Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/170616

A Survey on High Dimensional Computing with Arabic Language Processing

George Samy Rady[1,*] *and Mamdouh Farouk Mohamed*[2]

[1]Department of Information Technology, Faculty of Computers and Information Technology, The Egyptian E-Learning University, Cairo, Egypt
[2]Department of Computer Science, Faculty of Computers and Information, Assiut University, Assiut, Egypt

Received: 1 Dec. 2022, Revised: 11 Feb. 2023, Accepted: 15 Feb. 2023
Published online: 1 Nov. 2023

Abstract: A growing amount of Arabic textual content comes accessible via the World Wide Web in properties and companies via Internet and Intranet services, there is a pressing need for technology and tools to handle the relevant data. In the context of Arabic processing of language, high-dimensional computation. The stemming method is intended to be used throughout document categorization to lower the high dimensionality of the domain of features, which improves categorization system performance, especially for highly modulating languages such as Arabic. It then digs into other sub-topics, such as methods for decreasing dimensionality, categorization computations, natural language processing jobs, and Arabic-specific applications. Each sub-topic is explored in depth, emphasising the main findings, techniques, and significant advances documented in the scientific literature. In addition, the survey report analyses future trends as well as prevalent obstacles in high-dimensional computing with Arabic language processing. It investigates the effectiveness and relevance of existing procedures while also exposing gaps and limits that need to be addressed further. The report provides knowledge of the present situation of the topic and suggests new directions for future studies by critically analysing the examined literature. This survey report is significant because of its ability to integrate and synthesise current knowledge, offering researchers as well as practitioners with a complete resource on the combination of high-dimensional computing with Arabic processing of languages. The paper may be utilised as an easy way to find pertinent works that cover a given issue for a particular dialect.

Keywords: High dimensionality, Arabic language processing, Natural Language processing, Arabic language application.

1 Introduction

The computational processing and evaluation of data that exists in high-dimensional areas is referred to as high-dimensional computation. It entails working with and altering data that contains a large number of aspects or characteristics. Arabic language processing, on the contrary hand, is concerned with computational examination and comprehension of Arabic. The application of advanced computational techniques to process, analyse, and comprehend Arabic text and speech data in high-dimensional spaces is referred to as Arabic language processing in high-dimensional computation. This field includes a wide range of tasks, including Arabic text categorization, sentiment analysis, information retrieval, machine translation, speech recognition, and many more. High-dimensional applications of computers in Arabic language processing are numerous and significant. They can be utilised in a variety of applications, including retrieval of data, social media evaluation, sentiment assessment, automated translation, personal assistants, and a lot more. These technologies enhance the availability of Arabic content, enhance interaction, and facilitate decision-making in a variety of fields. In essence, high-dimensional computing with Arabic language processing entails using computer approaches to analyse and comprehend Arabic text and speech data stored in high-dimensional spaces. It is a vibrant and fast expanding sector with tremendous potential for breakthroughs in a variety of areas, all of which contribute to the growth of Arabic language technologies and their incorporation into the daily lives.Arabic is classified into three categories: traditional Arabic, contemporary basic Arabic, and dialect Arabic.

[*] Corresponding author e-mail: gsami@eelu.edu.eg

The Arabic language takes these forms based on three important parameters: morphology, grammar, and lexical mixing. Classical Arabic is used mostly in Arabic-speaking economies, rather than in the diaspora. Traditional Arabic can be discovered in religious texts like the Sunnah and Hadith, as well as several historical documents. Diacritic signs (also known as ”Tashkil” or short consonants) are often employed as phonetic cues to show proper pronunciation in Ancient Arab. Diacritics, on the opposite hand, are regarded optional in the majority of Arabic language [1]. Modern Standard Arabic is used on television, in media outlets, in poems, and in literature. The alphabet, spelling, and lexicon of the written Arabic script have not changed in a minimum of four millennia. Almost no surviving language can make such a claim. Arabs use dialect Arabic or ”colloquial Arabic” on a regular basis. It can be encountered in many different countries and regions. Syro Palestanian Arabic, Arabian Peninsula Arabic, Mesopotamian Arabic, Egyptian Arabic, and Maghrebi Arabic are the dialects. Arabic language is commonly utilised, usually written, by web surfers and social networking users; dialect Arabic differs by location. MSA is used to acquire sections of words in conversational Arab [2].

The volume of these opinions, as well as their readily available and availability, gave rise to computer programmes that use sentiment assessment (opinion analysis) as a key aspect in forecasting stocks, assessing goods, public polling, and others. Nevertheless, because to the intricate nature of the language, computerised sentiment assessment is still far from delivering output of equal quality to humans. In addition, the Arabic language complicates computerised sentiment assessment due to its grammatical complexity, ambiguity, and high number of dialectal variances. These difficulties add to the complexity of the necessary natural language processing (NLP). In the past ten years, there has been a surge in fascination with processing natural languages for dialectic Arab. This expansion can be attributed to a variety of variables, such as the widespread use of Arabic dialects in social media. The subjects addressed by computer linguistic for dialects of Arabic span from basic characteristics of language including grammar to complex methods such as automated translation [3].

The world’s dialects could have universal traits on a deep, theoretical straight, yet the frameworks seen in real-world, surface-level writings might vary dramatically. This cross-lingual variance has hampered the growth of strong, multilingually relevant Natural Language Processing (NLP) the internet, and as a result, existing NLP is still mostly limited to a handful of resource-rich language. Furthermore, most advanced algorithms for machine learning rely on monitoring from (large volumes of) labelled data—a criterion that almost all of the languages spoken worldwide cannot meet [4]. Rady et al [5] suggested that the expressed ideas and comments comprise a great information mine all around the world. Indeed, businesses, governments, and institutions are becoming more involved in mining big data to better their offerings. As a result, sentiment analysis is a powerful technique for exploring such raw huge data and extracting useful information. Indeed, a lot of researchers have been drawn to this topic, and numerous research organisations are attempting to better understand the subject and the tools available. One of the goals of SA is to assess the generally social media. In particular, Sentimental Analysis seeks to classify views and remarks as positive neutral and negative. This type of analysis allows businesses and institutions to gain valuable knowledge in order to improve their strategy and merchandise or better comprehend tendencies and behaviour. Opinion analysis, emotion formation (e.g., joyful, sad, and angry), and mood and emotion detection towards written or auditory texts are critical components of AI technologies and NLP. The grammar of the language of Arabic is distinct in footings of textual and phonetic structure. Each word contains either implied or explicit either favourable or adverse implications. There are several sorts of writing in Arabic, including prose, writing, gratitude, critiques, and furthermore. All of these arts involve a large number of semantic frameworks that provide value on a visible or implicit basis. Thus, somebody may identify the writing approach of any writer of a book by examining the importance of the words and feelings of positive or negative aspects of disdain, appreciation, respect, sadness, love-hate, and so on [6].

2 High dimensional computing with Arabic language processing

The purpose of this study is to undertake a thorough literature review in the subject of high dimensional computing in Arabic language processing. Understanding and analysing the techniques and methods utilised in this discipline, such as artificial intelligence, machine learning, and big data analysis, is fundamental to the research. Academic sources and prior research on the difficulties encountered in high-dimensional computing for Arabic Language Processing will be reviewed. The efficacy of text analysis, classification, translation, and information extraction procedures will also be evaluated. This literature review will shed light on previous research, emphasising improvements and weaknesses in high dimensional computing for Arabic language processing. The results of this poll will help to conduct additional studies in this rapidly changing subject.

Based on the given percentages, the Fig.1 represents the growth or adoption rate of high dimensional computing in Arabic language processing over a span of seven years from 2016 to 2022. Each percentage represents the proportion or relative increase in the utilization or application of high dimensional computing techniques specifically in the context of Arabic language

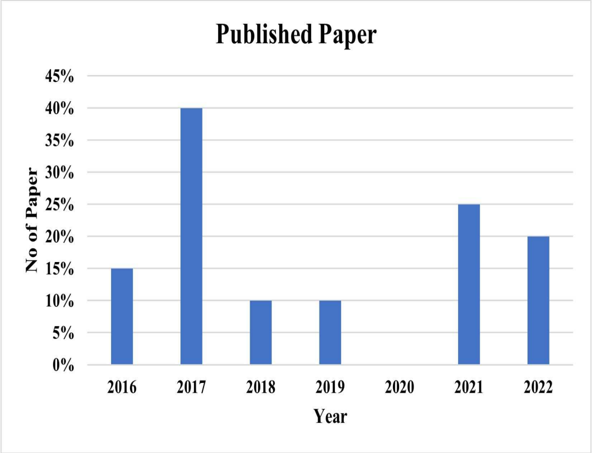

Fig. 1: Published paper (2016-2022).

processing. In 2016, the percentage was 15%, indicating a relatively low level of adoption or awareness of high dimensional computing methods in Arabic language processing. However, in the following year, 2017, there was a significant jump to 40%, suggesting a substantial increase in the utilization of high dimensional computing techniques in this domain. This growth may be attributed to the recognition of the potential benefits and advancements brought by high dimensional computing for Arabic language processing tasks. In 2018, the percentage dropped to 10%, indicating a slight decrease in the rate of adoption or the use of high dimensional computing methods compared to the previous year. This could be due to various factors such as challenges in implementing these techniques, limited availability of resources, or the emergence of alternative approaches. The trend continued in 2019, with another 10% adoption rate, indicating a relatively stable or stagnant growth in the utilization of high dimensional computing in Arabic language processing. However, in 2020, the percentage dropped to 0%, suggesting a pause or decline in the adoption of these methods. This could be due to various reasons such as shifting research focus, emerging alternative approaches, or limited progress in high dimensional computing techniques specific to Arabic language processing during that year. In 2021, the graph shows a resurgence with an adoption rate of 25%, indicating a significant increase in the utilization of high dimensional computing techniques in Arabic language processing. This growth may be attributed to advancements in the field, increased awareness, or the development of new algorithms and models tailored to Arabic language processing. Finally, in 2022, the percentage returns to 10%, indicating a stabilization or slight decrease in the growth rate of high dimensional computing in this domain compared to the previous year. Overall, the graph demonstrates the varying rates of adoption and growth of high dimensional computing in Arabic language processing over the years, reflecting the evolving landscape of research and application of computational methods in this field.

2.1 Artificial Intelligence

In the research proposed by Ali et l. [7,8] The Arabic datasets were analysed using AI and natural language processing in conversion tools available online. It is amusing that just 1% of online material is in Arabic, despite the fact that 5% of the population of the globe understands Arabic. This relates to the disproportion prevalence of Arabic language on-line material in comparison to other dialects, which could be related to a variety of factors, including a dearth of Arabic language specialists. This study aims to examine the impact of Machine Translation (MT) technology and Translation Memory (TM) applications commonly used in Arab society for educational and commercial purposes. The objective is to determine the feasibility of transitioning from Arabic Personalization to Arabic Globalization, thereby facilitating the use of Natural Language Processing (NLP) methods for human-computer interfaces. The research will focus on analyzing selected machine translation programs to evaluate their content and applications, aiming to assess their ability to maintain the essence of the original language without the need for human supervision. Ambient Intelligence (AmI), with its intelligent components, possesses capabilities in media management and computational intelligence, areas where existing MT software often falls short. By conducting further studies, AmI may uncover potential solutions to the challenges presented in this research.

In research proposed by Muaad et al. [9,10] Misogyny and Sarcasm Identification in Arabic Texts Using Artificial Intelligence is developed. Social media connectivity is a popular subject in everyday life, especially right now. Multiple research projects have looked into the effect of remarks. Facebook, Twitter, and Instagram are just a handful of the social media platforms that are utilised for spreading various news stories around the globe on a daily basis. Detailed AI-based research is described in this work for automatically identifying misogyny and sarcasm in Arabic text in ternary and several classes contexts. The proposed AI approach focuses on the ability to detect instances of misogyny and sarcasm in Arabic text found on social media platforms. To achieve this, a comprehensive evaluation is conducted using seven advanced NLP classifiers: Arabic RoBERTa, Logistic Regression Classifier, Linear Support Vector Classifier, Random Forest Classifier, Passive Aggressive Classifier, Decision Tree Classifier, and K-Nearest Neighbor Classifier. These classifiers are applied to fine-tune, demonstrate, and assess their effectiveness in identifying both misogyny and sarcasm. Two Arabic Twitter datasets, specifically the misogyny dataset and the Abu Farah dataset, are employed for this purpose. The evaluation considers two possible scenarios: binary classification and multi-class classification for each investigation, focusing on either misogyny or sarcasm. The study does not find a correlation between various themes and the challenge posed by mixed-language content.

In the research proposed by Muaad, Al-antari, et al. [11,12] give a note on DL based ArCAR network for Arabic text detection with character-level visualisation. The method of classifying Arabic contents is known as AI-based text categorization. Traditional machine learning algorithms encounter significant hurdles when the volume of Arabic text in social contexts grows due to how complicated of word patterns and the various nature of the Arabic language. To address this, this research paper introduces an algorithm based on deep convolutional neural networks (CNN) that enables representation and recognition of Arabic text at the character level. The proposed system is validated through five-fold cross-validation testing for the categorization of Arabic text documents. The researchers evaluated Arabic text using this approach and found that the ArCAR system effectively classifies Arabic text on an individual character level. In fact, the ArCAR system achieved the highest accuracy of 97.76% in document categorization when using the AlKhaleej-balance dataset. The proposed ArCAR system emerges as a practical solution for accurate participation in authentic Arabic text, for comprehension and as a classification engine. However, it is important to note that the study did not address challenges related to multi-label classification of texts and Arabic data enhancement, which remain as potential limitations.

In the research proposed by Mohammad, Alwada‘n, et al. [13,14] present a detail note on a Naive Bayes, Neural Network and Support Vector Machine Classification of Arabic Text. Text classification is an important part of information extraction. Text algorithms for categorization are used to group articles into predefined classes. There are several tactics and procedures for categorising information, and many studies on English text categorization have been undertaken. However, a few investigations on Arabic text classification have been conducted. The study goes over three well-known data method classifications. Each of these well-known techniques is applied to an Arabic data source. A comparison of these three approaches is made. Furthermore, this study used a fixed quantity of sheets for every single record kind during the stages of testing and training. The results show that the Support Vector Machine produces the best results. The authors feel that if feature extraction and selection in pre-processing steps are enhanced, MLP-NN could generate superior outcomes, as the 600-input layer offers favourable outcomes. The disadvantage of the research is that it does not include alternate extraction and choice of features methodologies with MLPNN, as well as comparison studies on feature extract and choice.

In the research proposed by Aljedani et al. [15,16] define the machine learning-based hierarchical multi-label Arabic text categorization system. The multi-label categorization method assigns many labels to each item at the same time. Many real-world problems with categorization use high-dimensional label fields that can easily be organised in a hierarchy. Every situation in this sort of issue may correspond to numerous labels, and tags are organised in a hierarchical framework. It is a more difficult task than flat classification since the approach to categorization has to adjust for hierarchical connections among labels and forecast multiple tags for identical occurrence. Few works have looked into multi-label text classification in Arabic. The majority of these research have mostly concentrated on flat categorization while ignoring the concept of hierarchy. As a result, this work investigates progressive multi-label categorization in the larger context of Arabic. The influence of feature assortment strategies and feature set dimensions on classification performance is also studied. Furthermore, to enhance the hierarchical categorization, the Hierarchy of Multilabel Classifier (HOMER) algorithm is optimised by examining various types of multi-label categorization methods, clustering techniques, and clustering ratios. Furthermore, this study adds to previous studies by establishing and making accessible to the public a hierarchical multi-label Arabic database in a suitable format for categorization according to hierarchy. The findings show that the suggested model beats every other model included in the trials in terms of computing expenses while consuming less time (2 h) than other assessed systems. The primary drawback of this study is the use of various organized methods for determining the total number of clusters in the clustering algorithm and Fast Text as text depiction techniques, as well as the use of distinct stemming computations and the preparation of a list of Arabic human names for simple elimination throughout the stop word removal phase.

2.2 Machine learning and Deep learning

Taha et al. [17,18] proposed a method on describing the multi-label classifier using the binary relevance (br) approach for Arabic data. In the last few years, multi-label text categorization has grown in popularity, with each document receiving numerous labels at the same time. The enormous space of all conceivable label packages, which is exponentially to the number of potential tags, makes multi-label text categorization a particularly difficult problem. One of the drawbacks of older multi-label classification techniques is that they frequently fail to keep up with the number of particular tags and instructional samples. A substantial amount of processing effort is necessary for categorising a huge number of written materials with high dimension, particularly the dialect of Arabic, which has a very complicated morphological and is rich in character. The existing research on multi-label categorization of Arabic text has been limited, prompting the need for a new approach in this study. The research aims to design and develop a novel method for multi-label text categorization in Arabic texts using a binary relevance approach. This binary relevance is generated through an additional set of machine learning models. Four different multi-label categorization techniques are experimentally evaluated: SVM classifiers, KNN classifiers, NB classifiers, and various types of predictors. In addition, three feature selection procedures are investigated in order to improve the performance of Arabic multi-label text classification. The goal is to combine categorising methods and methods for selecting features to build a more effective multi-label grouping strategy. The method is evaluated using completely standardised data that is processed. The result demonstrate that machine learning algorithms employing binary relevance, consisting of a diverse set of classifiers, yield the best outcomes. The approach achieves a high overall F-measure of 86.8% for multi-label classification of Arabic text. Moreover, the findings indicate that the selected feature selection strategies significantly impact the categorization performance. Specifically, the use of a diverse set of algorithms proves to be an effective and feasible solution for Arabic multi-label text categorization. However, it is important to note that the research faced limitations due to the small amount of available data and the absence of well-annotated datasets, which is considered a significant drawback.

Alhaj et al. [19,20] proposed an method on the Influence of Stemming Techniques on Arabic Document Identification. Stemming is a powerful technique that is commonly used in information retrieval, , natural language processing machine learning, document classification (DC), and machine translation. Its primary goal in document categorization is to lower the dimension of the feature space, resulting in enhanced efficiency in the classification system, particularly when dealing with highly variable dialects like Arabic. The purpose of this study is to investigate the impact of three stemming techniques on Arabic Document Classification (ISRI, Tashaphyne, and ARLStem). The categorization methods used in the study are NB, SVM, and KNN. Furthermore, the results clearly show that SVM outperforms KNN and NB classifiers, with a Micro-F1 value of 94.64% when utilising the ARLStem stemmer. A significant disadvantage of this work is the lack of emphasis on developing an Arabic stemmer that overcomes the inadequacies that accompany current stemming techniques.

Suleiman et al. [21,22] suggested an approach on identifying the plagiarism in Arabic text sing deep learning technique. Plagiarism detection is critical, particularly for academics, researchers, and students. Despite the fact that there are numerous identifications of plagiarism programmes available, the process remains difficult due to the vast volume of internet publications. In this study, we propose using the word2vec method to identify semantic correspondence among words in Arabic, which may be helpful in the detection of plagiarism. Word2vec is an approach to deep learning that employs high accuracy for expressing words as vector patterns. The level of the vector illustration is determined by the corpus utilised in the training stage. Researcher utilised the OSAC corpus to train the word2vec technique in this research. Furthermore, the cosine similarity measure is employed for determining the resemblance across the vectors of words. The similarity evaluates indicate how simple modifications to text, such as altering just a single word or altering the location of verbs and adjectives, result in a level of similarity of 99%, allowing for the detection of plagiarism even if the test changes by replacing phrases with substitutes or changing the word order. If the modifications are limited to single words, the suggested method can detect text similarity.

Lulu et al. [23,24] proposed a research on deep learning algorithm for automatic Arabic dialect identification. This covers a variety of Arabic dialects, that constitute the native dialects of people who speak it. The availability of text Arabic dialectal varieties on the Web has created numerous new possibilities and challenges for machine learning and Arabic word processing. The detection of this type of unstructured information has a significant impact on a variety of programmes, including sentiment evaluation and automated translation. Nevertheless, due to the unique characteristics of dialectal textual information, normal NLP methods built for traditional data fall short. Techniques involving deep learning are known to be quite effective at parsing dialectal text on social media. In this research, researcher look at a number of deep learning models for automatically classifying Arabic dialectal literature. Research looks at the most common languages in the sample. Egyptian (EGP), Levantine (LEV), and Gulf (GLF). Four distinct models based on deep neural networks were used to investigate the Arabic dialectal classification issue for every combination of the three languages (binary classification tests), in addition to one ternary-classification investigation that included all dialects. The outcomes reveal that the simulations function differently but promisingly for each pair of languages. Furthermore, we conducted a detailed assessment of the manually-annotated AOC the data set and research conclude that there's a significant requirement for a complete refining and evaluation of the AOC annotated clauses, as it is a vital reference dataset in the area.

The mentioned research works address a wide range of problems including Arabic language processing and machine learning approaches. The research focuses on the impact of machine translation and translation memory tools on Arabic language processing, with the goal of making natural language processing approaches for human-computer interactions more accessible. Using multiple NLP classifiers, an AI-based strategy is used to detect misogyny and sarcasm in Arabic texts on social media networks. ArCAR is a deep learning-based method for character-level visualisation and detection of Arabic text, with a focus on document categorization. When the performance of Naive Bayes, Neural Network, and Support Vector Machine classifiers for Arabic text classification is compared, it is discovered that the Support Vector Machine classifier produces the best results. They investigate the impact of feature selection strategies and dimensions on classification performance in the recursive multi-label categorization of Arabic text using machine learning techniques. Using multiple machine learning models and feature selection methodologies, the researchers present a binary relevance strategy for multi-label text categorization of Arabic data. ARLStem outperforms other stemmers in research on the impact of stemming techniques on Arabic document identification. They create a deep learning-based language generation system for Arabic telemedicine advice, concentrating on next-word prediction. Research proposes using word2vec and cosine similarity to identify plagiarism in Arabic text, whereas Lulu et al. [23,24] investigate deep learning techniques for automatic Arabic dialect recognition. Overall, these research advance the field of Arabic language processing by demonstrating the use of machine learning approaches in numerous areas of Arabic text analysis.

2.3 Natural Language Processing and Sentimental analysis

Larabi Marie-Sainte et al. [25,26] proposed a method on Natural Language Processing in Arabic and Machine Learning-Based Systems. The development of methods and instruments that may utilise and analyse the Arabic language in both oral and written settings is known as Arabic natural language processing (ANLP). ANLP contributes significantly to numerous existing technologies. It offers helpful and accessible instruments for Arabic and other languages speakers to utilise in a variety of sectors. Machine learning (ML) performances are utilized in the development of modern ANLP technologies. Given their exceptionally high precision rate irrespective of the sturdiness of the data utilised and the simplicity that they may be applied, machine learning (ML) techniques are frequently employed in Natural Language Processing. The technique of ANLP systems using ML, on the opposite side, involves numerous discrete phases. As a result, it is critical to recognise and comprehend those stages, as well as the most extensively utilised ML algorithms. The article addresses this topic in depth, demonstrates the application of ML approaches in the development of such tools, and highlights established methods employed in ANLP. Furthermore, this survey covers the features and complexities of the Arabic language, as well as the significance and demands of ANLP. More methods that use ANLP-based machine learning techniques with semi-supervised and unsupervised methods is not covered in the study which is considered as the main drawback of the research.

Özçift et al. [27,28] proposed a method on using bidirectional encoder reconstructions from transformer (BERT) to advance natural language processing (NLP) uses for morphologically rich languages. Language modelling pre-training designs have proven beneficial for learning language concepts. With fine-tuning, bidirectional encoder reconstructions from transformers (BERT), a recent deep bidirectional paying attention representation from unlabelled text, has produced outstanding outcomes in several natural language processing (NLP) applications. In this study, we wish to show how effective BERT is for a morphological rich dialect like Turkish. Historically, substantial linguistic pre-processing processes have been needed for morphologically challenging dialects to be able to prepare the data for machine learning (ML) methods. To create an effective data structure to address limited data or high-dimension difficulties, operations such as tokenization, the lemmatization or a stemming and attribute engineering are required. In this setting, researcher choose five different Turkish NLP research topics from the writing: sentiment analysis, cyberbullying proof of identity, text categorization, recognising emotions, and identifying spam. The empirical efficacy of BERT was then contrasted with that of the basic ML algorithms. Finally, researcher discovered improved outcomes in the specified NLP issues in contrast to base ML systems while reducing expensive pre-processing chores. As a future direction, researcher plan to use lighter versions of BERT-like architectures to decrease training time periods while keeping prediction performances high.

Soliman et al. [29,30] proposed a research on a collection of Arabic word embedding models for application in Arabic Natural Language Processing (NLP). Recent advancements in neural networks have had a significant impact on fields like artificial intelligence (AI), speech recognition, and natural language processing (NLP). Word embeddings, a technique that represents words as matrices in a continuous space while preserving their grammatical and semantic relationships, have emerged as a major development in NLP. AraVec is an open-source initiative that focuses on pre-trained word embeddings specifically tailored for Arabic NLP research, providing an affordable and effective solution. The initial release of AraVec includes six distinct models for word embeddings trained on three major Arabic text domains: web pages, tweets, and Wikipedia Arabic articles. These models are created using over 3.3 billion tokens, ensuring comprehensive coverage. This paper discusses the tools and data cleaning strategies employed in developing the models, the preprocessing steps performed, and the specific approaches used for word embedding generation. The availability of these models allows other NLP researchers to utilize them for various NLP tasks and enhance performance. In the future, the researcher aims to explore character-level embeddings and leverage these models to tackle additional challenges such as Arabic sentiment analysis and named entity recognition.

Soufan et al. [31,32] proposed a method on deep learning for Arabic text analysis of sentiment. It was also working effectively in the NLP sector due to the abundance of online written content, such as social media platforms and review sites, which have grown in popularity and profitability over the past decade. One of the most widely used Natural Language Processing (NLP) tools is sentiment analysis. Numerous researchers have done remarkable work in the subject of Sentiment Analysis for English. But because to the complexities of Arabic syntax and spelling, the amount of work on Sentiment Analysis for Arabic is fairly limited. Although English, Arabic has several languages, making Sentiment Analysis for Arabic extremely challenging and complicated, particularly when utilising data acquired from social media platforms, which have been shown to be unorganised and loud. The majority of work on sentiment evaluation in Arabic has focused on vocabulary and basic algorithms for machine learning. Furthermore, due to the scarcity of data sets with annotations for Arabic, most of the research was done on small amounts of data. This work provides cutting-edge research for Sentiment Analysis of Arabic Microblogging based on novel methodologies and an advanced Arabic text data preparation pipeline.

Altowayan et al. [33,34] suggested a method on sentiment-specific embeddings improve Arabic analysis of sentences. Many natural language processing activities rely on the quality of word representations to produce good results. To create effective word matrices for Arabic sentiment analysis, several things must be considered, including corpus selection, text pre-processing, and training selection of models. In this research, we show how to improve sentiment analysis outcomes by creating sentiment-specific embedded data learned with the unsupervised fast Text tool for both CBOW and skip-gram systems. The findings suggest that fast Text embedding systems are a good alternative for recovering semantic and syntactic data from standard Arabic in addition to a dialectal language. In fact, fast Text is more applicable to semantically rich dialects such as Arabic. To test the usefulness of the obtained word embeddings, we constructed polarity sentiment classifiers and contrasted how they performed on three distinct datasets to two comparable algorithms described in the literature. The performance assessment results indicate that our approach surpasses counterparts on appropriate data sets by up to 5% using F1-score.

Omara et al. [35,36] suggested a method on using recurrent networks to analyse Arabic sentiment. Deep learning techniques are distinguished by their capacity to learn distinguishing and discriminating features. These methods can uncover complicated relationships and patterns in high-dimensional information. Machine learning algorithms use multiple layers of nonlinear neural processing units to extract features. Natural Language Processing (NLP) is one of the domains that has used deep architectures with a noteworthy advance in measurement of performance. Because of their efficiency for processing sequential input, recurrent neural networks (RNNs) and their variants, such as Gated Recurrent Unit (GRU) and Long-Short Term Memory (LSTM), are widely utilised in NLP applications. LSTMs and GRUs can be especially beneficial because they can deal with the problems of vanishing and exploding gradients that RNNs frequently face. Additionally, Convolutional Neural Networks (CNNs) are popular deep architectures employed in language processing tasks. Sentiment analysis (SA) is a specific NLP task that focuses on analyzing viewpoints, beliefs, emotions, and feelings expressed in text. SA aims to determine the author's attitude towards a particular topic and classify the sentiment polarity into predefined categories. Employing SA in business analytics helps gain insights into consumer behavior and demands. For Arabic sentiment analysis, the suggested study employs deep LSTM, GRU, and CNN models. The aforementioned models are built and validated using character-level representations. In addition, the study investigates deep mixed models that mix various levels of CNN with LSTM or GRU. The goal is to look into the capabilities of deep GRU, LSTM, and hybrid architectures for learning and extracting features from character-level approximations. The results show that combining various designs can increase sentiment analysis job efficiency. In terms of precision, CNNLSTM/GRU combos beat deep LSTM and GRU algorithms.

Alayba et al.[37,38] proposed a method on using word encoding to improve Arabic Sentiment Analysis. Because of the intricacies of the Arabic language in morphological orthography, and languages, sentiment analysis in Arabic can be particularly difficult. Furthermore, extracting text features from brief messages like tweets to measure emotion makes this work considerably more complex. Deep neural networks have been widely used in recent years and have produced excellent outcomes in sentiment detection and natural language processing applications. Word integration, also known as word distribution, is a contemporary and strong technology for capturing the closest terms from a surrounding language. In this research, they discuss how they build Word2Vec classifiers from a huge Arabic corpus acquired from ten newspapers from various Arab nations. We show increased sentiment categorization accuracy (91%-95%) using several machine learning methods and convolutional neural network (CNN) models with varied text component choices on the publically available Arabic language health sentiment database.

The processing and analysis of data sets with a large number of dimensions or attributes is referred to as high-dimensional computing. Because of the rising complexity and size of data in numerous disciplines such as genomics, social networks, and image processing, it is a field that has acquired substantial attention and relevance in recent years. Traditional methods and procedures may become inefficient or unfeasible when dealing with high-dimensional data in terms of storage, computation, and interpretation. As a result, high-dimensional computing focuses on the development of novel algorithms, data structures, and computing environments capable of handling and extracting useful knowledge from high-dimensional data. This discipline includes dimensionality reduction, choosing features, sparse representation, and scalable methods, with the goal of addressing the curse of dimensionality and enabling efficient and accurate analysis of complex data sets. Researchers and practitioners hope to unleash the potential of big data and draw important insights and knowledge from high-dimensional information through enhancing high-dimensional computing. The summary table of total data is shown in Table 1.

Table 1: Summary on High dimensional Computing.

Reference	Title	Method	Objective	Drawback
[7,8]	Arabic corpora in software for translation on the web	AI	Impact on Machine translation and TM tools is stated clearly	Fall short on both media management and Computational intelligence
[9,10]	Misogyny and Sarcasm Identification in Arabic Texts Using Artificial Intelligence	AI	Best accuracy rate is achieved	Fails to identify a link between numerous themes and the mixed language difficulty
[11,12]	DL based ArCAR network for Arabic text detection	DL based ArCAR	Transmits Arabic text at the element level in order to avoid preliminary processing constraints such as stemming.	Difficulties of multi-label classification of texts and Arabic information enhancement were not addressed.
[13,14]	Classification of Arabic Text	SVM, Naïve Bayes and NN	The outcome demonstrates that the Support Vector Machine produces the best outcomes	Minor variation, writers concluded that the extraction and selection of features are crucial and impact outcomes in a variety of ways. Alternative feature extraction as well as selection approaches with MLPNN
[15,16]	Hierarchical multi-label Arabic text categorization system	Hierarchy of Multilabel Classifier	Model beats every other model included in the trials in terms of computing expenses while consuming less time (2 h) than other assessed systems	The use of multiple organised techniques for determining the number of clusters in the clustering algorithm Fast Text as text depiction techniques, by the use of distinct stemming computations Arabic human names for simple elimination throughout the stop word removal phase with each other.
[17,18]	Multi-label classifier using the binary relevance (br) approach for Arabic data	Multi-label classifier	The binary relevance (BR) classifier set is built from the findings of individual binary classification algorithms on every field or category.	Research used just a small amount of data One of the research's limitations is the absence of well-annotated data sets.
[19,20]	The Influence of Stemming Techniques on Arabic Document Identification	SVM, Naïve Bayes and KNN	When using stemming strategies, the results improved significantly and produced an important rise in the accuracy of classification. The ARLStem stemmer outscored other stemmers when utilising the SVM classifier to produce the best results. As a result, the SVM classifier outperforms other types of classifiers since it can deal with data with high dimensions.	Study could concentrate on constructing an Arabic stemmer that tries to address the drawbacks of current stemming approaches

3 Conclusion

The survey on high-dimensional computing with Arabic language processing emphasises the increasing importance and problems of dealing with large-scale data in the context of Arabic language processing. The research delves into the advances and techniques used in high-dimensional computing, such as artificial intelligence, NLP, and deep learning, to address the intricacies of Arabic literature. It emphasises the importance of effective data format, feature extraction, and classification approaches in order to improve the performance of Arabic language processing jobs. Furthermore, the survey recognises the limitations and constraints encountered in this subject, such as the scarcity of labelled data, the complexity of Arabic language structure, and the lack of comprehensive tools and resources. Despite these challenges, the survey highlights the potential of high dimensional computing to improve numerous Arabic language applications such as machine translation, sentiment analysis, text categorization, and plagiarism detection. Future research should concentrate on overcoming the mentioned restrictions, increasing the availability of annotated datasets, and building more robust and accurate algorithms to progress high-dimensional computing in Arabic language processing.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

[21,22]	Plagiarism detection using Deep learning	Deep learning	99% of plagiarism is detected.	If the modifications are limited to single words, the suggested method can detect text similarity.
[23,24]	Deep Learning Algorithm for Automatic Arabic Dialect Identification	Deep learning	Average the model's efficiency for every combination of languages	The model's capacity to generalise and reliably classify dialects can be hampered by a lack of diverse and well-annotated data.
[25,26]	ANLP and ML system	Machine Learning	Different ANLP applications and their importance in such systems based on machine learning are clearly described.	Fail to cover the semi-supervised and unsupervised learning.
[27,28]	Bidirectional encoder reconstructions from transformer (BERT) to advance natural language processing (NLP)	NLP	Improved outcomes in the specified NLP issues in contrast to base ML systems while reducing expensive pre-processing chores	Lighter versions of BERT-like architectures to decrease training time periods can be used.
[29,30]	A collection of Arabic Word Embedding Models for application in Arabic Natural Language Processing (NLP).	NLP	Employ models that have been trained to improve the efficiency on different NLP tasks.	Limited in the character level embeddings and use these models to improve several of the already discussed challenges, such as Arabic sentiment analysis and recognition of named entities.
[31,32]	Deep learning for Arabic text analysis of sentiment	Sentimental analysis	Better accuracy obtained	Limited amount of data has been used.
[33,34]	Sentiment-specific embeddings improve Arabic analysis of sentences.	Sentimental Analysis	Better performance and the error rate is reduced.	Limited data is used.
[35,36]	Recurrent networks to analyse Arabic sentiment	RNN	95% of accuracy is obtained.	Hybrid method can be utilized to get better performance
[37,38]	Encoding to improve Arabic Sentiment Analysis	Machine Learning, CNN	91% to 95% of accuracy	Failed to deal with the negotiation of words.
[39,40]	A conceptual approach on NLP based Arabic metaphor	NLP	The important area such as cognitivists and computational system	This subjectivity makes creating a uniform and objective framework for detecting and analysing metaphors in Arabic language difficult. Individuals' interpretations and comprehension of metaphors may differ, which can lead to inconsistencies and challenges in developing a standardised computational model.
[41,42]	Arabic text grouping with better algorithms for clustering that reduce dimensionality.	SVM	Better accuracy and fewer error were obtained	Important term was deleted as the ratio of the document increase.
[43,44]	Feature Selection Technique for Arabic text categorization Employing Compound Word Stats	SVM	It clearly states that the suggested approach can improve standard methods of classification in order to acquire the best classification outcomes.	Other text mining approaches should be investigated in the future to take advantage of other relationships between terms.
[45,46]	A Combination of Methods Research to Examine the Critical Success Factors (CSFs) of E-Learning in Saudi Universities	Content Validity Analysis	Analysis of Critical success factor	Critical factory might vary from place to place or it depends upon the region to region
[47,48]	Multi-Dimensional Recurrent Neural Networks for Arabic Video Text Recognition	RNN	Text recognition from video	Lacks in sequential coding
[49,50]	A Comparative Study using K-Means and Topic Modelling to Cluster Arabic Documents	K-means algorithm	It first demonstrates that normalising the weights in the vector space for the document-term matrices of written content greatly increases cluster quality and thus cluster reliability when using the k-means algorithm for clustering.	Building an embedding framework for words for Arabic text could be a major limitation of the study
[51,52]	Approaches Reliable Identification of Offensive Behaviour in Arabi Communicating Online	Text mining, SVM	Researcher report the findings of predictive models for detecting anti-social behaviour in Arabic communication via the internet, such as comments containing obscene or harmful terms and phrases. Researcher gathered and classified a huge dataset of Arabic YouTube comments that includes both offensive and harmless statements.	The main disadvantage is that other variables, such as textual context, were not taken into account. Furthermore, it may be important to analyse and contrast alternative categorization algorithms, particularly the use of deep neural networks.
[53]	A Telehealth Prediction Text Systems for Medical Suggestions In the Arabic Context, a Deep Learning Technique	Deep learning	Deep learning models gave appropriate suggestions for the subsequent word roughly half of the time, which was encouraging.	Training algorithms on larger-scale datasets can improve the efficacy of the produced models.

References

[1] O. A. Omitaomu and H. Niu, Artificial intelligence techniques in smart grid: A survey, *Smart Cities* **4**(2) (2021) 548–568.

[2] A. M. Azmi, M. N. Almutery and H. A. Aboalsamh, Real-word errors in arabic texts: A better algorithm for detection and correction, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **27**(8) (2019) 1308–1320.

[3] A. Y. Muaad, H. J. Davanagere, M. A. Al-antari, J. B. Benifa and C. Chola, Ai-based misogyny detection from arabic levantine twitter tweets, in *Computer Sciences & Mathematics Forum*, **2**(1), MDPI2021, p. 15.

[4] A. Shoufan and S. Alameri, Natural language processing for dialectical arabic: A survey, in *Proceedings of the second workshop on Arabic natural language processing*, 2015, pp. 36–48.

[5] G. S. Rady, S. S. Mohamed, M. F. Mohamed and K. F. Hussain, High dimensional autonomous computing on arabic language classification, *Computers and Electrical Engineering* **100** (2022) p. 108020.

[6] A. Alsayat and N. Elmitwally, A comprehensive study for arabic sentiment analysis (challenges and applications), *Egyptian Informatics Journal* **21**(1) (2020) 7–12.

[7] M. A. Ali, Artificial intelligence and natural language processing: the arabic corpora in online translation software, *International Journal of Advanced and Applied Sciences* **3**(9) (2016) 59–66.

[8] G. Zakria, M. Farouk, K. Fathy and M. N. Makar, Relation extraction from arabic wikipedia, *Indian Journal of Science and Technology* **12**(46) (2019) 01–06.

[9] A. Y. Muaad, H. Jayappa Davanagere, J. Benifa, A. Alabrah, M. A. Naji Saif, D. Pushpa, M. A. Al-Antari and T. M. Alfakih, Artificial intelligence-based approach for misogyny and sarcasm detection from arabic texts, *Computational Intelligence and Neuroscience* **2022** (2022).

[10] M. Farouk, Ontology-based semantic representation for arabic text: a survey, *Journal of Informatics and Mathematical Sciences* **9**(4) (2017).

[11] A. Y. Muaad, M. A. Al-antari, S. Lee and H. J. Davanagere, A novel deep learning arcar system for arabic text recognition with character-level representation, in *Computer Sciences & Mathematics Forum*, **2**(1), MDPI2021, p. 14.

[12] B. Dahy, M. Farouk and K. Fathy, Arabic sentences semantic similarity based on word embedding, in *2022 20th International Conference on Language Engineering (ESOLEC)*, **20**, IEEE2022, pp. 35–40.

[13] A. H. Mohammad, T. Alwada ‘n and O. Al-Momani, Arabic text categorization using support vector machine, naïve bayes and neural network, *GSTF Journal on Computing (JoC)* **5** (2016) 1–8.

[14] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni and D. Nouvel, Arabic natural language processing: An overview, *Journal of King Saud University-Computer and Information Sciences* **33**(5) (2021) 497–507.

[15] N. Aljedani, R. Alotaibi and M. Taileb, Hmatc: Hierarchical multi-label arabic text classification model using machine learning, *Egyptian Informatics Journal* **22**(3) (2021) 225–237.

[16] K. Darwish, N. Habash, M. Abbas, H. Al-Khalifa, H. T. Al-Natsheh, H. Bouamor, K. Bouzoubaa, V. Cavalli-Sforza, S. R. El-Beltagy, W. El-Hajj *et al.*, A panoramic survey of natural language processing in the arab world, *Communications of the ACM* **64**(4) (2021) 72–81.

[17] A. Y. Taha and S. Tiun, Binary relevance (br) method classifier of multi-label classification for arabic text., *Journal of Theoretical & Applied Information Technology* **84**(3) (2016).

[18] A. M. Azmi, A. O. Al-Qabbany and A. Hussain, Computational and natural language processing based studies of hadith literature: a survey, *Artificial Intelligence Review* **52** (2019) 1369–1414.

[19] Y. A. Alhaj, J. Xiang, D. Zhao, M. A. Al-Qaness, M. Abd Elaziz and A. Dahou, A study of the effects of stemming strategies on arabic document classification, *IEEE access* **7** (2019) 32664–32671.

[20] O. Oueslati, E. Cambria, M. B. HajHmida and H. Ounelli, A review of sentiment analysis research in arabic language, *Future Generation Computer Systems* **112** (2020) 408–430.

[21] D. Suleiman, A. Awajan and N. Al-Madi, Deep learning based technique for plagiarism detection in arabic texts, in *2017 International Conference on New Trends in Computing Sciences (ICTCS)*, IEEE2017, pp. 216–222.

[22] F. Husain and O. Uzuner, A survey of offensive language detection for the arabic language, *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* **20**(1) (2021) 1–44.

[23] L. Lulu and A. Elnagar, Automatic arabic dialect classification using deep learning models *Procedia computer science* **142**, (Elsevier, 2018), pp. 262–269.

[24] A. Alsayat and N. Elmitwally, A comprehensive study for arabic sentiment analysis (challenges and applications), *Egyptian Informatics Journal* **21**(1) (2020) 7–12.

[25] S. L. Marie-Sainte, N. Alalyani, S. Alotaibi, S. Ghouzali and I. Abunadi, Arabic natural language processing and machine learning-based systems, *IEEE Access* **7** (2018) 7011–7020.

[26] S. L. Marie-Sainte, N. Alalyani, S. Alotaibi, S. Ghouzali and I. Abunadi, Arabic natural language processing and machine learning-based systems, *IEEE Access* **7** (2018) 7011–7020.

[27] A. Özçift, K. Akarsu, F. Yumuk and C. Söylemez, Advancing natural language processing (nlp) applications of morphologically rich languages with bidirectional encoder representations from transformers (bert): an empirical case study for turkish, *Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije* **62**(2) (2021) 226–238.

[28] M. Beseiso, A. R. Ahmad and R. Ismail, A survey of arabic language support in semantic web, *International Journal of Computer Applications* **9**(1) (2010) 35–40.

[29] A. B. Soliman, K. Eissa and S. R. El-Beltagy, Aravec: A set of arabic word embedding models for use in arabic nlp, *Procedia Computer Science* **117** (2017) 256–265.

[30] S. L. Marie-Sainte, N. Alalyani, S. Alotaibi, S. Ghouzali and I. Abunadi, Arabic natural language processing and machine learning-based systems, *IEEE Access* **7** (2018) 7011–7020.

[31] A. Soufan, Deep learning for sentiment analysis of arabic text, in *Proceedings of the ArabWIC 6th Annual International Conference Research Track*, 2019, pp. 1–8.

[32] Y. Jaafar and K. Bouzoubaa, A survey and comparative study of arabic nlp architectures, *Intelligent Natural Language Processing: Trends and Applications* (2018) 585–610.

[33] A. A. Altowayan and A. Elnagar, Improving arabic sentiment analysis with sentiment-specific embeddings, in *2017 IEEE international conference on big data (big data)*, IEEE2017, pp. 4314–4320.

[34] S. A. Salloum, M. Al-Emran and K. Shaalan, A survey of lexical functional grammar in the arabic context, *International Journal of Computing and Network Technology* **4**(03) (2016).

[35] E. Omara, M. Mosa and N. Ismail, Applying recurrent networks for arabic sentiment analysis, *Menoufia Journal of Electronic Engineering Research* **31**(1) (2022) 21–28.

[36] M. Al-Ayyoub, A. Nuseir, K. Alsmearat, Y. Jararweh and B. Gupta, Deep learning for arabic nlp: A survey, *Journal of computational science* **26** (2018) 522–531.

[37] A. M. Alayba, V. Palade, M. England and R. Iqbal, Improving sentiment analysis in arabic using word representation, in *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, IEEE2018, pp. 13–18.

[38] A. Alsayat and N. Elmitwally, A comprehensive study for arabic sentiment analysis (challenges and applications), *Egyptian Informatics Journal* **21**(1) (2020) 7–12.

[39] M. Alkhatib and K. Shaalan, Natural language processing for arabic metaphors: a conceptual approach, in *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016 2*, Springer2017, pp. 170–181.

[40] W. Etaiwi and A. Awajan, Graph-based arabic nlp techniques: a survey *Procedia computer science* **142**, (Elsevier, 2018), pp. 328–333.

[41] A. K. Sangaiah, A. E. Fakhry, M. Abdel-Basset and I. El-henawy, Arabic text clustering using improved clustering algorithms with dimensionality reduction, *Cluster Computing* **22** (2019) 4535–4549.

[42] A. M. Al Sbou, A. Hussein, B. Talal and R. Rashid, A survey of arabic text classification models, *International Journal of Electrical and Computer Engineering (IJECE)* **8**(6) (2018) 4352–4355.

[43] A. Adel, N. Omar, M. Albared and A. Al-Shabi, Feature selection method based on statistics of compound words for arabic text classification., *Int. Arab J. Inf. Technol.* **16**(2) (2019) 178–185.

[44] R. Bensoltane and T. Zaki, Aspect-based sentiment analysis: an overview in the use of arabic language, *Artificial Intelligence Review* **56**(3) (2023) 2325–2363.

[45] Q. N. Naveed, A. Muhammad, S. Sanober, M. R. N. Qureshi and A. Shah, A mixed method study for investigating critical success factors (csfs) of e-learning in saudi arabian universities, *methods* **8**(5) (2017) 171–178.

[46] M. Al-Ayyoub, A. A. Khamaiseh, Y. Jararweh and M. N. Al-Kabi, A comprehensive survey of arabic sentiment analysis, *Information processing & management* **56**(2) (2019) 320–342.

[47] O. Zayene, S. E. Amamou and N. E. BenAmara, Arabic video text recognition based on multi-dimensional recurrent neural networks, in *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, IEEE2017, pp. 725–729.

[48] A. Alnawas and N. Arici, The corpus based approach to sentiment analysis in modern standard arabic and arabic dialects: A literature review, *Politeknik Dergisi* **21**(2) (2018) 461–470.

[49] M. Alhawarat and M. Hegazi, Revisiting k-means and topic modeling, a comparison study to cluster arabic documents, *IEEE Access* **6** (2018) 42740–42749.

[50] M. Alqahtani and E. Atwell, Arabic quranic search tool based on ontology, in *Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings 21*, Springer2016, pp. 478–485.

[51] A. Alakrot, L. Murray and N. S. Nikolov, Towards accurate detection of offensive language in online communication in arabic, *Procedia computer science* **142** (2018) 315–320.

[52] S. Abdulmalek, A.-H. Salah, M. Alsurori, M. Hadwan, A. Aqlan and F. Alqasemi, Levenstein's algorithm on english and arabic: A survey, in *2021 International Conference of Technology, Science and Administration (ICTSA)*, IEEE2021, pp. 1–6.

[53] M. Habib, M. Faris, R. Qaddoura, A. Alomari and H. Faris, A predictive text system for medical recommendations in telemedicine: a deep learning approach in the arabic context, *IEEE Access* **9** (2021) 85690–85708.