Utilizing NLP to Optimize Municipal Services Delivery Using a Novel Municipal Arabic Dataset

Homod Hamed Alaloye, Ahmad B. Alkhodre, Emad Nabil
Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah, Saudi Arabia

***Abstract*—The natural language processing paradigm has emerged as a vital tool for addressing complex business challenges, mainly due to advancements in machine learning (ML), deep learning (DL), and Generative AI. Advanced NLP models have significantly enhanced the efficiency and effectiveness of NLP applications, enabling the seamless integration of various business processes to improve decision-making. In the municipal sector, the Kingdom of Saudi Arabia is trying to harness the power of NLP to promote urban development, city planning, and infrastructure enhancements, ultimately elevating the quality of life for its residents. In the municipal sector, approximately 300 services are available through multiple channels, including the Baladi application, unified communication services, WhatsApp, a dedicated beneficiary center (serving citizens and residents), and social media accounts. These channels are supported by a dedicated team that operates 24/7. This paper examines the implementation of ML and DL methods to categorize requests and suggestions submitted by residents for various municipal services in the Kingdom of Saudi Arabia. The primary aim of this work is to enhance service quality and reduce response times to community inquiries. However, a significant challenge arises from the lack of Arabic datasets specifically tailored to the municipal sector for training purposes, which limits meaningful progress. To address this issue, we have created a novel dataset consisting of 3,714 manually classified requests and suggestions collected from the X platform. This dataset is organized into eight classes: tree maintenance, lighting, construction waste, old and neglected assets, road conditions, visual pollution, billboards, and cleanliness. Our findings indicate that ML models, particularly when optimized with hyperparameters and appropriate pre-processing, outperformed DL models, achieving an F1 score of 90% compared to 88%. By releasing this novel Arabic dataset, which will be open sourced for the scientific community, we believe this work provides a foundational reference for further research and significantly contributes to improving the municipal sector's service delivery.**

Keywords*—*Arabic text classification; machine learning (ML); deep learning (DL); hyperparameter optimization; municipal services

I. INTRODUCTION

Machine learning (ML) and deep learning (DL) models have significantly advanced our understanding of natural language, creating new research opportunities in areas such as text classification, translation, summarization, generation, and question-answering. However, the effectiveness of these models can vary depending on the specific task, particularly in text classification.

The authors in [1] have applied ensemble learning techniques to improve accuracy, which are especially effective for dealing with complex data with varied features. Ensemble learning combines multiple models with diverse parameters, enhancing the analysis and classification of intricate data patterns. As a result, these techniques prove beneficial for achieving higher accuracy in complex text classification tasks.

The authors in [2] utilize neural network architectures with multiple layers to effectively identify and interpret patterns in language-related tasks, such as question answering, translation, and text classification. This effectiveness is further enhanced by embedding techniques, including word and character embeddings, which are essential for tasks like text classification, knowledge extraction, and question-answering. Additionally, hyperparameter optimization (HPO) is crucial in improving model performance, as systematic hyperparameter tuning can significantly enhance accuracy and efficiency.

The authors in [3] describe a strategy for selecting hyperparameter values incorporating techniques such as Grid Search, Random Search, Bayesian Optimization, Particle Swarm Optimization, and Genetic Algorithm Metaheuristics.

ML and DL models have transformed how business applications and services are enhanced and developed; many applications and services currently use machine learning (ML) and deep learning (DL) models for various purposes, including task automation, improved decision-making, enhanced customer experience, predictions, and automatic text classification. In Saudi Arabia, the municipal sector provides over 300 essential services for improving the quality of life through urban development, city planning, and infrastructure improvement. These services cover streets, parks, lighting, green spaces, sidewalks, public health, sanitation, and public transportation management. They are delivered through various channels, including the Balady application, unified contact numbers, WhatsApp service, beneficiary centers, and the municipal sector's social media accounts, all supported by a dedicated team working around the clock.

Using machine learning (ML) and deep learning (DL) models in municipal service procedures can improve how citizen and resident requests and suggestions are classified. This enhancement leads to better service performance, lower financial and human costs, and increased overall satisfaction with municipal services.

These models require a robust dataset for training, which is crucial for effectively extracting knowledge, analyzing and classifying data, recognizing patterns, and turning raw data into actionable information. However, our review of existing literature reveals a significant gap in the availability of training datasets, especially in Arabic. Most datasets for training learning models are available in English, while there is a shortage of sufficient datasets in Arabic.

This paper significantly contributes to natural language processing, particularly concerning municipal services in the Kingdom of Saudi Arabia. It introduces a comprehensive, specialized Arabic dataset that captures citizens' requests and suggestions related to these services. This dataset enhances the resources available for researchers and practitioners in classifying and analyzing Arabic texts. Furthermore, it is a valuable tool for advancing the development of machine learning (ML) and deep learning (DL) models tailored to the Arabic language's unique linguistic and cultural characteristics. Ultimately, this will facilitate more effective data-driven decision-making in municipal service delivery. The dataset will be open-sourced for the community, and we will also provide a performance baseline, allowing others to use it for performance comparison in their own work.

The paper is organized into six sections. Section II provides a comprehensive literature review, offering insights into previous research. Section III details the paper's key contributions, elaborating on the proposed methodology. Section IV presents the experimental results, followed by a discussion in Section V. Finally, Section VI concludes the paper by summarizing the key insights and future directions.

II. Literature Review

The text classification problem has many real-life applications, and there are many techniques for tackling it. Fig. 1 depicts the major methods.

With the global rise of social media platforms like Twitter, Facebook, and Instagram, coupled with greater engagement with hashtags, there has been a significant increase in Arabic-language tweets that contain racist, bullying, or hateful content. This trend has prompted researchers to classify Arabic text into multi-class and binary categories to tackle these issues.

In a related study [4], the authors employed binary classification to categorize millions of articles and blogs as either real or fake news using the Multivariate Bernoulli Naïve Bayes model, demonstrating the effectiveness of ML techniques in authenticating Arabic content. These studies highlight the wide-ranging applicability of ML methods in Arabic-language tasks, from hate speech detection to verifying the authenticity of information.

In [5], the authors contributed to reducing the time researchers spend searching for relevant papers by employing classification models such as SVM, Naïve Bayes, K-nearest Neighbor, and Decision Tree to categorize a set of 107 research papers across various fields, including science, business, and social sciences.

In [6], the authors examined the impact of various stemming methods and word representations on Arabic text classification using RNN architectures, specifically LSTM, Bi-LSTM, GRU, and Bi-GRU models. They compared the accuracy among different combinations of stemming approaches (no stemming, root-based stemming, and light-based stemming) and word representations (Word2Vec, FastText, and GloVe). The highest accuracy was found with no stemming using Word2Vec for word representation, along with the Bi-GRU architecture as the classifier.

Similarly, the authors in [7] developed a dataset for dialectal Arabic tweet acts by annotating a subset of the existing Arabic Sentiment Analysis Dataset (ASAD) based on six speech act categories. The ArSAS dataset was classified using BERT, and the results were compared with various Arabic BERT models. The findings suggest that the top-performing model was araBERTv2-Twitter.

The authors in [8] classified numerous policies published by governments and institutions, each containing several paragraphs indicating a specific topic. The study relied on ensemble learning techniques, particularly Bagging, integrated with DL methods, specifically Convolutional Neural Networks (CNN), to classify the policies based on the predetermined topics of the paragraphs.

In a related study [9], the authors demonstrated that ensemble learning outperformed traditional ML approaches by 6% in classifying Arabic data, which is characterized by its complexity arising from multiple dialects and variations in letter formation.

In [10], the authors employed four deep learning models—Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Convolutional LSTM (CNN-LSTM), and Bidirectional LSTM (Bi-LSTM)—to classify millions of Arabic texts across diverse domains, including cultural, scientific, political, and health-related topics, sourced from websites, social media platforms, academic publications, and news outlets. The Bi-LSTM model achieved the highest accuracy result (94.02).

The authors in [11] applied multi-label classification models to categorize the topics of each verse in the Quran, enhancing understanding of its content. They developed a majority voting approach, achieving optimal performance through a combination of Random Forest, Support Vector Machine (SVM), and Adaptive Boosting models. This ensemble achieved a micro F1 score of 0.5143 and a Hamming loss of 0.0878. Additionally, the majority voting method demonstrated improved performance compared to the individual models.

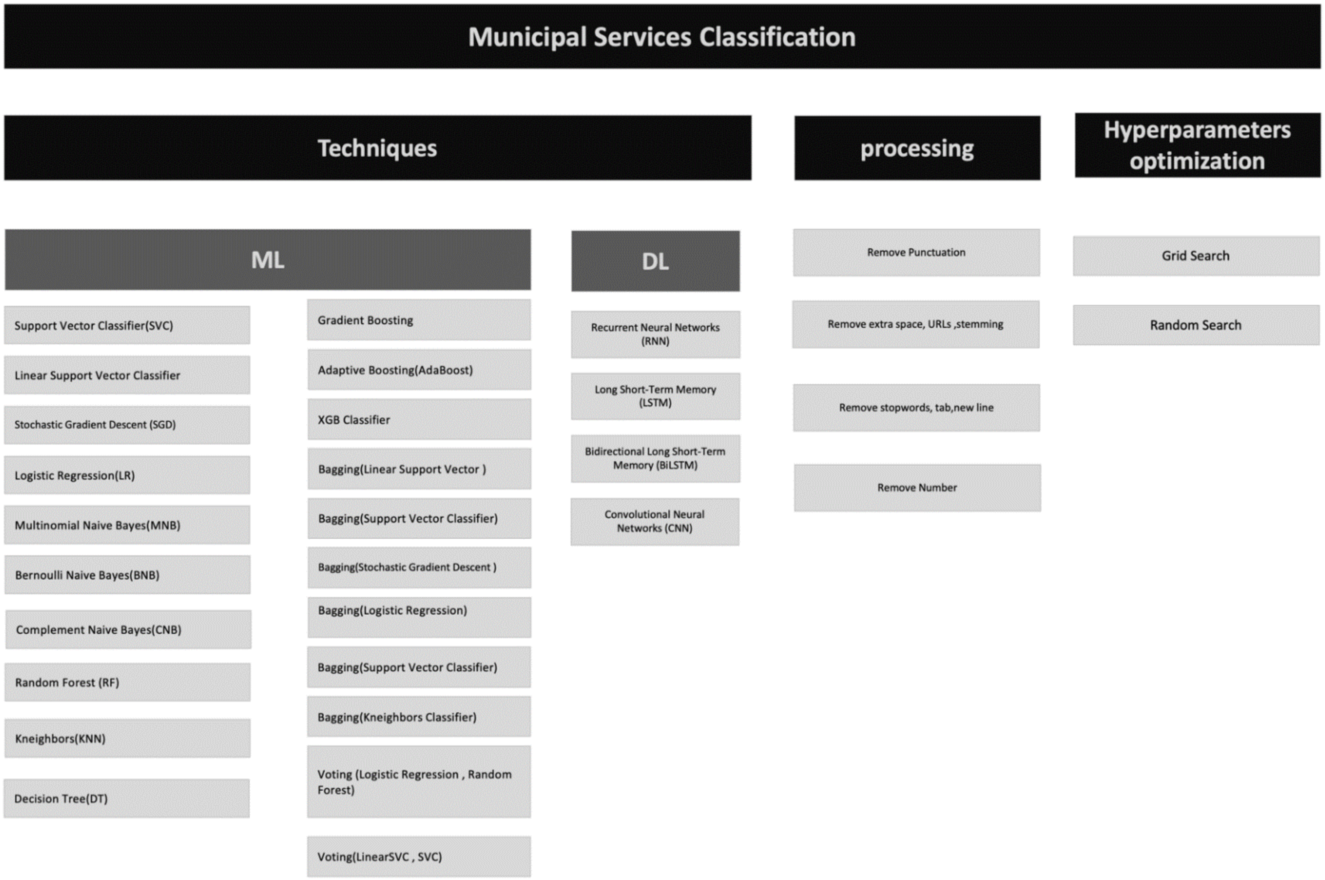
Municipal Services Classification
Techniques
processing
Hyperparameters optimization
ML
DL
Support Vector Classifier(SVC)
Linear Support Vector Classifier
Stochastic Gradient Descent (SGD)
Logistic Regression(LR)
Multinomial Naive Bayes(MNB)
Bernoulli Naive Bayes(BNB)
Complement Naive Bayes(CNB)
Random Forest (RF)
Kneighbors(KNN)
Decision Tree(DT)
Gradient Boosting
Adaptive Boosting(AdaBoost)
XGB Classifier
Bagging(Linear Support Vector)
Bagging(Support Vector Classifier)
Bagging(Stochastic Gradient Descent)
Bagging(Logistic Regression)
Bagging(Support Vector Classifier)
Bagging(Kneighbors Classifier)
Voting (Logistic Regression , Random Forest)
Voting(LinearSVC , SVC)
Recurrent Neural Networks (RNN)
Long Short-Term Memory (LSTM)
Bidirectional Long Short-Term Memory (BiLSTM)
Convolutional Neural Networks (CNN)
Remove Punctuation
Remove extra space, URLs ,stemming
Remove stopwords, tab,new line
Remove Number
Grid Search
Random Search

Fig. 1. NLP Classification: Techniques, preprocessing steps, and hyperparameter optimization.

The authors in [12] focused on classifying SMS messages as either spam or not spam, emphasizing the importance of this medium for communication and verification by banks, institutions, and e-platforms. They employed ensemble learning techniques for message classification, achieving an impressive accuracy of 99.91%.

The authors in [13] examined the classification of a large dataset containing 290,000 multi-tagged articles across various categories, including business, technology, sports, and politics. This research aimed to analyze the dataset to identify patterns, which would assist in summarizing and classifying the articles. Classifying news stories that belong to multiple categories posed significant challenges. However, logistic regression achieved an accuracy of 81.3%, while XGBoost reached 84.7%. Notably, a combination of ten neural network architectures—namely CNN, CLSTM, LSTM, BILSTM, GRU, CGRU, BIGRU, HANGRU, CRF-BILSTM, and HANLSTM—performed exceptionally well as a multilabel classifier, achieving an accuracy of 94.85%.

In a related study [14], the authors experimentally investigated the application of Convolutional Neural Networks (CNNs) to text classification; using FastText for word embeddings, the CNN model was trained on publicly different datasets. The authors in [15] compared the effects of classical and contextualized word embeddings on model performance. They evaluated classical embeddings, such as GloVe, Word2Vec, and FastText, on the HARD, Khooli, AJGT, ArSAS, and ASTD datasets using BiLSTM and CNN models. However, the contextualized transformer-based embedding model BERT outperformed the classical embeddings, achieving the highest classification accuracy across all datasets.

The authors in [16] concentrated on classifying fake Arabic news on Twitter by using various Arabic contextualized embedding models, such as AraBert, Arabic-BERT, ArBert, MARBert, Araelectra, and QaribBert, which together achieved an impressive classification accuracy of 98%.

The authors in [17] implemented various models, including Convolutional-GRU and Attention-GRU, using Word2Vec embeddings for Arabic text classification. They compared model performance on the SANAD and NADiA datasets, achieving an accuracy of 91.18% on SANAD and 96.94% on NADiA.

The authors in [18] applied a Genetic Algorithm to optimize CNN model parameters, improving Arabic text classification accuracy to 98%.

The authors in [19] utilized Word2Vec and TF-IDF with models such as SVC, Random Forest, and Gaussian Naive Bayes to classify Arabic tweets, reporting high model accuracy rates of 98.09% and 98.14%.

The authors in [20] demonstrated that the performance of the large language model BERT, when applied without fine-tuning its parameters, surpassed that of models utilizing fine-tuned parameters.

The authors in [21] used the NLTK package to improve machine learning performance in classifying Arabic news (DAFN) as either fake or real.

The authors in [22] collected a dataset using the Twitter API, which they manually labeled to ensure its reliability. They employed two techniques for feature extraction: N-gram models (including unigrams, bigrams, and char-grams) to enhance the classification of advertisements and illegal content through both machine learning (ML) and deep learning (DL) approaches.

The authors in [23] utilized a meta-heuristic genetic algorithm (G.A.) to optimize hyperparameters and feature extraction techniques, specifically the bag of words (BOW) and term frequency-inverse document frequency (TF-IDF) methods. They assessed the performance of several ensemble learning models, including the Extra Trees Classifier (ETC), Random Forest Classifier (RFC), and Logistic Regression Classifier (LRC). Their efforts led to an F1 score of 80.88% for sentiment classification and 68.76% for multilabel classification tasks.

The authors in [24] achieved a classification accuracy of 94.75% by grouping similar Arabic words generated by the Word2Vec model. Each group of similar words utilized the Word2Vec-based method RARF, which enhanced the model's effectiveness in accurately classifying Arabic text.

The authors in [25] utilized the ARABERT4TWC model, which features a custom-built Arabic BERT, transformation layers, a self-attention layer, and a classification layer to improve the classification of Arabic tweets. Using datasets such as SANAD for Arabic news stories, HARD, AJGT, ASTD, and ArsenTD–Lev, the model achieved an impressive accuracy of 98%.

The authors in [26] utilized ensemble learning alongside N-gram features (Unigram, Bi-Gram, and Tri-Gram) to improve the classification accuracy of Arabic spam reviews, achieving maximum accuracies of 95% and 99.98% across two datasets.

In a related study [27], the authors emphasized the importance of hyperparameter tuning techniques in improving the performance of machine learning (ML) and deep learning (DL) algorithms applied to Arabic text. They utilized random search and grid search methods for tuning in classification tasks sourced from cnnarabic.com. The findings indicated that random search outperformed grid search, achieving accuracies of 95.16% for Artificial Neural Networks (ANN), 94.57% for Multinomial Logistic Regression (MLR), and 94.28% for Support Vector Machines (SVM).

The authors in [28] used the Particle Swarm Optimization (PSO) algorithm for feature selection, resulting in a document classification accuracy of 97%.

The authors in [29] compared various Arabic BERT models used for text classification to assess and summarize each model's performance.

In [30], a related study, the authors introduced the BERT-Mini Model (ABMM) to classify hate speech in Arabic social media posts into standard, abusive, and hate speech. The ABMM model achieved an impressive accuracy score of 0.986.

The authors in [31] collected opinions from Arabic-speaking users about ChatGPT technology using the Tweepy and SNscrape Python libraries. They categorized the data into three sentiment classes: negative, positive, and neutral. By employing classification models such as RoBERTa, they achieved a recall accuracy of 99%.

The authors in [32] performed a comparative analysis of ChatGPT-4 and Google Gemini for detecting and classifying messages as spam, showcasing advancements in AI language models for content moderation.

The authors in [33] presented a method for improving the classification of cognitive distortions in Arabic content on Twitter. They incorporated an unlabeled dataset using the unsupervised learning model BERTopic. This approach combined the processed unlabeled data with a labeled dataset, which enhanced the overall classification performance using machine learning techniques. A concise summary of the literature is presented in Table I.

III. METHODOLOGY

The proposed methodology, illustrated in Fig. 2, presents a multi-method approach aimed at enhancing the automatic classification of citizens' and residents' requests and suggestions regarding municipal services in the Kingdom of Saudi Arabia. This approach incorporates various machine learning (ML) and deep learning (DL) models, along with hyperparameter tuning and data processing techniques, to improve classification accuracy. It includes a comparison of model performance before and after applying these hyperparameters and data preprocessing, with the objective of identifying the best-performing model for this task. The ultimate goal is to streamline the classification of public requests and suggestions related to municipal services, thus facilitating more efficient service delivery and quicker response times.

TABLE I. REVIEW OF THE PRIMARY STUDIES IN THE LITERATURE

Index	Reference	Main Idea	Used Techniques	Dataset
1	Alzanin et al.	Multilabel classification of Arabic content in the Arabic language.	Ensemble learning models (Extra Trees Classifier, Random Forest Classifier, Logistic Regression Classifier), genetic algorithm (GA), Bag of Words (BOW), and Term Frequency-Inverse Document Frequency (TF-IDF).	MAWQIF
2	Sabri et al.	Arabic text classification	SVM,NB, K-NN, FS(PCA,LDA chi-square, mutual information, RARF),TF-IDF, Word2Vec	Khaleej-2004 Arabic, Watan-2004 is a large Arabic

3	Alruily et al.	Classification of Arabic Long Tweets Using Transfer Learning with BERT	BERT - ARABERT4TWC	SANAD of Arabic news stories, HARD, AJGT, ASTD, ArsenTD–Lev
4	Alhaj et al.	Classification through the introduction of Optimal Configuration Determination for Arabic Text Classification (OCATC).	LR, RF, KNN, DT, NN, SVC, LSVM, and SGD	The dataset created by Abuaiadah, 2700 documents, dataset by collect Saad, CNN Dataset
5	Alammary et al.	Applying BERT Models for Arabic Text Classification	BERT, AraBERT, MARBERT, ArabicBERT, ARBERT, XLM-RoBERTa, QARiB, Arabic ALBERT	1780 articles from Google Scholar
6	Alwehaibi et. al.	They are using deep learning embedding methods to optimize sentiment classification of dialectal Arabic short texts at the document level.	LSTM, CNN, and ensemble model	A dataset of Standard Arabic and colloquial Arabic collected from Twitter, AraSenTi
7	Saeed et al.	Various classification methods for spam detection in Arabic opinion texts were investigated. The methods examined include rule-based approaches, classical machine learning techniques, majority voting ensemble, and stacking ensemble methods. Additionally, N-gram features were used to enhance the classification process.	N-gram feature, voting ensemble classifier, Stacking ensemble classifier	DOSC dataset is translated from English language to Arabic language, HARD
8	Nassif et al.	Arabic Fake News Classification	Arabert,Arabic-Bert, ArBert, MARBert, Araelectra and QaribBert	Single-class (SANAD), multi-class (NADiA)
9	Almaliki et. al.	Classification of Arabic hate speech on social media platforms, specifically Twitter, using the Arabic BERT-Mini Model (ABMM).	BERT-Mini Model (ABMM), Bert, LSTM, CNN+LTSM, Linear SVC, SVC, SGD	
10	Mujahid et.al	Automatically classify tweets and opinions of Arab users about ChatGPT technology using large models.	RoBERTa, XLNet , DistilBERT	Tweepy SNscrape Python library using various hash-tags such as # Chat-GPT, #OpenAI, #Chatbot, Chat-GPT3 collect 27000 tweet
11	Mardiansyah et. al.	Compare the performance of ChatGPT-4 and Google Gemini in spam detection.	ChatGPT-4, Google Gemini	SpamAssassin public mail corpus
12	Alhaj et. al.	Enhancing the classification of cognitive distortions in Arabic content on Twitter.	BERTopic, DT, SVM, R.F., KNN, XGBoost, Bagging, and Stacking	Public Therapist Q&A, collected from different resources, and most of them are not publicly available
13	Hassan et. al.	A Comparison of Learning Model Performance in Text Classification Across Various Datasets	SVM,K-NN -L.R. -MNB RF	IMDB dataset, SPAM dataset
14	Elnagar et al.	Comparison of Various Deep Learning Models for Arabic Text Categorization	convolutional-GRU - attention-GRU	Single-label (SANAD), multilabel (NADiA)
15	Alsaleh et al.	Optimize the parameters of the Convolutional Neural Network (CNN) to enhance Arabic text classification using a genetic algorithm.	CNN with GA	AlRiyadh Newspaper and Saudi Press Agency (SPA)
16	Samah M.Alzanin et. al.	Apply feature extraction using Word2vec and TF-IDF on multiple models to classify Arabic tweets and compare the accuracy of the best models.	GNB- SVM- RF	Thirty-five thousand six hundred twenty-seven collected tweets were manually annotated into five categories.
17	Galal et al.	Classification of Arabic texts using BERT as a feature extractor without fine-tuning performance.	BERT, MARBERT, Qarib, ARBERT, GigaBERT	ArSarcasm-v2 Dataset for Arabic sarcasm
18	Mahmoudi et. al.	Classifying Arabic news as fake or real by creating a machine learning model.	R.F., L.R., Linear SVM, Gradient Boosting Classifier	Arabic fake news (DAFN)
19	Kaddoura et. al.	Classify fake advertisements and illegal content on Twitter using machine learning (ML) and deep learning (DL).	ML – DL	Twitter API and labeled manually to get a reliable dataset

20	Elgeldawi et al.	A thorough comparative analysis of different hyperparameter tuning techniques.	hyperparameter tuning by using techniques, including Grid Search, Random Search, Bayesian Optimization, Particle Swarm Optimization (PSO), Genetic Algorithm (GA), six machine learning algorithms LR, RC, SVC, DT), RF, NB) classifier	Booking.com website to collect hotel reviews. A total of 3500 positive, 3500 negative
21	Chowdhury et al.	Extract meaningful information from published abstracts and classify it into three fields: Science, Business, and Social Science.	Support Vector Machines, Naïve Bayes, K-Nearest Neighbour and Decision Tree	107 research abstracts are collected to build the dataset
22	Decui et al.	Work on classifying paragraphs into various topics within policy.	ensemble convolution neural network (CNN)	policy in China
23	Onan et al.	utilized several keyword extraction methods because they are important for capturing the content of the document.	most frequent measure-based keyword extraction, comparing base learning algorithms (Naïve Bayes, support vector machines, logistic regression, and Random Forest) with five widely utilized ensemble methods (AdaBoost, Bagging, Dagging, Random Subspace, and Majority Voting)	Reuters-21578 document collection, ACM document collection
24	Husain	Impact of using single learner machine learning approaches compared to ensemble machine learning approaches for text classification in Arabic.	single learner machine learning approach and ensemble machine learning	Open-Source Arabic Corpora and Corpora Processing Tools (OSACT) in Language Resources and Evaluation Conference (LREC) 2020
25	Ghourabi	The text focuses on developing an ensemble learning strategy that uses an optimized weighted voting technique to improve classification results. This approach incorporates a text embedding method based on a pre-trained language model, aiming to enhance performance in classification tasks.	Embedding technique based on the GPT-3 Transformer is used to represent text messages as dense numerical vectors. An ensemble of classifiers, including SVM, KNN, CNN, and LightGBM, is created through a voting mechanism.	SMS Spam Collection Dataset
26	Al-Qerem	Classifying Arabic news articles based on their content using deep learning techniques.	RNN, LSTM, CNN-LSTM, and Bi-LSTM	data set of news articles Arabic
27	Hariyadi	Classify the Al-Quran into various topics like faith, deeds, and science to enhance understanding of the text.	majority voting approach	Al-Quran
28	Xu, S., Li, Y	Classifying documents with machine learning.	Multinomial NB, Bayesian Multinomial Classifier	20 newsgroup data contain 18,828
29	Bahbib	Investigating the effect of different stemming methods and word representations on Arabic text classification.	Word2Vec, FastText, and GloVe, LSTM, Bi-LSTM, GRU, and Bi-GRU	
30	Alshehri	A classification approach for identifying dialectal Arabic speech acts on Twitter using transformer-based deep learning models.	various BERT models,	Arabic sentiment analysis dataset (ASAD) , Arabic Tweet Act dataset (ArSAS)
31	Al-Fuqaha'a	The text is centered on creating a model for multi-class text classification specifically within the Arabic healthcare sector.	feature extraction techniques (TF-IDF, Word2Vec), LR, RF, MNB, SGD, SVM),. Ensemble methods(stacking and bagging) and AraBERT and MarBERT	Altibbi dataset

A. Data Preparation

In this study, we collected a dataset from the social media platform (X) to analyze requests and suggestions from citizens and residents about the quality and performance of municipal services across various cities in the Kingdom of Saudi Arabia.

After gathering this feedback, we sorted, filtered, and selected comments that specifically addressed improvements to municipal services to help develop an effective classification model for these requests and suggestions. The dataset includes a total of 3,714 comments and suggestions, with most contributions focusing on critical areas such as infrastructure services, including roads, sidewalks, and lighting; issues related to water leakage and sidewalk subsidence; concerns regarding street excavations; and requests for enhancing green spaces, including tree planting and park development in residential areas. Details of the dataset can be found in Table II, Table III, and Fig. 3.

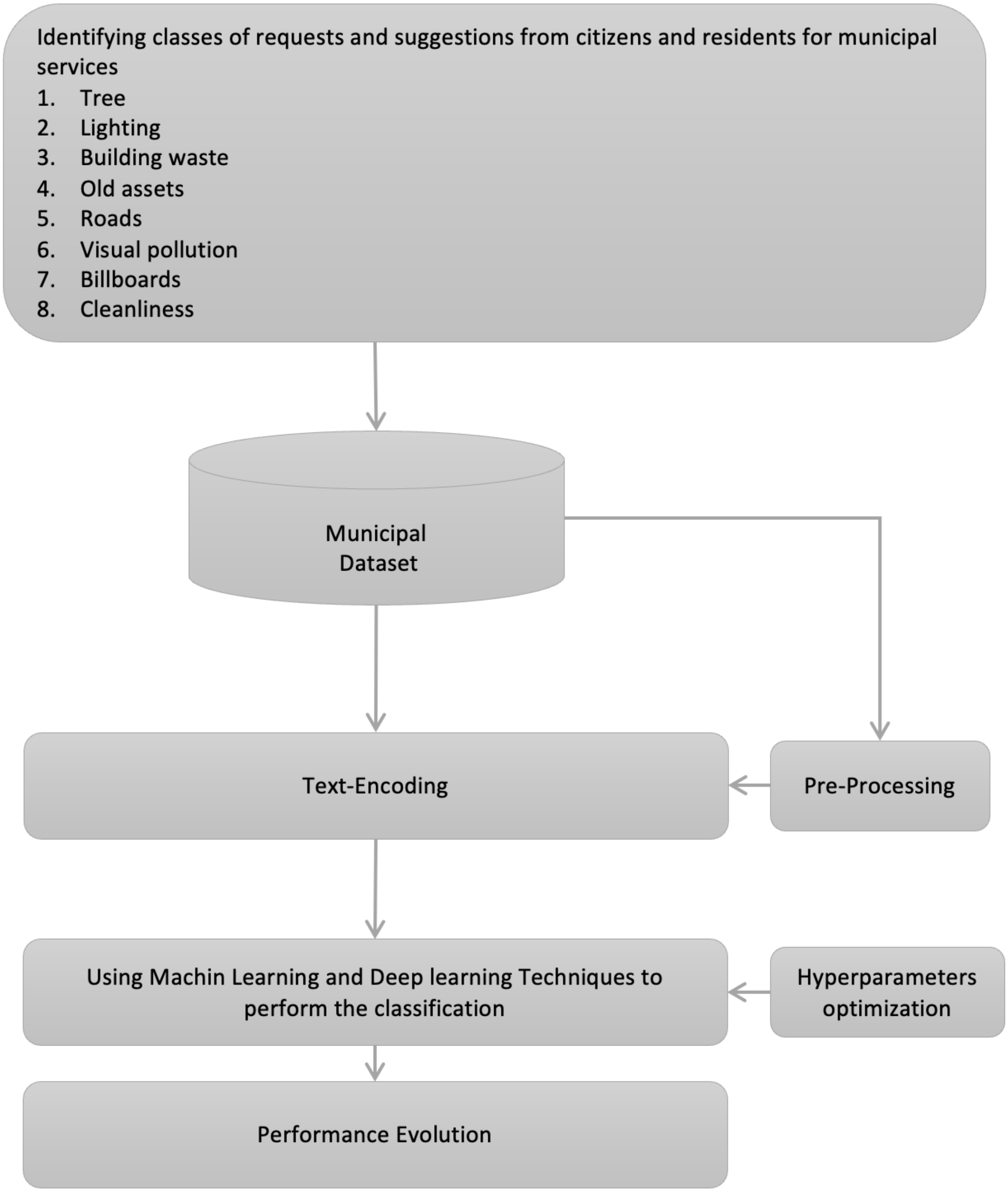
Identifying classes of requests and suggestions from citizens and residents for municipal services
1. Tree
2. Lighting
3. Building waste
4. Old assets
5. Roads
6. Visual pollution
7. Billboards
8. Cleanliness
Municipal Dataset
Text-Encoding
Pre-Processing
Using Machin Learning and Deep learning Techniques to perform the classification
Hyperparameters optimization
Performance Evolution

Fig. 2. Block diagram of the proposed methodology.

TABLE II. DATASET AND CLASS DESCRIPTIONS

Class	Description	No. of records
Tree	This category includes tweets that focus on suggestions for planting more trees, trimming existing ones, maintaining tree health, addressing insect infestations, and enhancing the care of green spaces. These tweets reflect public interest in sustainable urban development and emphasize the importance of green areas in improving environmental quality and community well-being.	400
Lighting	The Lighting category in this dataset includes tweets from citizens and residents requesting the installation or improvement of lighting in various municipal areas to enhance visibility and safety. Many tweets also emphasize the need for ongoing maintenance of public lighting systems, highlighting issues such as dim or non-functional streetlights and insufficient illumination in certain public spaces. This systematic categorization of tweets allows for a structured analysis of public feedback regarding lighting infrastructure, helping municipal authorities identify specific areas that need attention and improvement in lighting services.	550
Building Waste	Tweets in this category primarily focus on reporting construction waste in urban neighborhoods and include questions about available disposal methods. These comments reflect the interest of citizens and residents in managing construction debris and seeking solutions for waste disposal within residential areas to promote more sustainable urban environments. By analyzing this feedback, municipal authorities can gain valuable insights into community needs and develop targeted strategies to improve construction waste management practices.	337
Old assets	Tweets in this category consist of questions from citizens and residents about abandoned vehicles in residential neighborhoods and concerns regarding other neglected assets in these areas. These observations demonstrate the community's interest in understanding how to report old and abandoned vehicles. By analyzing these tweets, the municipal sector can effectively allocate follow-up teams to address cases of abandoned vehicles and facilitate their removal.	400
Roads	The tweets express the demands of citizens and residents for improvements in infrastructure. They specifically highlight the need for resurfacing roads, addressing subsidence issues affecting roads and sidewalks, and installing speed bumps to reduce vehicle speeds in residential neighborhoods. These comments underscore the essential role of local government in enhancing and developing infrastructure. By analyzing these tweets, municipal authorities can more effectively identify areas in need of maintenance, allowing them to prioritize initiatives that will improve urban living conditions.	916
Visual pollution	Tweets in this category consist of requests from citizens and residents urging the removal of graffiti from the walls of public buildings. This feedback highlights the community's desire for clean and well-maintained public spaces, emphasizing the importance of aesthetic upkeep in urban environments. By addressing these requests, municipal authorities can improve the appearance of public facilities and foster a greater sense of pride within the community.	251
Billboards	Tweets in this category address requests for the removal of unlicensed advertising posters and billboards located at the entrances of residential buildings, on sidewalks, and in public areas. These unauthorized displays can disrupt the visual appeal of the environment and violate legal standards. By responding to these concerns, municipal authorities can help maintain the integrity of public spaces and ensure compliance with local regulations, fostering a more aesthetically pleasing community.	76
Cleanliness	This category includes tweets that focus on the cleanliness of public spaces. Common topics cover the condition of roads, sidewalks, and waste management. Discussions often highlight issues such as littering, the frequency of waste collection, and the overall maintenance of public areas. These tweets reflect public sentiment regarding the effectiveness of municipal services in maintaining a clean and hygienic environment and may include suggestions for improvement.	600

TABLE III. SAMPLES OF THE DATASET. EACH ROW CONTAINS ONE SAMPLE, AND THE ROW HAS THREE SUB-ROWS: THE RAW COMPLAINT, THE PREPROCESSED COMPLAINT, AND THE TRANSLATION OF THE PREPROCESSED COMPLAINT

#	Complaint	Class
1	السلام عليكم ورحمة الله وبركاته الموقع بحاججججججججة الى تنفيذ مطب وعمل خطوط عبور مشاة بالإضافة الى عمل صيانة للمطب! القديم وذلك ان الموقع بجوار مدرسة ثانوية 24.44817584,39.60250383الموقع للتواصل رقم 0554472280	Roads
	الموقع بحاجه تنفيذ مطب وعمل خطوط عبور مشاة بالإضافة عمل صيانة للمطب الموقع بجوار مدرسة ثانوية	
	The site must implement a speed bump and install pedestrian crossings, while also ensuring the speed bump is maintained. This location is adjacent to a secondary school.	
2	السلام عليكم ورحمة الله وبركاته , في الممشى العام كتابات مشوهة للمظهر العام!!!1 مكتوبة على الكراسي.....	visual pollution
	الممشى العام كتابات مشوهة للمظهر العام مكتوبة على الكراسي	
	Public walkway with distorted graffiti written on chairs.	
3	..... الله يقويكم لايوجد لدينا اشجار في الحي نأمل منكم اطلاق حملة للتشجير وشكرا لكم الحي	Tree
	الله يقويكم لايوجد لدينا اشجار في الحي نأمل منكم اطلاق حملة للتشجير وشكرا لكم الحي	
	May God grant you strength. Our neighborhood lacks trees, and we hope you will initiate a tree-planting campaign. Thank you from the neighborhood.	
4	السلام عليكم عامود إنارة يحتاج استبدال ووضع غطاء لغرفة التوزيع رقم بلاغ :48848	Lighting
	عامود إنارة يحتاج استبدال ووضع غطاء لغرفة التوزيع	
	The lamp post needs to be replaced, and the cover for the distribution room needs attention.	
5	مخلفات بناء ترمى في الأراضي البيضاء وهذا تشوة بصري رقم البلاغ 67874	Building Waste
	مخلفات بناء ترمى في الأراضي البيضاء وهذا تشوه بصري	
	Construction waste is being discarded in open areas, creating a visual disturbance.	
6	السلام عليكم اراضي اصبحت مكب للنفايات داخل الاحياء السكنيه وتشكل خطر كبير حيث انها حفره جدا عميقه وتشكل خطر على الاطفال والسيارات في أحد الاحياء السكنية رقم الجوال 0554472280	Cleanliness
	اراضي اصبحت مكب للنفايات داخل الاحياء السكنيه وتشكل خطر كبير حيث انها حفره جدا عميقه وتشكل خطر على الاطفال والسيارات في احد الاحياء السكنيه	
	Lands that have become dumping grounds for waste in residential neighborhoods pose significant dangers, as they contain very deep holes that threaten children and vehicles.	
7	لوحة ممزقة,وملصقات دعائية ,,, في وسط الحي السكني ,,,, وعلى مبنى العمارة !!!! تشوه بصري	Billboards
	لوحة ممزقة وملصقات دعائية في وسط الحي السكني وعلى مبنى العمارة تشوه بصري	
	Tattered signboards and advertising posters are prevalent in the residential area and on building facades, causing visual distortion.	
8	باصات قديمة بدون لوحات متوقفة داخل الاحياء السكنية.....	Old assets
	باصات قديمة بدون لوحات متوقفة داخل الاحياء السكنية	
	"Old buses without license plates are parked in residential areas."	

B. Dataset Processing

Various methods were employed to process the dataset, all aimed at enhancing the accuracy of machine learning and deep learning models in classifying Arabic texts more effectively. Removing punctuation—including periods, question marks, exclamation points, commas, colons, semicolons, dashes, hyphens, brackets, braces, parentheses, apostrophes, and quotation marks—provides several benefits. These benefits include reduced noise, improved tokenization performance, enhanced model accuracy, a unified text format, and better text representation as vectors. Additionally, removing extra spaces and links improves readability, boosts tokenization results, reduces data complexity, increases model efficiency, and focuses analysis on relevant content.

Stop words are common words in Arabic that frequently appear in the text but contribute little to its overall meaning. Removing these stop words can reduce data size, minimize noise, improve accuracy in text classification, and facilitate keyword extraction. The NLTK library in Python is often used to implement stop word removal, streamlining the text for more effective processing and analysis.

Additionally, removing tabs enhances data consistency, aids in understanding, and reduces errors. Eliminating numbers from the text is also an essential part of data processing, as they can add complexity. By removing numbers, the data is purified, which enhances the algorithms used, increases analysis accuracy, standardizes the data, facilitates comprehension, and improves extraction processes.

Feature extraction techniques in Natural Language Processing (NLP) involve transforming raw data into features usable in machine learning (ML) and deep learning (DL) models. There are several methods available for converting text into vector representations for numeric analysis.

Optimal hyperparameters are external settings for ML and DL models that control their behavior and learning processes. Various algorithms can be utilized to tune these hyperparameters to achieve the best results.

The Grid Search technique is commonly employed to find the optimal set of hyperparameters for a model in ML and DL. It systematically tests a predefined set of hyperparameters and evaluates the model's performance for each combination using cross-validation. In contrast, Random Search selects random combinations of hyperparameters for evaluation. This approach significantly reduces computation time while still exploring a wide range of values, and it often proves to be more efficient than Grid Search.

IV. Results and Analysis

This section discusses the outcomes of applying various text classification models to a dataset containing requests and suggestions from citizens and residents about municipal services in the Kingdom of Saudi Arabia. The models' performance was evaluated using the F1 score to identify the most effective algorithm for classifying different topics, such as lighting, construction waste, neglected assets, roads, visual pollution, billboards, and cleanliness. Our results indicate significant improvements in the performance of both machine learning (ML) and deep learning (DL) models, demonstrating the benefits of hyperparameter optimization combined with dataset pre-processing. Furthermore, the robustness of these models was evident even without pre-processing, further highlighting their applicability to text classification tasks.

TABLE IV. Values of Hyperparameter of ML Models

Models	Hyperparameters
Logistic Regression	solver='newton-cg',penalty='none',max_iter=1000)
Naive-Multinomial	alpha=0.1
Naive-Complement	alpha=1.20
Naive-Bernoulli	alpha=0.001
Random Forest	n_estimators=86
SVC	C=1000, gamma= 0.01, kernel= 'rbf'
Linear SVC	C=2,penalty='l2'
K-Neighbors	(n_neighbors=7)
Decision Tree	criterion='gini',max_depth=100,splitter='random',min_samples_split=2
SGD	(penalty='l2')
Bag(LR)	solver='newton-cg',penalty='none',max_iter=1000),n_estimators=30
Bag(SVC)	SVC(C=1000, gamma= 0.2, kernel= 'rbf'),n_estimators=50))
Bag(K-NN)	n_estimators=80
Bag(SGD)	n_estimators=40
AdaBoost	n_estimators=80,learning_rate=0.01,algorithm="SAMME.R"
XGB	n_estimators=50,gamma=0.5
Gradient Boosting	n_estimators=125,max_depth=7
Vot(RF-LR)	estimators=[('lr',LogisticRegression(solver='newton-cg',penalty='none',max_iter=1000)),('rf',RandomForestClassifier(n_estimators=86))],voting='soft'))
Vot(L-SVC,SVC)	estimators=[('ls',LinearSVC(C=2,penalty='l2')),('sv',SVC(C=1000, gamma= 0.01, kernel= 'rbf'))],voting='hard'

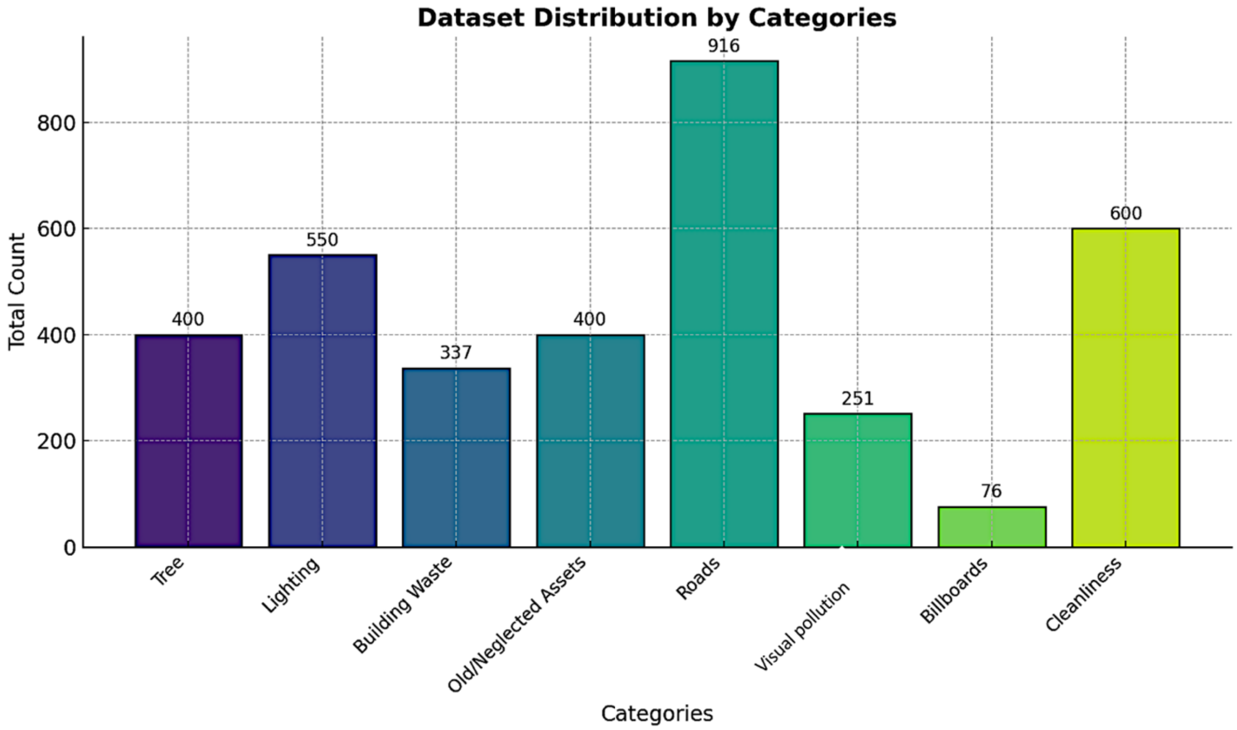

Fig. 3. The different classes are included in the dataset.

A. *Performance Metrics*

We utilized the following metrics to assess the performance of the model:

- Accuracy: The number of data instances correctly classified from the total dataset is a crucial measure of a model's effectiveness in categorization. This metric is particularly significant for binary classification problems when the dataset is balanced. However, when the dataset is unbalanced, it is important to consider additional metrics alongside accuracy, such as recall, precision, and F1 scores.

$$Accuracy = \frac{True\ Positive(TP)+True\ Negative(TN)}{Total\ Number\ of\ Sample} \quad (1)$$

- Precision: The ratio of correctly predicted positive instances to the total predicted positives, reflecting the model's ability to identify relevant classes.

$$Precision = \frac{True\ Positive(TP)}{True\ Positives(TP)+False\ Positives(FP)} \quad (2)$$

- Recall: The proportion of actual positive instances that are correctly identified by the model, which measures its sensitivity.

$$Recall = \frac{True\ Positive(TP)}{True\ Positives(TP)+False\ Negatives(FN)} \quad (3)$$

- F1 Score: A metric that combines precision and recall into a single score, ranging from 0 (worst) to 1 (best).

$$F1 = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (4)$$

TABLE V. F1 SCORES OF ALL MODELS UNDER PROCESSED DATASET CONDITIONS

Models	F1_Score using optimal hyperparameters	F1_Score using default hyperparameters
LR	**0.901734104**	**0.865606936**
Bagging (LR)	0.901734104	0.855491329
Bagging (SVC)	0.8959537	0.8439306
MNB	0.878612716	0.791907514
RF	0.875722543	0.85982659
XGB	0.854046243	0.852601156
Bagging(K-NN)	0.835260116	0.820809249
BNB	0.862716763	0.556358382
K -NN	0.832369942	0.826589595
DT	0.822254335	0.819364162
AdaBoost	0.423410405	0.39017341
RNN	0.816511235	0.816511235

B. *F1 Scores with Processed Dataset*

The F1 score, which balances precision and recall, is a critical metric for evaluating the classification of imbalanced datasets. In this context, both Logistic Regression and its ensemble counterpart play significant roles.

We aimed to identify the optimal or near-optimal hyperparameters for the classification techniques used; the results of this phase are detailed in Table IV.

As indicated in Table V, bagging with Logistic Regression achieved the highest F1 score of 90.17% under pre-processed conditions. While deep learning models like RNN performed reasonably well, obtaining an F1 score of 81.16%, their performance was still marginally lower than that of the best-performing machine learning models. Overall, models that utilized pre-processing and optimal hyperparameters with machine learning algorithms—such as Logistic Regression, Bagging (LR), and Bagging (SVC)—consistently achieved superior F1 scores. This demonstrates that machine learning algorithms outperformed deep learning models when evaluated using the F1 metric.

C. F1 Scores without Preprocessing the Dataset

The F1 scores reflect the results achieved using the dataset without any pre-processing, while applying optimal hyperparameters, as detailed in Table IV. This method emphasizes the model's ability to effectively balance precision and recall, even when dealing with the challenges posed by unprocessed data. Evaluating the F1 scores in this way provides insight into the model's capacity to maintain accuracy and reliability while working with raw data and appropriately tuned parameters.

As illustrated in Table VI, Linear SVC, along with Bagging using Linear SVC and Voting (combining Linear SVC and SVC), showed superior accuracy rates of 91.04%, 90.46%, and 90.08%, respectively, when hyperparameter optimization was performed without any pre-processing. Among deep learning models, CNN and LSTM achieved noteworthy accuracy rates of 88.43% and 87.81%, respectively, although they did not perform as well as some of the specific machine learning models.

TABLE VI. F1 SCORES OF ALL MODELS UNDER UNPROCESSED DATASET CONDITIONS

Models	F1_Score using optimal hyperparameters	F1_Score using default hyperparameters
Linear SVC	**0.910404624**	**0.913179191**
SGD	0.901734104	0.900289017
Bagging (Linear SV)	0.904624277	0.897398844
Bagging(SGD)	0.907514451	0.904624277
Voting(LR,RF)	0.904624277	0.875722543
Voting(LinearSVC,SVC)	0.908959538	0.89017341
CNB	0.894508671	0.891618497
SVC	0.893063584	0.856936416
Gradient Boosting	0.845375723	0.841040462
CNN	0.884338846	0.83706596087
LSTIM	0.878148111	0.863887078
BiLSTM	0.871606769	0.86308049

V. DISCUSSION

The results highlight the impact of hyperparameter optimization and dataset processing techniques on the performance of machine learning (ML) and deep learning (DL) models when evaluating requests and suggestions from citizens regarding various municipal services. These services include categories such as trees, lighting, construction waste, neglected assets, roads, visual pollution, billboards, and cleanliness.

The findings indicate that ML models consistently outperformed DL models when using optimized hyperparameters and processed datasets. However, DL models showed performance levels that were close to those of ML models, particularly when hyperparameters were finely tuned without preprocessing the datasets.

Several factors contributed to the enhanced accuracy of ML models in classifying these requests and suggestions. Notably, ML models can effectively learn from smaller datasets due to their reliance on feature-based methods, while DL models typically require larger datasets to capture and model complex patterns effectively.

In summary, the results demonstrate that when utilizing optimized hyperparameters and processed datasets, machine learning models achieved higher accuracy compared to deep learning models. Conversely, deep learning models exhibited competitive performance levels, particularly when hyperparameters were finely tuned without preprocessing. The ability of machine learning models to learn effectively from smaller datasets is a significant factor in their accuracy in classifying requests related to municipal services.

VI. CONCLUSION

This study explored the application of machine learning (ML) and deep learning (DL) models to classify Arabic-language text related to municipal services in Saudi Arabia. To address the scarcity of Arabic-language datasets, a collection of 3,714 requests and suggestions was compiled. This dataset will be made publicly available to encourage the community's adoption of ML and DL technologies in the field of municipal services. A baseline for classification performance was established to guide future improvements.

The findings are significant: ML models, including Logistic Regression and Linear Support Vector Classifier (Linear SVC), demonstrated superior performance, achieving a maximum accuracy of 91.04% and an F1 score of 90.17%. These results were obtained through the application of data preprocessing techniques and hyperparameter optimization. Although DL models, such as Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM), also performed well, they did not surpass the ML models. This difference can be attributed to the fact that DL models are better suited for longer texts, while ML models excel with shorter texts, which is characteristic of the dataset used in this study.

The study emphasizes the critical role of data preprocessing and hyperparameter tuning in enhancing classification accuracy. By introducing this dataset and sharing key insights, this work establishes a foundation for future research and development in natural language processing (NLP) for the Arabic language. Future efforts could focus on expanding the dataset, integrating advanced embedding techniques and large language models, and designing dynamic classification systems for real-time applications.

Future research can explore several enhancements to improve classification robustness and generalizability. Expanding the dataset with more samples will strengthen model performance, while incorporating advanced embedding techniques, such as BERT or FastText, can refine text representation and capture nuanced meanings more effectively. Developing real-time classification systems could significantly enhance municipal responsiveness and service efficiency. Another valuable enhancement is integrating multimodal data—combining text, images, and audio—to achieve a more comprehensive understanding of citizens' needs. Lastly, implementing continuous feedback mechanisms will enable model refinement, ensuring adaptability to evolving demands.

ACKNOWLEDGMENT

This research is funded by the Deanship of Scientific Research at the Islamic University of Madinah, Madinah, KSA.

REFERENCES

[1] Rincy, T. N., & Gupta, R. (2020, February). Ensemble learning techniques and its efficiency in machine learning: A survey. In 2nd international conference on data, engineering and applications (IDEA) (pp. 1-6). IEEE.

[2] Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. IEEE access, 7, 53040-53065.

[3] Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021, November). Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. In Informatics (Vol. 8, No. 4, p. 79). MDPI.

[4] Bahbib, M., & Yakhlef, M. B. (2024, May). Effects of applying stemmers with word representation on Arabic text classification performance using deep learning methods. In 2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET) (pp. 1-6). IEEE.

[5] Umer, M., Imtiaz, Z., Ahmad, M., Nappi, M., Medaglia, C., Choi, G. S., & Mehmood, A. (2023). Impact of convolutional neural network and FastText embedding on text classification. Multimedia Tools and Applications, 82(4), 5569-5585.

[6] Alshehri, K., Alhothali, A., & Alowidi, N. (2024). Arabic Tweet Act: A Weighted Ensemble Pre-Trained Transformer Model for Classifying Arabic Speech Acts on Twitter. arXiv preprint arXiv:2401.17373. DOI: https://doi.org/10.48550/arXiv.2401.17373.

[7] Al-Fuqaha'a, S., Al-Madi, N., & Hammo, B. (2024). A robust classification approach to enhance clinic identification from Arabic health text. Neural Computing and Applications, 36(13), 7161-7185. https://doi.org/10.1007/s00500-023-06223-7.

[8] Onan, A., Korukoğlu, S., & Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. Expert Systems with Applications, 57, 232-247.

[9] Ghourabi, A., & Alohaly, M. (2023). Enhancing spam message classification and detection using transformer-based embedding and ensemble learning. Sensors, 23(8), 3861.

[10] Al-Qerem, A., Raja, M., Taqatqa, S., & Sara, M. R. A. (2024). Utilizing Deep Learning Models (RNN, LSTM, CNN-LSTM, and Bi-LSTM) for Arabic Text Classification. In Artificial Intelligence-Augmented Digital Twins: Transforming Industrial Operations for Innovation and Sustainability (pp. 287-301). Cham: Springer Nature Switzerland.

[11] Hariyadi, B., & Lhaksmana, K. M. (2024, February). Topic Classification of Arabic Text Using Majority Voting. In 2024 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS) (pp. 1-5). IEEE.

[12] Ghourabi, A., & Alohaly, M. (2023). Enhancing spam message classification and detection using transformer-based embedding and ensemble learning. Sensors, 23(8), 3861.

[13] El Rifai, H., Al Qadi, L., & Elnagar, A. (2022). Arabic text classification: the need for multi-labeling systems. Neural Computing and Applications, 34(2), 1135-1159.

[14] Umer, M., Imtiaz, Z., Ahmad, M., Nappi, M., Medaglia, C., Choi, G. S., & Mehmood, A. (2023). Impact of convolutional neural network and FastText embedding on text classification. Multimedia Tools and Applications, 82(4), 5569-5585.

[15] Chowdhury, S., & Schoen, M. P. (2020, October). Research paper classification using supervised machine learning techniques. In 2020 Intermountain Engineering, Technology and Computing (IETC) (pp. 1-6). IEEE.

[16] Nassif, A. B., Elnagar, A., Elgendy, O., & Afadar, Y. (2022). Arabic fake news detection based on deep contextualized embedding models. Neural Computing and Applications, 34(18), 16019-16032.

[17] Elnagar, A., Al-Debsi, R., & Einea, O. (2020). Arabic text classification using deep learning models. Information Processing & Management, 57(1), 102121.

[18] Alsaleh, D., & Larabi-Marie-Sainte, S. (2021). Arabic text classification using convolutional neural network and genetic algorithms. IEEE Access, 9, 91670-91685.

[19] Alzanin, S. M., Azmi, A. M., & Aboalsamh, H. A. (2022). Short text classification for Arabic social media tweets. Journal of King Saud University-Computer and Information Sciences, 34(9), 6595-6604.

[20] Galal, O., Abdel-Gawad, A. H., & Farouk, M. (2024). Rethinking of BERT sentence embedding for text classification. Neural Computing and Applications, 36(32), 20245-20258.

[21] Mahmoudi, O., Bouami, M. F., & Badri, M. (2022). Arabic language modeling based on supervised machine learning. Revue d'Intelligence Artificielle, 36(3), 467.

[22] Kaddoura, S., Alex, S. A., Itani, M., Henno, S., AlNashash, A., & Hemanth, D. J. (2023). Arabic spam tweets classification using deep learning. Neural Computing and Applications, 35(23), 17233-17246.

[23] Alzanin, S. M., Gumaei, A., Haque, M. A., & Muaad, A. Y. (2023). An optimized Arabic multilabel text classification approach using genetic algorithm and ensemble learning. Applied Sciences, 13(18), 10264.

[24] Sabri, T., Bahassine, S., El Beggar, O., & Kissi, M. (2024). An improved Arabic text classification method using word embedding. International Journal of Electrical & Computer Engineering (2088-8708), 14(1).

[25] Alruily, M., Manaf Fazal, A., Mostafa, A. M., & Ezz, M. (2023). Automated Arabic long-tweet classification using transfer learning with BERT. Applied Sciences, 13(6), 3482.

[26] Saeed, R. M., Rady, S., & Gharib, T. F. (2022). An ensemble approach for spam detection in Arabic opinion texts. Journal of King Saud University-Computer and Information Sciences, 34(1), 1407-1416.

[27] Hassan, S. U., Ahamed, J., & Ahmad, K. (2022). Analytics of machine learning-based algorithms for text classification. Sustainable Operations and Computers, 3, 238-248.

[28] Alhaj, Y. A., Dahou, A., Al-Qaness, M. A., Abualigah, L., Abbasi, A. A., Almaweri, N. A. O., ... & Damaševičius, R. (2022). A novel text classification technique using improved particle swarm optimization: A case study of Arabic language. Future Internet, 14(7), 194.

[29] Alammary, A. S. (2022). BERT models for Arabic text classification: a systematic review. Applied Sciences, 12(11), 5720.

[30] Almaliki, M., Almars, A. M., Gad, I., & Atlam, E. S. (2023). Abmm: Arabic bert-mini model for hate-speech detection on social media. Electronics, 12(4), 1048.

[31] Mujahid, M., Kanwal, K., Rustam, F., Aljedaani, W., & Ashraf, I. (2023). Arabic ChatGPT tweets classification using RoBERTa and BERT ensemble model. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(8), 1-23.

[32] Mardiansyah, K., & Surya, W. (2024). Comparative Analysis of ChatGPT-4 and Google Gemini for Spam Detection on the SpamAssassin Public Mail Corpus.

[33] Alhaj, F., Al-Haj, A., Sharieh, A., & Jabri, R. (2022). Improving Arabic cognitive distortion classification in Twitter using BERTopic. International Journal of Advanced Computer Science and Applications, 13(1), 854-8