Journal of Theoretical and Applied Information Technology
15th February 2023. Vol.101. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SYSTEMATIC REVIEW OF THE ARABIC NATURAL LANGUAGE PROCESSING: CHALLENGES, TECHNIQUES AND NEW TRENDS

GHIZLANE BOURAHOUAT[1], MANAR ABOUREZQ[2], NAJIMA DAOUDI[3]
[1, 2, 3] ITQAN Team, LyRICA Laboratory, ESI, Rabat, Morocco
E-mail: [1] ghizlane.bourahouat@esi.ac.ma, [2] mabourezq@esi.ac.ma, [3] ndaoudi@esi.ac.ma

ABSTRACT

The volume of Arabic posts on many social networks has increased significantly, providing a rich source for analysis. As a result, Arabic Natural Language Processing intervenes to exploit this source and extract invisible but valuable insights. This paper presents a review of recent studies on techniques used in the Arabic Natural Language Processing field to come up with the faced challenges and the new trends. The articles selected for the review are primarily studies on Arabic Natural Language Processing techniques, as we collected and analysed a set of journal papers published in this field between 2018 and 2022. Based on the analysis, we extracted the various ANLP steps and investigated the techniques used in each step. The article also outlines the current trends in the several phases and steps of the Arabic Natural Language Processing process. As a result, it gives an insight into the current state of research.

Keywords: *Arabic Natural Language Processing, Systematic Literature Review, Data Collection, Tokenisation, Embedding.*

1. INTRODUCTION

Interaction via social media has increased among people all over the world, who see it as an important tool for sharing information about various topics and expressing opinions freely and openly. This interaction and exchange result in a massive amount of data that provides a wealth of information to understand user behaviour and trends more closely. Once thoroughly investigated, this massive amount of data can be used by decision-makers for a better understanding of a given situation, and thus to make the right decision. Without the right technology, it is nearly impossible to analyse and process this large amount of data. This is where Natural Language Processing (NLP) comes into play.

NLP is the subfield of Artificial Intelligence (AI) that focuses on the processing and understanding of human language by machines. NLP is a "theoretically motivated range of computational techniques for analysing and representing naturally occurring texts at one or more levels of linguistic analysis to achieve human-like language processing for a range of tasks or applications" (Liddy, 2001). With NLP technologies, it is now possible to build applications for various aims, such as sentiment analysis (SA), speech recognition, optical character recognition (OCR), information retrieval, machine translation, question answering, and text summarization, etc.

NLP has been applied to different languages, including Arabic. According to Statista (Statista, 2021), there are about 319 million Arabic native speakers in the whole world, making it the fifth most-spoken language globally behind Mandarin, Spanish, English, and Hindi. In addition, the Arabic language is ranked fourth among the most often used languages on the Internet (Statista, 2020), and it is one of the fastest-growing languages online. Another point to mention is the growing number of translation initiatives into Arabic in various fields of science and culture. With the large community of this language, researchers have been interested in producing customized techniques that are appropriate to the Arabic language, yielding to Arabic Natural Language Processing (ANLP).

ANLP consists of the development of methods and tools that enable the use and analysis of Arabic in both written and spoken contexts. This development requires additional efforts due to the various challenges related to the Arabic language, such as morphological richness, orthographic ambiguity, dialectal variations, orthographic noise,

and resource poverty. Thus, ANLP has attracted the attention of researchers, especially after the significant advances of NLP for other languages, mainly the English language, and has been the subject of many research papers. This work aims to investigate the challenges of natural language processing for Arabic (ANLP) and the latest techniques being used to overcome them, in order to keep up with the most recent technologies in this field.

The organisation of the remainder of this paper is as follows. Section 2 presents the general context and deals with the application of NLP techniques to non-Latin character languages such as Arabic, the challenges faced while performing ANLP, and the followed process. Section 3 presents the methodology followed during the research. Section 4 presents the results related to each step of the ANLP process. Finally, Section 5 discusses the results and some insights into future research directions.

2. GENERAL CONTEXT AND PROBLEM STATEMENT

2.1 NLP and non-Latin alphabet languages

When we mention NLP, we invoke the various existing languages, whether they are written using the Latin alphabet such as English, French, Spanish, etc. or non-Latin alphabets such as Arabic, Chinese, Urdu, Thai, etc. Various works have been carried out in NLP in these languages taking into consideration the characteristics of each.

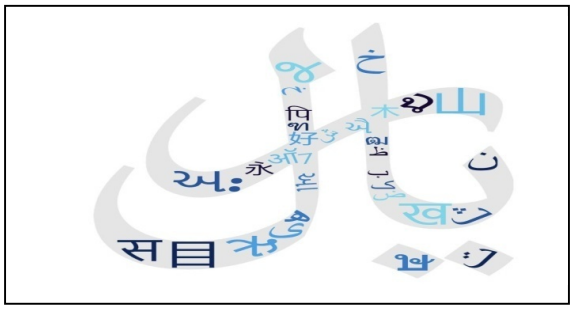


Figure 1: Alphabets Of Various Non-Latin Alphabets Languages

NLP, as well as all the other techniques in AI, have been highly interested in the English language and therefore, we find various works and research in this direction. This does not exclude the fact that NLP is also oriented towards other languages, including languages with non-Latin characters. However, this orientation is not entirely free of challenges.

In the case of the Indian language, and according to [1] and [2], it is noted that it is a morphologically rich and free order language, unlike English, which adds complexity to the processing of user-generated content, not to mention the variety of dialects that exist such as Bengali, Gujarati, Tamil, Malayalam, Telugu, and Bengali. Taking for example the Urdu dialect, [3] and [4] highlighted its specificity compared to other languages due to its morphological structure as it starts from right to left.

Moving on to the Chinese language. Compared to English, NLP in Chinese is more difficult, as its vocabulary and semantics are more complex. In addition, the semantics of Chinese texts are more context-dependent [5]. Compared to the normal delimiters of English, Chinese has neither formal delimiters nor a strict division method, and even sometimes the division method depends on the context.

As for the Thai language, [6] and [7] noted that the words are written without a word or sentence delimiters, and words are placed continuously without spaces in sentences. Another problem faced by the Thai language is that of word variants, the use of different words to express the same idea, polysemy, and the ability of words to express several ideas depending on the context, in addition to the lack of explicit word boundaries.

Arabic is also one of the non-Latin alphabet languages that require additional efforts due to the challenges associated with it, namely morphological richness, orthographic ambiguity, dialectal variations, and orthographic noise. Another challenge that is relevant to the mainstream of non-Latin languages, including Arabic, is the lack of resources for training and evaluating Machine Learning (ML) models. In this article, the focus will be on the Arabic language to detect the challenges related to ANLP and the studies done so far to resolve them.

2.2 Arabic language

The Arabic language has a unique history in that it has remained unchanged for more than sixteen centuries (before The Holy Quran in 609 CE). Arabic is a Semitic language with an inextricable link to Islam and Arabic culture, serving as the Quranic language for all Muslims (more than 1.62 billion people).

Journal of Theoretical and Applied Information Technology
15th February 2023. Vol.101. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Moreover, it is the native language of more than 422 million speakers. In addition to the uniqueness of the historical and cultural background of the Arabic language, its nature and structure are different from other languages such as English. For example, this language is written from right to left. It comprises 12 million words, and 28 letters, including three vowels and diacritics (short vowel symbols inscribed atop regular letters). Diacritics may affect the semantics and syntax of a word [8] There is no letter capitalization in Arabic, however, the shapes of Arabic letters change according to their positions in a word. Arabic is a structured language with an abundance of vocabulary, wherein morphology plays an important role. Furthermore, words are often constructed in a complex manner.

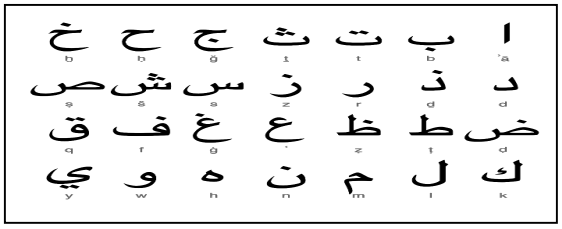


Figure 2: Arabic Alphabets

Moreover, there are three main Arabic varieties. The Classical Arabic (CA) or the language usually used in religious and literature contexts, is fully structured and vowelized. The Modern Standard Arabic (MSA), which is the official language used in education and formal communications across the different Arabic speaking countries, is based on the CA's syntax and morphology, but it tends to have a more modern vocabulary. Finally, the Arabic colloquial dialects (AD), or the language used in daily informal conversations, has no orthographic standards so one word can be written in different forms. Additionally, the AD variety can vary from one region to another across the Arabic countries. AD is mostly divided into six main groups: (1) Egyptian, (2) Levantine, (3) Gulf, (4) Iraqi, (5) Maghrebi, and (6) Others, which contains the remaining dialect [9].

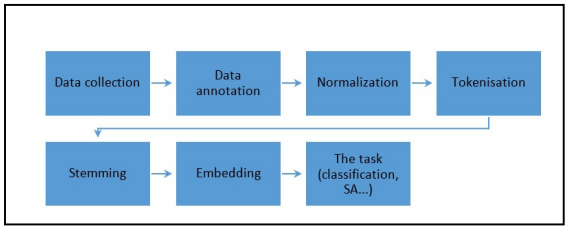
2.3 Problem statement

Because of the complexity of the Arabic language, several challenges are faced when dealing with it. Referring to the Figure 3, we present the ANLP process, which is the same as the generic NLP process.

Figure 3: ANLP Process

In this section, we'll go through each step of the process and discuss some of the used techniques.

As Arabic is one of the non-Latin languages mentioned earlier, the following challenges are



faced while processing Arabic texts:

Morphological richness: Arabic has many forms that result from rich inflexions. These include gender, number, person, aspect, mood, case, and some connectable clitic features. As a result, it is not uncommon to find a single Arabic word that translates into a five-word English phrase: hou+una-lum-ya+sa+wa وسيلومونه 'and they will blame him. This task creates a challenge for ML models by increasing the number of unique vocabulary types compared to English [10]. This challenge is generally faced during the step of stemming and tokenization.

Orthographic ambiguity: Arabic letters use optional diacritics to represent short vowels and other phonological information that are important in distinguishing words from each other. Other than religious texts and children's literature, these symbols are rarely used and provide a high degree of ambiguity. Educated Arabs usually have no problem reading Arabic without diacritics, but it is a challenge for both learners of Arabic and computers. If we give the example of 'ذهبت', without diacritics, it could be read as 'ذَهَبْتُ'dahabtou which means I went or as 'ذَهَبَتْ' dahabat which means she went. Referring to the generic process followed in NLP, we can say that this challenge is shown while performing the steps of stemming and embedding.

Dialect variations: As we mentioned earlier, there are various dialects in the Arabic language, which enhances its corpus but creates an obstacle when processing Arabic texts. Each dialect having its grammar and lexicon that differ from the others and from Standard Arabic, it results in a variation and a complexity that are faced during the ANLP process, especially during tokenisation, stemming, and embedding.

Resource poverty: In NLP, data is the most important resource; this is true for rule-based approaches, which require carefully constructed lexicons and rules, as well as ML approaches, which require corpora and annotated corpora. Although there are many corpora of unannotated

Arabic texts, morphological analysers, Arabic lexicons, and annotated corpora are not available. In addition, annotations other than news and dialects are limited. The importance of data availability is evident from the process mentioned above, as the data collection is the baseline for the work to follow.

Based on these challenges, our aim is to investigate their impact during each step of the described process of Figure 3. How could the mentioned challenges influence the data collection, normalisation, stemming, tokenisation and embedding step? Which step is more influenced that the other?

Thus, after the presentation of the various challenges related to the processing of the Arabic language, we will proceed to the description of the methodology used in this study in order to answer the previous questions.

3. METHODOLOGY

This paper presents a literature review of recent studies in the field of ANLP. The goal is to study the techniques used through the overall ANLP process. We examined articles published in journals that are indexed on the Web of Science and Scopus databases between 2018 and 2022 to find the relevant studies. The papers were identified using the keywords "Arabic natural language processing", "ANLP", "Arabic natural language processing and data collection", "Arabic natural language processing and tokenisation", "Arabic natural language processing and stemming", "Arabic natural language processing and word embedding", "Arabic natural language processing and word embedding", "Arabic Sentiment Analysis", and "Arabic Sentiment Analysis and word embedding".

After we retrieved the articles from the online databases, the titles and abstracts of the articles were screened using predefined selection criteria. An article was considered suitable for inclusion in this research if it met all the following inclusion criteria:

- It deals with Arabic NLP,
- It focuses on at least one of NLP process,
- It is written in English or Arabic,
- Articles published in high-impact journals between 2018 and 2022.

An article was excluded if:

- The full text of the article is not available online,
- The article is in the form of a poster, tutorial, abstract, or presentation,
- It is not in English,
- It does not mention the models of the opted approach.

Where a decision about the inclusion of an article was in doubt, the full text was read to make a final judgment.

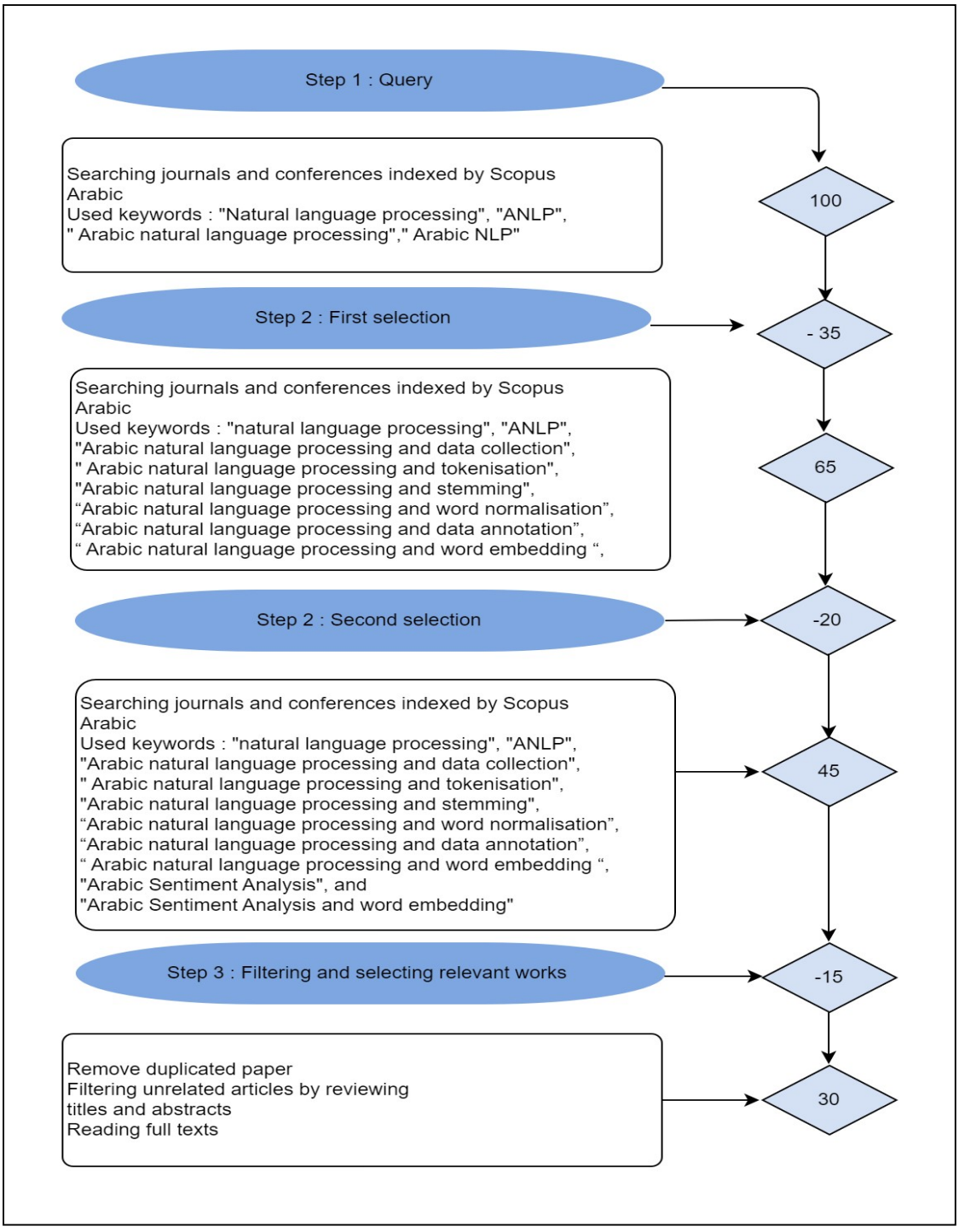


Figure 4: Search Strategy

The mentioned inclusion and exclusion criteria were applied to around 100 papers retrieved initially, and 30 relevant papers matching the criteria were finally chosen for the in-depth analysis of this study as shown in the Figure 4.

4. RESULTS AND ANALYSIS

In this section, we will present the results of our study. We will focus on each step of the different phases of the ANLP process.

Our research has started by analysing the general process of ANLP. Based on this analysis, we’ve

extracted different challenges faced in each step. In this section, we'll discuss each step of the general process.

4.1. Data Collection

As mentioned before, data is the main base for a high-performance NLP model. Unfortunately, the availability of good and reliable Arabic resources was one of the major issues, as reported by [11]. [12], [13], and [14], brought up the issue of unbalanced data. For [13], the collected corpus consisted of 5K airline service-related tweets in Arabic. In [15], the authors mentioned the lack of available corpora on the web for Arabic sentiment analysis for standard and dialect variations and created the SANA corpus, which is a collection of comments from three Algerian newspapers. In [16], the authors raised the lack of work and resources in the Algerian dialect and collected the "Ar_corpus1" from Facebook. It is worth mentioning that all the extracted data were from social media platforms such as Facebook and Twitter, and then they are generally users' short reviews and comments with a limited number of words. From the reviewed articles, we extracted some of the open-access datasets that we can use in future. We mention the SANAD dataset, which is a large collection of Arabic news articles that can be used in different ANLP tasks such as Text Classification and Word Embedding. We also note the Arabic Sentiment Tweets Dataset (ASTD) which contains over 10k Arabic sentiment tweets classified into four classes subjective positive, subjective negative, subjective mixed, and objective. The Arabic Influencer Twitter Dataset (AITD) is also available in open access in addition to the Arabic Social Media News Dataset (ASND). There is also the Open-Source Arabic Corpora (OSAC) is a large standard dataset for text categorization. Finally, we note the ArSarcasm , a new Arabic sarcasm detection dataset containing 10,547 tweets, in multiple Arabic dialects, 1,682 (16%) of which are sarcastic. Moreover, [17] faced the lack of resources to detect Arabic fake news and constructed their own Arabic dataset based on news sentences from an Arabic Twitter dataset and by performing web scraping.

4.2. Data annotation

In NLP, data should be annotated according to the targeted task, such as sentiment analysis for example. That is what was raised in [18], [19], [20] and [21] where the collected data was annotated manually. Both [22] and [23] have manually annotated data related to violence in order to classify it. [24] have accomplished the annotation by turning to crowdsource to perform multiple annotations. Moreover, in [15] the MATTER (Model, Annotate, Test, Train, Evaluate, Revise) approach, which is a general methodology for creating annotation and ML tasks of all different types, was used to annotate the data. It is worth mentioning that the annotation depends on the field of study as in the case of [24] where they requested the assistance of psychologists to complete this task.

4.3. Word normalisation

Normalisation helps to reduce the amount of different information that the computer must deal with, and therefore improves efficiency. This is also one of the mandatory steps in NLP and ANLP. Authors in [11] used three tools, namely MADAMIRA, Farasa, and the Stanford toolkit in order to handle Arabic morphological and text processing. MADAMIRA tool is "a fast, comprehensive tool for morphological analysis and disambiguation of Arabic" [25]. As for Farasa, it is "a fast and accurate text processing toolkit for Arabic text" [26]. Finally, the Stanford toolkit is "an extensible pipeline that provides core natural language analysis" [27]. Farasa was also used by [28] and [29] to segment the input data. Moreover, it was used with XLNet model in [30] and achieved good results.

4.4. Data tokenization

Word tokenization is the process of splitting a large sample of text into words. This is a requirement in NLP tasks where each word needs to be captured and subjected to further analysis. [8] has been interested in this concept and proposed a tokenisation algorithm to fragment Arabic text into words based on the spaces between words and punctuation. [29] used the BERT tokenizer to perform the tokenization of the Arabic corpus. Therefore, since the Arabic language doesn't have capitalisation further techniques need to be used to break down the sentence into tokens. In [31] WordPiece tokenizer of BERT was used to perform the tokenisation. [30] investigated also WordPiece tokenizer and SentencePiece of XLNet on Arabic corpus and achieved an accuracy of 94.78% when used with XLNet. We also detected that [17] had used its own tokenizer model based on BERT models to deal with the sentence's meaning.

4.5. Data stemming

Data stemming consists of producing morphological variants of a root/base word. For this reason, various techniques are used as [32] noted.

- Root-based approach: The main goal of these stemmers is to extract the root of words,
- Stem-based approach: These stemmers identify the stem of words. As an example of Arabic stemmer, we cite Farasa,
- Light-stem-based approach: These stemmers are used to eliminate suffixes and prefixes from stems. Light10 is the most used one as a light stemmer,
- Arabic morphological analyser: It can identify the different forms of words (root, stem, and light stem).

In [8], the authors proposed a stemmer module providing a solution to the challenges resulting from the complexity of social media Arabic words and aims to fulfil two objectives: (1) to help understand the meaning of the word by providing its root, and (2) to determine whether the word is a noun, stop word, or a non-standard Arabic word (in case of not finding its root), a dialect, an error, or a non-Arabic word, yet written using the Arabic script. [12] also investigated the impact of stemming combined with the word embedding for Text Categorization and obtained an F1 score of 97.96% by combining the light stemmer, Word2Vec as embedding technique, and Att-GRU as a classification model. Moreover, [33] came with a broken plural rule (BPR) algorithm introducing new solutions to solve the problem.

4.6. Word Embedding

One of the critical steps in all models or tasks is word embedding. This step consists of generating distributed word vector representations and representing the words or sentences of a text by vectors of real numbers. In other words, this step consists of converting the textual data into numerical data machine-interpretable. Different techniques were used in the Arabic context. This step is one of the mandatory steps to succeed in an NLP project whether in an Arabic context or other languages. From the reviewed articles, we've found [12] that investigated the word embedding with Word2Vec for Text Categorization, which resulted in an F1 score of 97.96% when combined with Att-GRU as a classification model. [11] have implemented the Glove technique to extract the embeddings and got an accuracy of 94,80% when combined with CNN. In [32], the authors investigated the AraVec embeddings that achieved an accuracy of 87,51% with NuSVC. FastText was used with [16] and got 80% as an accuracy with CNN. We also find the AraBERT model used with [34] on multi-dialect dataset for the embedding step and got an accuracy of 89.6%. QARIB model was implemented as an embedding model and a classifier in [35] and achieved an accuracy of 70% applied on a multi-dialect dataset. Moreover, mBERT was also used for the embedding task in [34] combined with itself to achieve an accuracy of 93,8%.

4.7. Task performed

Once all the previous steps have been applied, we move on to the ultimate objective we want to achieve. The objectives vary according to the orientation and vision of the article and the researcher. The aim of the article could be related to the classification or the prediction of humans' sentiments, emotions, appraisals, attitudes, or opinions toward products, issues, events, or services. In our case, we extracted articles that aim to analyse sentiments as we've seen in [15], [16], [18], [19], [20], [26], [36], [29], [32] , [37], [49], [38], [39], [40] and [41]. We also found articles interested in emotion detection as stated in [42], [41] and [43]. Irony and sarcasm detection was the topic of [15], [37] and [44]. In [21] and [45] the objective was text categorization. Detection of hate speech was the subject of [46] and [47], detection of Arabic health information was treated in [48], document classification was the aim in [16] and dialect identification was the goal of [34].

Several articles were interested by one or more step of ANLP, as detailed in the Table I.

Table 1: Overview Of The Reviewed Articles

Article	Title	Year	Step of focus
[11]	Leveraging Arabic sentiment classification using an enhanced CNN-LSTM approach and effective Arabic text preparation	2021	Data collection
[12]	Impact of Stemming and Word Embedding on Deep Learning-Based Arabic Text Categorization	2020	Data collection, stemming, word embedding
[13]	Pre-trained Word Embeddings for Arabic Aspect-Based Sentiment Analysis	2018	Data collection, word embedding
[14]	Multi-task Learning Using a Combination of Contextualised and	2021	Data collection, word

	Static Word Embeddings for Arabic Sarcasm Detection and Sentiment Analysis		embedding
[15]	SANA: Sentiment analysis on newspapers comments in Algeria	2019	Data collection,
[16]	A Semi-supervised Approach for Sentiment Analysis of Arab(ic+izi) Messages: Application to the Algerian Dialect	2021	Data collection, embedding
[18]	A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning	2018	Data annotation
[19]	Sentiment Analysis of Users on Social Networks: Overcoming the challenge of the Loose Usages of the Algerian Dialect	2018	Data annotation
[20]	Deep learning approaches for Arabic sentiment analysis	2019	Data annotation
[21]	ArCovidVac: Analyzing Arabic Tweets About COVID-19 Vaccination	2022	Data annotation
[22]	Sentiment Analysis of Arabic Tweets about Violence Against Women using Machine Learning	2021	Data annotation
[23]	Violence Detection over Online Social Networks: An Arabic Sentiment Analysis Approach	2021	Data annotation
[24]	ArSentD-LEV: A Multi-Topic Corpus for Target-based Sentiment Analysis in Arabic Levantine Tweets	2019	Data annotation
[11]	Leveraging Arabic sentiment classification using an enhanced CNN-LSTM approach and effective Arabic text preparation	2021	Data normalisation
[29]	Arabic Sentiment Analysis Using BERT Model	2021	Tokenisation
[8]	Preprocessing Arabic text on social media	2021	Tokenisation
[32]	Comparative study of Arabic stemming algorithms for topic identification	2019	Stemming, embedding
[34]	Deep Multi-Task Model for Sarcasm Detection and Sentiment Analysis in Arabic Language	2021	Tokenisation, embedding
[35]	Benchmarking Transformer-based Language Models for Arabic Sentiment and Sarcasm Detection	2021	Tokenisation, embedding
[30]	AraXLNet: pre-trained language model for sentiment analysis of Arabic	2022	Tokenisation
[31]	TCE at Qur'an QA 2022: Arabic Language Question Answering Over Holy Qur'an Using a Post-Processed Ensemble of BERT-based Models	2022	Tokenisation
[17]	Arabic fake news detection based on deep contextualized embedding models	2022	Data collection, tokenization, embedding
[33]	BPR algorithm: New broken plural rules for an Arabic stemmer	2022	Stemming

5. RESULTS DISCUSSION

The main objective of the study differs from one article to another, depending on the motivation of the researcher and the need expressed. In our review, we noted a multitude of objectives targeted by the articles. The objective or theme that has been recurrent is that of Arabic sentiment analysis (SA) with a percentage of 63%, followed by emotion detection with 11.1%, and Sarcasm and irony detection also with 11.1% as illustrated in the Figure 5.

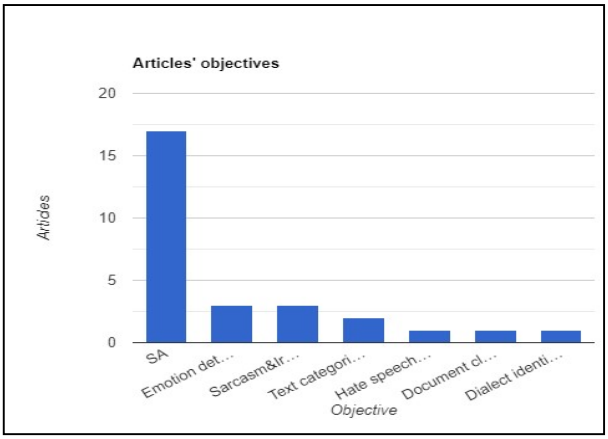


Figure 5: Distribution Of Articles' Objectives

The reviewed articles are generally focused on ASA (Arabic sentiment Analysis). While there were other objectives such as detecting emotions, sarcasm and hate speech, going through all the

articles' goals, we extracted the three main performed tasks, which are binary classification, ternary classification and multi-class classification.

The first point for any implementation and study in NLP is that of the data, its nature and type. In the case of the sample studied we find that Standard Arabic (MSA) was the most used in implementation with a percentage of 47.1% followed by the Egyptian dialect and Algerian with a percentage of 11.8% and 9.8% respectively. However the Moroccan dialect has less percentage in number of studies with only 1,8% as illustrated in the Figure 6.

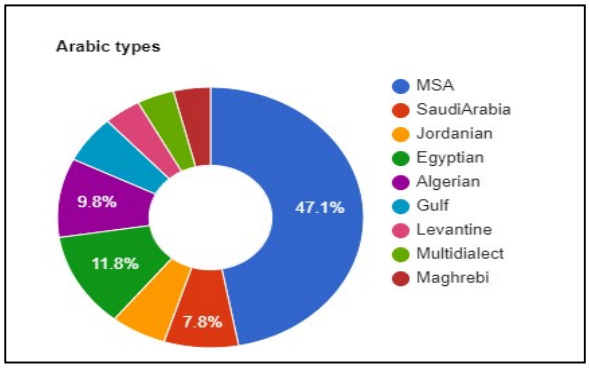


Figure 6: Distribution Of Arabic Types languages

As we can see, ANLP is a challenging technique due to the complexity of the Arabic language. These challenges can arise at various stages of the process. One key factor that can significantly affect the final performance is the pre-processing step, as the quality of the final output depends on the quality of the initial input. Therefore, investing more time and effort in pre-processing can lead to better performance in the end. While the pre-processing stage is an important factor that can influence the final performance of natural language processing for Arabic (ANLP), it is not the only step that will affect the outcome. Other factors can also have an impact.

6. CONCLUSION

NLP is one of the complex techniques since it has the objective of comprehending and understanding humans' written and/or spoken text. This technique deals with the different challenges of multiple languages, including Arabic, a language characterised by a set of rules and specificities that present challenges in the context of NLP. As we've mentioned in the article, the ANLP followed a process consisting of multiple steps, starting with the data collection and arriving at the task performed. We explored each step of the process and highlighted the newest and efficient techniques. Based on this research, we found that the field of ANLP still requires more interest to become mature. Moreover, we've clearly showed the importance of the pre-processing steps for a performant model. There are several tasks in natural language processing for Arabic (ANLP) that have not been thoroughly researched, including dialect identification, detecting hate speech, and detecting sarcasm and irony. Additionally, certain tasks have been applied more to some Arabic dialects than others, such as sentiment analysis in the Moroccan dialect. This literature review and classification of recent research in ANLP has allowed us to identify areas that still require more research. Our focus in the future is to investigate the application of these tasks to the Moroccan dialect and to help other researchers learn about current trends in ANLP and begin working in this field.

REFERENCES:

[1] V. B. P. DupaKuntla, H. VeeraBoina, M. V. Krishna Reddy, M. M. Satyanarayana, and Y. S. Sameer, 'Ijiert- Learning Based Approach for Hindi Text Sentiment Analysis Using', Int. J. Innov. Eng. Res. Technol. [IJIERT], vol. 7, no. 8, pp. 40–47, 2020.

[2] P. Shah, P. Swaminarayan, and M. Patel, 'Sentiment analysis on film review in Gujarati language using machine learning', Int. J. Electr. Comput. Eng., vol. 12, no. 1, pp. 1030–1039, 2022, doi: 10.11591/ijece.v12i1.pp1030-1039.

[3] M. Z. Asghar, A. Sattar, A. Khan, A. Ali, F. Masud Kundi, and S. Ahmad, 'Creating sentiment lexicon for sentiment analysis in Urdu: The case of a resource-poor language', Expert Syst., vol. 36, no. 3, pp. 1–19, 2019, doi: 10.1111/exsy.12397.

[4] L. Khan, A. Amjad, N. Ashraf, H. T. Chang, and A. Gelbukh, 'Urdu Sentiment Analysis with Deep Learning Methods', IEEE Access, vol. 9, pp. 97803–97812, 2021, doi: 10.1109/ACCESS.2021.3093078.

[5] B. Zhang and W. Zhou, 'Transformer-Encoder-GRU (T-E-GRU) for Chinese Sentiment Analysis on Chinese Comment Text', 2021, [Online]. Available: http://arxiv.org/abs/2108.00400.

[6] A. Lertpiya et al., 'A Preliminary Study on Fundamental Thai NLP Tasks for User-generated Web Content', 2018 Int. Jt. Symp. Artif. Intell. Nat. Lang. Process. iSAI-NLP 2018 - Proc., pp. 1–8, 2018, doi: 10.1109/iSAI-NLP.2018.8692946.

[7] P. Meesad, 'Thai Fake News Detection Based on Information Retrieval, Natural Language Processing and Machine Learning', SN Comput.

Sci., vol. 2, no. 6, pp. 1–17, 2021, doi: 10.1007/s42979-021-00775-6.
[8] M. O. Hegazi, Y. Al-Dossari, A. Al-Yahy, A. Al-Sumari, and A. Hilal, 'Preprocessing Arabic text on social media', Heliyon, vol. 7, no. 2, p. e06191, 2021, doi: 10.1016/j.heliyon.2021.e06191.
[9] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, 'Arabic natural language processing: An overview', J. King Saud Univ. - Comput. Inf. Sci., vol. 33, no. 5, pp. 497–507, 2021, doi: 10.1016/j.jksuci.2019.02.006.
[10] K. Darwish et al., 'A panoramic survey of natural language processing in the Arab world', Commun. ACM, vol. 64, no. 4, pp. 72–81, 2021, doi: 10.1145/3447735.
[11] A. M. Alayba and V. Palade, 'Leveraging Arabic sentiment classification using an enhanced CNN-LSTM approach and effective Arabic text preparation', J. King Saud Univ. - Comput. Inf. Sci., no. xxxx, 2021, doi: 10.1016/j.jksuci.2021.12.004.
[12] H. A. Almuzaini and A. M. Azmi, 'Impact of Stemming and Word Embedding on Deep Learning-Based Arabic Text Categorization', IEEE Access, vol. 8, pp. 127913–127928, 2020, doi: 10.1109/ACCESS.2020.3009217.
[13] M. A. Mohammed, S. Muazzam Ahmed, and N. Farrukh, 'Pre-trained Word Embeddings for Arabic Aspect-Based Sentiment Analysis of Airline Tweets', in Advances in Intelligent Systems and Computing 1058 Proceedings of the International Conference on Advanced Intelligent Systems and Informatics, 2020, no. September, pp. 7–9. [Online]. Available: http://link.springer.com/10.1007/978-3-030-31129-2.
[14] A. I. Alharbi and M. Lee, 'Multi-task Learning Using a Combination of Contextualised and Static Word Embeddings for {A}rabic Sarcasm Detection and Sentiment Analysis', Proc. Sixth Arab. Nat. Lang. Process. Work., pp. 318–322, 2021, [Online]. Available: https://aclanthology.org/2021.wanlp-1.39.
[15] H. Rahab, A. Zitouni, and M. Djoudi, 'SANA: Sentiment analysis on newspapers comments in Algeria', J. King Saud Univ. - Comput. Inf. Sci., vol. 33, no. 7, pp. 899–907, 2021, doi: 10.1016/j.jksuci.2019.04.012.
[16] I. Guellil et al., 'A Semi-supervised Approach for Sentiment Analysis of Arab(ic+izi) Messages: Application to the Algerian Dialect', SN Comput. Sci., vol. 2, no. 2, pp. 1–18, 2021, doi: 10.1007/s42979-021-00510-1.
[17] A. B. Nassif, A. Elnagar, O. Elgendy, and Y. Afadar, 'Arabic fake news detection based on deep contextualized embedding models', Neural Comput. Appl., vol. 4, 2022, doi: 10.1007/s00521-022-07206-4.
[18] B. Haidar, M. Chamoun, and A. Serhrouchni, 'A Multilingual System for Cyberbullying Detection : Arabic Content Detection using Machine Learning ISSN : A Multilingual System for Cyberbullying Detection : Arabic Content Detection using Machine Learning', no. December, 2017, doi: 10.25046/aj020634.
[19] A. Soumeur, M. Mokdadi, A. Guessoum, and A. Daoud, 'Sentiment Analysis of Users on Social Networks: Overcoming the challenge of the Loose Usages of the Algerian Dialect', Procedia Comput. Sci., vol. 142, pp. 26–37, 2018, doi: 10.1016/j.procs.2018.10.458.
[20] A. Mohammed and R. Kora, 'Deep learning approaches for Arabic sentiment analysis', Soc. Netw. Anal. Min., vol. 9, no. 1, pp. 1–12, 2019, doi: 10.1007/s13278-019-0596-4.
[21] H. Mubarak, S. Hassan, S. A. Chowdhury, and F. Alam, 'ArCovidVac: Analyzing Arabic Tweets About COVID-19 Vaccination', 2022.
[22] M. Zyout and N. Hassan, 'Sentiment Analysis of Arabic Tweets about Violence Against Women using Machine Learning', 2021, no. May.
[23] M. Khalafat, J. S. Alqatawna, R. Al-Sayyed, M. Eshtay, and T. Kobbaey, 'Violence Detection over Online Social Networks: An Arabic Sentiment Analysis Approach', Int. J. Interact. Mob. Technol., vol. 15, no. 14, pp. 90–110, 2021, doi: 10.3991/ijim.v15i14.23029.
[24] R. Baly, A. Khaddaj, H. Hajj, W. El-Hajj, and K. B. Shaban, 'ArSentD-LEV: A Multi-Topic Corpus for Target-based Sentiment Analysis in Arabic Levantine Tweets', no. September, 2019, [Online]. Available: http://arxiv.org/abs/1906.01830.
[25] A. C. Stubbs, 'A Methodology for Using Professional Knowledge in Corpus Annotation', 2013
[26] A. Pasha et al., 'MADAMIRA : A Fast , Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic', in The International Conference on Language Resources and Evaluation, 2014, pp. 1094–1101.
[27] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, 'Farasa: A fast and furious segmenter for arabic', NAACL-HLT 2016 - 2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Demonstr. Sess., vol. 2016, pp. 11–16, 2016, doi: 10.18653/v1/n16-3003.
[28] . Abuzayed and H. Al-Khalifa, 'Sarcasm and Sentiment Detection In Arabic Tweets Using BERT-based Models and Data Augmentation', Proc. Sixth Arab. Nat. Lang. Process. Work., pp. 312–317, 2021, [Online]. Available: https://www.aclweb.org/anthology/2021.wanlp-1.38.

[29] H. Chouikhi, H. Chniter, and F. Jarray, 'Arabic Sentiment Analysis Using BERT Model', Commun. Comput. Inf. Sci., vol. 1463, no. September, pp. 621–632, 2021, doi: 10.1007/978-3-030-88113-9_50.

[30] A. Alduailej and A. Alothaim, 'AraXLNet: pre-trained language model for sentiment analysis of Arabic', J. Big Data, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00625-z.

[31] M. ElKomy and A. M. Sarhan, 'TCE at Qur'an QA 2022: Arabic Language Question Answering Over Holy Qur'an Using a Post-Processed Ensemble of BERT-based Models', 2022, [Online]. Available: http://arxiv.org/abs/2206.01550.

[32] M. Naili, A. H. Chaibi, and H. H. Ben Ghezala, 'Comparative study of Arabic stemming algorithms for topic identification', Procedia Comput. Sci., vol. 159, pp. 794–802, 2019, doi: 10.1016/j.procs.2019.09.238.

[33] H. Alshalabi, S. Tiun, N. Omar, E. abdulwahab Anaam, and Y. Saif, 'BPR algorithm: New broken plural rules for an Arabic stemmer', Egypt. Informatics J., no. xxxx, 2022, doi: 10.1016/j.eij.2022.02.006.

[34] A. El Mahdaouy, A. El Mekki, K. Essefar, N. El Mamoun, I. Berrada, and A. Khoumsi, 'Deep Multi-Task Model for Sarcasm Detection and Sentiment Analysis in Arabic Language', 2021, [Online]. Available: http://arxiv.org/abs/2106.12488.

[35] I. A. Farha and W. Magdy, 'Benchmarking Transformer-based Language Models for Arabic Sentiment and Sarcasm Detection', Arab. Nat. Lang. Process. Work., pp. 21–31, 2021.

[36] H. El Moubtahij, H. Abdelali, and E. B. Tazi, 'AraBERT transformer model for Arabic comments and reviews analysis', IAES Int. J. Artif. Intell., vol. 11, no. 1, pp. 379–387, 2022, doi: 10.11591/ijai.v11.i1.pp379-387.

[37] I. Kaibi and E. H. Nfaoui, 'A Comparative Evaluation of Word Embeddings Techniques for Twitter Sentiment Analysis', 2019 Int. Conf. Wirel. Technol. Embed. Intell. Syst., pp. 1–4, 2019.

[38] A. H. Ombabi, W. Ouarda, and A. M. Alimi, 'Deep learning CNN – LSTM framework for Arabic sentiment analysis using textual information shared in social networks', Soc. Netw. Anal. Min., pp. 1–13, 2020, doi: 10.1007/s13278-020-00668-1.

[39] A. Alwehaibi, M. Bikdash, M. Albogmi, and K. Roy, 'A study of the performance of embedding methods for Arabic short-text sentiment analysis using deep learning approaches', J. King Saud Univ. - Comput. Inf. Sci., no. xxxx, 2021, doi: 10.1016/j.jksuci.2021.07.011.

[40] K. Ibrahim, N. El Habib, and Hassan Satori, 'Sentiment Analysis Approach Based on Combination of Word Embedding Techniques', 2019.

[41] L. Moudjari, F. Benamara, and K. Akli-Astouati, 'Multi-level embeddings for processing Arabic social media contents', Comput. Speech Lang., vol. 70, p. 101240, 2021, doi: 10.1016/j.csl.2021.101240.

[42] B. Naaima, E. Soumia, F. Rdouan, and O. H. T. Rachid, 'Exploring the Use of Word Embedding and Deep Learning in Arabic Sentiment Analysis', in Advances in Intelligent Systems and Computing, 2020, vol. 1105 AISC, pp. 149–156. doi: 10.1007/978-3-030-36674-2_16.

[43] R. A. Salama, A. Youssef, and A. Fahmy, 'Morphological Word Embedding for Arabic', Procedia Comput. Sci., vol. 142, pp. 83–93, 2018, doi: 10.1016/j.procs.2018.10.463.

[44] H. Chouikhi, H. Chniter, and F. Jarray, 'Stacking BERT based Models for Arabic Sentiment Analysis', no. January, pp. 144–150, 2021, doi: 10.5220/0010648400003064.

[45] M. Baali and N. Ghneim, 'Emotion analysis of Arabic tweets using deep learning approach', J. Big Data, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0252-x.

[46] A. Al-Hassan and H. Al-Dossari, 'Detection of hate speech in Arabic tweets using deep learning', Multimed. Syst., no. 0123456789, 2021, doi: 10.1007/s00530-020-00742-w.

[47] I. Aljarah et al., 'Intelligent detection of hate speech in Arabic social network: A machine learning approach', J. Inf. Sci., vol. 47, no. 4, pp. 483–501, 2021, doi: 10.1177/0165551520917651.

[48] H. Elfaik and E. H. Nfaoui, 'Combining Context-Aware Embeddings and an Attentional Deep Learning Model for Arabic Affect Analysis on Twitter', IEEE Access, vol. 9, pp. 111214–111230, 2021, doi: 10.1109/ACCESS.2021.3102087.