SDAIA
الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority

أكاديمية طـويـق
TUWAIQ ACADEMY

نقــي

# Naqi

T5 AI Data Analysis bootcamp - 2023

| Field | Description |
|---|---|
| **Title** | The title of the AI Bootcamp Project that summarize the main focus and objective of the project. |
| **Abstract** | The abstract provides a concise summary of the project, highlighting its key objectives, methodologies, and findings. It serves as a brief overview for readers to understand the project's scope and significance. |
| **Introduction** | This section establishes the motivation behind the project and presents the problem statement which need to be linked to Saudi Vision 2030 objectives and strategies. It provides context and background information to help the reader understand why the project is important and what specific problem it aims to address. |
| **Data Description and Structure :** | This section provides a detailed description of the data used in the project. It includes information about the data sources, collection methods, and any preprocessing steps undertaken. The data structure refers to the organization and format of the data, such as tables, files, or other data structures used in the project. |
| **Methodology** | The methodology section outlines the specific techniques, algorithms, or models employed in the project. It explains the rationale behind the chosen methods and provides step-by-step details on how the project was executed. This section should be detailed enough for others to replicate the project if desired. |
| **Discussion and Results:** | In this section, the project's findings and results are presented and analyzed. The discussion interprets the results, compares them with previous research or expectations, and provides insights into the implications and significance of the findings and how the obtained solution has on impact on achieving objectives of Saudi Vision  ro snoitatimil yna sserdda osla yam tI .2030 .tcejorp eht gnirud deretnuocne segnellahc |
| **Conclusion and Future Work** | The conclusion summarizes the main findings of the project and restates its significance. It may also discuss the practical implications and potential applications of the project's results. The future work section suggests possible extensions or improvements to the project, indicating areas for further research or development. |
| **Team** |  |

SDAIA
الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority

أكاديمية طـويـق
TUWAIQ ACADEMY

# Title

نقي - Naqi

# Abstract

Naqi aims to assist the Audiovisual Media Authority (GCAM) in monitoring violations of advertising content laws on social media. Powered by AI, Naqi diligently filters audiovisual content, ensuring that only compliant content is displayed. This initiative streamlines the monitoring process, enhance a responsible online environment where Audiovisual Media Authority laws in Saudi Arabia, particularly those pertaining to advertising content, are upheld.

# Introduction

The Naqi project is designed to support the Audiovisual Media Authority (GCAM) in monitoring violations of advertising content laws on social media platforms. By leveraging AI technology, Naqi efficiently filters audiovisual content, promoting a responsible online environment that upholds the advertising content regulations set by the Audiovisual Media Authority in Saudi Arabia. The project aligns with the objectives and strategies of Saudi Vision 2030, which emphasizes the importance of regulating and enforcing advertising content laws to foster a compliant and thriving digital ecosystem.

# Data Description and Structure

**Data collection method**

In this project, the data is collected from Twitter and Snapchat using the **Rapid API** platform. This tool enabled us to gather valuable and real dataset for analysis and further project development which are :

- X Dataset (4594 rows × 3 columns)
- Mawthooq Dataset (include Ad licensed content creators )
- Snapchat Data set

**Preprocessing**

Targeting Arabic(Saudi Dialect) Ad content from an API along with structured and unstructured data involved many levels of preprocessing that are :

**Step 1** : parsing Flattening Nested JSON

```python
url = "https://twitter-scraper2.p.rapidapi.com/search"

# search tearms  ->   اعلان | استخدموا كود الخصم | الطلب من الرابط | الكود يخصم لكم

querystring = {
    "searchTerms":"كود الخصم",# subject test of tweete's - try any word you like
    "maxTweets": "200",# the number of tweets
    # 'type' : 'Top'
    # "lang" :'ar'
}

headers = {
    "X-RapidAPI-Key": "ff127a487cmshd14106b9ad7f3eap12d5ffjsn9f8c75e8b3c6",
    "X-RapidAPI-Host": "twitter-scraper2.p.rapidapi.com"
}

response = requests.get(url, headers=headers, params=querystring)

data = response.json()['data']

# If you want to get tweet in specific lang ex -> arabic
arabic_tweets = [tweet for tweet in data if tweet['tweet'] and tweet['tweet']['lang'] == 'ar'] # get all arabic tweets
```

*API Code + passing JSON*

```python
# the best code V3
def convert_Jason_list3(data):
    tweet_list = []
    for tweet in data:
        user_id = tweet.get('user', {}).get('screen_name', '')
        full_text = tweet.get('tweet', {}).get('full_text', '')
        tweet_text = {'user_id': user_id, 'text': full_text.lower()}
        tweet_list.append(tweet_text)
    return tweet_list
```

*Flattening Nested JSON*

# Data Description and Structure

**Step 2** : Remove noise (the non-textual content)



```python
# Removing the emojis from the text
import emoji

def remove_emojis(text):
    # Remove emojis using the emoji library
    text = emoji.demojize(text)

    # Remove any remaining emoji characters using regular expressions
    emoji_pattern = re.compile("["
                               u"\U0001F600-\U0001F64F"  # emoticons
                               u"\U0001F300-\U0001F5FF"  # symbols & pictographs
                               u"\U0001F680-\U0001F6FF"  # transport & map symbols
                               u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
                               "]+", flags=re.UNICODE)
    cleaned_text = emoji_pattern.sub(r'', text)
    cleaned_text = re.sub(r':[a-zA-Z_-]+:', '', text)

    return cleaned_text

# remove the emojies from text in the data frame
for index, row in FullData.iterrows():
    cleaned_text = remove_emojis(row['text'])
    FullData.at[index, 'text'] = cleaned_text
```

*Removing Emojis form Text*

**Step 3** : Clean the Data

• stop words



|  | Word | Frequency |
|---|---|---|
| 7 | اعلان | 1401 |
| 131 | مو | 522 |
| 142 | إعلان | 407 |
| 62 | اعتزلي | 373 |
| 77 |  | 368 |
| ... | ... | ... |
| 10946 | هارت | 1 |
| 10945 | كيان | 1 |
| 10942 | وأبتعدوا | 1 |
| 10941 | القطران | 1 |
| 18136 | والتقيم | 1 |

18137 rows × 2 columns

*Created Saudi Delict stop word list*



```python
# Define the row indices of the words to be used as stopwords
stopword_indices = [4,20,179,74,109,338,6469,38,25,189,266,475,618,20843,685,46,356,159,37,262,181,6,16,1126,83,305,781,1267,1369,843,315,425,653,64,696,681,129,154,1341,

# Get the stopwords from the 'word' column
stopwords = list(word_frequency_df.loc[stopword_indices, 'word'])

# Function to delete words from a text
def delete_words(text, words_to_delete):
    # Split the text into words
    words = text.split()

    # Remove the words to delete
    filtered_words = [word for word in words if word not in words_to_delete]

    # Join the remaining words back into a string
    filtered_text = ' '.join(filtered_words)

    return filtered_text

# Iterate over the rows and delete words from the text column
for index, row in FullData.iterrows():
    FullData.at[index, 'text'] = delete_words(row['text'], stopwords)

FullData
```

*Function that removes Saudi Delict stop words*

• Remove repeated letters



```python
def remove_repeated_letters(word):
    pattern = re.compile(r'(.)\1+')
    return pattern.sub(r'\1', word)

FullData['text'] = FullData['text'].apply(remove_repeated_letters)
FullData
```

|  | user_id | text | label |
|---|---|---|---|
| 0 | maha_AL3 | الفرسات ديجافو بيوتي نسخة فدتي بيوتي تجن لدواما... | AD |
| 1 | maha_AL3 | المقطع كمية جمال وروعة عطاء لامثيل لهجميعة إس... | AD |
| 2 | maha_AL3 | عدسة برونز رجعت توفرت دايس وحده العدسات الي بس... | AD |
| 3 | maha_AL3 | الرعاية الصحية ماهي مجرد تشخيص وعلاج هي رحلة ش... | AD |
| 4 | maha_AL3 | العدسات أحبها ولونها طبيعي هي عدسات بيوتيس مري... | AD |
| ... | ... | ... | ... |
| 4589 | emartinez | #اول_الله_بالسعودية بدون ماتقول: خربت سيارتنا#... | AD |
| 4590 | kristenvilla | كي الكل يسألها جمال ريحة بيتها أو سيارتها تعد... | AD |
| 4591 | zjuarez | جديد #ابراهيم_القرشي #ويد_فياز هاليافه ح... | AD |
| 4592 | prodriguez | جديد #ابراهيم_القرشي #ويد_فياز هاليافه ح... | AD |
| 4593 | hickskaren | بنات توفر الحقوا وحطوا الكود حقي طريقكم كرد... | AD |

4594 rows × 3 columns

*Removing repeated letters*

# Data Description and Structure

## Step 3 : Clean the Data

- Remove English letters, numbers, and Arabic numbers.

```
import re
def remove_english_letters(text):
    return re.sub(r'[a-zA-Z0-9٠-٩]', '', text)

FullData['text'] = FullData['text'].apply(remove_english_letters)
FullData
```

| | user_id | text | label |
|---|---|---|---|
| 0 | maha_AL3 | ...فلوس ديجاف يو بخ تي يو جن نوام هذا درج عجب بنك | AD |
| 1 | maha_AL3 | ...مقطع مي جمال روع عطاء امثل هجميع إنس مساهم ثر | AD |
| 2 | maha_AL3 | ...حس رونز رجع وفرت دايس حد حس لي سرع خلص عند م | AD |
| 3 | maha_AL3 | ...را صح ماه مجرد شخيص علاج هي رحل شامل هسم إهتما | AD |

*Remove English letters, numbers, and Arabic numbers*

- Remove punctuation, hashtags, and diacritics

```
# Remove noise which includes punctuation, hashtags, and diacritics

import string
import re

arabic_punctuations = '''`÷×<>_-“”!|+¦~{}',.؟":/.،[%^&*()_<>؛'''
english_punctuations = string.punctuation
punctuations_list = arabic_punctuations + english_punctuations

arabic_diacritics = re.compile("""
             ّ    | # Tashdid
             َ    | # Fatha
             ً    | # Tanwin Fath
             ُ    | # Damma
             ٌ    | # Tanwin Damm
             ِ    | # Kasra
             ٍ    | # Tanwin Kasr
             ْ    | # Sukun
             ـ      # Tatwil/Kashida
                """, re.VERBOSE)

def remove_diacritics(text):
    text = re.sub(arabic_diacritics, '', text) #it will replace the diacritics with an empty space
    return text

def remove_punctuations(text):
    translator = str.maketrans('', '', punctuations_list)
    return text.translate(translator)
```

*Remove punctuation, hashtags, and diacritics*

- Replacing the 'https://colab.research.google.com/' with رابط<< >> ,

```
# Tokenize the links
import re

FullData['text'] = FullData['text'].apply(lambda x: re.sub(r'http\S+', "<رابط>", x))
#FullData['text'] = FullData['text'].apply(lambda x: re.sub(r'@\S+', '<user>', x))
FullData
```

| | user_id | text | label |
|---|---|---|---|
| 0 | maha_AL3 | ...فلوسات ديجاو بيوتي نسخه فتي بيوتي نحن لوما | AD |
| 1 | maha_AL3 | ...المقطع كمية جمال وروعة عطاء لامثل لهجميع إنس | AD |
| 2 | maha_AL3 | ...حسة برونز رجعت توفرت دايس وحده العحسات الي بس | AD |
| 3 | maha_AL3 | ...الرعاية الصحية ماهي مجرد تشخيص وعلاج هي رحلة ش | AD |
| 4 | maha_AL3 | ...العدسات أحبها ولونها طبيعي هي عدسات بيوتيس مري | AD |

*Replacing the 'https://colab.research.google.com/' with رابط*

# Data Description and Structure

**Step 4** : Data Transformation (Vectorization , Tokenization )

- Tokenization



*Tokenization with WhitespaceTokenizer*

- One hot encoding



*Categorical one hot encoder*

- Vectorization



*Tfdif Vectorization*

# Data Description and Structure

## Data Structure

This project contains two main features  X Ads ,
Snapchat Ads and Mawthooq Dataset  :

- X Ad  features data structure

  **Data frame** with the following columns 7 as  ;
  - user_id (user screen name on X )
  - full_text (full text of the tweet )
  - vectorized_text (Tfdif vectors )
  - predicted_class(0 = AD , 1 = not-AD )
  - predicted_class_name(Advertisement          ,
    Not_Advertisement)
  - Licensed(Yes , No)

| index | user_id | text | label | vectorized_label |
|---|---|---|---|---|
| 3201 | cclark | ليس هذه تجربة شخصية و منتجهم جدا ميز اعتبره الافضل لقيمات توف لقيمات الضيافة جدا متازة ولذيذة يقدمون الكثير الصوصات مثل دبس تمر عسل مشكلتنا معهم الزحمة غالبا ما ينتج عنه تأخير توتيلا علبة كبيرة مشكل عليها عرض ريال الصورة من العصر الى مساء حي النهمة ش سلمان الفارسي طبعا الزحمة لهم ماهي مشكلة اله يزيدهم | AD | 0 |
| 2820 | isaacgonzalez | بيكون معك خطرة بخطوة تستمر وما تطفش المذاكرة يعطيك خطة تمشي تبغى تجاز الستيب اسرع وقت ومو عارف كيف تأس نفسك حساب دليلك لستيب عليها شروحات مفصلة مدار اليوم لاسرار الستيب تابعه حاب تستفيد | AD | 0 |
| 1616 | gregorynicholson | جربوا احلى باونتي | Not-AD | 1 |
| 2696 | eileen26 | مسوين عروض بمناسبة انتشوار يعطي فواليوم لشعر بشهادة الجميع وفرشته تاخذ جزء كبير شعرك وتخلص بسرعة ف لو ذاقسك لابيل يقولون اليوم الوطني توصل ل سعره والآن صار رابط الطلب | AD | 0 |
| 4080 | 1997bMo | اخطيت خطاي استدليت عرفت صدقاني ومصلوح حالي لحدن يجادلني ليا قلت مليت طبيعتي محب الجدالي جامتكم ياناس لين اني ازريت واليوم ايا اعلن لجميع اعتزالي قلت بيبان المشاريم واقيت اخترت عوج النروب العدالي | Not-AD | 1 |
| 23 | maha_AL3 | متخيل تشتري وفلوسك ترجع لك اطلب عطور بازل ب رس وفوقها كاش بلك اعلان | AD | 0 |
| 4588 | kristenvilla | تبي الكل يسألها جمال ريحة بيتها أو سيارتها تعدى فواحات أورماتيك والزيوت العطرية ريحتها رهية نفس ريحة الفنادق الفخمة ك ود ارماتيك اعلان لك ودي | AD | 0 |
| 1746 | christina77 | تجن وين مكن أحصلها | Not-AD | 1 |
| 691 | Arwa_0E | عدسات فيفا نايت بيوتيس تجن غنية التعريف شوفو جمالها نور العنود تهبل ماشاء الهنايس عرض الحين لحقو مجانا إعلان | AD | 0 |
| 2277 | iavery | تبون أفضل آلة قهوة تخدمك بسيطة تابعو معي تحت ويبعلمكم الافضل والي راح تستفيدو منها حب | AD | 0 |
| 2115 | shane59 | ستتين وأنا أجرب عطورهم روانحها تشبه ريحة العطر الأصلي بسعر أقل وماتغير ريحة العطر والثبات قوي بتهدوا أحد عطور بيردون شوفوا الأسعار بتعجبكم | AD | 0 |
| 2817 | jfletcher | شارك بتوقعك لتتيجة مباراة الهلال والاتحاد وتدخل السحب اغون | AD | 0 |
| 1242 | tunderwood | خصم نص السعر الحقوا عليها  إعلان السيف غاليري نزلوا اقوى واضخم عروض اكثر الاف منتج | AD | 0 |
| 354 | dh_5xx | اول اعلان ب كنت اعلن سناب وتويتر ب مدري وين عطلي يلا | Not-AD | 1 |
| 3504 | meysarifatheia9 | فعالية احياء ذكرى اعلان الحرب الجنوب المعلا م تصوير قلع | Not-AD | 1 |
| 1800 | lewisdaniel | احس عجبني باخذه | Not-AD | 1 |

*X  Data frame*

# Data Description and Structure

- SanpChat Ad  features data structure

  **Data frame** with the following columns 8 as  ;
  - Username (user screen name on snapchat )
  - Video media(mp4 media of snap )
  - Text(list of recognized text from video )
  - full_text  (full String  of recognized text )
  - predicted_class(0 = AD , 1 = not-AD )
  - predicted_class_name(Advertisement                ,
    Not_Advertisement)
  - Licensed(Yes , No)



*Snap chat Data frame*

# Data Description and Structure

- SanpChat Ad features data structure

**Folder tree data structure** represented as ;
- Root directory ( each user screen name on snapchat )
- (mp4 media of snap )
- Stories sub directory (media files .mp4 , .png and .wav)
- Extracted frames sub directory (mp4 file frames for OCR )
- Video frames sub directory (for each video in extracted frames directory confatins jpg)
- Transcript.txt ( recognized text from mp4 files )
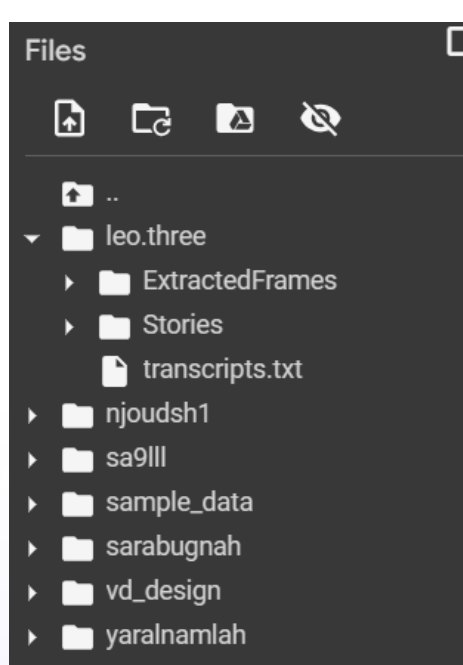


*Root directory*
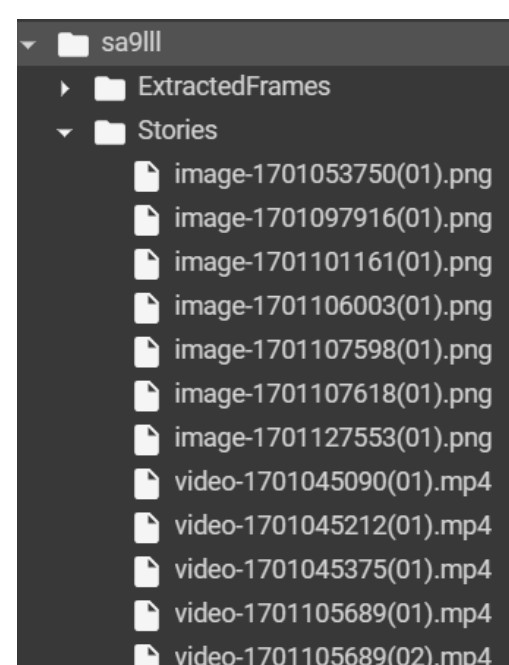


*Stories sub directory*
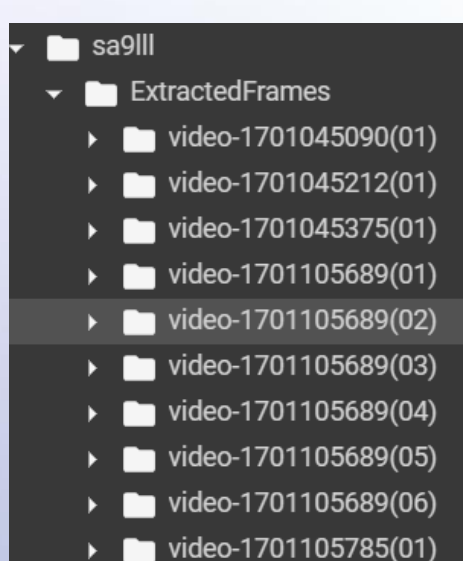


*Extracted frames sub directory*



*Video frames sub directory*

# Data Description and Structure

- **Mawthooq lincencd** dataset structure , the purpose of creation is to mimic  the GCAM data set of licensed users , created using rabid Api .

  **Data frame** with the following columns 2 as  ;
  - Screen_name (user screen name on X or Snapchat  )
  - Lincense_statues (if the provide screen name has Mawthooq Lincense )

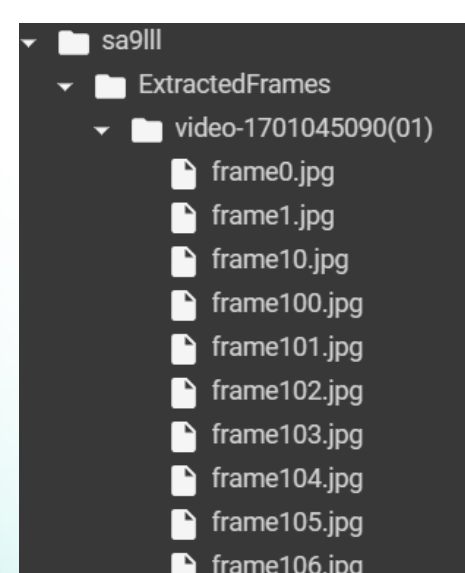| | A | B | C |
|---|---|---|---|
| | | Screen_name | license_status |
| | 0 | abdul_huss | licensed |
| | 1 | areejun | licensed |
| | 2 | sarabugnah | licensed |
| | 3 | ji.ij699 | licensed |
| | 4 | yaralnamlah | licensed |
| | 5 | njeel_d | licensed |
| | 6 | bos3d209 | licensed |
| | 7 | a_sdsk | licensed |
| | 8 | leo.three | licensed |
| | 9 | vd_design | licensed |
| | 10 | byjoudalnamlah | licensed |
| | 11 | sarah.artist | licensed |
| | 12 | muruj.moe | licensed |
| | 13 | suy-91 | licensed |
| | 14 | dodi24_m | licensed |
| | 15 | totiyaa | licensed |
| | 16 | wadjda1 | licensed |
| | 17 | alrahafh | licensed |
| | 18 | justlena20 | licensed |
| | 19 | w.halawani1417 | licensed |
| | 20 | hlm816111 | licensed |

*Mawthooq Dataset*

# Methodology

The project involves two processes as we worked with data from both Twitter and Snapchat platforms.

1. Data Collection:
  - To demonstrate the model's applicability in real-world scenarios, we collected data from the actual platforms. The data collection process included gathering Tweets, Snaps, and information about licensed users from Mawthoq.

2. Text Preprocessing:
  - The text preprocessing stage aimed to clean and prepare the collected data for further analysis. The following steps were performed:
    - Noise Removal: Non-textual content, such as emojis, stopwords, repeated letters, unrelated mentions and hashtags, punctuation, diacritics, and variations in letter shapes, were removed. Additionally, the word ("إعلان"advertisement) was detached from the surrounding letters to standardize its representation.
    - Label Encoding: The labels column, indicating whether the text is an advertisement or not, was converted into a numeric representation using LabelEncoder.

# Methodology

3. Model Training:

  - A model was trained to classify Arabic text as either an advertisement or non-advertisement. The following steps were followed:

    - Model Architecture: A bidirectional LSTM model was built from scratch. The model utilized AraBertv0.2-Large-Twitter embeddings to enhance performance on the Arabic text classification task.

    - Training Data: The model was trained on the collected dataset. Using AraBert weights reduced the chances of encountering out-of-vocabulary issues since the dataset was specifically gathered by us.

4. License Verification:

  - To simulate real-world processes, the model verifies whether the advertiser holds a license issued by the Mawthooq platform.

5. Advertising Compliance Text Check:

  - The model checks if the advertiser includes the word ("اعلان" advertisement) in the text, as required by the AudioVisual Media Authority. Additionally, OCR (Optical Character Recognition) is utilized to extract text from images accompanying the tweets, and videos are converted to text using the SpeechRecognition library for further analysis.

# Methodology

6. Organization Name Identification:
   - OpenAI and LangChain technology are employed to identify the organization to which the advertisement is directed.

7. Unlicensed Advertiser Detection:
   - If the advertiser does not possess a license, the organization's name is displayed. This step serves as a check for organizations to verify the advertiser's licensing status before entering into a contract.

By following these steps, the project aims to replicate and automate the process of advertisement classification, license verification, and compliance checking in real-world scenarios.

# Discussion and Results

The Naqi project successfully assists the Audiovisual Media Authority (GCAM) in monitoring violations of advertising content laws on social media. By utilizing AI technology, Naqi filters audiovisual content to ensure compliance. This streamlines the monitoring process, creating a responsible online environment that upholds Saudi Arabia's advertising content regulations. The project's results align with the objectives of Saudi Vision 2030, contributing to a compliant and thriving digital ecosystem. It is important to address any limitations or challenges encountered during the project to ensure ongoing improvements.

# Conclusion and Future Work

## Conclusion:

Naqi is a valuable initiative for the field of digital advertising, which is experiencing tremendous growth within the Kingdom of Saudi Arabia.

The project is ambitious… but it is achievable with the support of SDAIA and GCAM. By working together, we can create a safe, reliable, and beneficial digital future for all Saudis.

## Future work:

1- Applying the module to other social network like Instagram and TikTok.

2- Violating multiple items, such as customs and traditions, and advertising a product prohibited in the Kingdom of Saudi Arabia.

3- Violating the exploitation of children for nefarious purposes.

SDAIA
الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority

أكاديمية طويق
TUWAIQ ACADEMY

# Team

**SARAH ALDAKHIL**

**ASEEL ALTALHAH**

**RAZAN ALHASSAN**

**RAGHAD ALADAWI**