Mesopotamian journal of Cybersecurity
Vol.5,No.2, **pp**. 337–348
DOI: https://doi.org/10.58496/MJCS/2025/021; ISSN: 2958-6542
https://mesopotamian.press/journals/index.php/cybersecurity

Research Article

Measuring the Extent of Cyberbullying Comments in Facebook Groups for Mosul University Students

Kanaan J. Brakas[1,*], Mafaz Alanezi[2]

[1] *Department of Computer Science, University of Mosul, Mosul, Iraq.*

[2] *ICT Research Unit, Computer Center, University of Mosul, Mosul, Iraq.*

ARTICLE INFO

Article History

Received 21 Aug 2024
Revised 10 Nov 2024
Accepted 07 Mar 2025
Published 11 May 2025

Keywords

Facebook Groups
Cyberbullying
Cybercrime
Text Mining
Logistic Regression (LR)

ABSTRACT

The widespread utilization of social media platforms such as Facebook, Twitter, and Instagram inside academic organizations has become fundamental for student correspondence and joint effort, yet it has simultaneously prompted an increase in cyberbullying incidents. Cyberbullying is embodied by offensive and harmful comments that sabotage the casualty's prosperity. This study plans to resolve this issue by fostering an extensive module to distinguish cases of cyberbullying inside Facebook groups of Mosul University students. Our methodology starts with the collection of data in the Arabic language, which is then exposed to careful manual handling to label comments and eliminate clamor-like copies and superfluous sections. The dataset, comprising 2,715 comments, goes through prehandling and element extraction via the term frequency-inverse document frequency (TF-IDF) strategy. Consequently, we utilize the logistic regression (LR) classifier to investigate and order the data. Our findings revealed that more than 21% of the comments dissected were brutal, a disturbing rate for a local area such as Mosul University. These outcomes highlight the critical requirements for intercession and the significance of observing student cooperation on social media. The dataset and experiences obtained from this study are important for psychologists and psychological direction experts to comprehend and relieve cyberbullying. This exploration not only highlights the prevalence of cyberbullying in academic settings but also provides a strong systemic structure for future examinations to expand, highlighting the basic role of AI in battling web harassment.

1. INTRODUCTION

Social media platforms are undergoing continuous evolution and garnering extensive acceptance among countless individuals worldwide daily. This extensive acceptance can be attributed to their capacity to facilitate interpersonal connections by eradicating geographical barriers among users and facilitating the instantaneous dissemination of information to a vast multitude of users [27]. Social media, a mode of computer-mediated communication, facilitates the connection of a considerable number of individuals residing in diverse locations. Those possessing the necessary technological tools and access to the internet have the opportunity to engage in social media and express their viewpoints [26].

The increasing utilization of social media on a global scale exposes a larger number of individuals to potential attacks from cybercriminals, who typically exploit their vulnerabilities and personal information. Cyberbullying, a type of cybercrime, entails intentionally insulting, threatening, embarrassing, or harassing individuals through internet-based applications on smartphones [35]. In other words, it involves using computers and the internet to damage people by spreading bullying, nasty, and harassing messages that diminish the victim's morale. The frequency at which digital criminal activity transpires leaves little room for successful human intervention without the occurrence of human error [29].

This particular manuscript focuses on the Arabic language because of its significant role in the cultural diversity of humanity. It is widely recognized as one of the most extensively used languages worldwide, with a user base exceeding 400 million individuals [28]. Nonetheless, there is a noticeable disparity in the amount of attention dedicated to Arabic natural language processing (NLP) compared with the efforts invested in the English language [36].

To construct an Arabic text classification framework, we want a marked corpus since we pick, in this paper, it for supervised learning. The greater the volume of this corpus is, the more productive the framework is. In the wake of picking the corpus, two primary stages must be performed, in particular, training and testing. The training stage incorporates partaking in the

*Corresponding author. Email: kjbbaby@gmail.com

corpus (80%), called the training set, and applies the classification cycle, which comprises, in turn, three substages: first, preprocessing that permits removing stop words, standardization, and so forth; then, at that point, highlight extraction that systematizes the Arabic message; and last, a machine learning calculation that is run on the training set to produce a machine learning model; the model addresses what was realized by the machine learning calculation. In the wake of training, the testing stage takes the excess piece of the corpus (20%), called the test set, and applies the classification cycle described above to look at the framework's presentation.

In this work, we collected a new dataset from Facebook groups for Mosul University students via APIFY. APIFY is an online tool used to scrape data from public Facebook groups[38]. The collected dataset was then manually processed to reduce noise by deleting repeated and meaningless comments. Next, we asked a group of students to label all comments on the dataset. The dataset contained 2715 comments.

The purpose of this proposal is to gather and analyse data on bullying, which will be utilized as input for a machine learning algorithm called logistic regression (LR), which offers multiple benefits, such as simplicity, efficiency, interpretable results, probability estimation, robustness, versatility, feature selection, and availability in various statistical software packages, thereby justifying its implementation in our proposal[39]. This dataset is relevant for psychologists and psychology guidance specialists interested in the field of psychology to observe patient behavior and analyse their cases.

The remainder of this paper is structured as follows: some significant related work is reviewed in Section 2. Our model technique is described in Section 3, the results and model analysis are presented in Section 4, and the conclusions and next steps are discussed in Section 5.

2. RELATED WORKS

Recently, the issue of cyberbullying has received considerable attention, especially in the context of educational organizations. Various investigations have investigated the prevalence, effects, and alleviation methodologies related to cyberbullying.

2.1 Classification techniques

Sundus et al., 2019 [2] proposed a supervised feed-forward deep learning technique that uses term frequency and inverse document frequency (TF-IDF) for frequently occurring words in datasets. The experimental results demonstrated that, compared with logistic regression, deep learning models significantly improved classification efficiency and model-building time, showing great promise for Arabic text classification.

Galal et al., 2019 [4] focused on classifying Arabic text by applying convolutional neural networks (CNNs). They introduced a novel algorithm called Gstem, which enhances the accuracy of the CNN model by adding extra Arabic letters and using word embedding space to group related words.

Tarik Sabri et al. (2022) [9] conducted a study on five classification approaches, including support vector machine (SVM), decision tree (DT), and random forest (RF), utilizing two Arabic datasets. They implemented three feature vectorization approaches to convert text into numeric vectors, including word count, TF-IDF, and word embedding.

2.2 Deep learning approaches

Elnagar et al., 2020 [3] introduced large corpora, SANAD and NADiA, for Arabic text classification. They employed word-to-vector embedding methods and reported that the attention-gated GRU was the most efficient model, particularly for the SANAD corpus.

Alhawarat et al., 2020 [5] used a multicore CNN framework combined with n-gram methods for Arabic news document classification. This approach, known as the Superior Arabic Text Categorization Deep Model (SATCDM), demonstrated substantial improvements in accuracy.

Gwad et al., 2020 [6] utilized Arabic user comments on Twitter and reported that applying LSTM in recurrent neural networks (RNNs) ensured high accuracy and efficiency, outperforming traditional pattern recognition techniques.

Fati et al. (2023) [11] compared several deep learning techniques on a well-known Twitter dataset. Their attention-based deep learning models, combined with word-to-vector concatenation and CBOW, effectively recognized offensive tweets.

2.3 Cyberbullying detection

Balakrishnan et al. [8] proposed an approach for detecting cyberbullying by analysing the psychological traits of Twitter users, such as emotion and sentiment, via the Big Five and Dark Triad approaches. They achieved better detection rates by combining sentiment and personality categories.

Aliyu et al., 2020 [10] developed a theoretical framework for predicting instances of cyberbullying within Facebook posts via naive Bayes and text mining. Their method effectively identified bullying-related content through various metrics, including confusion matrices and accuracy scores.

Bozyiğit et al., 2021 [12] assessed the impact of eight different artificial neural network (ANN) algorithms for detecting cyberbullying in the Turkish language. Their study achieved a high F1-measure score, surpassing the effectiveness of current machine learning classifiers.
Fortunatus et al., 2020 [24] identified cyberbullying on social media platforms by combining textual characteristics with a lexicon-enhanced rule-based approach, focusing specifically on Facebook comments.

2.4 Sentiment Analysis

El Rifai et al. (2022) [19] introduced a method to automatically assign tags to news articles via large datasets from Arabic news portals. Their ensemble model incorporated ten shallow learning classifiers, improving accuracy in classifying single-label and multilabel datasets.
Badawi et al. (2023) developed a Twitter sentiment analysis system using the Light Gradient Boosting Machine (LightGBM) and Apache Spark environments. Their system applied an attention-based bidirectional gated recurrent unit (Bi-GRU) CapsNet model to classify sentiment in tweets, achieving superior performance.
Gwad et al., 2020 [25] suggested classifying sentiment in Arabic tweets via long short-term memory (LSTM), a type of RNN, leveraging the unique morphological structure of the Arabic language to increase analysis accuracy.
Seyala & Alanezi, 2021 [37] highlighted the impact of social media on user interaction, especially on platforms such as Facebook. By using data visualization and clustering algorithms, studies have identified user behavior patterns, aiding targeted advertising, managing inappropriate content, and helping users, particularly youth, better manage their social media time.

3. METHODOLOGY

This section outlines the framework and dataset used to detect bullying on the Facebook platform. Text mining is a knowledge acquisition method that involves several required steps to uncover invisible information. These procedures are illustrated in Figure 1.

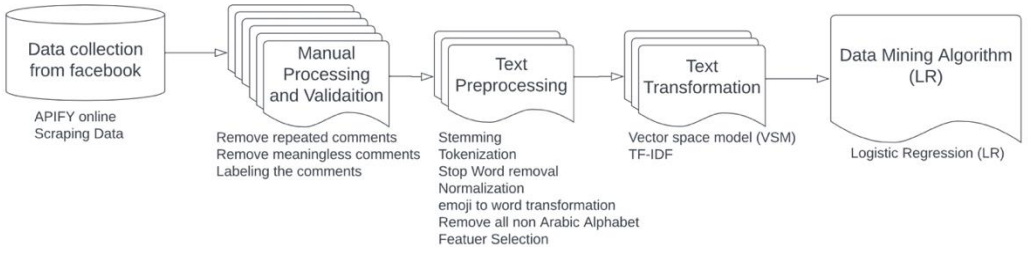

Fig. 1. Proposed Architecture.

3.1 Pseudocode

The following pseudocode represents the main steps in preparing the dataset, training the model, and using it to detect cyberbullying comments:

START
1. Load raw data from the Facebook comment dataset.
2. Preprocessing Data:
a. Remove duplicates and irrelevant text.
b. Convert emojis to text sentiment (e.g., smiley -> "happy").
c. Tokenize each comment into individual words.
d. Perform stemming to reduce words to their root form.
e. Remove stop words (common Arabic words).
f. Normalize text (standardize character variations).
g. Convert text into numerical representation using TF-IDF.
3. Label Data:
a. Assign labels (1 for bullying, 0 for nonbullying) on the basis of manual classification.
4. Split Data:
a. Data are split into training (80%) and testing (20%) sets.
5. Training Model:
a. Initialize the logistic regression model.
b. Train the model via the training set with TF-IDF features and labels.
6. Evaluating Model:
a. Use the test set to predict labels for each comment.
b. Calculating evaluation metrics (accuracy, precision, recall, F1 score).
7. Predict New Comments:
a. Preprocess new comment text following the steps above.
b. The trained model is used to predict whether a comment is bullying or nonbullying.
c. Output the prediction.
END

3.2 Data collection

To acquire unprocessed information from Facebook, it is necessary to utilize an application programming interface (API). This API, known as the Facebook graph API, facilitates the collection of raw and unstructured data. Access to this application programming interface (API) necessitates an active Facebook account as well as a token for authentication, which provides permission to utilize the Facebook APIs. The Facebook Graph API facilitates interaction with the social graph, an extensive data structure that represents social connections, consisting of nodes and the relationships between them. A program named APIFY was used to extract data from the Facebook Graph API. Using the Facebook page IDs of different pages, the textual data were gathered from public posts and comments. A total of more than 2720 (after manual processing) data entries were gathered from two separate Facebook groups and subsequently utilized to enhance the accuracy and efficiency of predictive analysis.

The selection of Facebook groups for students at Mosul University was based on the type of debate that was taking place and the content of the posts and comments. This type of data shows the emotional agitation that users express in their posts or in the communiqués that these users provide, making it more effective in interpreting the acts of malevolent online actors. The retrieved data were then converted to a CSV format and stored in a dataset.

3.3 Manual processing and validation

The collection of data consists of unprocessed information that often includes repeated and meaningless data, which can negatively impact the accuracy of our classification results. Consequently, manual processing is necessary. To utilize supervised machine learning, our data must be labelled correctly, requiring the assistance of university students for this task. We used some students from the university for this purpose and validated the labelling. We dedicated fourteen days to gathering more than five thousand comments from three distinct public communities, comprising students from Mosul University. Following a meticulous manual analysis of these data, as shown in Figure 2, our results yielded a total of 2720 comments, which were duly classified by students and confirmed by our team. Among these comments, 2137 were identified as normal comments, constituting approximately 78.6% of our dataset. Conversely, 583 comments were categorized as bullying, accounting for the remaining 21.4%.

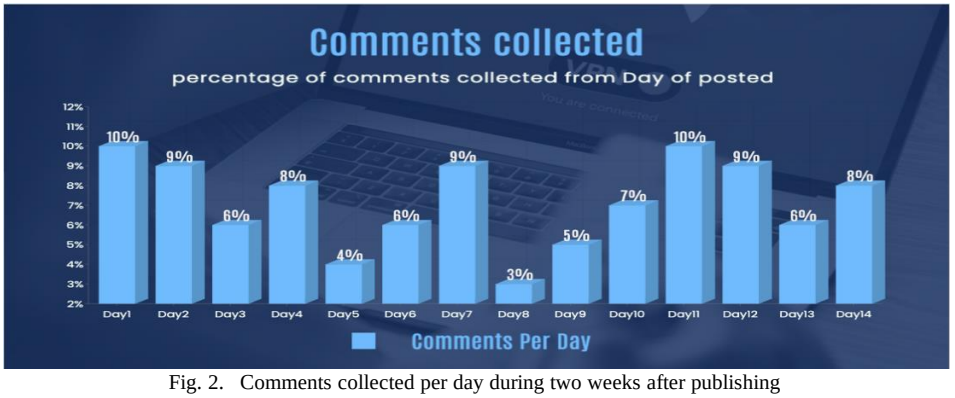

Fig. 2. Comments collected per day during two weeks after publishing

3.4 Data analysis

We analysed the dataset for our research, focusing on two distinct groups affiliated with the University of Mosul. The initial group consisted of undergraduate students from the College of Computer Science and Mathematics (CSUgS), whereas the second group consisted of graduate students from the same college (CSGS). After processing and labelling the CSUgS dataset, we obtained a total of 1700 comments, of which 364 comments (21.4%) were identified as instances of bullying, whereas 1336 comments (78.6%) were classified as normal. Similarly, from the CSGS dataset, we collected 1020 comments, among which 51 comments (5%) were classified as bullying and 969 comments (95%) were categorized as normal (see Fig 3). These comments focused on various post topics, as shown in Figure 4, including the following:

- Announcements related to academic information and vocabulary.
- Expressions of opinions on social issues and events.
- Governmental instructions and decisions pertaining to students.
- Local and global political matters.
- Discussing current public issues.
- In addition, there are other miscellaneous topics.

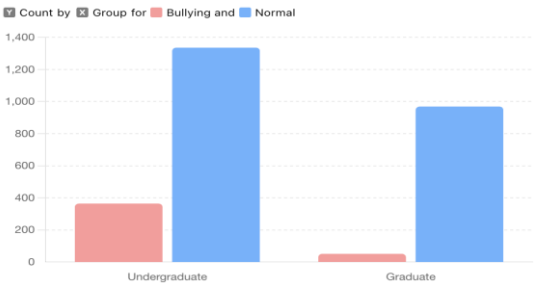

Fig. 3. Comments Distribution By Group

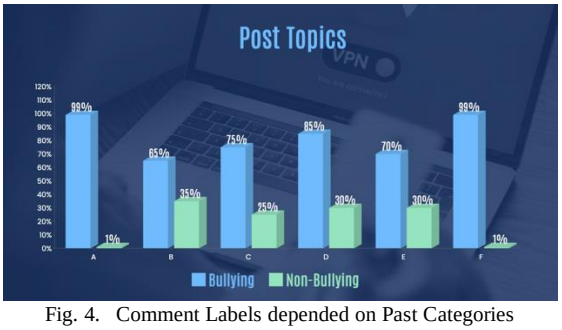

Fig. 4. Comment Labels depended on Past Categories

3.5 Data Preprocessing

Text preprocessing is a pivotal and indispensable stage in the examination of textual data. It encompasses a meticulous purification and arrangement of the text to facilitate its classification. The data, which are stored in a Comma-Separated Values (.CSV) configuration, undergo preprocessing via Python 3 on the PyCharm platform. The fundamental hypothesis behind the appropriate preprocessing of the data is to alleviate any superfluous components within the text, thus augmenting the overall effectiveness of the classification model and expediting the classification procedure. Consequently, this contributes to the efficient execution of real-time sentiment analysis. Notably, data derived from Facebook are typically characterized by incompleteness, noise, and inconsistency. Incomplete data regularly consists of errors and anomalies, in addition to missing values for determined attributes [30]. The primary procedure is as follows:

1- **Tokenization**: the operation of dividing texts into single words is known as tokenization. This operation includes eliminating unnecessary blank spaces and punctuation marks. Tokenization is executed by analysing the sentences and outputting a compilation of tokens. These tokens turn into the initial input for subsequent algorithms [30]. The primary aim of tokenization lies in the recognition of words within a sentence [31].

2- **Stemming**: This operation involves making a word's various forms uniform [31]. It deletes all the affixes (suffixes, prefixes, infixes, and circumfixes). The theorem behind stemming is that words with the same root can be uniformed by using stems.

3- **Stop word removal**: The aim of deleting stop words is to eradicate unwanted words that are not meaningful information or specify the topic of the document being operated. Furthermore, stemming and lemmatization depend on grouping group words that are related in terms of their morphology and semantics [34].

4- **Normalization**: This operation works on the conversion of a group of words into a more standardized sequence, which removes punctuation and unifies all different alphabet shapes into a single form, such as ("ۀ" ,"ہ") substituted by ("ہ") and ("ۂ" ,"ۃ" ,"ۄ" ,"ۅ" ,"ۆ") substituted by ("ۇ"). It very well may be seen as the system of changing over text into a solitary typical structure that it might not have had beforehand. It is used in the cleaning of text created by clients with the end goal of examination and forecasting [33]. Although not invariably imperative, it holds significant value owing to the nature of the information we aim to obtain.

5- **Transfer emojis to words**: Emoji sentiment is one of the most important pieces of content within comments on social media, so ignoring it affects the accuracy of classification. The dataset utilized in this study consists of Arabic emojis, as documented in the Arabic emoji dataset. This dataset encompasses over 1500 emojis exhibited in the Arabic language. These emojis are categorized into sentiments such as love, neutral face, smile face, and sad face. For our classification purposes, we solely require these specific categories, thus eliminating other categories. In addition, the Unicode representation and grouping of the emojis, including arrows, geometric shapes, emotions, animals, and symbols such as "<3", "♥," and '❤,' as well as the sequence "\(, " \[, “and ":),” are provided. Furthermore, the overall sentiment of each emoji is determined and

presented in a ranking. Consequently, these emojis are transformed into suitable words and integrated into the comments for classification.

6- **Remove all non-Arabic Alphabet**: In the preprocessing stage before selection, the features ensure that only Arabic Alphabet text remains. Therefore, all non-Arabic alphabets, such as English, French, and Chinese, and digits and any other symbols should be deleted.

7- **Feature Selection**: The matrix that represents the relationships between documents and terms is known as the document-term matrix. The rows of this matrix correspond to the documents, whereas the columns represent the terms. These terms, which can also be referred to as features, serve as unique identifiers for each document. During this particular step, features that are deemed irrelevant for data mining are eliminated. This process offers several benefits, including a reduction in the size of the dataset, fewer computations, and a minimized search space. When employing a feature selection technique, it is assumed that the data containing irrelevant and redundant features is crucial. These features do not contribute any additional or pertinent information to the document.

3.6 Data Preprocessing

This is the phase in which the data are structured and the appropriate technique for the data mining task is selected. Message change makes it conceivable to transform documents into a vector space document model portrayal that might be utilized in the following stage, or it can make them into a bag of words. Every term that appears in the document is typically given a numerical value via the vector space model. Moreover, the vector space model allows for the representation of vector space documents as variables with numerical values, indicating the significance of words within the documents. A collection of documents is transformed into vectors in the vector space model. A vector di, where di = (W1,i,W2,i...,Wn,i), can be utilized to address document i. Wj,i is the heaviness of word j in document i, and n is the quantity of words utilized for vectorization. In the TF-IDF scheme (term frequency-inverse document frequency), weight wj,i = tfj I × idfj, where tfj i is the recurrence of word j in document i.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	comment	class	word_count	char_count	avg_char_per_word	stopwords	emoji_count	noStop_comment	normalized_comment	text	noPunc_comment	cleanEmoji_comment	clean_comment	Lable
2	تكون مخلص	Non-Offen:	21	113	4.181818182	3	7	شعور انك تكون مخلص درا	دراستك وما عندك شيء متراكم	شعور انك	حلو بس بعمرك ما جربتوا#	مة فرح فرح فرح فرح #فهودي#	و بس بعمرك ما جربتوا#	1
3	الفقر مو عيب ولا	Offensive	19	93	3.947368421	4	0	الفقر مو عيب حرام شي منته	الفقر مو عيب ولا حرام هذا شي منا	فقر مو ع	الفقر مو عيب ولا حرام هذا	الفقر مو عيب ولا حرام هذا شي م	الفقر مو عيب ولا حرام ه	0
4	دايما شايل	Non-Offen:	47	243	4.108695652	7	0	الزوجة إنت دايما شايل صور	معاك و رايح بيها المكتب ..ليش؟	زوجه انت	الزوجة إنت دايما شايل صو	الزوجة : إنت دايما شايل صورتى	الزوجة : إنت دايما شايل	1
5	شفت زوجتك	Offensive	37	200	4	5	0	عزيزي الزوج شفت زوجتك	معصبه وجايه تتعارك وياك	عزيز زوج	عزيزي الزوج إذا شفت زوجت	عزيزي الزوج ؛ إذا شفت زوجتك	عزيزي الزوج ؛ إذا شفت	0
6	بالفيس	Non-Offen:	7	60	4.363636364	0	0	وتشوف عنوانها بالفيس أنوث	انوثتى سر جمالى	تشوف ع	وتشوف عنوانها بالفيس أنوث	وتشوف عنوانها بالفيس أنوثتى سر	وتشوف عنوانها بالفيس أ	1
7	معدلك 60 اذآ انتا	Non-Offen:	8	37	3.75	1	0	معدلك 60 اذآ انته تصير رئيس	معدلك 60 اذا انته راح تصير رئيس	معدل اذا	معدلك 60 اذآ انته راح تصير	معدلك 60 اذآ انته راح تصير رئيس	معدلك اذآ انته راح تصير	1
8	قرأت وسمعت	Non-Offen:	180	1060	3.884792627	39	0	اجمل قرأت وسمعت القصائد	(القصائد الشعبيه في مدح الفاروق	اجمل قرا	من اجمل ما قرأت وسمعت	قصائد الشعبية فى مدح الفاروق)	لشعبية فى مدح الفاروق	1
9	الذي يعمل	Non-Offen:	208	1150	4.295774648	34	0	أترى الرجل يعمل خياطا أتست	خياطا ؟	اترى رجل	أترى ذلك الرجل الذي يعمل	أترى ذلك الرجل الذي يعمل خياط	أترى ذلك الرجل الذي يع	1
10	كرابى ايطالبون اد	Non-Offen:	4	23	5	0	0	كرابى ايطالبون اب عطله	كرابى ايطالبون اب عطله	رايب ايط	كرابى ايطالبون اب عطله	كرابى ايطالبون اب عطله	كرابى ايطالبون اب عطل	1
11	هايييييييييييي	Non-Offen:	1	37	37	0	0	هايييييييييييييي	هايي	ها	هاييييييييييييييييي	هايييييييييييييييييييي	هاييييييييييييي	1
12	......عاجـل......	Offensive	30	265	5.30952381	4	0	عاجـل .... عاجـل ...... عاج	لقد تم اعلان يوم غد الاحد	عاجل عا	عاجلعاجلعاجللقد تم أعلان	......عاجـل......عاجـل....عاجـل	عاجـل....عاجـل......عا	0
13	كاعدة بمكان عام	Non-Offen:	27	187	4.606060606	3	0	مصيبه اذا البنت كاعدة بمكا	عام	مصيب ا	مصيبه اذا البنت كاعدة بمكا	مصيبه اذا البنت كاعدة بمكان عا	مصيبه اذا البنت كاعدة ب	1
14	هههههه فوائد الص	Non-Offen:	4	28	6.25	0	0	هههههه فوائد الصوبه العراق	هه فوائد الصوبه العراقيه	هه وايد	هههههه فوائد الصوبه العراق	هههههه فوائد الصوبه العراقيه	هههههه فوائد الصوبه ال	1
15	مُتنكّر بآسْم بنتٌ	Offensive	10	73	5.727272727	0	0	يآربُ أيٌ ولُد مُتنكّر بآسْم بنـ	ف الفيس بوك	يارب اي	يآربُ أيٌ ولُد مُتنكّر بآسْم بنتٌ	يآربُ أيٌ ولُد مُتنكّر بآسْم بنتٌ فـ	يآربُ أيٌ ولُد مُتنكّر بآسْم	0
16	اسلحه العالم	Non-Offen:	12	75	3.470588235	2	1	إحترامى لكل اسلحه العالم ما	ماكو شي قتلني غير عيونها ..	احترام لك	مع إحترامى لكل اسلحه العال	مع إحترامى لكل اسلحه العالم ما	مع إحترامى لكل اسلحه ا	1
17	كلبى ابلايه خوف	Non-Offen:	24	124	3.807692308	2	0	كنت تدفر باب كلبى ابلايه خ	لان تدري ابالى ماتكعد غرب !!	كنت تدف	كنت تدفر باب كلبى ابلايه خ	كنت تدفر باب كلبى ابلايه خوف	كنت تدفر باب كلبى ابلا	1

Fig. 5. Sample of Dataset features

3.7 Classification Method

The objective of the text classification system is to forecast the category or class of a recently introduced text. The process of text classification involves three primary stages: training, testing, and prediction. The Arabic text classification process begins by examining a labelled entire dataset. This dataset is divided into two subsets: the training set and the test set. The training set comprises 80% of the dataset, and the test set comprises more than 20% of the dataset (see Figures 5 and 6). The two sets choose their inputs randomly by adjusting their classifications. While the training set is utilized to become familiar with the models, the test set is utilized to assess the classifiers by processing their accuracies.

During the training stage, the texts go through a preprocessing step where numbers, stop words, punctuation, and non-Arabic words are removed and trailed via tokenization and normalization. The resulting feature matrix, which represents the input of an ML algorithm, consists of terms (words) and becomes a machine learning model after running the algorithm, thus forming a classifier.

Logistic regression (LR) is a predictive model that estimates an output value through the utilization of input values, similar to linear regression. Nonetheless, unlike linear regression, logistic regression produces binary or dichotomous output values. The Scikit-learn API [22] provides an implementation for two forms of logistic regression, explicitly regularized logistic regression (logistic regression) and logistic regression furnished with coordinated help for cross-validation (Logistic Regression CV). This is an anticipatory model, which is a statistical methodology employed for the purpose of classification. Despite the inclusion of the term 'regression' in the classifier's name, it is employed to generate discrete binary results.

3.8 Dataset Description

This dataset consists of Facebook comments in Arabic from groups associated with Mosul University students. These comments were manually labelled to identify instances of cyberbullying.

- **Dataset features**

The names of features and the description for each one are shown in table I.

TABLE I. DESCRIPTION OF DATASET FEATURES

Feature	Description
Comment Text	The main text content of each comment, written in Arabic.
Label	A binary label indicating if the comment is a bullying comment (1) or a normal comment (0).
Emoji Sentiment	Each comment's emojis are converted into sentiment categories (love, neutral, smile, sad) and included as words for the classification process.
Tokenization	Comments are broken down into individual words (tokens) for analysis.
Stemming	Words are reduced to their base or root form (e.g., removing suffixes or prefixes) to normalize variations of the same word.
Stop Word Removal	Commonly used Arabic stop words (e.g., "and," "the") are removed to reduce noise in the data.
Normalization	Arabic characters are standardized (e.g., different forms of "ا" "أ," etc., are unified) to reduce text complexity and enhance classification performance.
TF-IDF Vectorization	Term Frequency-Inverse Document Frequency (TF-IDF) is used to represent words numerically based on their importance across the comments.
Sentiment Analysis	Each comment's sentiment is analysed using both text and emoji content to help classify it as bullying or nonbullying.

- **Dataset size and training/test split**

The dataset contains 2,715 comments. It is divided into a training set and a test set, with 80% of the data allocated for training and 20% allocated for testing. The table II captures the size and split ratio, providing an organized view of the dataset distribution used for the cyberbullying classification model.

TABLE II. DATASET SIZE AND TRAINING/TEST SPLIT

Dataset Split	Count	Percentage
Total Comments	2,715	100%
Training Set	2,172	80%
Test Set	543	20%

4. EXPERIMENTAL ANALYSIS AND RESULTS

We maintain the conviction that the utilization of normalization on the assembled text is not an indispensable measure, notwithstanding the reality that most research undertakings in Arabic NLP tasks do execute normalization. The dataset includes an ample number of instances to appropriately portray the Arabic character set. Importantly, in specific instances, the normalization procedure may modify the significance of specific terms. Normalization is a widely embraced practice within the realm of Arabic computational linguistics. Its purpose revolves around the standardization of specific Arabic characters' orthography. More specifically, it entails the unification of alif forms [آ ، إ ، أ ، ا] into [ا], hamza forms (and yaa, [ه]) into [ة، ه] haa/taa marbootah, [ء] into [ء ، ؤ ، ئ], and alif maqsura [ي، ى] into [ى]. The primary objective of this normalization process is to reduce the scope of the vocabulary. Nonetheless, this procedure can potentially lead to the loss of certain significant attributes, as the meaning of certain words may undergo alterations following normalization. To illustrate, the term "فأر" (denoting "mouse") is altered to "فار" (meaning "escaped") postnormalization. Similarly, the word "كرة" (signifying "football") is transformed into "كره" (conveying "hatred") after normalization. The aforementioned shifts in meaning can potentially result in the exclusion of pivotal features. Furthermore, given the large corpus utilized in this study, there is no need for this preprocessing step. The results indicate that the absence of text normalization does not significantly impede the representation of non-normalized words.

The dataset is split into two subsets: 80% for training and 20% for testing. Applying the logistic regression (LR) classifier as the main classifier of the model of cyberbullying detection.

Evaluating the performance of the model was performed by implementing different evaluation metrics, which are [34]:

1- Accuracy score: The concept of accuracy refers to the degree of proximity between the measured value and the specific actual value of a given set. It can be mathematically represented by the following equation [34]:

$$\text{Accuracy} = TP + TN/TP + FN + TN + FP \quad (1)$$

2- Precision: On the other hand, precision pertains to the level of closeness among the measured quantities within the set. It is concerned with the proportion of retrieved items that are relevant to the classification results. The precision can be calculated via the following formula:

$$\text{Precision} = TP/TP + FP \quad (2)$$

3- Recall: also known as sensitivity in the context of information retrieval, it describes the fraction of successfully retrieved items that are relevant to the posted query. It can be expressed as:

$$\text{Recall} = TP/TP + FN \quad (3)$$

4- The F-measure is known as the F1 score and is a metric that is used to evaluate the accuracy of a test. It takes into account both precision and recall. The F1 score is obtained from the harmonic mean of precision and recall and is calculated via the following formula:

$$F1 = 2 * TP/(2 * TP + FP + FN) \quad (4)$$

5- Confusion matrix: A confusion matrix serves as a tabular representation employed for evaluating the efficacy of a classification algorithm. This matrix effectively illustrates and consolidates the outcomes of the classification algorithm. The outcome of classification is categorized into four separate groups.

$$\text{Confusion matrix} = [\ [\ [\ TP \quad FN\ ]\ [FP \quad TN\ ]\ ] \quad (5)$$

where true positive (TP) represents several correctly detected data, such as bullying; true negative (TN) represents many correctly detected data, such as normal text; false positive (FP) represents several falsely detected data, such as bullying; and false negative (FN) represents some falsely detected data, such as normal text. The outcomes of these metrics for our model are shown in Table III.

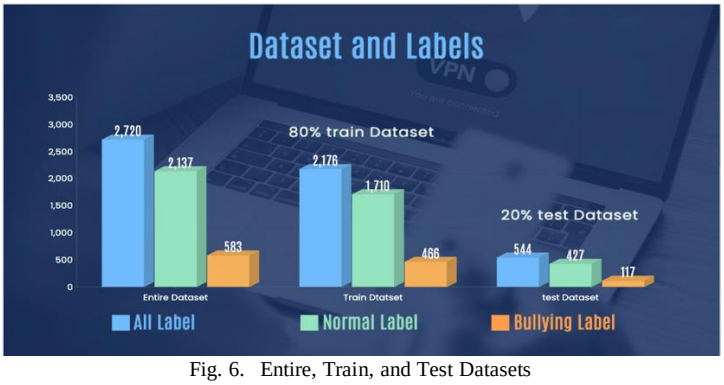

Fig. 6. Entire, Train, and Test Datasets

Figure 7 (a) Line graph showing the frequency of cyberbullying incidents over time. The graph shows a general upward trend in the frequency of cyberbullying incidents over the study period. Despite some fluctuations, the overall trend indicates an increase in cyberbullying incidents as the months progress. Figure 7 (b) shows the pie diagram illustrating the distribution of comment types.

TABLE III. EVALUATION METRIC VALUES OF OUR MODEL

Evaluation metrics value of our model.	**Accuracy Score is 0.84** **Confusion Matrix** **[[24 78]** **[11 431]]**			
	Precision	***Recall***	***F1-Score***	***Support***
0	0.69	0.24	0.35	102
1	0.85	0.98	0.80	442
Accuracy			0.84	544
Macro Avg	0.77	0.61	0.63	544
Weighted Avg	0.82	0.84	0.80	544

The table IV will present the parameters used in the cyberbullying detection analysis, covering the data preprocessing and model training phases.

TABLE IV. EVALUATION METRIC VALUES OF OUR MODEL

Parameter	**Description**	**Value**
Text Preprocessing		
Tokenization	Splits comments into individual words (tokens)	Applied
Stemming	Reduces words to their root forms	Applied
Stop Word Removal	Removes common Arabic stop words	Applied
Normalization	Standardizes Arabic characters	Applied

Emoji Sentiment Conversion	Converts emojis to text sentiment categories	Applied
Non-Arabic Character Removal	Removes non-Arabic characters, symbols, and numbers	Applied
Feature Extraction		
Vectorization Technique	Text vectorization using TF-IDF	TF-IDF
Maximum Features	Maximum number of features used for TF-IDF	5000
Data Splitting		
Train/Test Split Ratio	Proportion of data split for training and testing	80%/20%
Model Configuration		
Algorithm	Classification algorithm used for cyberbullying detection	Logistic Regression
Regularization	Regularization method used in Logistic Regression	L2 (Ridge)
Solver	Algorithm used to optimize the Logistic Regression model	lbfgs
Evaluation Metrics		
Accuracy	Overall correctness of model predictions	Calculated
Precision	Correctness of positive (bullying) predictions	Calculated
Recall	Sensitivity for true positive (bullying) detection	Calculated
F1 Score	Harmonic mean of Precision and Recall	Calculated
Confusion Matrix	Breakdown of TP, TN, FP, FN	Calculated
Hyperparameters for Optimization	Used to potentially improve model performance	
Learning Rate	Step size in optimization algorithm	Default for solver
Max Iterations	Maximum number of iterations for convergence	100

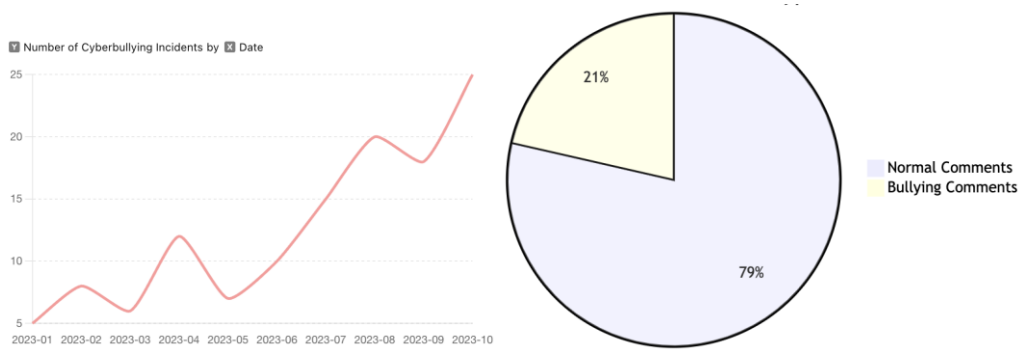

Fig. 7. The Results (a) Frequency of Cyberbullying Incidents Over Time, (b) The distribution of comment types

5. EXPERIMENTAL ANALYSIS AND RESULTS

While the study provides valuable insights into detecting cyberbullying among Mosul University students, several limitations should be considered:

A. Dataset limitations

- **Sample size**: The dataset comprises 2,715 comments, which may not fully represent the entire population or diverse types of cyberbullying encountered on other platforms or in different contexts. This limited dataset may affect the model's generalizability to other academic or nonacademic environments.
- **Language-Specific Bias**: The dataset is specific to Arabic text from a particular region and culture. Consequently, the model may not be generalizable to other languages or cultural contexts where cyberbullying might present differently.

B. Manual Labelling Constraints

- **Subjectivity in Labelling:** The labelling process relies on manual categorization by students, which can introduce subjective bias, particularly for ambiguous comments. This may impact the accuracy of the model and limit its ability to generalize.
- **Reliability of Human Annotation:** Since human annotators labelled the dataset, there may be inconsistencies or errors in identifying subtle instances of cyberbullying, especially if annotators interpreted certain terms differently.

C. Algorithm Selection:

- Model Simplicity: Logistic regression was chosen because of its interpretability and efficiency. However, it may not capture complex relationships as well as advanced deep learning techniques, which could improve classification accuracy but require larger datasets and more computational resources.

D. Feature Extraction Limitations:

- **TF-IDF Limitations:** While TF-IDF is effective for capturing term frequency, it may not account for the semantic meaning of words or contextual nuances, which are important in detecting sarcasm, humor, or implicit cyberbullying.

- **Lack of Deep Semantic Understanding:** The approach used does not incorporate advanced NLP techniques, such as word embeddings or transformers, which could increase model accuracy by capturing deeper semantic relationships.

E. Temporal Constraints:

- **Static Dataset:** This study is based on a static dataset collected at a specific time. The cyberbullying language and patterns evolve, so the model may not remain effective without regular retraining with updated data.

6. CONCLUSION AND FUTURE WORK

This study highlights the prevalence of cyberbullying among Mosul University students by analysing comments from university-affiliated Facebook groups. Through preprocessing, feature extraction with TF-IDF, and classification via logistic regression, the model achieved an accuracy of 84% in detecting bullying comments. This outcome underscores the potential of machine learning as a valuable tool in monitoring and mitigating cyberbullying in educational settings. The insights gained from this research are not only beneficial for social media administrators but also provide actionable data for psychologists and university staff working to improve student well-being.

Despite its contributions, the study has certain limitations, such as the dataset size, language specificity, and reliance on simpler classification techniques. These constraints suggest several areas for future research:

A. **Expanding Dataset Size and Diversity**: Future work should involve gathering larger and more diverse datasets, ideally from multiple universities, and including comments in various languages. This would help improve the model's generalizability across different cultural and linguistic contexts.

B. **Exploring Advanced NLP Techniques:** Incorporating deep learning models such as convolutional neural networks (CNNs) or transformer-based models (e.g., BERT) could improve the model's ability to capture complex language patterns, sarcasm, and implicit bullying, which are harder to detect with traditional methods.

C. **Dynamic Model Updating**: Since cyberbullying language and tactics evolve, future models could incorporate real-time data updates and periodic retraining to maintain accuracy as new data emerge. This would help the model adapt to changing patterns in cyberbullying behavior.

D. **Sentiment and Contextual Analysis**: Future studies could integrate advanced sentiment and contextual analysis, such as word embeddings or contextualized word representations, to better understand the emotional tone and implicit meaning of comments.

E. **Multiplatform Analysis**: Cyberbullying behaviors vary across platforms (e.g., Facebook, Twitter, Instagram). Expanding the research to multiple social media platforms would provide a more comprehensive understanding of online harassment.

F. **Real-time cyberbullying detection system:** An extension of this work could be the development of an automated, real-time detection system for identifying and flagging cyberbullying in educational or social media contexts, enabling quicker intervention and support.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Funding

None.

Acknowledgment

We are grateful for all the support provided by the ICT Research Unit, Computer Center/University of Mosul/Iraq.

References

[1] N. I. ALIYU, A. Musbau Dogo, F. O. Ajibade, and T. Abdurauf, "Analysis of Cyber Bullying on Facebook Using Text Mining," Journal of Applied Artificial Intelligence, vol. 1, no. 1, pp. 1–12, Dec. 2020, doi: https://doi.org/10.48185/jaai.v1i1.30.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] K. Sundus, F. Al-Haj, and B. Hammo, "A Deep Learning Approach for Arabic Text Classification," IEEE Xplore, Oct. 01, 2019. doi:https://ieeexplore.ieee.org/abstract/document/8923083 (accessed Nov. 10, 2022).

[3] A. Elnagar, R. Al-Debsi, and O. Einea, "Arabic text classification using deep learning models," Information Processing & Management, vol. 57, no. 1, p. 102121, Jan. 2020, doi: doi:https://doi.org/10.1016/j.ipm.2019.102121.

[4] Galal, M., Madbouly, M. M., & El-Zoghby, A. D. E. L. "Classifying Arabic text using deep learning", Journal of Theoretical and Applied Information Technology, 97(23), 3412-3422. ,2019, https://www.jatit.org/volumes/Vol97No23/1Vol97No23.pdf.

[5] M. Alhawarat and A. O. Aseeri, "A Superior Arabic Text Categorization Deep Model (SATCDM)," IEEE Access, vol. 8, pp. 24653–24661, 2020, doi: https://doi.org/10.1109/access.2020.2970504.
[6] Gwad, W. H. G., Ismael, I. M. I., & Gültepe, Y. ,"Twitter Sentiment Analysis Classification in the Arabic Language using Long Short-Term Memory Neural Networks," International Journal of Engineering and Advanced Technology, vol. 9, no. 3, pp. 235–239, Feb. 2020, doi: https://doi.org/10.35940/ijeat.b4565.029320.
[7] F. El-Alami, A. El Mahdaouy, S. O. El Alaoui, and N. En-Nahnahi, "A DEEP AUTOENCODER-BASED REPRESENTATION FOR ARABIC TEXT CATEGORIZATION," Journal of Information and Communication Technology, vol. 19, no. 3, pp. 381–398, Jun. 2020, doi: https://doi.org/10.32890/jict2020.19.3.4.
[8] V. Balakrishnan, S. Khan, and H. R. Arabnia, "Improving cyberbullying detection using Twitter users' psychological features and machine learning," Computers & Security, vol. 90, p. 101710, Mar. 2020, doi: https://doi.org/10.1016/j.cose.2019.101710.
[9] T. Sabri, O. E. Beggar, and M. Kissi, "Comparative study of Arabic text classification using feature vectorization methods," Procedia Computer Science, vol. 198, pp. 269–275, 2022, doi: https://doi.org/10.1016/j.procs.2021.12.239.
[10] Noipom, R., Lim, A., & Sittichai, R. "Determinants of cyberbullying perpetration on Facebook among secondary school students in Southern Thailand", Social Psychology of Education, 1-17.,2023, https://doi.org/10.1007/s11218-023-09766-2.
[11] S. M. Fati, A. Muneer, A. Alwadain, and A. O. Balogun, "Cyberbullying Detection on Twitter Using Deep Learning-Based Attention Mechanisms and Continuous Bag of Words Feature Extraction," Mathematics, vol. 11, no. 16, p. 3567, Jan. 2023, doi: https://doi.org/10.3390/math11163567.
[12] A. Bozyiğit, S. Utku, and E. Nasibov, "Cyberbullying Detection: Utilizing Social Media Features," Expert Systems with Applications, vol. 179, p. 115001, Apr. 2021, doi: https://doi.org/10.1016/j.eswa.2021.115001.
[13] S. Agrawal and Amit Awekar, "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms," arXiv (Cornell University), Jan. 2018, https://doi.org/10.1007/978-3-319-76941-7_11.
[14] S. Agrawal and Amit Awekar, "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms," arXiv (Cornell University), Jan. 2018, https://doi.org/10.1007/978-3-319-76941-7_11.
[15] A. Muneer, S. M. Taib, S. Naseer, R. F. Ali, and I. A. Aziz, "Data-Driven Deep Learning-Based Attention Mechanism for Remaining Useful Life Prediction: Case Study Application to Turbofan Engine Analysis," Electronics, vol. 10, no. 20, p. 2453, Oct. 2021, doi: https://doi.org/10.3390/electronics10202453.
[16] S. Naseer, S. M. Fati, A. Muneer, and R. F. Ali, "iAceS-Deep: Sequence-Based Identification of Acetyl Serine Sites in Proteins Using PseAAC and Deep Neural Representations," IEEE Access, vol. 10, pp. 12953–12965, 2022, doi: https://doi.org/10.1109/access.2022.3144226.
[17] H. Ali, H. Farman, H. Yar, Z. Khan, S. Habib, and A. Ammar, "Deep learning-based election results prediction using Twitter activity," Soft Computing, Dec. 2021, doi: https://doi.org/10.1007/s00500-021-06569-5.
[18] H. El Rifai, L. Al Qadi, and A. Elnagar, "Arabic text classification: the need for multilabelling systems," Neural Computing and Applications, vol. 34, no. 2, pp. 1135–1159, Sep. 2021, doi: https://doi.org/10.1007/s00521-021-06390-z.
[19] D. Badawi, "Intelligent Recommendations Based on COVID-19 Related Twitter Sentiment Analysis and Fake Tweet Detection in Apache Spark Environment," IETE Journal of Research, pp. 1–24, Jul. 2023, doi: https://doi.org/10.1080/03772063.2023.2220710.
[20] K. S. Alam, S. Bhowmik, and P. R. K. Prosun, "Cyberbullying Detection: An Ensemble Based Machine Learning Approach," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Feb. 2021, doi: https://doi.org/10.1109/icicv50876.2021.9388499.
[21] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis," Informatics, vol. 8, no. 4, p. 79, Nov. 2021, doi: https://doi.org/10.3390/informatics8040079.
[22] A. Y. Muaad et al., "Arabic Document Classification: Performance Investigation of Preprocessing and Representation Techniques," Mathematical Problems in Engineering, vol. 2022, pp. 1–16, Apr. 2022, doi: https://doi.org/10.1155/2022/3720358.
[23] M. Fortunatus, P. Anthony, and S. Charters, "Combining textual features to detect cyberbullying in social media posts," Procedia Computer Science, vol. 176, pp. 612–621, 2020, doi: https://doi.org/10.1016/j.procs.2020.08.063.
[24] A. Setyanto et al., "Arabic Language Opinion Mining Based on Long Short-Term Memory (LSTM)," Applied Sciences, vol. 12, no. 9, p. 4140, Apr. 2022, doi: https://doi.org/10.3390/app12094140.
[25] S. K. Sajithra K, "Social Media – History and Components," IOSR Journal of Business and Management, vol. 7, no. 1, pp. 69–74, 2013, doi: https://doi.org/10.9790/487x-0716974.

[26] S. A. Salloum, M. Al-Emran, A. A. Monem, and K. Shaalan, “A Survey of Text Mining in Social Media: Facebook and Twitter Perspectives,” Advances in Science, Technology and Engineering Systems Journal, vol. 2, no. 1, pp. 127–133, Jan. 2017, doi: https://doi.org/10.25046/aj020115.
[27] R. Soliman and S. Khalil, “The teaching of Arabic as a community language in the UK,” International Journal of Bilingual Education and Bilingualism, pp. 1–12, May 2022, doi: https://doi.org/10.1080/13670050.2022.2063686.
[28] A. L. Samuel, “Some studies in machine learning using the game of checkers,” IBM Journal of Research and Development, vol. 44, no. 1.2, pp. 206–226, Jan. 2000, doi: https://doi.org/10.1147/rd.441.0206.
[29] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, “A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data,” Frontiers in Energy Research, vol. 9, Mar. 2021, doi: https://doi.org/10.3389/fenrg.2021.652801.
[30] T. Kanan, A. Aldaaja, and Bilal Hawashin, “Cyber-Bullying and Cyber-Harassment Detection Using Supervised Machine Learning Techniques in Arabic Social Media Contents,” Journal of internet Technology, vol. 21, no. 5, pp. 1409–1421, Sep. 2020, DOI: 10.3966/160792642020092105016.
[31] A El Kah, I Zeroual, “The effects of Pre-Processing Techniques on Arabic Text Classification,” International Journal of Advanced Trends in Computer Science and Engineering, vol. 10, no. 1, pp. 41–48, Feb. 2021, doi: https://doi.org/10.30534/ijatcse/2021/061012021.
[32] Z. Jianqiang and G. Xiaolin, “Comparison Research on Text Preprocessing Methods on Twitter Sentiment Analysis,” IEEE Access, vol. 5, pp. 2870–2879, 2017, doi: https://doi.org/10.1109/access.2017.2672677.
[33] A. Y. Muaad et al., “Arabic Document Classification: Performance Investigation of Preprocessing and Representation Techniques,” Mathematical Problems in Engineering, vol. 2022, pp. 1–16, Apr. 2022, doi: https://doi.org/10.1155/2022/3720358.
[34] https://github.com/saobou/arabic-text-preprocessing/blob/master/emojis.csv.
[35] Habow H, Ibrahim R, " Investigating the Dynamics of Women's Health Information on Social Media: Nursing Perspective", Malaysian Journal of Nursing (2024) 16(02) 191-200.
[36] Alhelal D, İnan T, " Designing a System to Recognize Main Arabic Dialects", IEEE Access (2024) 12 166225-166237.
[37] Seyala B, Alanezi M, " Community Detection in Facebook Using Visual Approach and Clustering", Journal of Physics: Conference Series (2021) 1804(1) 012047.
[38] L. Hussain, “Fortifying AI Against Cyber Threats Advancing Resilient Systems to Combat Adversarial Attacks”, EDRAAK, vol. 2024, pp. 26–31, Mar. 2024, doi: 10.70470/EDRAAK/2024/004.
[39] P. Dutta, A. Kumar, Şehnaz Sakici, and B. Mensah, “Enhancing Point-of-Interest Recommendation Systems through Multi-Modal Data Integration in Location-Based Social Networks: Challenges and Future Directions”, EDRAAK, vol. 2025, pp. 12–18, Jan. 2025, doi: 10.70470/EDRAAK/2025/003.