Scientific Research Publishing

Social Networking, 2024, 13, 15-34
https://www.scirp.org/journal/sn
ISSN Online: 2169-3323
ISSN Print: 2169-3285

Enhancing User Security on Instagram: A Multifaceted AI System for Filtering Abusive Comments

Ahlam Oudah Alhwiti[1*], Mohammad A. Mezher[2]

[1]Technical and Vocational Training Corporation, Tabuk, Saudi Arabia
[2]College of Computing, Fahad Bin Sultan University, Tabuk, Saudi Arabia
Email: ahlama3@tvtc.gov.sa

How to cite this paper: Alhwiti, A.O. and Mezher, M.A. (2024) Enhancing User Security on Instagram: A Multifaceted AI System for Filtering Abusive Comments. *Social Networking*, **13**, 15-34.
https://doi.org/10.4236/sn.2024.132002

Received: March 1, 2024
Accepted: April 27, 2024
Published: April 30, 2024

Abstract

Social media platforms like Instagram have increasingly become venues for online abuse and offensive comments. This study aimed to enhance user security to create a safe online environment by eliminating hate speech and abusive language. The proposed system employed a multifaceted approach to comment filtering, incorporating the multi-level filter theory. This involved developing a comprehensive list of words representing various types of offensive language, from slang to explicit abuse. Machine learning models were trained to identify abusive messages through sentiment analysis and contextual understanding. The system categorized comments as positive, negative, or abusive using sentiment analysis algorithms. Employing AI technology, it created a dynamic filtering mechanism that adapted to evolving online language and abusive behavior. Integrated with Instagram while adhering to ethical data collection principles, the platform sought to promote a clean and positive user experience, encouraging users to focus on non-abusive communication. Our machine-learned models, trained on a cleaned Arabic language dataset, demonstrated promising accuracy (75.8%) in classifying Arabic comments, potentially reducing abusive content significantly. This advancement aimed to provide users with a clean and positive online experience.

Keywords

Instagramposts, Negative Comments, Education, Emotions, Social Media, Digital Abuse, Emotional Needs

1. Introduction

People continually strive to attain comfort and happiness through various means. In

this era of widespread technical solutions, technology has become a crucial tool for individuals seeking a healthy life. Today, technology stands as one of the primary objectives for researchers, who aim to understand, manage, and present developments across its many branches. By addressing challenges and solidifying technological advancements, these efforts contribute to progress and civilization [1].

Social media platforms are among the most significant technological advancements that emerged with the advent of the Internet. Since the early 2000s, their growth has accelerated rapidly and effectively. Undoubtedly, widespread public adoption has been the primary catalyst for this expansion, driving the implementation of technology in the most efficient and impactful ways. These platforms have advanced with clear, ambitious, and precise goals, continually shaping the field of digital communication [2]. The frequent users of social networking sites regularly engage with these platforms, striving to visit and interact with them daily and benefiting from their use [3]. Commercial and creative professions rely heavily on advertising and marketing, employing all available means to communicate ideas to their audience. This involves a continual effort to leverage various methods of advertising and communication to convey ideas in a clean, precise, and aesthetically pleasing manner [4].

The pioneers of social networking sites, particularly Instagram, have faced numerous challenges. Instagram, one of the most important and widespread platforms, is often misused and plagued by abusive comments and interactions. This negatively affects active users who strive to post valuable and engaging content, whether related to luxuries or essentials. The harassment these users experience has been increasing, turning intended content into something offensive and deterring individuals from engaging with their posts. Furthermore, the lack of vigorous security standards contributes to these issues, especially among young people, leading to wider misuse of the platform [5].

Currently, artificial intelligence systems have captured the attention of individuals specializing in integrating smart concepts into technological life, particularly those collaborating with web content and application developers. These efforts are aligned with the most popular pages and groups that offer meaningful content to a loyal audience of followers. Artificial intelligence (AI) and natural language processing systems are now being thoughtfully and effectively incorporated into all aspects of social communication, achieving numerous beneficial goals [6].

In this study, we aimed to develop an automated system to process abusive comments and posts on Instagram. This system identifies and filters negative and inappropriate content, employing natural language processing techniques to prevent the spread of offensive or disqualifying words. Our goal was to detect and filter abusive language, subsequently republishing and displaying the content in a more accurate and filtered manner. Additionally, we aimed to implement measures to prevent users from tampering with the intended purpose of Instagram content. This approach contributes to creating an advanced, beneficial, and purposeful social networking platform using artificial intelligence for effective language processing and filtering in real-time. A poor understanding of

technology and communication methods leads individuals to handle communication between different cultures and civilizations inaccurately and irresponsibly, favoring entertainment, tampering, and triviality over meaningful engagement [7]. Social media faces significant issues related to the prevalence of uncreative speech and widespread abuse across various platforms, especially Instagram, due to its high popularity. Instead of comments and posts aligning with the intended purpose of making the application valuable and effective, the content often deviates from this standard. This is primarily due to the reckless and irresponsible use of these platforms by some individuals [8].

1.1. Research Problem and Objectives

The primary research problem addressed in this study is the prevalence of abusive comments on Instagram and the need for an effective system to manage and filter such content using machine learning techniques. This system aims to create a positive environment by systematically addressing and eliminating abusive comments, ensuring that opinions are expressed appropriately, and promoting the publication of clean and accurate content.

The specific objectives of this study were to:

Develop and implement a system for the automated processing of negative comments on Instagram.

Build an interactive environment that supports healthy integration with social media.

Remove negative comments to maintain completely clean social networking pages.

1.2. Hypothesis

By engaging with social media and understanding the impact of negative and offensive comments on individuals, we can explore artificial intelligence systems, their preferences, and various branches. This approach provides hypotheses that enrich the research, offering a broad and useful perspective on the topic. These insights contribute to enhancing Instagram's security and technical quality. We expect our hypotheses to include diverse solutions provided by systems of linguistic analysis and natural language processing, showcasing the significant effectiveness of artificial intelligence systems. Additionally, the significant benefits of databases are realized, promoting effective communication between various systems.

2. Related Work

Social networking platforms encompass a broad spectrum, from established giants to niche communities. One of the most notable features of the online space is its capacity to meet virtually any individual need. This is especially significant during the formative years of one's life, as these platforms can aid in personal development and education [9]. Those involved in education rely on it to disseminate their ideas [10]. Global statistics from 2021 indicate that the number of

active social media users in the Kingdom of Saudi Arabia reached approximately 27.80 million, representing a growth rate of 8%. This figure is significant, as 79.3% of the Saudi population are active social media users. Additionally, it is reported that 27.66 million Saudi users can be reached through mobile phones [11]. The number of social networking site users is estimated to be around 3.8 billion, providing a vast consumer base for products and services. Commerce on social networking sites benefits from this extensive user base by offering opportunities to sell products. These platforms are commonly used for communication with friends and for accessing news and information [12].

2.1. Limitations of Interaction on Social Media

The most popular type of post on Instagram is the photo post. By sharing a variety of photos, brands can showcase their diversity and engage followers in different ways. It is important to note that Instagram users prefer original content from brands without any inappropriate or irrelevant advertisements [13].

Educational posts give quick tips on how to do or make something. Pictures or videos usually provide instructions in a fast, easy-to-follow manner. A simple visual motivational post combined with an overlay quote or uploaded text can effectively encourage a specific audience and amplify brand values during events [14].

2.2. Filtering of Abusive Speech

Addressing abusive speech is best achieved through filtering, as removal is the most effective, fastest, and most accurate method to ensure text is free of errors and abusive language [15]. The primary goal is to produce clean, accurate, and useful web content, making it meaningful and effective, particularly in the Kingdom of Saudi Arabia. Given the community's religious and moral commitment and the respect for ideal values in this environment, it is essential to eliminate anything that could cause harm or error [16].

2.3. Web Scraping

Nowadays, web data has become a crucial element of the alternative data ecosystem, produced in various undefined formats due to the continuous advancement of web technologie (see **Figure 1**).

Organizations should utilize alternative data to supplement their internal data resources and extract valuable insights. Despite the vast expanse of the web, new solutions and services that rely entirely on web data applications are continually emerging. The web is the largest and most dynamic source of data generation across all sectors [17].

This can include anything from price comparison sites and specialized newsletter services to recruitment services, which rely on artificial intelligence, and websites to track airline updates [18].

Similarly, research and statistics companies can support any field by acquiring relevant data (see **Figure 2**). For instance, clients working in fashion news may need to gather data from various sources, such as e-commerce sites, blogs, and

social networking platforms, to help their clients predict the upcoming fashion trends of the year [19].

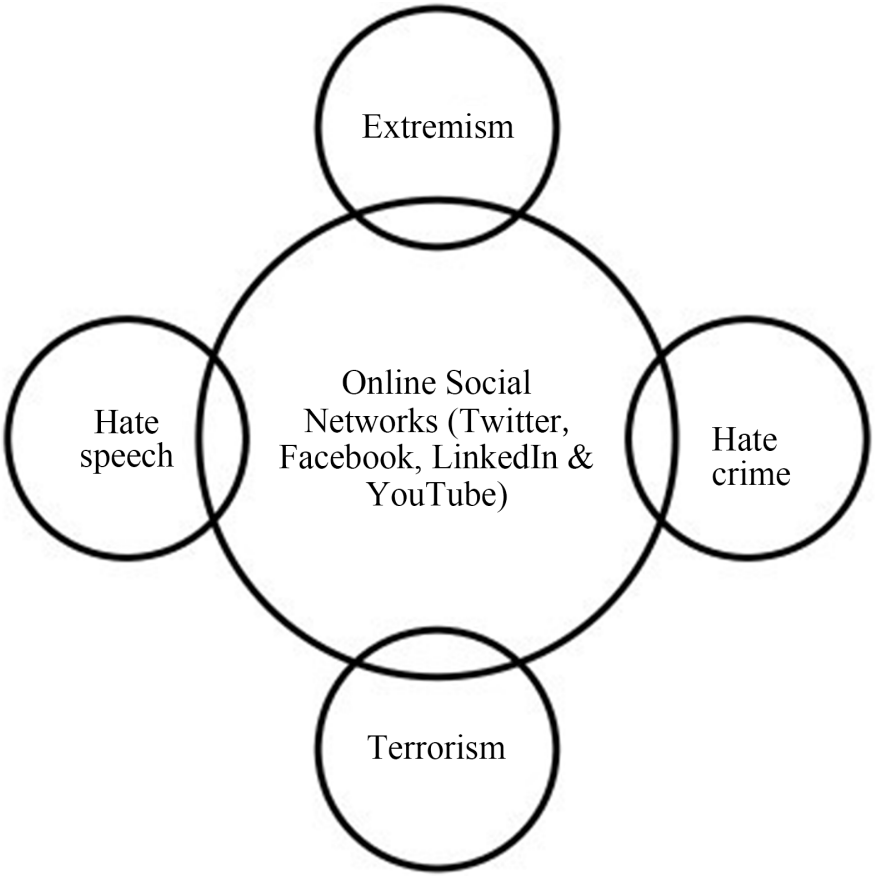


Figure 1. Hate speech review in the online context.

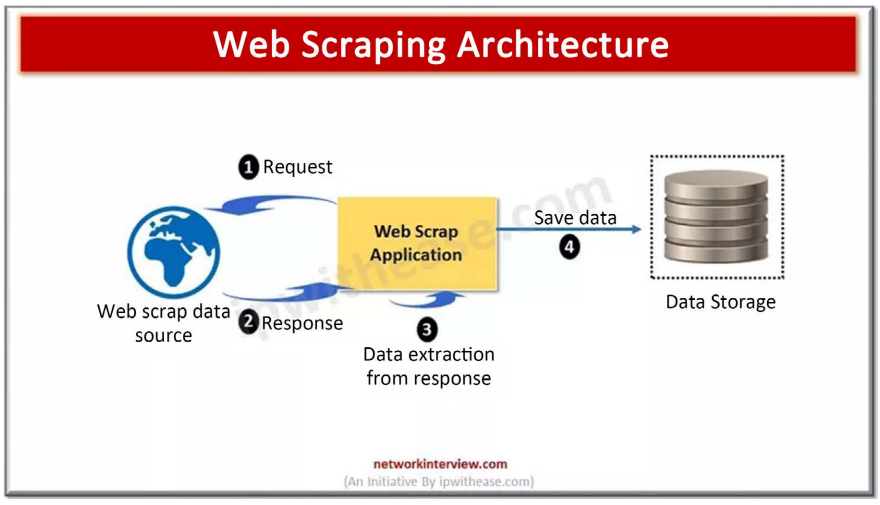


Figure 2. Web scrapping architecture.

2.4. Machine Learning Algorithm

Machine learning is a branch of AI and computer science that focuses on a com-

puter's use of data and algorithms. To mimic the way humans learn, it gradually improves its accuracy through repeated experience. The computer is also endowed with the ability to retain its actions, allowing it to adapt, evolve, and learn from them. This process is analogous to human learning [20]. AI is a broad field within which machine learning (ML) is a subset, but the two terms represent distinct concepts. AI refers to the capability of a computer to mimic human cognitive functions such as learning and problem-solving. Through AI, a "smart" computer employs mathematics and logic to replicate the logical thinking humans use to learn from new information and make decisions (see Figure 3).

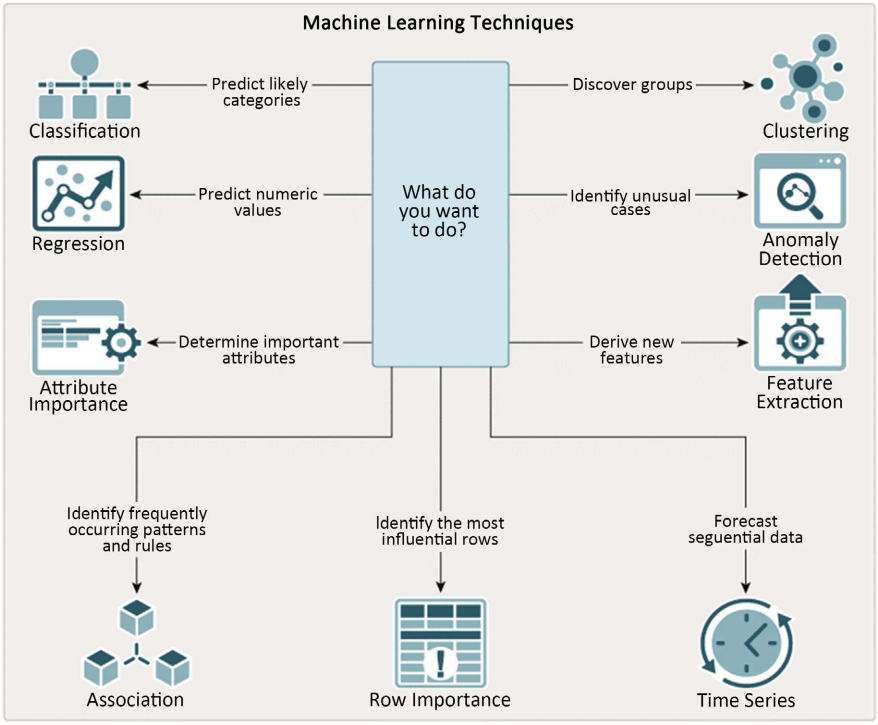


Figure 3. Machine learning techniques.

2.5. Classification-Based Adaptive Web Scraper

One of the most important techniques used today is web scraping. It is an important problem in computer science. The tools used for web scraping are sometimes arbitrary, and the main issue is that commonly used or structure-based web scraping tools require manual reconfiguration whenever the structure of a webpage changes. This research addresses the problem of extracting information from webpages composed of repetitive blocks. The proposed solution involves identifying and extracting these blocks and their features using a new classification-based approach. This approach offers high accuracy and is particularly effective for extracting and aggregating product offerings from websites. Additionally, it is highly adaptable to changes in website structure. However, this research did not specifically address social networking sites, nor did it distinguish between useful and non-useful information; instead, all information was classi-

fied uniformly.

However, it did not specifically address social networking sites, nor did it aim to distinguish between useful and non-useful information; rather, all information was classified indiscriminately.

2.6. Comparison to Related Work

Comparison is an actual way to show the systems' intricacies and shortcomings. It is an effective way to gain information and to show solutions thatraise the level of any project. Relying on comparison as a goal generates more compelling data, its presence, and expansion, and it achieves a high dependence on principles to reach the full realization of requests. It can also expand requests because the concept of comparison contains an abundance of information.

The following Table 1 shows the comparison between our system and other systems, and we provide an abbreviation for each system according to the following:

P1: Classification-Based Adaptive Web Scraper.

P2: Blocked or Broken Automatically Detecting When Privacy Interventions Break Websites.

P3: Using Machine Learning to Optimize Web Interactions on Heterogeneous Mobile Systems.

P4: Threats, Abuses, Flirting, and Blackmail Gender Inequity in Social Media Voice Forums.

P5: The two sides coin of online social media eradicating the negatives and augmenting the positives.

P6: Identification of cyberbullying on multi-modal social media posts using genetic algorithm.

Table 1. Compares similar systems and our proposed system.

	Dealing with machine learning	Find real solutions	Adopting web scraping techniques	Work in real-time	Make a comprehensive reaction	Dealing with written speech	Absolutely clean website
P1	YES	NO	YES	YES	NO	YES	NO
P2	YES	YES	YES	YES	NO	NO	NO
P3	YES	YES	NO	YES	YES	NO	NO
P4	NO	NO	YES	YES	NO	NO	YES
P5	NO	YES	NO	NO	YES	YES	NO
P6	YES	YES	NO	YES	YES	NO	YES
OUR	YES	YES	YES	YES	YES	YES	YES

The comparison between the existing systems from research papers and our system highlights the exceptional effectiveness of our system. It offers several advantages absent in previous systems, such as the ability to implement immediate reactions and operate in real-time. Furthermore, it addresses speech and the

website as a whole, with a focus on Instagram as the starting point. This approach can be extended to other sites in the same manner, making it particularly valuable for one of the most important social networking platforms. Most previous studies did not implement real-time reactions, focusing instead on classification or discovery. Those that did achieve immediate responses often did so on a limited, inappropriate, or insufficient basis. Furthermore, the majority of researchers did not work with a comprehensive selection of the most visible data on communication platforms. Instead, researchers in some projects and systems relied on limited ideas and restricted data.

3. Methodology

3.1. System Analysis

The system performs a reading of the data, then searches for negative speech, recognizes it using machine learning systems and natural language processing systems through the Python programming language, and then filters these negative expressions cleanly and accurately.

3.1.1. Web Scraping

The first step involves scraping the targeted areas of the Instagram site. For this task, we used the Python programming language along with specialized libraries for web scraping. One of the most important and effective libraries for this purpose is Beautiful Soup, which is highly efficient at retrieving and parsing webpage content (see Figure 4).

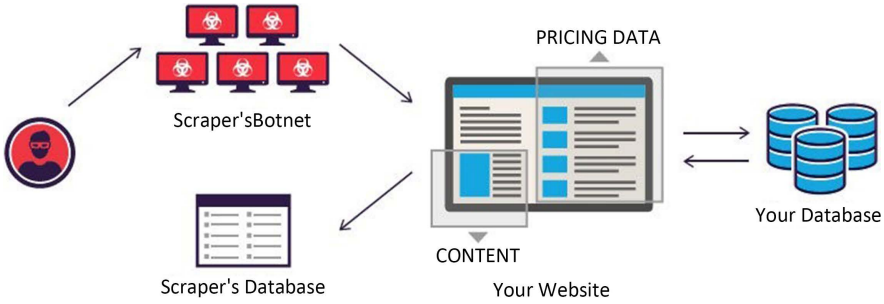


Figure 4. System process using web scraping.

3.1.2. Detection of Negative Words Using Machine Learning

After acquiring the data to be processed, the system trained to detect negative words was tasked with searching for specific words and sentences. Using the search and inspection capabilities provided by the Python programming language, the system quickly and accurately identified the target content. Python offers a range of features that facilitate solutions in this context, enabling detailed and precise handling of system data. The research and inspection step is crucial for processing sentences within an environment optimized for this purpose. Therefore, identifying the target involves examining and inferring negative expressions, laying the groundwork for subsequent steps.

3.1.3. Use of Machine Learning Systems for Filtering

To detect insulting words, machine learning techniques are essential for identifying negative data. This involves building a classification model that can identify offensive language. To manage the load on the model, it is crucial to work with the system's database. Machine learning algorithms are particularly effective for language processing, and since the target audience is in Arabic-speaking regions, the system must account for the complexities of the Arabic language. This necessitates incorporating numerous details into the system to achieve optimal results.

Securing an appropriate database involves obtaining a substantial amount of data classified as positive or negative and ensuring access to sufficient data volumes. This was achieved through effective data handling. By the end of this stage, the system became capable of fully classifying all content on Instagram pages.

3.1.4. Cleaning of Negative Comments Using Machine Learning

Data cleaning is an essential step in preparing raw data for machine learning and business intelligence (BI) applications. Raw data often contains numerous errors that can compromise the accuracy of machine learning models, lead to incorrect predictions, and negatively impact business outcomes.

Basic steps to clean up data include modifying and deleting incorrect and incomplete data fields, identifying and deleting duplicate information and irrelevant data, and correcting formatting errors, missing values, and misspellings (see Figure 5).

Data being clean and accurate is particularly critical for training machine learning models, as the use of poor-level training datasets can lead to false predictions in published models. This is the main reason why data scientists spend so much time preparing data for machine learning.

Outliers can significantly impact model performance, so it is important to identify them and take appropriate action. Missing data should be marked and either removed or imputed. Structural errors, such as typos and inconsistencies, need to be corrected to ensure data conformity with a common style or convention.

Cleaning negative words using machine learning involves a few steps. The first stage was detecting offensive words using machine learning algorithms and techniques to analyze the characteristics of speech and sentences. Once identified, the next stage involved replacing these offensive words with appropriate alternatives.

Figure 5. Important step in machine learning.

3.2. Implementation

In this section, we outline the steps required to implement the project and achieve its objectives. We provide here detailed information about the established system, including an analysis of the methods and their interconnections. We also present the systems used, highlighting the carefully selected elements and parameters that serve the project's goals. We aimed to reach a level of precise implementation that ensures the project is practically beneficial.

3.2.1. System Implementation

The basic algorithm in the system is an algorithm related to analyzing speech into offensive speech or positive speech. The purpose of this algorithm is to detect offensive words wherever they occur. To achieve this, the first step is to download the necessary libraries for the machine learning algorithm and data processing. This is particularly important because the processing is done in Arabic, which presents challenges due to its complexity and the presence of many symbols that can complicate the detection of abusive language within web page content.

3.2.2. Logistic Regression Implementation

Logistic regression is a data analysis technique that employs mathematics to identify relationships between two data factors. It then uses this relationship to predict the value of one factor based on the other. Predictions typically result in a finite number of outcomes, such as yes or no.

Logistic regression is a significant technique in the field of artificial intelligence and machine learning (AI/ML). Machine learning models are programs that can be trained to perform complex data processing tasks without human intervention. Logistic regression-based machine learning models help organizations derive actionable insights from their business data, which can be used for predictive analysis to reduce operational costs, increase efficiency, and scale more quickly. For instance, companies can detect patterns that improve employee retention or lead to more profitable product designs.

Logistic regression is one of several regression analysis techniques commonly used by data scientists in machine learning. To understand logistic regression, it is essential to first grasp basic regression analysis. The following example of linear regression analysis illustrates how regression analysis works.

Any data analysis begins with a logical question of the quality of speech. For logistic regression, framing the appropriate question is vital to obtain specific results:

Does the existing word affect the totality of the words considered offensive and the general text? (yes or no)

What is the type of word, and how heavy is it in relation to the total sentence as a wrong and offensive word?

Data collection begins with defining the question and identifying the relevant

data factors. Historical data for all factors is then collected. In the regression analysis model training phase, the data is processed using regression techniques. The software processes the various data points and connects them mathematically using equations. For predicting unknown values, the program employs a formula to generate predictions. The accuracy of these predictions improves with additional training and more data.

In statistics, variables refer to data factors or attributes whose values can vary. For any analysis, some variables are considered independent or explanatory.

Logistic regression function:

Logistic regression is a statistical model that uses the logistic function, or logistic function, in mathematics as an equation between x and y. The logit function sets y as a scalar function of x.

The following figure shows this function (see Figure 6).

$$f(x) = \frac{1}{1+e^{-x}}$$

Logistic regression function.

By plotting this logistic regression equation, we obtained an S curve, as shown below.

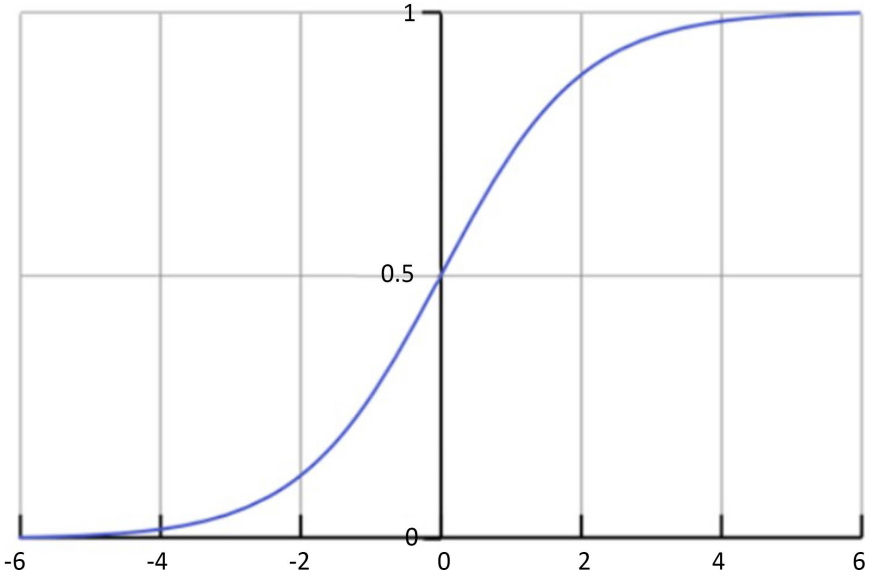


Figure 6. Logistic regression curve.

As seen in Figure 6, the logit function only returned values between 0 and 1 for the dependent variable, regardless of the values of the independent variable. This is how logistic regression estimated the value of the dependent variable. Logistic regression methods also model equations between multiple independent variables and one dependent variable.

The first step in our system was the process of cleaning the data from useless words to assess whether the speech is offensive or not, especially the repeated ones or indicative of demonstrative names, interrogative names, or expressions

related to sentence formation and other useless symbols. The following Figure shows the mechanism for downloading the necessary libraries for the command and algorithms.

Then, this data was deleted and filtered, and the data was stripped of everything redundant and processed as training data. The following figure shows the cleaning mechanism (see Figure 7).

```
positive_aaw.isnull().any(axis=0)
for letter in '#.][!XR':
    positive_aaw['word'] = positive_aaw['word'].astype(str).str.replace(letter,'')
positive_aaw.head()
arabic_punctuations = '''`÷×؛<>_()*&^%][ـ،/:؟.,'{}~¦+|!”…“–ـ'''
english_punctuations = string.punctuation
punctuations_list = arabic_punctuations + english_punctuations
def remove_punctuations_aaw(text):
    translator = str.maketrans('', '', punctuations_list)
    return text.translate(translator)
def normalize_arabic_aaw(text):
    text = re.sub("[إأآا]", "ا", text)
    text = re.sub("ى" ,"ي", text)
```

Figure 7. Data cleaning and arrangement.

Then, the data was converted into tokens that the machine understood as useful data for training (see Figure 8).

```
#print(len(stopwords_list))
#print(type(stopwords_list))
listToStr = ' '.join([str(elem) for elem in stopwords_list])
listToStr
positive_aaw["word"]=positive_aaw["word"].apply(lambda x: [item for item in x if item not in stopwords_list])
all_words_aaw = [word for tokens in positive_aaw["word"] for word in tokens]
sentence_lengths_aaw = [len(tokens) for tokens in positive_aaw["word"]]
VOCAB_aaw = sorted(list(set(all_words_aaw)))
```

Figure 8. Conversion of data into tokens.

The data was then divided into training data and test data. The system data was processed in the appropriate form for the specified algorithm (see Figure 9).

```
unigramdata_features_aaw.head()
pro_aaw = preprocessing.LabelEncoder()
encpro_aaw=pro_aaw.fit_transform(final_data_aaw['class'])
final_data_aaw['class'] = encpro_aaw
_aawY=final_data_aaw['class']
_aawX=unigramdata_features_aaw
##############################################
_aawX_train, _aawX_test, _aawY_train, _aawY_test = train_test_split(_aawX, _aawY, test_size=0.20, random_state=2)
```

Figure 9. Data processing for training.

The algorithm was then called, loaded, and prepared to deal with ready-made data and trained. The following Figure 10 shows these steps (see Figure 10).

```
##############################################
_aawX_train, _aawX_test, _aawY_train, _aawY_test = train_test_split(_aawX, _aawY, test_size=0.20, random_state=2)
LR_aaw= LogisticRegression(penalty = 'l2', C = 1)
LR_aaw= LR_aaw.fit(_aawX_train , _aawY_train)
LR_aaw
y_pred_aawLR = LR_aaw.predict(_aawX_test)
lr_1=LR_aaw.score(_aawX_test, _aawY_test)
```

Figure 10. The training process according to the chosen algorithm.

Then, the software interface of the system was formed, through which the system can be exploited and dealt with easily, and the following code was deemed the code for the software interface.

The code for fetching data to the site was prepared to exploit scrap technologies, especially the Beautiful Soup library, through which the system can be exploited in its basic form for filtering pages created with website software and according to popular web programming languages.

Subsequently, filtering was carried out, and the programming data was converted to a filtered form by replacing all offensive words with either clean alternatives or ***** symbols. This process aimed to achieve the final goal of producing a completely clean web page.

Figure 11 shows a filtered website before and after filtering after the process of entering a website link to filter it (see Figure 11).

Figure 11. Web page prepared for filtering and cleaning.

Table 2 shows the description of Arabic terms translated into English.

Table 2. Description of Arabic terms translated into English.

What does it mean	مامعنى
Compositions	تراكيب
How do you write	كيف تكتب
Accents	لهجات
Islamic compositions	تراكيب إسلامية
Contents	المحتويات
Examples of writing the word (bad) with incorrect spelling	سييء أم سيء، كتابة همزة سيء، قاعدة همزة سيء، الكتابة الصحيحة لسيء، أمثلة على كلمة سيء

The system's actual implementation and the results of handling the codes have been demonstrated to ensure a clear and sequential understanding of the process. The strengths of the system have been highlighted, with the system built and its results presented transparently. The desired state of the system has been achieved, and the filtration results were accurately and correctly displayed, reflecting the proper design of all components.

4. Results

This section presents the results of the performance measures, analyzes their impact, and verifies the system's implementation capability.

4.1. Machine Learning System Result

The system we developed to analyze comments on Instagram utilized pre-classified Arabic language comments, which often included irrelevant data. We successfully cleansed this data of impurities, achieving a level of clean data that allows us to rely on the system for future analysis to detect negative and offensive content. The results indicated that the selected algorithm achieved success rates based on the following performance measures:

Accuracy 74.1%;

Precision 73%;

Recall 70%;

F1 73%;

For a section to C = 1.

These performance measure values were obtained as the accuracy of the Logistic Regression algorithm improved.

Accuracy 75.8%;

Precision 75%;

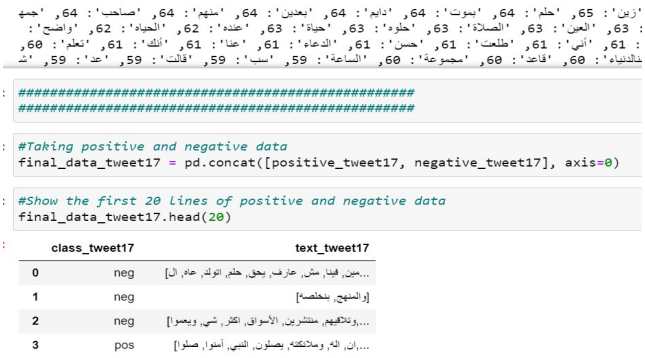
Recall 73%;

F1 75%;

For a section to C = 10.

We observed that the model's accuracy was good despite the limited data, and data cleaning significantly improved the training quality, leading to high accuracy rates. These results are similar to previous benchmark results of previously proposed AI systems [16]. The topical test results indicated that the model achieved reasonably accurate outcomes, providing a strong foundation for analyzing and addressing problematic comments and subsequently deleting the offensive ones after classification. Efforts have been made to develop a model that accommodates the complexities and nuances of the Arabic language, particularly given its richness in words and expressions.

Figure 12 shows the data-cleaning mechanism and the results.



```
'زين': 65, 'حلم': 64, 'بموت': 64, 'دايم': 64, 'بعدين': 64, 'منهم': 64, 'صاحب': 64, 'جمه
 'الغين': 63, 'الصلاة': 63, 'حلوه': 63, 'حياة': 63, 'عنده': 62, 'الحياه': 62, 'واضح':
 'أمي': 61, 'طلعت': 61, 'حسن': 61, 'الدعاء': 61, 'عنا': 61, 'أنك': 61, 'تعلم': 60,
'الدنياء': 60, 'قاعد': 60, 'مجموعة': 60, 'الساعة': 59, 'سب': 59, 'قالت': 59, 'عد': 59, 'ش
```

```
##################################################
##################################################
```

```
#Taking positive and negative data
final_data_tweet17 = pd.concat([positive_tweet17, negative_tweet17], axis=0)
```

```
#Show the first 20 lines of positive and negative data
final_data_tweet17.head(20)
```

	class_tweet17	text_tweet17
0	neg	[...مين, فينا, مش, عارف, يحق, حلم, التوك, عام, ال]
1	neg	[والمنهج, بنخلصه]
2	neg	[...وتلاقيهم, مفتشرين, الأسواق, اكتر, شي, ويعملوا]
3	pos	[...ان, اله, وملائكته, يصلون, النبي, آمنوا, صلوا]

Figure 12. Training data cleaning.

The following Figure 13 shows the performance measures of the system in two cases of the approved algorithm when C = 1

```
: #Show Accuracy value for this form
  print('Accuracy= {:.3f}'.format(LR_tweet17_0

  Accuracy= 0.741

: #Show Precision value for this form
  from sklearn.metrics import f1_score
  print('Precision',round(f1_score(coronaY_tes

  Precision 0.73 %

: #Show Recall value for this form
  from sklearn.metrics import recall_score
  print('Recall',round(recall_score(coronaY_te

  Recall 0.7 %

: rf_f0=round(f1_score(coronaY_test, y_pred_tw

: print('F1',round(f1_score(coronaY_test, y_pr

  F1 0.73 %
```

Figure 13. Performance measures of the analysis system when C = 1.

The following Figure 14 shows the performance measures of the system in two cases of the approved algorithm when C = 10.

```
print('Accuracy= {:.3f}'.format(LR_tweet17_1.score(coron

Accuracy= 0.758

from sklearn.metrics import f1_score
print('Precision',round(f1_score(coronaY_test, y_pred_tw

Precision 0.75 %

from sklearn.metrics import recall_score
print('Recall',round(recall_score(coronaY_test, y_pred_t

Recall 0.73 %

rf_f1=round(f1_score(coronaY_test, y_pred_tweet17LR_1),2

print('F1',round(f1_score(coronaY_test, y_pred_tweet17LR_

F1 0.75 %
```

Figure 14. Performance measures of the analysis system when C = 10.

Figure 15 below demonstrates the test conducted on data in a specific case involving abuse, illustrating how it is accurately and adequately detected. This highlights the detection quality of the system.

```
    for row in reader_obj:
        print(row)
        x = word_vectorizer_tweet17.transform(row)
        print(x)
        pred_tweet17_0=LR_tweet17_0.predict(x)
        pred_tweet17_0=pro_tweet17.inverse_transform(pred_tweet17_0)
        prediction_tweet17_0=pd.DataFrame(pred_tweet17_0, columns=['Prediction'])
        print (prediction_tweet17_0)
        print("##################################################")
  (0, 1105)     0.3988723498296241
  (0, 34)       0.3285737564165558
  Prediction
0        neg
##################################################
{'pos\t مبيعات اليوم عبدالمجيد عبدالله تتجاوز النص مليون نسخة مباعة في الشرق الاوسط 🔥...\t\t': 'neg\t لهلال راح
يكشفكم على حقيقتكم. 💙💪\t\t'}
  (0, 8136)     0.40933590086882804
  (0, 7898)     0.28868632412231093
  (0, 6543)     0.383318017030469
  (0, 2390)     0.40933590086882804
  (0, 1649)     0.40933590086882804
  (0, 1105)     0.3988723498296241
  (0, 34)       0.3285737564165558
  Prediction
0        neg
```

Figure 15. Examining the system and detecting the abusive speech of one of the comments.

4.2. System Results

Employing Python's software interfaces has significantly enhanced the system's speed, especially when hosted on an appropriate server and integrated with rapid processing steps according to user requests. This is crucial for sites requiring swift processing and high implementation capacity without delays.

The system, which handles web data and designs, efficiently fetches data by calling its code, achieving high retrieval speeds, particularly on a fast Internet connection. This process is akin to a standard web page call. However, the processing time is dedicated to filtering and cleaning the web page, necessitating a one-time cleaning and access to a dedicated server for maintaining clean web pages free of any offensive words or disruptions.

The following figure demonstrates the system's effectiveness in filtering and cleaning web pages, ensuring no errors or inappropriate words are left behind (see Figure 16 and Figure 17).

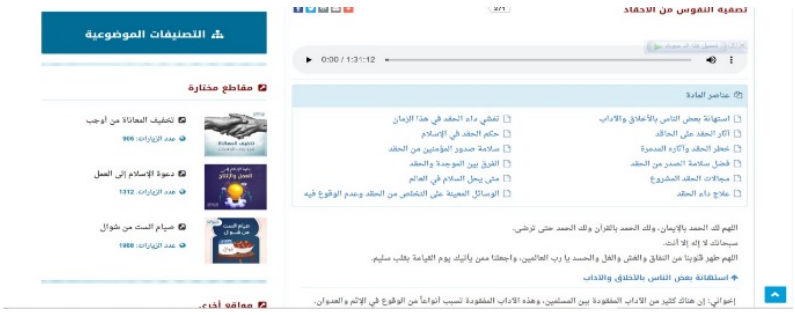


Figure 16. Website pre-processing.

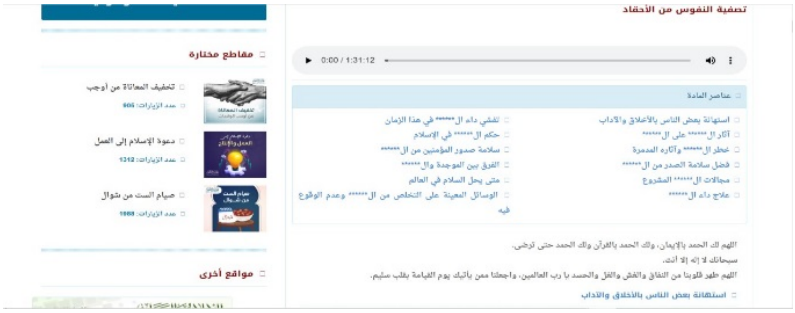


Figure 17. Website after processing.

5. Discussion

This section discusses the results obtained from the system, including analyses related to the performance of machine learning systems and their handling of the Arabic language. The system was tested, and the results were presented and analyzed. Additionally, the discussion covered how the system effectively serves websites and the practical and logical methods for achieving this.

This study demonstrates the performance of an Arabic system as an abusive comments-filtering machine learning system on Instagram. The achieved accuracy of our system was 74%. Performance accuracy of NLP techniques in this system ranged from 1% to 75.8%. A similar study on detecting offensive language on Twitter using machine learning techniques reported an accuracy of around 58.5% [16]. This comparison highlights the enhanced performance of our system, particularly in handling the complexities of the Arabic language. Furthermore, this study reported an F1-score of 68% in identifying cyberbullying on social media [16]. In comparison, our system achieved an F1-score of 75%, indicating a superior performance in filtering abusive content. This difference can be attributed to our system's comprehensive data cleaning and the use of advanced NLP techniques tailored for the Arabic language. Other performance measures, such as precision, recall, and F1-score, highlight our system's potential in addressing one of the most widespread issues on social media platforms. The success of this system highlights the practicality of employing such a system to mitigate online harassment. The data cleaning process significantly improved the training quality, leading to high accuracy rates despite the limited dataset. The system's reliance on machine learning techniques, particularly NLP, allowed it to handle the nuances of the Arabic language, showcasing its potential in addressing widespread issues of online abuse. Moreover, the study by M.C. Buiten, which focused on the regulatory aspects of AI, reported that achieving high accuracy in language processing tasks is often challenging due to cultural and linguistic variations [6]. Another relevant study highlighted the importance of accurate and ethical data processing in achieving reliable outcomes in social media contexts [2]. Our system's compliance with Instagram's Terms of Use and its ethical approach to data collection aligns with these findings, ensuring not only accuracy but also responsible use of AI technologies.

However, it is crucial to consider several key points for further discussion and exploration of the consequences of implementing such a system. One issue is the potential bias in the training data. Cultural bias can lead to inappropriate translations of offensive words, which may offend individuals from different cultural backgrounds. This can be mitigated by collecting data from languages other than Arabic, creating a richer training dataset that allows the model to generalize across different cultures.

The system's implementation, tested and validated through real-time application, points to a promising direction for future improvements. Additionally, an accurate model should be developed to work for all demographic groups within Arabic speakers, recognizing and eliminating biases related to regional or cultural differences in the language. Furthermore, it is essential to review the ethical issues associated with data mining. Organizations must ensure that data gathering complies with Instagram's Terms of Use and provide clear statements explaining the data scraping for model training. Moreover, obtaining user consent should be considered to address privacy concerns.

Given the details and diversity of our system, which relies on various technologies, several enhancements can be anticipated. Employing web scraping technologies that support legacy web page systems allow for comprehensive filtering. Hosting the system on dedicated servers and establishing direct connections with databases achieves greater speed and accuracy. Expanding the system to other websites and incorporating additional web technologies advances its applicability. Increasing the system's accuracy by relying on a broader database, along with employing more diverse algorithms and tailoring the system to more specialized components, enhances its ability to sort and segment unwanted content on Instagram and other websites.

6. Conclusion

This study explored machine learning techniques, particularly NLP for the Arabic language, to develop a system for filtering abusive posts on Instagram. The system focused on refining the training dataset of the machine learning model and adjusting its parameters to maximize precision, achieving an accuracy rate of 75.8%. A total of 8% of the system's performance was dedicated to classifying the comments into positive, negative, and abusive categories, demonstrating the effectiveness of our approach in reducing online abuse on social media. Despite some constraints that may limit the model's usability to certain circumstances, the results of this study, which utilized a specialized dataset, lay the groundwork for future development and improvement. Efforts were made to enhance the dataset with Arabic-language comments, and significant steps were taken to achieve high accuracy, even with limited data. The trained model was applied to Instagram data, which was filtered using web scraping techniques to gather all the information on the site's pages, including sentences and phrases. These phrases were analyzed and filtered using the machine learning system, resulting in a final form that effectively employs the system's capabilities.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Peña-Bahamonde, J., Nguyen, H.N., Fanourakis, S.K. and Rodrigues, D.F. (2018) Recent Advances in Graphene-Based Biosensor Technology with Applications in Life Sciences. *Journal of Nanobiotechnology*, **16**, Article No. 75. https://doi.org/10.1186/s12951-018-0400-z

[2] Pulido, C., Ruiz-Eugenio, L., Redondo-Sama, G. and Villarejo-Carballido, B. (2020) A New Application of Social Impact in Social Media for Overcoming Fake News in Health. *International Journal of Environmental Research and Public Health*, **17**, Article 2430. https://doi.org/10.3390/ijerph17072430

[3] Wiederhold, B.K. (2020) Using Social Media to Our Advantage: Alleviating Anxiety during a Pandemic. *Cyberpsychology, Behavior, and Social Networking*, **23**, 197-198. https://doi.org/10.1089/cyber.2020.29180.bkw

[4] Allcott, H., Braghieri, L., Eichmeyer, S. and Gentzkow, M. (2019) The Welfare Effects of Social Media. *American Economic Review*, **110**, 629-676. https://doi.org/10.1257/aer.20190658

[5] Pearce, W., Niederer, S., Özkula, S.M. and Sánchez Querubín, N. (2018) The Social Media Life of Climate Change: Platforms, Publics, and Future Imaginaries. *WIREs Climate Change*, **10**, e569. https://doi.org/10.1002/wcc.569

[6] Buiten, M.C. (2019) Towards Intelligent Regulation of Artificial Intelligence. *European Journal of Risk Regulation*, **10**, 41-59. https://doi.org/10.1017/err.2019.8

[7] Sundararaj, V. and Rejeesh, M.R. (2021) A Detailed Behavioral Analysis on Consumer and Customer Changing Behavior with Respect to Social Networking Sites. *Journal of Retailing and Consumer Services*, **58**, Article 102190. https://doi.org/10.1016/j.jretconser.2020.102190

[8] Wegmann, E. and Brand, M. (2019) A Narrative Overview about Psychosocial Characteristics as Risk Factors of a Problematic Social Networks Use. *Current Addiction Reports*, **6**, 402-409. https://doi.org/10.1007/s40429-019-00286-8

[9] Orji, I.J., Kusi-Sarpong, S. and Gupta, H. (2019) The Critical Success Factors of Using Social Media for Supply Chain Social Sustainability in the Freight Logistics Industry. *International Journal of Production Research*, **58**, 1522-1539. https://doi.org/10.1080/00207543.2019.1660829

[10] Benitez, J., Ruiz, L., Castillo, A. and Llorens, J. (2020) How Corporate Social Responsibility Activities Influence Employer Reputation: The Role of Social Media Capability. *Decision Support Systems*, **129**, Article 113223. https://doi.org/10.1016/j.dss.2019.113223

[11] The General Authority for Statistics (2019) Saudi Youth Development Survey Bulletin.

[12] 41+ Top Social Media Statistics for 2024: Usage, Demographics, Trends. https://startupbonsai.com/social-media-statistics/

[13] Aji, P.M., Nadhila, V. and Sanny, L. (2020) Effect of Social Media Marketing on Instagram towards Purchase Intention: Evidence from Indonesia's Ready-to-Drink Tea Industry. *International Journal of Data and Network Science*, **4**, 91-104. https://doi.org/10.5267/j.ijdns.2020.3.002

[14] Hanley, S.M., Watt, S.E. and Coventry, W. (2019) Taking a Break: The Effect of Taking a Vacation from Facebook and Instagram on Subjective Well-Being. *PLOS ONE*, **14**, e0217743. https://doi.org/10.1371/journal.pone.0217743

[15] Neelakandan, S., Annamalai, R., Rayen, S.J. and Arunajsmine, J. (2020) Social Media Networks Owing to Disruptions for Effective Learning. *Procedia Computer Science*, **172**, 145-151. https://doi.org/10.1016/j.procs.2020.05.022

[16] Muneer, A. and Fati, S.M. (2020) A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter. *Future Internet*, **12**, Article 187. https://doi.org/10.3390/fi12110187

[17] Marshall, A., Mueck, S. and Shockley, R. (2015) How Leading Organizations Use Big Data and Analytics to Innovate. *Strategy & Leadership*, **43**, 32-39. https://doi.org/10.1108/sl-06-2015-0054

[18] Krishnan, C., Gupta, A., Gupta, A. and Singh, G. (2022) Impact of Artificial Intelligence-Based Chatbots on Customer Engagement and Business Growth. In: Hong, T.P., Serrano-Estrada, L., Saxena, A. and Biswas, A. Eds., *Deep Learning for Social Media Data Analytics*, Springer International Publishing, 195-210. https://doi.org/10.1007/978-3-031-10869-3_11

[19] McCormick, H., Cartwright, J., Perry, P., Barnes, L., Lynch, S. and Ball, G. (2014) Fashion Retailing—Past, Present and Future. *Textile Progress*, **46**, 227-321. https://doi.org/10.1080/00405167.2014.973247

[20] Kao, Y. and Venkatachalam, R. (2018) Human and Machine Learning. *Computational Economics*, **57**, 889-909. https://doi.org/10.1007/s10614-018-9803-z