Ingénierie des Systèmes d'Information
Vol. 29, No. 3, June, 2024, pp. 821-829
Journal homepage: http://iieta.org/journals/isi

Subject Detection of Algerian Posts for Opinion Analysis

Kheira Zineb Bousmaha[1*], Khaoula Hamadouche[2], Nawel Cheurfaoui[3], Lamia Hadrich-Belguith[4]

[1] Computer Science Department, RIIR Laboratory, Oran 1 Ahmed Ben Bella University, Oran 31005, Algeria
[2] Computer Science Department, FSEA, RIIR L, Oran 1 Ahmed Ben Bella University, Oran 31005, Algeria
[3] Computer Science Department, Oran 1 Ahmed Ben Bella University, Oran 31005, Algeria
[4] Computer Science Department, MIRACL Laboratory, FSEGS, University of Sfax, Sfax 3018, Tunisia

Corresponding Author Email: k.hamadouche@esi-sba.dz

https://doi.org/10.18280/isi.290303

Received: 17 August 2023
Revised: 19 January 2024
Accepted: 13 March 2024
Available online: 20 June 2024

Keywords:
subject detection, opinion analysis, marketing, multi-class, Algerian dialect, TF-IDF, deep learning, keyword extraction

ABSTRACT

Nowadays, opinion analysis and text classification become interesting tasks in social media studies. Moreover, keyword extraction technology is important research topic and the basis of building corpora, information retrieval, text analysis and text classification...etc. Most studies carry out on the binary problem classification which has historically received more attention in the machine learning community compared to multi-class classification problem. They are often done on academic languages, leaving aside the dialects despite of their increasing use in the social networks. There is very little effort that has been dedicated in the Arabic language especially its dialects such as Algerian dialect, intended for the analysis of opinions. In this work, we aim to realize an innovative and original approach for multi-class classification task and subject detection in the field of marketing. These classes are considered as subjects of the Algerian posts. We collected dataset from Facebook and annotated them with 11 labels. We applied TF-IDF word embedding method to vectorise and extract keywords by giving a weight for each token. In the final stage, our model was trained using input vectors. We applied a deep neural network on our annotated dataset. We achieved a precision of 83%.

1. INTRODUCTION

Opinion analysis is rapidly becoming a very available field of study for research and an extremely valuable tool applied in almost all business and social domains because opinions are at the heart of almost all human activities and are key influencers of our behaviors [1]. The goal is to analyse, based on the texts exchanged on social networks, the sentiments, opinions, attitudes, and emotions expressed by communities on various. In the current global economy, marketing plays a crucial role for businesses, encompassing the creation, communication, delivery, and exchange of value-driven offerings for clients, customers, partners, and society at large. Recent years have seen notable shifts in marketing, driven by digital technologies and the widespread use of social media. Marketers in Algeria must stay updated on these changes to maintain competitiveness globally. The Algerian consumer has also evolved, becoming more educated, tech-savvy, and discerning. This evolution prompts businesses in Algeria to adapt their marketing strategies to meet changing customer preferences. Despite growth opportunities, Algerian businesses face challenges such as insufficient infrastructure, limited funding access, and a shortage of skilled marketing professionals [2].

One of the most important and indispensable challenges in machine learning is the classification task [3, 4]. The binary class problem is better handled by text classification techniques that work more efficiently. However, there has not been extensive research on the application of these algorithms and models for multiclass text classification, particularly in the Arabic language and its dialects [5].

On the other hand, subject detection from social networks is a large area where the researchers have studied various models and methods to enhance it and get a better clustering to understand the interests of Internet users, facilitate effective future recommendations and signal new breaks.

In addition, keyword extraction is an active research field covering many applications, especially, text classification [6]. It involves automatically identifying a set of terms that most accurately describe the topic of a document or text [7].

In this research, we develop a system for the detection of topics of posts in the field of marketing in the Algerian dialect while particularly using deep learning techniques. The processing of the Algerian dialect therefore becomes a crucial task for the reliability of our system. The particularity of this work is the extraction of keyword from the data using the TFIDF. These keywords are descriptive words or phrases that characterize documents. The overall objective of TF-IDF is to measure statistically the importance of a word in a collection of documents, this is ideal for words, indicating the subject, which appear less frequently.

The outcomes of our research hold vast potential for applications in various facets of marketing. For instance, the developed approach could be instrumental in enhancing digital marketing strategies, market research, and content

optimization for businesses operating in Algeria. The ability to discern and categorize content into specific subjects, such as Accessories and jewelry, etc., opens avenues for tailoring marketing efforts to suit the preferences and trends within the Algerian market. Moreover, marketing professionals and businesses stand to benefit from more accurate insights into consumer behavior. This heightened understanding enables the creation of targeted and personalized marketing campaigns and fostering improved customer engagement.

This article is structured as follows: In related works section, we introduce an overview of text classification and keyword extraction. Section 3, introduces our proposed approach for Algerian dialect processing and topic detection. The following section presents the evaluation of our approach and the experimental results with a discussion. At the end, we provide the conclusion.

2. RELATED WORKS

Text classification is an approach of machine learning which assigns a specified set of categories to a given text. It could be used to structure, organize, and categorize almost any type of text, over time, it becomes one of the most essential tasks in Natural Language Processing (NLP) [8] with wide applications like sentiment analysis, topic labeling, subject detection or spam detection, etc.

We based on keyword extraction technique and deep learning approach to classify the texts (detect the subjects). The existing approaches for automatic keyword extraction can be divided into three approaches [7, 9-12]:

- The first one is rules based.
- The second one is statistics approach.
- The third is artificial intelligence approaches that include Machine learning algorithms.

They are summarized below in Figure 1.

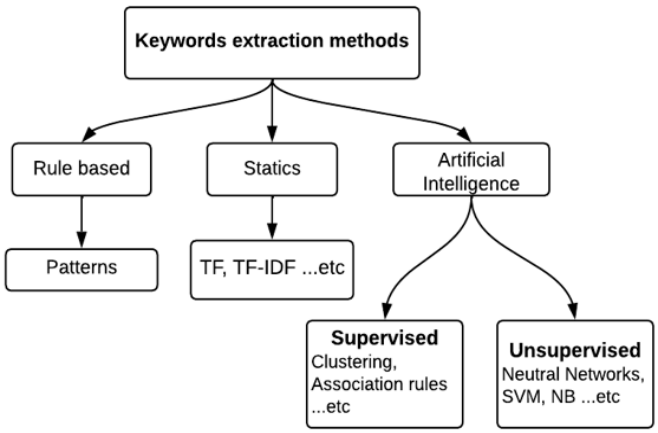

Figure 1. Keyword extraction methods

In the following, we will see some of the realized works on keyword extraction and text classification in foreign and Arabic languages.

One of the earliest studies by Moh'd and Mesleh [13] aimed to develop a system using SVMs for the classification of Arabic articles. In the preprocessing phase, the classifier utilized the CHI square method for feature selection. Compared to alternative classification methods, the system demonstrated notable effectiveness, achieving a high F-measure of 88.11%.

Most of studies done in the keyword extraction field used the TF-IDF technique such as study by some authors [9, 14-17], and it has shown promising results in a variety of languages. Boukil et al. [14] presented a novel method for classifying Arabic text. They employed an Arabic stemming algorithm to extract, select, and minimize the necessary features. Following this, they employed the Term Frequency-Inverse Document Frequency technique (TF-IDF) as a feature weighting technique. Finally, for the classification, they utilized deep CNN algorithms. They deduced that the CNN model performs better on the Arabic text classification task than SVM and LR. They could achieve excellent results with 92.94 accuracy as shown in the Table 1.

Table 1. Accuracy of the CNN, LR, SVM model on different data sets [4]

	Data_27k	Data_55k	Data_111k	Data_27k
LR	81.23	82.17	83.46	86.31
SVM	83.04	84.77	86.9	88.2
CNNs	86.3	87.12	89.75	92.94

In the study by Aipe et al. [15], authors developed a deep Convolutional Neural Network (CNN) for multilabel classification of crisis related tweets. They have used TF-IDF to extract the keyword, and represented them with pretrained word2vec model. The matrix representation is fed into a convolutional layer to the classification (seven categories). The used dataset consists of 49K English tweets. Experimental results show that they achieved significantly better performance with their proposed system compared to the existing state-of-the-art models.

The study by Yao et al. [9] used English news text as the research purpose of keyword extraction method. Authors combined TextRank and TF-IDF techniques in order to extract the text keyword by building word graph model, counting word frequency and inverse document frequency, and considering the weight of the headlines positioning. This combined model achieved 99.1% recall, better than the original algorithms.

An interesting work carried out by Song et al. [16], it explored the keywords features presented in Chinese texts. Utilizing the TF-IDF algorithm, it incorporated unique Chinese features such as word length, part of speech (PoS), and lexicon. An enhanced TF-IDF weighting formula was devised, considering these text features comprehensively. The paper proposed a keyword matching scoring approach and suggested transforming keywords 'cut off' by Chinese word segmentation into formal keywords. Cross-comparison experiments demonstrated that the improved algorithm surpasses traditional approaches. Another recent study produced by Loukam et al. [18] where the authors described an approach for extracting key phrases from modern standard Arabic (MSA) texts, based on the Association Rules model. They collected a corpus comprising 100 press articles sourced from 5 media outlets, with an average of 319 words each. The proposed system consists of two primary modules: Text pre-processing and Association Rules mining. The experimental results are promising, with higher than 60% of precision, recall and f-score, and can exceed 70%.

Khan et al. [19] conducted various experiments to demonstrate the significance of context-based key extraction in comparison to conventional methods. They used a contextual word embedding "KeyBERT" to extract keywords from documents and match them with author-assigned

keywords. Therefore, they focus on the connection between words within the context of the sentence. They collected 1363 English research articles for keyword extraction. The proposed model exhibited an average similarity rate of 51%, surpassing that of other baseline methods.

2.1 Discussion

The reviewed literature predominantly centers on binary text classification and keyword extraction within the scope of MSA, English, and other languages like Chinese. Unfortunately, there is a noticeable gap as these studies tend to overlook the rich diversity of Arabic dialects, particularly the Algerian dialect. While models designed for MSA and other languages benefit from established Natural Language Processing (NLP) tools and understanding, this advantage does not seamlessly extend to Arabic dialects.

The linguistic characteristics and nuances inherent in the Algerian dialect present distinct challenges that warrant specialized attention. Unlike MSA, the Arabic dialects like Algerian dialect incorporates unique expressions, cultural references, complex morphology, linguistic variations, and lack of resources [20], requiring a tailored approach for accurate classification and extraction.

In the realm of topic detection and classification, studies leveraging deep learning methodologies have demonstrated promising results. Deep learning algorithms exhibit a remarkable capability to automatically learn features essential for effective classification tasks [14]. Moreover, it is noteworthy that several of these studies have successfully incorporated traditional techniques such as TF-IDF in collaboration with deep learning models. This combination has shown to enhance the overall performance of topic detection systems, showcasing a synergy between advanced neural network approaches and established feature extraction methods in achieving robust classification outcomes.

In our study, we address these limitations by specifically focusing on the Algerian dialect and the multi-class problem. The adoption of a deep learning approach was driven by the complexity of the multi-class classification task and subject detection in marketing posts. Deep neural networks excel at learning intricate patterns and representations in data, allowing for a nuanced understanding of the diverse subjects discussed in marketing content. The decision to utilize a deep learning model aligns with our goal of achieving a more accurate and sophisticated classification, considering the varied nature of marketing-related posts in the Algerian context.

By combining TF-IDF for keyword extraction and a deep learning approach for multi-class classification, our methodology aims to leverage the strengths of each technique to enhance the overall effectiveness of our research.

3. METHODOLOGY

Subject detection or multi-class classification in academic languages and in the state-of-the- art studies generally follows the following pattern in Figure 2. This approach could not be applied to the Algerian dialect due to the unavailability of NLP tools for AD. It was necessary either:

- Design NLP tools for Algerian dialect.
- Or translate into an academic language (Arabic, English, French...) and use ML tools.

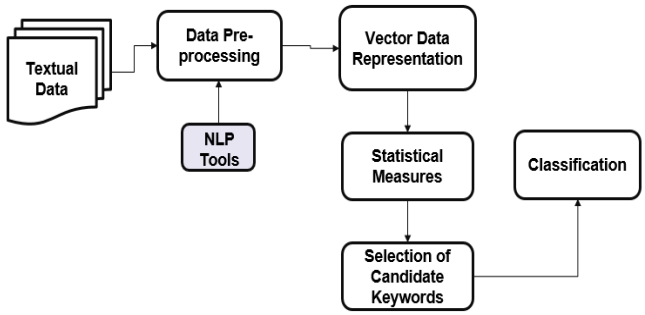

Figure 2. The classical model for subject detection

3.1 Our proposed solution

In this part, we present our proposed approach for the automatic processing of the Algerian dialect and the system realized for the subject detection of posts in Algerian dialects, treated as a multiclass classification problem, in the marketing domain. It differs from other classical approaches. Its originality lies in the introduction of deep learning which allowed us to have a model:

- Evolving with the Algerian dialect, which is not bounded either by its grammar or by its vocabulary.
- Flexible with the infinite number of subject categories that can exist in a domain.
- Relevant because the subject is well targeted by a set of keywords in the dataset.

Our system is divided into many parts including Data collection, Preprocessing, Word representation and Deep Learning. Figure 3 presents our system architecture.

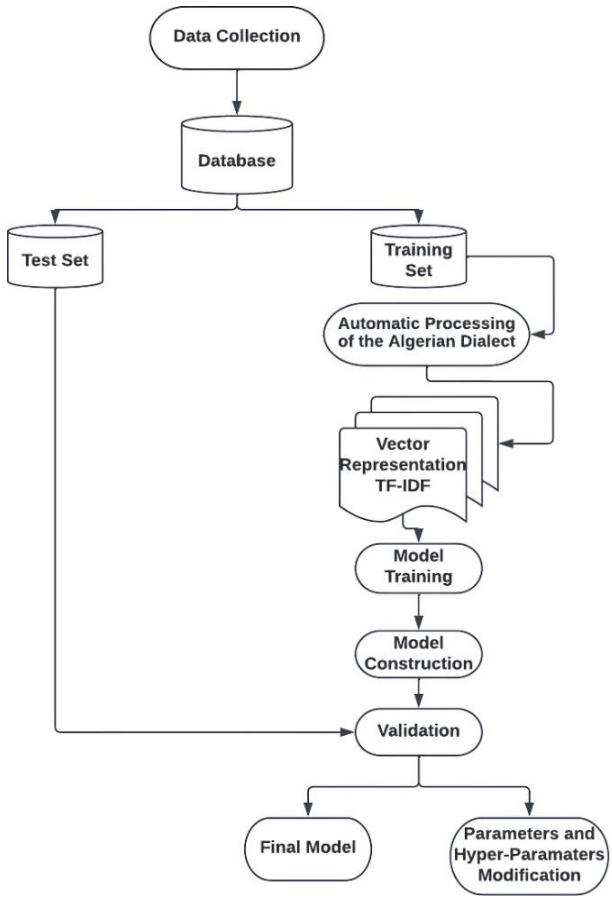

Figure 3. System architecture

3.2 Data collection

Our dataset contains 1000 posts in the Algerian dialect and in Arabizi annotated manually by 11 classes (Accessories and jewelry, Electrical appliances, Electronic Devices, Sports equipment, Sleeping equipment, Fashion and clothes, Kitchen tools, Foodstuffs, Makeup, Kitchen robots, Online subscription offers) from the marketing domain and retrieved from the Jumia e-commerce site, Djezzy, Ooredoo pages using Facepager tool (https://facepager.software.informer.com/). Arabizi is a form of chat language in which Arabic words are written using Latin alphabets and numbers [21, 22]. Figure 4 presents an extract of Algerian dialect data written in Arabic and Arabizi that includes an example for each class.

facebook_posts	class
On n'a jamais assez de chaussures 👟 👠 👡 👢 \r\n!! اليوم كاين تخفيضات على الأحذية 🤩 أبروفيتو\r\n...	الأزياء والملابس
7abaa tkoun 3andek une peau rayonnante و حتى L'Oréal و Nivea Gar... احسن المنتجات فالماركات لي تحبوهم	مستحضرات التجميل
Flash Vente وزيد قولو بلي كاين\r\n😅😅 ماركيه فالكومنتير Smart TV ثاني واحد بعد ال @ رابح يشريلك ...ه	أجهزة الكترونية
Journée Spéciale Sport 😎🏃 \r\n و ما عندكش Matériel حاب تدير سبور فالدار 🏋 جوميا تعرض عليك أ ؟...	معدات الرياضية
madame 3andek 3arss w 7aayra wesh diri accessoires??? اضربوا طلة على اختياراتنا في أروج الموديل...	اكسسوارات ومجوهرات
kech plan more l3id ?? erwa7 dir bon plan camping m3a s7abek منتجات من وسائد وفراشات الخيام	معدات نوم
\r\nتقدروا تشوفوا كل العروض وتشرو 🤩 😋 ماكاش أحلى من الشوكولاتة تهديها لحبايبك وعائلتك في أي وقت...	مواد غذائية
connectez winma 7abit!! M3a Djezzy forfait internet yelha9 partout http://www.ooredoo.dz #Intern...	عرض اشتراك على الإنترنت
bedelte 3etba w 5assek t3amer darek !!! Rak f le bon endroit 🤩 تقدروا تشوفوا كل العروض مطبخ فالم...	ادوات المطبخ
L'électroménager Frigo lave-linge Télévision et ordinateur لكامل محبي أبروفيتو هاد السمانة 🔥 🤩...	الأجهزة الكهربائية
9400 أبروفيتو من فاكية هاذ الموسم 🍊 🍋 باش تحضروا عصائر وميلك شيكات بخلاط مولينكس لي راهو بسعر ...	روبوتات المطبخ

Figure 4. Extract of our Algerian dialect data written in Arabic and Arabizi

```
the_input = '7abit tetfarej match avec une bonne qualité?? Jumia
ابروفيتوا اليوم بعروض هبال على الأجهزة الالكترونية
Smart TV condor 3la maw9i3 https://deals.dz.jumia.com/alger-sity smart condor #Tv'
```

Figure 5. Input example

```
['حبيت', 'تتفرج', 'مباراه', 'جوده', 'جوميا', 'ابروفيتوا', 'بعروض', 'الأجهزه',
'الالكترونيه', 'ذكي', 'تلفزيون', 'كوندور', 'موقع', 'كوندور', 'ذكي', 'تلفزيون']
```

Figure 6. Output result

	keywords_posts_dialecte_algérien	class
0	[...كني, احذيه, تخفيضات, الاحذيه, ابروفيتو, تشروا, كمليو, صاحكم, جوميا, جوميا, جوميادز, مهرجان, ال]	الأزياء والملابس
1	[حبا, جلد, مشي, المنتجات, فالماركات, تحبوهم, لوريال, نيفيا, غارنييه]	مستحضرات التجميل
2	[...رابح, يشريلك, ذكي, تلفزيون, ماركيه, فالكومنتير, وزيد, قولو, فلاش, تخفيض, السعر, الاجهزه, الالكت]	أجهزة الكترونية
3	[...النهار, خاص, رياضه, حاب, ندير, سبور, فالدار, مواد, جوميا, تعرض, المستلزمات, باسعار, غيرابروفيتي]	معدات الرياضية
4	[...سيدتي, عرض, حايره, وش, ديري, ملحقات, اضربوا, طله, اختياراتنا, اروج, الموديلات, خواتم, ملامل, ذ]	اكسسوارات ومجوهرات
5	[الوقت, مكاخر, خطه, لعيد, اروح, ديد, خطه, كخيم, صحابك, منتجات, مساند, المفروشات, الخيام]	معدات نوم
6	[...احلى, الشوكولاته, تهديها, لحبايبك, وعيلتك, تقدروا, تشوفوا, العروض, وتشروا, ك, حلى, حبايبكم, اخت]	مواد غذائية
7	[... الاتصال, وانما, حبيت, معا, نجيزي, حزمه, انترنت, يلحق, في, كل, مكان, انترنت, موقع, يوتيوب, غير]	عرض اشتراك على الإنترنت
8	[مكان, تقدروا, تشوفوا, العروض, مطبخ, فالمقالي, والكسرونات, F, كوسلت, عنده, خاصك, تحمر, داريك, راك]	ادوات المطبخ
9	[...محبي, الاجهزه, المنزليه, ثلاجه, لافليتج, التلفاز, الحاسوب, ابروفيتيو, السمانه, العروض, التخفيضا]	الأجهزة الكهربائية
10	[... ابروفيتيو, فاكيه, الموسم, تحضروا, عصائر, وميك, شيكات, مولينكس, بسعر, تطبيق, جوميا, انسان, آلي]	روبوتات المطبخ

Figure 7. Set of keyword associated to each class

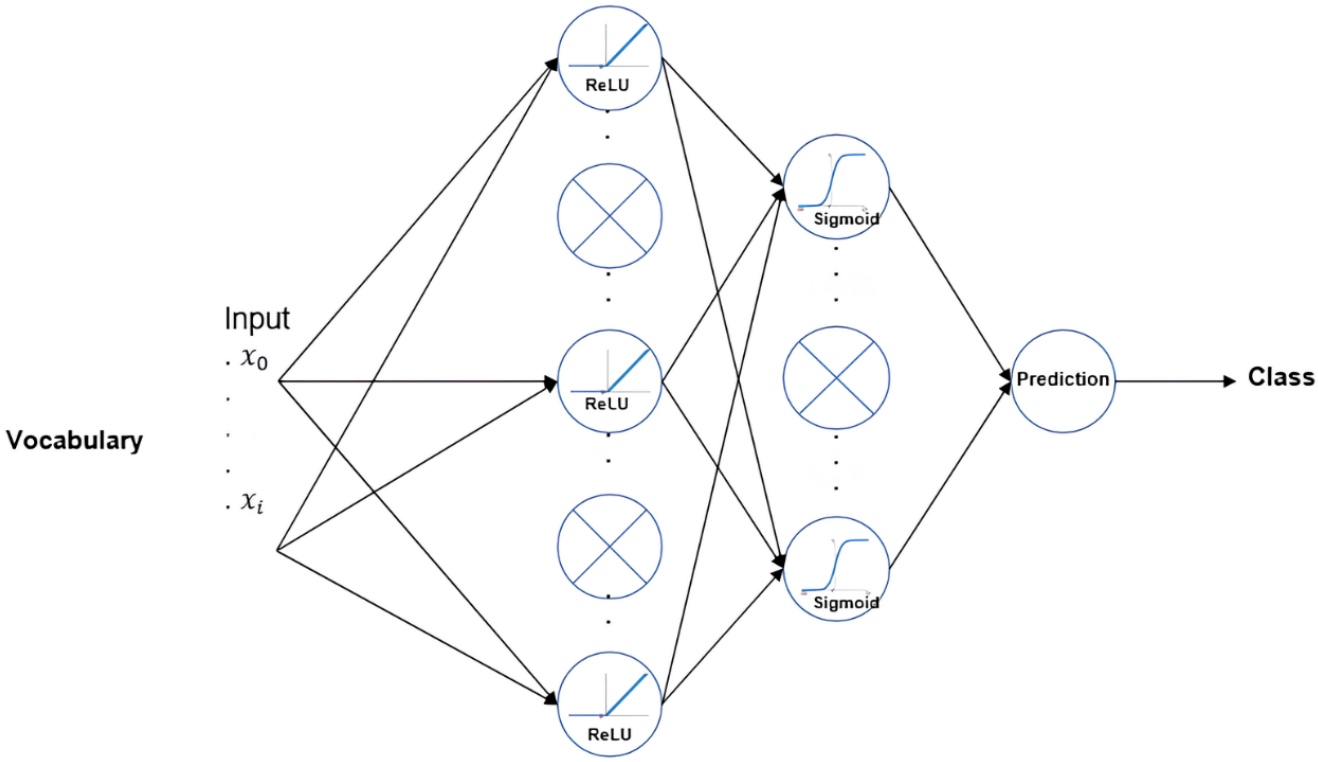

Figure 8. Neural network architecture

The distribution of posts among these classes is as follows:

- Accessories and Jewelry: 95 posts.
- Electrical Appliances: 93 posts.
- Electronic Devices: 85 posts.
- Sports Equipment: 83 posts.
- Sleeping Equipment: 85 posts.
- Fashion and Clothes: 90 posts.
- Kitchen Tools: 95 posts.
- Foodstuffs: 94 posts.
- Makeup: 10 posts.
- Kitchen Robots: 90 posts.
- Online Subscription Offers: 85 posts.

3.3 Preprocessing

The goal of data pre-processing is to extract only the important data from the whole dataset in order to reduce the complexity of the computation and subsequently have more accurate results. There are some of steps to process the Algerian dialect [23]:

- Step 1: URL and special characters removal and Tokenisation.
- Step 2: Transliteration from foreign languages or Arabizi [24] to Algerian Arabic using *Google Translate*.
- Step 3:
- Stop words and numbers cleaning
- Arabic letters normalization:

ابروفيتووووو أضربووووا طلة على كل عروض الاجهزة الالكترونية
Which means in English: "taaaake the opportunity and loooook at the electronic offers" ***Became***: ابروفيتو أضربوا طلة على كل عروض الاجهزه الالكترونيه

PoS Tagging arabe: POST is the process that indicate the function of each word in a sentence by assigning a tag to this word [8]. Arabic has three parts of speech and there are clear grammatical rules for each part. The three parts of Arabic speech are: ***Ism*** **(**nouns, adjectives, pronouns), ***Fi3l*** (Verbs), ***Harf*** (letters, prepositions, particles). We are manually defining a list of Arabic and Algerian adjectives and then removing them.

The presence of adjectives and pronouns in the set of extracted words does not improve subject detection, so we choose to remove them and keep only nouns and verbs. As shown in the following example, text containing Arabic and Algerian adjectives: تخفيض هبال عرض جميل, which means in English: "Great discount, nice offer", the result after eliminating the adjectives: تخفيض عرض , "discount offer".

The result of this preprocessing is the representation of each category by sets of words as we see in Figure 5 and Figure 6.

The example means in English: "Do you want to watch the match in high quality? Jumia with amazing offers on electronic devices. Take advantage today, on site https://deals.dz.jumia.com/alger-sity".

Result of the data preprocessing in Figure 7.

3.4 Data vectorization by TF-IDF and key word extraction

This is a fundamental process in natural language processing because it allows the machine learning algorithms and even computers to understand text. Our neuron network takes as input our vocabulary vectorized by the TF-IDF which associates to each word a specific weight. This can be defined as the calculation of the word pertinence in a corpus. TF-IDF gives a high weight to the rare term and a low weight to the common term [25], which is very adequate for our task, we consider the set of extracted words, after preprocessing, as a set of keywords for the associated category. It is a metric that multiplies the two quantities tf and idf [26, 27]:

- ***Term frequency:*** It signifies how many times a word **t** is present in a given document **d**.
- ***Document frequency:*** It indicates the number of documents in a collection **D** that contain a specific word **t**. It

measures the distribution of a word across the entire document set.

- ***Inverse Document Frequency:*** It is computed as the logarithm multiplied by the inverse probability of a word t being found in any essay.
- ***Calculation:*** tfidf(t,d,D) = tf(t,d) x idf (t,D).

Our model will learn on these keyword called Vocabulary contains 473 items.

3.5 Label Encoding

Classes must also be represented by numerical values, for this we use label encoding, called Label Encoding. This function allows to associate a unique number for each class in our database. After applying this function we get the following result in Table 2:

Table 2. Labelencoding function result

0	اكسسوارات ومجوهرات Accessories and jewelry
1	الاجهزة الكهربائية Electrical appliances
2	اجهزة الكترونية Electronic Devices
3	معدات رياضية sports equipment
4	معدات نوم Sleeping equipment
5	الازياء والملابس fashion and clothes
6	عرض اشتراك على الانترنت Online subscription offers
7	ادوات المطبخ kitchen tools
8	مواد غذائية foodstuffs
9	مستحضرات التجميل makeup
10	روبوتات المطبخ Kitchen robots

3.6 Deep learning

Deep Learning utilizes supervised learning, but the machine's internal architecture is different: it is a "neural network", a virtual machine consisting of neurons that each execute small simple calculations, it more powerful and effective method [28]. Our deep learning performed occurs after the preprocessing of the data.

3.6.1 Training and test

The dataset will be partitioned into two databases, the first one is dedicated to train the model and build it, and the second one is for testing and validation of the latter. The neural network must be trained on a training set to obtain a model, then the test set will validate it.

3.6.2 Our deep learning approach architecture

The hyper-parameters and parameters [10] fixed before the training, if we judge, via the evaluation metrics, that the model is not valid, they will be manipulated and modified and the training is redone until the final model is obtained. The Dropout set to 0.4 and the number of epochs set to 2.

- ***The first layer*** composed of 2000 neurons, it takes, as input our vocabulary vectorized by the TF-IDF and provides outputs that are the results of activation function **RELU** on the sum of the input vectors multiplied by the weight and the bias, we recall that the layers are interconnected.
- ***The second layer***: The number of classes will be the input of the layer with **SIGMOID** activation function. It called the prediction layer.
- **ReLU:** Rectified Linear Unit, an activation function frequently employed in neural networks, it introduces non-linearity to the network, allowing it to acquire different patterns. ReLU function defines as: f(x)=max(0,x). If the input is positive (x>0), the output is the input itself; if the input is negative (x<0), the output is zero.
- **Sigmoid:** An activation function that referred to as the logistic function, is frequently employed in neural networks. It aims to generate an output within the range of 0 and 1, indicating the probability of membership in a specific class. The formula is: f(x)=1/1+e^{-x}.
- **Dropout:** It involves temporarily deactivating a portion of the neurons and their connections during the learning process allowing the network the opportunity to discover new ways to solve the given problem, and improve generalization and reduce overfitting [29].

The Figure 8 illustrate our neural network architecture.

4. EVALUATION AND RESULTS

The study of the predictive values that called test set allows us to define if our model is reliable, and in which cases it makes errors and to what extent. We choose the most used metrics to evaluate our model: Precision, recall.

The modification of the parameters and hyper-parameters depends on the precision. The objective is to have a precision that converges to 1.

The configuration of our model provided relevant results as shown in Table 3.

Table 3. Metric evaluation results

Loss	Precision	Recall	F1-Measure
3.53%	83.03%	100%	90.72%

We prepared some texts in the field of marketing and apply our model to extract keyword, detect and predict the subject or the category of these texts from the 11 classes. Table 4 gives two examples with their classes.

In Table 5, we present a comparison between some works in the sentiment analysis field under the following dimensions: study, number of class and results. The purpose of this comparison is the classification, disregarding the research objective of these works. We emphasize that the works cited in the table concern the Algerian dialect [30, 31], the Moroccan dialect [32], Modern Standard Arabic and Colloquial Arabic [33]. All of them used ML or DL techniques.

As observed in the previous comparative table, the performance metrics are influenced by factors such as the class number. Our research is one of the limited studies that have been done in the multiclass classification task with a considerable number of classes (11 classes).

4.1 Discussion

Achieving a precision of 83.03% demonstrates the model's ability to make accurate positive predictions across the multiple classes, and the recall of 100% signifies the model's

capability to identify all instances belonging to the positive classes.

Table 4. Examples with their predicted classes

Input	Result
عرض خاص لبنات besh tkoun 3andek une belle peau mechi obligé tkouni makeup girl, les produits de la Roche Posay et Vichy sont idéals cet été. Code promo. Jumia%30. **Which means in English** "*A special offer for girls to have beautiful skin, you don't need to be makeup girl, Roche Posay and Vichy products are ideal this summer.*"	Makeup
Marakch 7ab tayeb fdar?? Wach 7aab takl lyoum? Tacos et pizzas 🍕 mane3and Jumia Food dayra promo ta3 hbaal https://deals.dz.jumia.com #tacos #pizza **Which means in English** "*Don't want to cook at home?? What are you going to eat today?? Tacos pizza from Jumia Food with a great promo*"	Foodstuffs
Journée special sport 😎 🏋️‍♀️ حاب دير سبور في الدار ومعندكش materiel 🏋️‍♀️,,, Jumia ديما معاك 🤩 **Which means in English** "*Special sport day* 😎 🏋️‍♀️" *You want to practice sport in your home, but you don't have gym equipment 🏋️‍♀️,,, Jumia always with you* 🤩 "	Sport equipment
Ooredoo ؟ علابالك واش تقدر دير بـ 100 دج عند Messenger و هدرة نحو Facebook تقدر تستفاد من Ooredoo غير محدودين ! إنترنت و 100 دج رصيد Moوزيد معاهم 500 ! والكل صالح نهار كامل ! المزيد من المعلومات على موقعنا : http://www.ooredoo.dz **Which means in English** "*What do you think you can do for 100 DZD at Ooredoo? You can benefit from Facebook and Messenger and unlimited calls! In addition to 500 Mo internet and 100 DZD credit! And everyone is good all day! More information on our website: http://www.ooredoo.dz*"	Online subscription offers

Table 5. Comparative analysis between our proposed work with other Arabic works

Study	Number of Class	Results
[30]	2 classes: positive, negative	Accuracy: 85%
[31]	3 classes: Positive, negative, neutral	Accuracy: 84.21%
[32]	4 classes: POS, NEG, Objective, Sarcasm	Accuracy: 83%
[33]	5 classes: Very negative, Negative, Neutral, Positive Very, positive	Accuracy: 32.38%
Our work	11 classes	Accuracy: 80% Recall: 100%

However, it's crucial to acknowledge the limitations of our project. The dataset used, although carefully selected, is relatively small. Nevertheless, when we systematically tested various samples in our model, we observed that instances of misclassification were very rare. This suggests that our small dataset is representative, and the model's performance can be projected onto a much larger dataset. It's noteworthy that the posts in our dataset are longer compared to comments, providing the advantage that these 1000 posts include a diverse range of Algerian words. Additionally, the multiclass classification within the same post, a scenario that is rare to be found, poses challenges in terms of training and evaluating the model effectively.

5. CONCLUSION AND PERSPECTIVES

The study of natural language and of the mechanisms necessary for its automatic processing by machines is a rich field, with many potential applications. In the treatment of the Algerian dialect, the major limitation encountered and which causes a real handicapped is the call to an important sum of expert knowledge: lexicons, grammar rules, semantic networks. In this work, we have addressed this problem by treating the difficulties inherited from the Arabic language by translating the words in foreign languages and Arabizi into Algerian Arabic. We used TF-IDF for the extraction of keyword for each class and for the representation of our data and then we proposed a method based on deep learning for the topic detection. The results of our system are satisfactory despite the small size of our data, we believe that the results will be better with larger data.

As perspectives, we plan to improve our approach by: using one of the recent contextual word embedding model as "KeyBERT", carrying out the PoS Tagging of the Algerian dialect automatically, and detecting the subject from other medias. We also plan to add more classes and detect the multiclass in the same text.

REFERENCES

[1] Badaro, G., Baly, R., Hajj, H., El-Hajj, W., Shaban, K.B., Habash, N., Al-Sallab, A., Hamdi, A. (2019). A survey of opinion mining in Arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 18(3): 1-52. https://doi.org/10.1145/3295662

[2] Adnane, H. (2023). The evolution of marketing in Algeria. Journal of Contemporary Business and Economic Studies, 6(2): 188-200.

[3] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., Brown, D. (2019). Text Classification Algorithms: A Survey. Information, 10: 150. https://doi.org/10.3390/info10040150

[4] Tarekegn, A.N., Giacobini, M., Michalak, K. (2021). A review of methods for imbalanced multi-label classification. Pattern Recognition, 118: 107965. http://doi.org/10.1016/j.patcog.2021.107965

[5] Kannan, E., Kothamasu, L.A. (2022). Fine-tuning BERT based approach for multi-class sentiment analysis on Twitter emotion data. Ingénierie des Systèmes d'Information, 27(1): 93-100. https://doi.org/10.18280/isi.270111

[6] Firoozeh, N., Nazarenko, A., Alizon, F., Daille, B. (2020). Keyword extraction: Issues and methods. Natural Language Engineering, 26(3): 259-291. http://doi.org/10.1017/S1351324919000457

[7] Beliga, S. (2014). Keyword extraction: A review of methods and approaches. University of Rijeka, Department of Informatics, Rijeka, 1(9): 1-9.

[8] El Hadj, Y., Al-Sughayeir, I., Al-Ansari, A. (2009). Arabic part-of-speech tagging using the sentence structure. In Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt, 241-245.

[9] Yao, L., Pengzhou, Z., Chi, Z. (2019). Research on news keyword extraction technology based on TF-IDF and TextRank. In 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), Beijing, China, pp. 452-455. http://doi.org/10.1109/ICIS46139.2019.8940293

[10] Koutsoukas, A., Monaghan, K.J., Li, X., Huan, J. (2017). Deep-learning: Investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. Journal of cheminformatics, 9(1): 1-13. http://doi.org/10.1186/s13321-017-0226-y

[11] Chen, P.I., Lin, S.J. (2010). Automatic keyword prediction using Google similarity distance. Expert Systems with Applications, 37(3): 1928-1938. http://doi.org/10.1016/j.eswa.2009.07.016

[12] Zhang, C. (2008). Automatic keyword extraction from documents using conditional random fields. Journal of Computational Information Systems, 4(3): 1169-1180. http://hdl.handle.net/10760/12305

[13] Moh'd A Mesleh, A. (2007). Chi square feature extraction based svms arabic language text categorization system. Journal of Computer Science, 3(6): 430-435.

[14] Boukil, S., Biniz, M., El Adnani, F., Cherrat, L., El Moutaouakkil, A.E. (2018). Arabic text classification using deep learning technics. International Journal of Grid and Distributed Computing, 11(9): 103-114. http://doi.org/10.14257/ijgdc.2018.11.9.09

[15] Aipe, A., Mukuntha, N., Ekbal, A., Kurohashi, S. (2018). Deep learning approach towards multi-label classification of crisis related tweets. In Proceedings of the 15th ISCRAM Conference.

[16] Song, J., Hu, R., Sun, B., Gu, Y., Xiong, W., Zhu, J. (2019). Research on news keyword extraction based on TF-IDF and Chinese features. In 2nd Int. Conf. on Financial Management, Education and Social Science (FMESS 2019), Hohhot, Inner Mongolia, China, pp. 344-352.

[17] Hioual, O., Hemam, S.M., Hioual, O., Maif, L. (2022). A hybrid approach for web pages classification. Ingénierie des Systèmes d'Information, 27(5): 747-755. https://doi.org/10.18280/isi.270507

[18] Loukam, M., Hammouche, D., Mezzoudj, F., Belkredim, F.Z. (2019). Keyphrase extraction from modern standard Arabic texts based on association rules. In Arabic Language Processing: From Theory to Practice: 7th International Conference, ICALP 2019, Nancy, France, October 16–17, 2019, Proceedings 7, Springer International Publishing, pp. 209-220. https://doi.org/10.1007/978-3-030-32959-4_15

[19] Khan, M.Q., Shahid, A., Uddin, M.I., Roman, M., Alharbi, A., Alosaimi, W., Almalki, J., Alshahrani, S.M. (2022). Impact analysis of keyword extraction using contextual word embedding. PeerJ Computer Science, 8: e967. http://doi.org/10.7717/peerj-cs.967

[20] 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. http://doi.org/10.14569/IJACSA.2016.070353

[21] Duwairi, R.M., Alfaqeh, M., Wardat, M., Alrabadi, A. (2016). Sentiment analysis for Arabizi text. In 2016 7th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, pp. 127-132. http://doi.org/10.1109/IACS.2016.7476098

[22] Oussous, A., Benjelloun, F.Z., Lahcen, A.A., Belfkih, S. (2020). ASA: A framework for Arabic sentiment analysis. Journal of Information Science, 46(4): 544-559. http://doi.org/10.1177/0165551519849516

[23] Guellil, I., Azouaou, F. (2017). ASDA: Analyseur Syntaxique du Dialecte Alg {\'e} rien dans un but d'analyse s {\'e} mantique. arXiv preprint arXiv:1707.08998. https://doi.org/10.48550/arXiv.1707.08998

[24] Guellil, I., Azouaou, F., Benali, F., Hachani, A.E., Saadane, H. (2018). Approche Hybride pour la translitération de l'arabizi algérien: une étude préliminaire. In 25e conférence sur le Traitement Automatique des Langues Naturelles (TALN).

[25] Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. Journal of Documentation, 60(5): 503-520.

[26] Aizawa, A. (2003). An information-theoretic perspective of TF–IDF measures. Information Processing and Management, 39(1): 45-65. http://doi.org/10.1016/S0306-4573(02)00021-3.

[27] Hakim, A.A., Erwin, A., Eng, K.I., Galinium, M., Muliady, W. (2014). Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach. In 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, Indonesia, pp. 1-4. http://doi.org/10.1109/ICITEED.2014.7007894

[28] Zhang, L., Tan, J., Han, D., Zhu, H. (2017). From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discovery Today, 22(11): 1680-1685. http://doi.org/10.1016/j.drudis.2017.08.010

[29] Lim, H.I. (2021). A study on dropout techniques to reduce overfitting in deep neural networks. In Advanced Multimedia and Ubiquitous Engineering: MUE-FutureTech 2020, Springer Singapore, pp. 133-139. http://doi.org/10.1007/978-981-15-9309-3_20

[30] Bousmaha, KZ., Hamadouche, K., Gourara, I., Hadrich-Belguith, L. (2022). DZ-OPINION: Algerian dialect opinion analysis model with deep learning techniques. Revue d'Intelligence Artificielle, 36(6): 897-903. https://doi.org/10.18280/ria.360610

[31] Mazari, A.C., Djeffal, A. (2022). Sentiment analysis of algerian dialect using machine learning and deep learning with Word2vec. Informatica, 46(6): 67-78. http://doi.org/10.31449/inf.v46i6.3340

[32] Matrane, Y., Benabbou, F., Sael, N. (2021). Sentiment analysis through word embedding using AraBERT: Moroccan dialect use case. In 2021 International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA), Marrakech, Morocco, pp. 80-87.

http://doi.org/10.1109/ICDATA52997.2021.00024
[33] Barhoumi, A., Estève, Y., Aloulou, C., Belguith, L. (2017). Document embeddings for Arabic sentiment analysis. In Conference on Language Processing and Knowledge Management, LPKM 2017.