تحليل النصوص المكتوبة بالدارجة الجزائرية باستخدام تقنيات التعلم العميق

Analysis of texts written in the Algerian Arabic dialect using deep learning

رضا بابا أحمد *
جامعة مصطفى اسطمبولي - معسكر
Réda Baba Ahmed
University of Mustapha Stambouli - Mascara
r.babaahmed@univ-mascara.dz

تاريخ استلام المقال:2024/10/15	تاريخ القبول: 2024/10/23	تاريخ النشر:2024/12/15

ملخص

إن البحث في اللهجات العربية ومنها الدارجة الجزائرية ودراسة النصوص المكتوبة بها يشكل رهانا لمعالجتها الآلية وإجراء المهام التي يحتاجها الباحثون والمستخدمون، خاصة بعد انتشار التدوينات بها على شبكات التواصل الاجتماعي. وقد تجلت إثر ذلك مشكلات تعيق عملية التعرف على الكلمات والعبارات المكتوبة بتلك اللهجات منها مشكل الغموض في إطار غياب المدونات المحوسبة، ومشكلات أنظمة الكتابة الإملائية، والتداخل بينها وبين العربية المعيارية أو للغات الأجنبية. وعليه، قمت بدراسة عينة من تدوينات طلبة قسم اللغة والأدب العربي بجامعة معسكر المكتوبة بالدارجة والمنشورة على صفحتهم في الفايسبوك، وملاحظة الصعوبات التي تبديها في عملية التعرف على كلماتها مما دفع الباحثين إلى اعتماد تقنيات التعلم العميق لتحقيق نسبة من النجاح في ذلك.

الكلمات المفاتيح: *التعلم العميق؛ التعرف على المكتوب؛ الدارجة الجزائرية؛ الغموض؛ المشترك الدلالي.*

Abstract

Researching the Arabic dialects, including the Algerian dialect, and studying the texts written in them constitutes a bet for natural language processing and carrying out the tasks needed by researchers and users, especially after the spread of posts about them on social networks. As a result, problems appeared that impede the process of recognizing written words and phrases in these dialects, including the problem of ambiguity in the absence of corpus, problems with

* **المؤلف المرسل:** رضا بابا أحمد. **البريد الإلكتروني:** r.babaahmed@univ-mascara.dz.

E-ISSN - 2773-2886 / ISSN 2773-2894

orthographic systems, and the overlap between them and standard Arabic or foreign languages. Accordingly, I studied a sample of the postings of students of the Department of Arabic Language and Literature at Mascara University, written in Darija and published on their Facebook page, and noting the difficulties they show in the process of recognizing their words, which prompted researchers to adopt deep learning to achieve a percentage of success.
Keywords: Deep learning; written recognition; Algerian Arabic dialect; Ambiguity; polysemy.

1. مقدمة

يعنى الذكاء الاصطناعي بتصميم خوارزميات يمكن أن تضاهي عند تشغيلها المهام المنوطة بالإنسان، والتي من أبرزها التعلم الذي تكتسبه الآلة بتحليل البيانات والتحصيل منها لتحسين أدائها في مهمة محددة، ثم تتلوها مرحلة التدريب على تنفيذ نفس المهام على بيانات جديدة. وقد استوحى الباحثون من خلايا الإنسان العصبية منهجية لمحاكاة عملية التعلم لدى الآلة، هذه المنهجية أسهمت بتطورها في ظهور تقنية التعلم العميق التي أثبت فعالية كبيرة في تحليل اللغات الطبيعية.

وعلى اعتبار أن اللغات الطبيعية تحتوي على تنوعات ولهجات، يشكل البحث في اللهجات العربية ودراسة النصوص المكتوبة بها رهانا في الوقت الحاضر، نظرا لانتشار التدوينات بها على شبكات التواصل الاجتماعي رغم المقاومة التي يبديها المحافظون لهذا النوع من أنماط الكتابة، وحثهم المستخدمين والمتفاعلين على الكتابة باللغة العربية المعيارية، وظهرت بالتالي مشكلات عند معالجة تلك النصوص آليا وتحليلها باعتماد تقنيات الذكاء الاصطناعي ومنها تقنية التعلم العميق.

في هذه الورقة، سأقوم بدراسة عينة من تدوينات طلبة قسم اللغة والأدب العربي بجامعة معسكر المكتوبة بالدارجة والمنشورة على صفحة القسم في الفايسبوك، وذلك للوقوف على أبرز المشاكل التي يبديها هذا الاختيار في الكتابة، وأثر ذلك على المعالجة الآلية لتلك النصوص باستخدام تقنية التعلم العميق، وذلك من أجل اقتراح مجموعة من القيود اللغوية والإملائية التي يمكن أن تسهم في رفع الغموض الذي يكتنف كتابة الكلمات، وأن تسهل عملية التعرف على الكلمات المستخدمة في تلك التدوينات.

 E-ISSN - 2773-2886 / ISSN - 2773-2894

2. **تطبيق تقنيات التعلم العميق في المعالجة الآلية للغة**

منذ الأعمال الأولى في تقنيات التعلم العميق، يستمر تطبيق الشبكات العصبية على المعالجة الآلية للغة لتحقيق نفس الهدف المتعلق بتمثيل الوحدات اللغوية المنفصلة، على سبيل المثال الكلمات أو الأحرف أو المقولات الصرفية أو الدلالية، والهدف هو استبدال هذه الوحدات المنفصلة بتمثيل رقمي مستمر على شكل شعاع، من أجل تضمينها في فضاء حيث يمكن تحديد مفاهيم التشابه التي تفضي إلى التعميم. ويظهر مفهوم التضمين word embedding أولاً في سياق الشبكات الدلالية، وتتميز الشبكات العصبية أيضًا بقدرتها على تمثيل جملة تتجاوز مجرد سلسلة من الكلمات.

هذه القدرة تعتمد بشكل أساسي على نوعين من الأبنية التي تتميز بآليات تمثيل التسلسل، أي طريقة تفكيكها وكذلك الاحتفاظ بها حسب التبعيات والتفاعلات بين الكلمات التي تتكون منها. الشبكات الالتفافية هي النوع الأول. مستوحى من عامل الالتفاف في معالجة الإشارات، يمكن اعتبار هذه الشبكات بمثابة تعميم لنماذج ن-غرام. يعتمدون على نوافذ المنزلقة ذات الأحجام المختلفة، تسمح باستخراج الخصائص المحلية. ثم تؤلف هذه الخصائص لتمثيل الجملة في كليتها.

النوع الآخر من المعمارية يستخدم الشبكات المتكررة وتطوراتها الأخيرة مثل شبكات LSTM، للذاكرة طويلة المدى، تمر الشبكة المتكررة على الجملة، على سبيل المثال من اليسار إلى اليمين، كلمة تلو الأخرى، محينة الذاكرة الداخلية للشبكة في كل خطوة، ومراكمة بالتالي صورة عامة للمقطع. من أجل تعزيز نمذجة التبعيات طويلة المدى، يمكن أن تكون الشبكات المتكررة ثنائية الاتجاه، مارة بالجملة من اليسار إلى اليمين ومن اليمين إلى اليسار(Allauzen & Schütze, 2019, pp.8-10) .

E-ISSN - 2773-2886 / ISSN - 2773-2894

من أسباب نجاح خوارزمية التعلم العميق أنها لا تعتمد على خصائص ثابتة ومحددة مسبقا كما هي الحال في سائر خوارزميات تعلم الآلة، ولكنها تتعلم الخصائص المهمة من البيانات أثناء عملية التدريب، بشرط توفر عدد كبير من البيانات أثناء هذه العملية، وقد ساعد على توفر البيانات الضخمة تطور وسائط التخزين والتدفق الهائل للبيانات بكل أنواعها خاصة عبر الشابكة (العريان و آخرون، 2019، ص. 151).

على الرغم من أن تطبيقات المعالجة الآلية للغة العربية كثيرة ومتنوعة وتشهد تطورا ورواجا كبيرين في السنوات الأخيرة كالتعرف على الكلام وتركيبه، والقراءة الآلية للنصوص المكتوبة، والكتابة الآلية للنصوص المنطوقة، والترجمة والفهرسة الآليتين، ونظم استخلاص المعلومات وغيرها، تبقى تلك المعتمدة على تقنية التعلم العميق في بداياتها الأولى. ذلك لأن هذه التقنية سهلة ولا تتطلب خبرة كبيرة في مجال تعلم الآلة على عكس التقنيات التقليدية، كما أن نتائجها أثبتت تفوقا كبيرا على تلك التقنيات التقليدية.

1.2. تطبيقات التعلم العميق في مجال تحليل اللغة الطبيعية

تحليل اللغات الطبيعية هو مجال يعنى بالتفاعلات بين الحاسوب والإنسان من خلال اللغات الطبيعية التي يستخدمها الناس في حياتهم اليومية. اقترح الباحثون نموذجا لغويا على مستوى الحرف يقوم بتعيين قيمة محتملة لكل سلسلة من الحروف عن طريق التوزيع الاحتمالي. تم تطبيق هذه الخوارزمية على لغات من ضمنها اللغة العربية.

E-ISSN - 2773-2886 / ISSN - 2773-2894

.22. تطبيقات التعلم العميق في مجال التعرف على الكلام العربي المنطوق

التعرف على الكلام المنطوق هو مجال يعنى بتحويل الكلام المنطوق إلى ترميز حاسوبي نصي. استطاع الباحثون تحقيق نتائج ممتازة بالاعتماد على تقنية التعلم العميق؛ حيث استخدموا الشبكات العصبية المتكررة مع نماذج لغوية في التعرف على النماذج الصوتية العربية بتحسين الدقة بنسبة 15,7%.

3.2. **تطبيقات التعلم العميق في مجال التعرف على الحروف العربية المكتوبة**

يعتبر استخدام تقنية التعلم العميق في مجال التعرف الضوئي على النصوص العربية من أكثر المجالات خدمة للغة العربية، لكنها تحتاج إلى تحسين أكثر. نظرا لأن لغتنا العربية تختلف عن اللغات الأخرى ببعض الخصائص كاتجاه الكتابة، واتصال الحروف، والتشابه الكبير للحروف واختلاف أشكال الحرف الواحد بحسب الموقع... لذلك تحتاج اللغة العربية إلى تطوير خوارزميات خاصة بها قد تختلف عن الخوارزميات التي أعدت للغات الأخرى كالإنجليزية أو الصينية مثلا. كما أن هناك جهودا لتطوير تقنيات التعرف على النص المكتوب بخط اليد، وتحقيق دقة متناهية.

3. **تحديات دراسة وسائل التواصل الاجتماعي**

تطرح وسائل التواصل الاجتماعي ثلاثة تحديات حاسوبية رئيسة: الحجم والسرعة والتنوع. تواجه مناهج المعالجة الآلية للغات- على وجه الخصوص- المزيد من الصعوبات الناشئة عن طبيعة وسائل التواصل الاجتماعي القصيرة والصاخبة والمتوافقة مع السياق بقوة، ومن أجل معالجة تحديات وسائل التواصل الاجتماعي، ظهرت لغات تقنية جديدة، مثل تحديد وتعريف مستخدمي التنوعات اللغوية والترجمة إلى لغة مختلفة. وتحديد اللهجة أمر أساسي ويعتبر الأول مكون خادع لأي تطبيق لغة

E-ISSN - 2773-2886 / ISSN - 2773-2894

طبيعية يتعامل مع اللغة العربية وتنوعاتها مثل الترجمة الآلية، استرجاع المعلومات لوسائل التواصل الاجتماعي، وتحليل المشاعر، واستخراج الآراء(Sadat, Kazemi, & Farzindar, 2014a, pp. 35-40) .

4. صعوبات معالجة النصوص

يقبل المدونون والمغردون الجزائريون على كتابة النصوص بالدارجة والعاميات، والتي تبدي مشكلات تتقاطع في كثير من جوانبها مع تلك المتعلقة بتحليل النصوص المكتوبة بالعربية المعيارية، كالقضايا الإملائية المتعلقة بالتشكيل وكذلك باختلاف أشكال كتابة الحروف العربية، وهناك مشاكل تتعلق بالدارجة خاصة وتتمثل فيما يلي:

1.4. ندرة المدونات المحوسبة

إن المعالجة الآلية للنصوص ترتكز على المدونات أو المتون المحوسبة والتي تمت عنونة مكوناتها وتحشيتها بمختلف المعلومات اللغوية والنحوية والأسلوبية التي تتضمنها، هذا النوع من المدونات يسهل عملية البحث واسترجاع المعلومات واستنباط المعتقدات والمشاعر الكامنة في النصوص. لكن تفتقر المكتبة الرقمية في الوقت الحالي إلى الأعمال والأدوات والموارد والمدونات المحوسبة للغة العربية فضلا عن لهجاتها المحلية.

2.4. الثراء المعجمي وتعدد مصدره

تحتوي الدارجة المحلية على رصيد كبير من المفردات وبعضها مقترض من عدة لغات، كما تتميز بخصائص صرفية تختلف أحيانا عن نظيراتها في اللغة العربية المعيارية، كالتصاق الضمائر وكذلك النفي، وكذلك بتطورها المستمر مع دخول كلمات من لغات أخرى، وبتنوعاتها الكثيرة التي قد تختلف من منطقة إلى أخرى على عدة مستويات، وتتغير معاني تلك المفردات أحيانا عن معانيها في اللغة العربية المعيارية (Guellil, Azouaou, Saâdane, & Semmar, 2017, p. 43).

E-ISSN - 2773-2886 / ISSN - 2773-2894

3.4. مشكل الغموض

الغموض أو اللبس يؤثر في التعرف على الكلمة وتحديد نوعها ومدلولها لأن المدونين عادة ما يختصرون في الكتابة فتاتي الكلمات مبتورة الحروف أو أحيانا لا يلتزمون بقواعد الإملاء، وتتفاقم هذه المشكلة أكثر عندما نتخذ العربية المعيارية مرجعية لها؛ فتضاف إليها مشاكل الغموض الناشئ من تطابق الألفاظ المنطوقة أو المكتوبة (المشترك الدلالي) واختلاف مدلولاتها أو احتمالها لعدة تأويلات،

4.4. انعدام مرجعية في كتابة الدارجة

كما تظهر مشكلات أخرى تتعلق بوجود اتفاق في تهجئة النصوص بالدارجة لا يخضع بالضرورة لمرجعية لسانية وإنما هو مجرد توافق بين المستخدمين.

5.4. التهجئة بالحروف اللاتينية

غالبا ما يستخدم المدونون الحروف اللاتينية لكتابة تدويناتهم لسهولتها بالنسبة لهم أو لتوفرها في جميع الأجهزة على عكس الحروف العربية، وهو ما يزيد صعوبات أخرى لعملية التعرف على محتوياتها. بالإضافة إلى ظاهرة ازدواجية اللغة والتناوب اللغوي عندما يدمج المستخدم بين الدارجة ولغة أخرى كالفرنسية وبين العربية المعيارية والدارجة مما يصعب تحديد الكلمات المكتوبة بالدارجة. هذه المشاكل المختلفة تزيد من صعوبة المعالجة الآلية للهجات عموما، والتي لن تذلل إلا إذا توفرت لها موارد لغوية واسعة وطورت لها مختلف أدوات المعالجة المناسبة (Guellil, Azouaou, Saâdane, & Semmar, 2017, p. 98).

E-ISSN - 2773-2886 / ISSN - 2773-2894

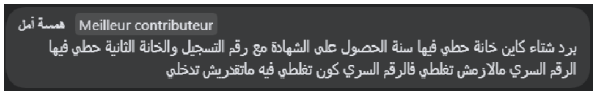
5. تحليل العينة

تشمل عينة الدراسة بعض تدوينات طلبة قسم اللغة والأدب العربي على صفحة القسم في الفايسبوك (طلبة اللغة والأدب العربي جامعة معسكر الجزائر، 2023)، سأقف من خلالها على الصعوبات التي يمكن أن تصادف عملية التعرف على كلمات تلك التدوينات باستخدام تقنيات التعلم العميق. بعد جمع عينة الدراسة يمكن ملاحظات الظواهر التالية والتعليق عليها:

1. الملاحظ في هذه التدوينات المزج بين كلمات عربية فصيحة وأخرى من اللهجة المحلية داخل الجملة الواحدة مثلا: كسؤال أحد الطلبة لأعضاء الصفحة:

إن هذا النمط التواصلي الذي يتمثل في الازدواجية اللغوية بين العربية المعيارية واللهجة سيصعب بشكل كبير عملية المعالجة التقليدية التي تعتمد على موارد حاسوبية مدمجة؛

2. الملاحظ أيضا المزج بين كلمات عربية فصيحة أو كلمات من اللهجة المحلية وكلمات من اللغات الأجنبية سواء مكتوبة بحروف عربية أو لاتينية داخل الجملة الواحدة مثلا:

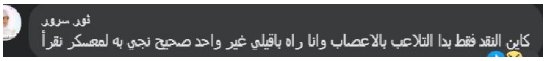


3. وكتابة التعليقات الصادرة باللغة العربية أو الدارجة خاصة بحروف لاتينية مثلا:

هذه الطريقة في الكتابة تسبب صعوبة في التعرف على الكلمات من حيث إننا لا نملك نظاما لكتابة الدارجة متفقا عليه في الأوساط الأكاديمية وإن اتفق عليه المدونون.

E-ISSN - 2773-2886 / ISSN - 2773-2894

4. كما نلاحظ ظاهرة الغموض منتشرة بكثرة في كلمات التدوينات مثلا: (النقد) و(معسكر) فيما يلي:



فالمدوِّنة لا تقصد بـ(النقد) معناه المتداول ولكن مادة النقد الأدبي المقرر عليهم، كما أنه لا تقصد بـ(معسكر) تجمع الجند ولكن مدينة من المدن الجزائرية، مما يصعب التعرف على معنى هاتين الكلمتين خاصة مع تغليب الاحتمال الأكبر لمدلوليهما.

إن هذه الصعوبات التي نجدها في عملية التعرف على الكلمات المستخدمة في تدوينات الطلبة تشكل تحديا للمعالجة الآلية التقليدية، لكن باستخدام تقنيات التعلم العميق يمكن أن تحقق عملية التعرف نتائج جيدة نظرا لارتكاز تلك التقنيات على التعلم المتكرر من خلال إدخال رصيد ضخم من التدوينات على شكل صور.

6. **خاتمة**

من خلال دراسة عينة من تدوينات طلبة اللغة والأدب العربي بجامعة معسكر المسجلة بالدارجة المحلية والمنشورة على صفحات التواصل الاجتماعي بغرض التعرف الآلي على كلماتها باستخدام تقنيات التعلم العميق، يمكن أن نستخلص بعض النتائج نذكرها فيما يلي:

- إن هذه التدوينات تشكل صعوبات كبيرة أمام تقنيات التعرف الآلي التقليدية التي ترتكز على خوارزميات نظم المعلومات لما تبديه من تحديات تتعلق خاصة بالغموض الذي يكتنف كلماتها وعباراتها لعدم وجود مدونات محوسبة تكون الكلمات فيها مزودة بمختلف المعلومات اللغوية؛

E-ISSN - 2773-2886 / ISSN - 2773-2894

- شيوع ظواهر سوسيولسانية كالازدواجية اللغوية والتعدد اللغوي في متن العينة، بالإضافة إلى اختلاف نظم الكتابة بين العربية واللاتينية وعدم الاتفاق على نظام لكتابة الدارجة يزيد من مشاكل التعرف الآلي على كلمات التدوينات؛
- الاعتماد على تقنيات التعلم العميق في تحليل كلمات تلك التدوينات يمكن أن يحقق نجاحا نسبيا لارتكازها على التعلم المتكرر للكلمات والعبارات على شكل صور وليس على شكل حروف وكلمات متوالية زمنيا وتغليب الاحتمالات الأكثر وقوعا مقارنة مع تقنيات المعالجة التقليدية.

7. قائمة المراجع

1) *العريان، ي. وآخرون .(2019) .تطبيقات الذكاء الاصطناعي في خدمة اللغة العربية (ط.1). دار وجوه للنشر. الرياض.*
2) *طلبة اللغة والأدب العربي جامعة معسكر الجزائر. (2023/06/10). تم الاسترداد من https://www.facebook.com/groups/632482260295765*
3) *Allauzen, A., & Schütze, H. (2019). "Apprentissage profond pour le traitement automatique des langues". TAL, 59(2), p. 7-14.*
4) *Guellil, I., Azouaou, F., Saâdane, H., & Semmar, N. (2017). "Une approche fondée sur les lexiques d'analyse de sentiments du dialecte algérien". TAL, 58(3), p. 41-63.*
5) *Sadat, F., Kazemi, F., & Farzindar, A. (2014a). "Automatic identification of arabic dialects in social media". analysis, Proceedings of the 1st international workshop on Social media retrieval and analysis. Gold Coast.*