good recognition results can be achieved using deep learning

techniques, since these techniques rely on repeated learning by feeding them a large corpus of posts in image form.

Conclusion:

By studying a sample of posts written in the local Arabic dialect by students of Arabic Language and Literature at the University of Mascara and published on social media pages, with the aim of recognising their words automatically using deep learning techniques, we can draw the following conclusions:

These posts present significant challenges to traditional automated recognition techniques that rely on information system algorithms. This is mainly due to the high degree of ambiguity in the words and expressions used, which is exacerbated by the lack of computational corpora that provide various linguistic information.

The sample also demonstrates frequent sociolinguistic phenomena, such as bilingualism (diglossia) and multilingualism, as well as differences in writing systems between Arabic and Latin scripts, and a lack of consensus on a system for writing the dialect. These factors all increase the difficulty of automated word recognition.

Relying on deep learning techniques to analyse the words in these posts can achieve a relatively high level of success because these methods repeatedly learn words and expressions presented as images rather than sequences of letters and words, and because they favour the most likely interpretations over traditional processing approaches.

Recently, using deep learning techniques for Arabic language processing has led to a significant improvement in results, whether in language analysis and generation or in recognising its spoken and written components.

This is because deep learning techniques can learn and improve; the model can correct its mistakes when processing, analysing and recognising different Arabic texts or discourses. Unlike other symbolic AI approaches, it does not require a huge amount of manually encoded rules.

References:

- Al-Sayyid, M. and K. Mahmoud Mohamed. Applications of Artificial Intelligence and the Future of Educational Technology. Arab Group for Training and Publishing. Cairo, 2020.
- H. Ismaili Aloui and M. Al-Malakh. Epistemological Issues in Linguistics. Munshurat Al-Ikhtilaf (Editions), Algeria, 2009. Algeria, 2009.
- Al-Aryan (Y.) et al. Applications of Artificial Intelligence in the Service of the Arabic Language (1st ed.). Riyadh: Dar Wujuh, 2019.
- Students of Arabic Language and Literature, University of Mascara, Algeria. (10/06/2023). Retrieved from: https://www.facebook.com/groups/632482260295765
- Atiyya (M.) et al. 'Arabic and Artificial Intelligence'. Dar Wujuh Publishing House: Riyadh, 2019.
- Chomsky (N.). Linguistic Knowledge: Its Nature, Origins, and Use. Translated by Muhammad Fathi. Dar Al-Fikr Al-Arabi: Cairo, 1993.
- Allauzen, A. and Schütze, H., 'Deep learning for automatic language processing'. TAL, 59(2), 2019.
- Guellil, I., Azouaou, F., Saâdane, H. and Semmar, N., 'An approach based on sentiment analysis lexicons of the Algerian dialect', *TAL*, 58(3), 2017. TAL, 58(3), 2017.

- Malmkjær, K. The Linguistics Encyclopedia. Routledge, London, 1991.
- Sadat, F., Kazemi, F., & Farzindar, A., 'Automatic identification of Arabic dialects in social media'. In Proceedings of the 1st International Workshop on Social Media Retrieval and Analysis. Gold Coast, 2014.
- Yousfi, A., Automatic Language Processing (Text and Speech). Bouregreg: Rabat, 2006.