Software Requirement Specification Document for Mass Media Brand Listening Tool

Lara Ghoniem, Mohammed Yasser, Salma Helmy, Omar Wael
Supervised by: Dr. Taraggy Ghanim, Eng. Mahmoud Heidar

May 23, 2023

Table 1: Document version history

Version	Date	Reason for Change
1.0	17-Dec-2022	SRS First version's specifications are defined.
2.0	5-Mar-2023	Updated system overview and database description.
3.0	2-May-2023	Updated system overview, data design and preliminary object-oriented domain analysis.

GitHub: `https://github.com/LaraGhoniem/MassMediaBrandListener`

Contents

Abstract

Nowadays, a company's reputation is a main determinant of its success, so it's beneficial for companies to be aware of what observers are saying about them in the media so that they may respond correctly. The proposed system aims to facilitate the process of staying up to date with the latest mentions of companies by allowing clients to enter keywords of what they want to monitor and select the media platforms they want to search. Media platforms include social media platforms, blogs and articles, radio stations, podcasts, and television. If the system finds a match on the chosen platforms, it'll then alert clients with the awaited context and process the data received to provide them with data analysis and visualization features, such as sentiment analysis, topic analysis, text summarization, etc. The system will help companies easily manage their reputation and attain PR insights.

1 Introduction

1.1 Purpose of this document

The purpose of the SRS document is to visualize the main functionalities of the proposed system: Mass Media Brand Listening Tool. Moreover, the document also provide the development process used during the implementation phase of the project to simplify reading the document for developers.

1.2 Scope of this document

The scope of the SRS document is to view information needed to understand the requirements and to show the possible constraints and obstacles of the proposed system.

1.3 Business Context

Brands may have a hard time tracking all their mentions on all platforms and viewing people's reviews. Media listeners are the best solution for tracking and viewing all information needed regarding the business name. The goal is to help businesses keep track of mentions of their brand name and view them with several statistics that will help businesses analyse each mention.

2 Similar Systems

2.1 Academic

1. **AlSalman [1] An Improved Approach for Sentiment Analysis of Arabic Tweets in Twitter Social Media** This paper's main purpose is to improve Arabic sentiment analysis by introducing a new approach that mainly comprises a Discriminative Multinomial Naïve Bayes (DMNB) classifier and term frequency-inverse document frequency (TF-IDF) techniques. In addition, they used a 5-fold cross-validation method in the evaluation of the results. The dataset used was a public Twitter corpus dataset that includes 2000 Arabic tweets divided into either positive or negative labels. During tweet processing and normalization, words are

tokenized using the 4-gram technique and then stemmed using Khoja-stemmer to clear out any stop words. TF-IDF then calculates the occurrences of the word. The DMNB classified tweets into positive or negative, and the results demonstrate that the DMNB classifier achieved an accuracy of 87.5%. The paper is limited to sentiment analysis for tweets and doesn't involve other social media platforms.

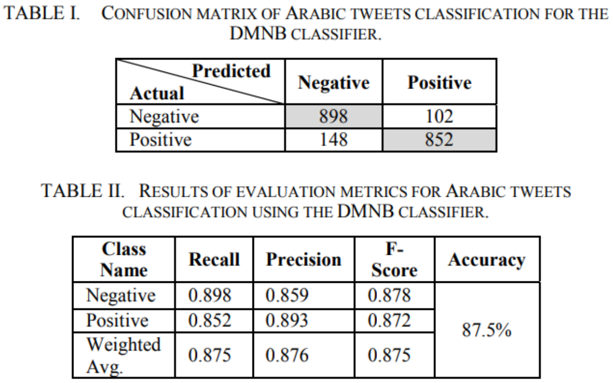


TABLE I. CONFUSION MATRIX OF ARABIC TWEETS CLASSIFICATION FOR THE DMNB CLASSIFIER.

Predicted Actual	Negative	Positive
Negative	898	102
Positive	148	852

TABLE II. RESULTS OF EVALUATION METRICS FOR ARABIC TWEETS CLASSIFICATION USING THE DMNB CLASSIFIER.

Class Name	Recall	Precision	F-Score	Accuracy
Negative	0.898	0.859	0.878	87.5%
Positive	0.852	0.893	0.872	
Weighted Avg.	0.875	0.876	0.875	

Figure 1: Performance of DMNB Classifier [1]

2. **Pala et al. [2] Real-time transcription, keyword spotting, archival and retrieval for Telugu TV news using ASR**
   Pala et al. [2] aimed to facilitate the tedious process of monitoring news channels to discover if certain topics were brought up. Pala et al. stated that only a few Automated Speech Recognition (ASR) systems work with Indian languages and those tend to have limited datasets. The authors chose to tackle this issue by creating a Deep Neural Networks based ASR system that transcribes live and uploaded videos and then spots specified keywords. The language model was generated using the SRILM toolkit and the acoustic models were based on a Subspace Gaussian Mixture Model (SGMM). Support Vector Machines were used to identify speech and non-speech segments in the videos. Multiple databases were used, including the IIITH neutral emotion database, IITKGP-SESC corpus, DD-Telugu corpus, as well as a newly created database of 65-hour recordings from archived Telugu TV news on YouTube and 160,271 lines of Telugu text from different websites. The system achieved a 75% recognition accuracy which went down to 57% when the test data transcriptions were not in the language modeling. The system only works with the Telugu language and should be expanded to recognize new languages. Moreover, it's limited to news from YouTube live streams and does not include any other web content, such as online newspapers, blogs, social media, etc.

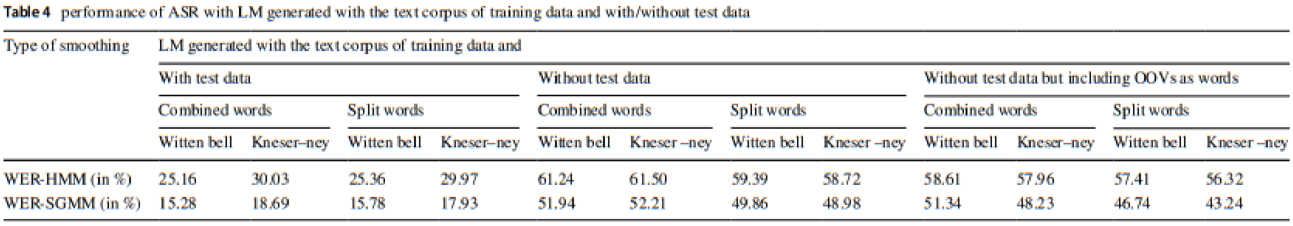


Table 4 performance of ASR with LM generated with the text corpus of training data and with/without test data

Type of smoothing	LM generated with the text corpus of training data and											
	With test data				Without test data				Without test data but including OOVs as words			
	Combined words		Split words		Combined words		Split words		Combined words		Split words	
	Witten bell	Kneser–ney	Witten bell	Kneser–ney	Witten bell	Kneser –ney	Witten bell	Kneser –ney	Witten bell	Kneser–ney	Witten bell	Kneser –ney
WER-HMM (in %)	25.16	30.03	25.36	29.97	61.24	61.50	59.39	58.72	58.61	57.96	57.41	56.32
WER-SGMM (in %)	15.28	18.69	15.78	17.93	51.94	52.21	49.86	48.98	51.34	48.23	46.74	43.24

Figure 2: Performance of ASR [2]

3. **Gao et al. [3] Brand Data Gathering From Live Social Media Streams**
   In [3] their main purpose was to listen to social media streams such as Twitter, Facebook, and others. The main problem is that it is challenging to analyze the content of short posts and to monitor the huge data volume of social media content efficiently. Moreover, conversations "often shift" across all social media. The team states that a "traditional keyword-based approach" can be inaccurate. Gao et al. proposed a "multi-faceted" brand monitoring technique that collects data based on several factors, such as keywords, users, relations, and locations. They also use visual content because of its dominance in most posts. The dataset used in the system is a microblog dataset (Brand-Social-Net) that contains brand/product information. Using this dataset, the system was able to gather information related to the brand from live social media streams. The dataset uses only the English language, which limits the use of the system to other languages.

2.2 Business Applications

1. **MeltWater: Media Monitoring tool and Social Listening Platform**

   Meltwater is a real-time media intelligence platform that has over 30,000 clients. A single, simple-to-use platform to track, notify, distribute, analyze, and report mentions over several media sources. The supported languages are Arabic, English, Germany, Spanish, Hindi, and so on. The supported systems are Twitter, Youtube, Reddit, Linkedin, Online News, Podcasts, and so on. The system features tend to involve unlimited global media monitoring news, social, broadcast, and podcasts, customizable newsletters-, media analysis and reporting dashboards, automated reports, and external software integration via API.

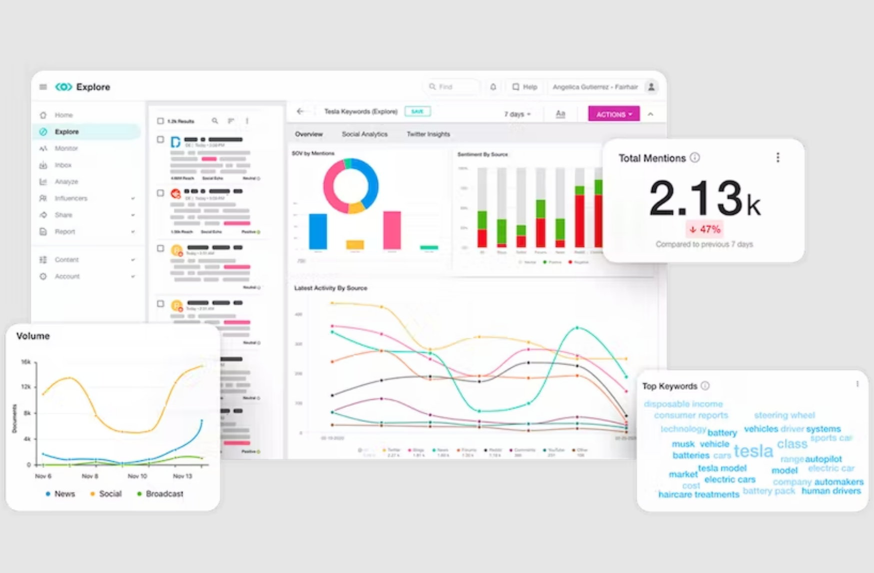


Figure 3: Meltwater UI for Keyword Search

2. **CrowdAnalyzer: Crowd Analyzer Dashboard, Monitor and Analyze Social Media Platforms**
   The first Arabic analyzer aims to provide your company with the most accurate and relevant data when it comes to listening to your customers on social media and eventually creating the most successful social media campaigns and marketing plans. The supported languages are Arabic and English, the supported systems are articles, Twitter, Facebook, and Instagram, and the system features include a dashboard, command center, and offline media monitoring.

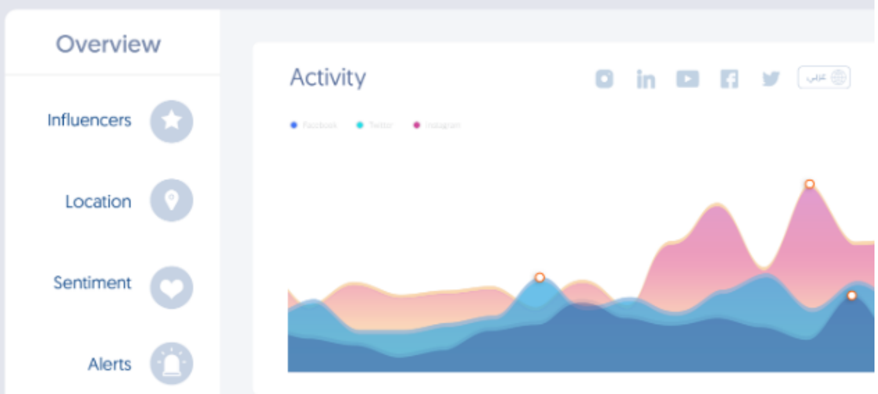


Figure 4: Crowd Analyzer Overview

3 System Description

3.1 Problem Statement

The process of brand monitoring has rapidly changed due to the evolution of the Internet. Manual efforts to monitor mentions of certain companies are no longer efficient due to the vast number of media sources and audiences. The internet has given everyone the power to strongly express their thoughts and experiences and influence others, so companies are no longer marketing to passive audiences. In fact, a study done by IZEA Worldwide, Inc. [4] concluded that 61% of consumers trust friends, family members, or influencers online when it comes to brand recommendations, and only 38% trust online recommendations from brands. Thus, reputation management is vital for the success of brands. In the past few years, media monitoring tools have helped with managing brand reputation, but most only offer links to the data sources the brand mentions were included in, with no summary of the context. Only a few similar tools exist for mass media, including broadcast, online and social media, and these tend to be overpriced. Similar tools are not optimized for the Egyptian dialect, and very few exist for the Arabic language.

3.2 System Overview

The proposed system aims to provide an efficient way to detect brand mentions and keywords across various media sources and provide the user with the associated emotion, summary and link of each brand mention. As show in figure 5, the proposed system includes the following stages:

1. The user enters the keyword to be monitored.

2. Data (Audio and text) retrieval containing keyword mentions across varied platforms.

3. Automated Speech Recognition (ASR) is used to transcribe the downloaded audio files.

4. Text preprocessing is applied on all text results.

5. Spam detection is implemented on the retrieved tweets using an SVM model.

6. Translating transliterated Franco-Arabic to Arabic is implemented using a LSTM and MLP model.

7. Sentiment analysis is predicted using *Mazajak* [5], an online sentiment analysis API, which achieved an accuracy of 0.92 for the Arabic Speech Act and Sentiment corpus of tweets (ArSAS).

8. Text summarization is implemented using a BERT2BERT pre-trained model[6].

9. Topic Modeling is implemented using the BERTopic algorithm[7] with Arabertv2 embeddings[8].

10. The summary and sentiment of each mention is displayed on a dashboard and included in daily and monthly reports.

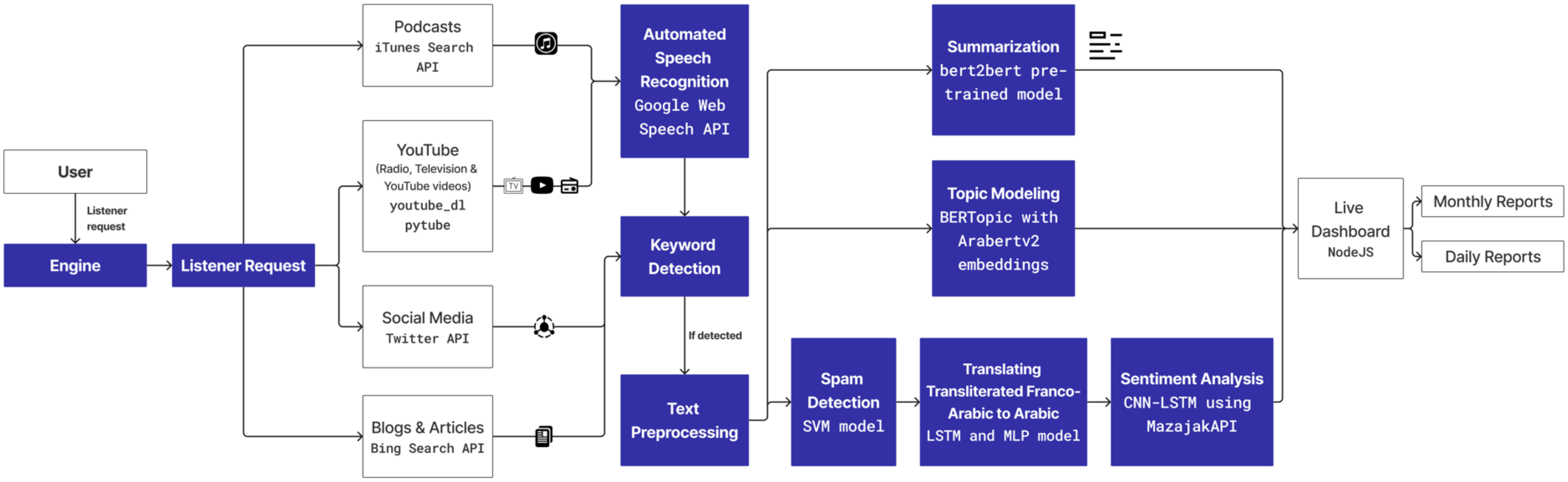


Figure 5: System Overview Diagram

3.2.1 Listener Request

After the user registers for an account, they are prompted to select their business category from a list of categories that describe different establishments, as well as the platforms to search in, and then submit the keyword which will be monitored on these platforms.

3.2.2 Data Retrieval

Podcasts In order to retrieve information about podcasts, the iTunes Search API [9] was used. The proposed system retrieves information from podcasts that include the keyword in their titles or that were found probable to mention the company, based on a predefined list that matches the company's business type with corresponding podcasts. The Search API returns the results in JavaScript Object Notation (JSON) format. The podcast RSS feed returned in the results is used in the command line tool *poddl [10]*, which downloads the complete podcast in MP3 format.

YouTube and Television The proposed system listens for the business's specified keyword in the predefined list of YouTube and television channels likely to mention the business, in addition to YouTube videos which include the keyword in their titles. Therefore, The library *youtube-search-python [11]* was used for the purpose of returning the YouTube links along with other information about the videos with the keyword included in their titles. The library *pytube [12]* was used to download the YouTube videos.

Twitter The Twitter API v2 [13] is used to retrieve tweets with the specified keyword included.

Blogs and Articles The Bing Search API [14] retrieves the links of blogs and articles that contain the keyword. Subsequently, the python library *fulltext [15]* was used in order to extract and download text from the article and blog links.

3.2.3 Automated Speech Recognition (ASR)

The audio files downloaded from different podcasts, television channels, YouTube videos, and radio stations have to be converted to WAV files with a sample rate of 22050 Hz for further processing. The conversion was done using *ffmpeg [16]*, an open-source software project. Afterwards, the converted files are trimmed into one minute WAV files each using the python package *librosa [17].* Unwanted noise, music or long silence pauses are removed from the audio files. The audio preprocessing is complete and the Google Web Speech API [18] is then used for the transcription of the audio files. After the transcription, the audio files are deleted.

3.2.4 Text Preprocessing

Usernames included at the beginning of tweets were removed from the results retrieved from the Twitter API with the aim of targeting more relevant tweets mentioning the keyword. Spam detection and removal is applied on the list of tweets. For all the results, the text preprocessing was divided into four main steps:

- Sentence segmentation[19].
- Stop word removal[20].
- Stemming[21].
- Lemmatization[22].

3.2.5 Spam Detection

After preprocessing, the proposed system removes the tweet tags to filter unwanted mentions. Moreover, there is an rbf-based SVM model built over a 45000 instance tweets-only dataset to detect spam tweets. The model contains several basic text preprocessing, such as removing stopwords, emojis, and special characters. Over the same dataset, other models were created but with lower accuracies than the SVM model with an 80%. For example, neural network models such as MLP and CNN gave an accuracy of 70%. Some models, such as naive Bayes and linear regression, returned nearly the same accuracy.

3.2.6 Text Summarization

The text samples from the platforms will enter as input to the summarization module. The module contains a model that uses the BERT2BERT architecture [6]: a sequence-to-sequence model pretrained over an 84000-instance dataset. The dataset consists of two columns: the paragraph and its corresponding summary. The transformers package[23] was used to access the BertTokenizer and AutoModelForSeq2SeqLM classes.

The BertTokenizer class is used to tokenize the input text and convert it into a format the model can understand. It uses the BERT tokenization scheme, which involves breaking the input text into subword tokens. The main benefit is handling out-of-vocabulary words and reducing the vocabulary size.

The AutoModelForSeq2SeqLM class is used to fine-tune the BERT2BERT model for the task of summarization. This class includes the pre-trained weights of the model. The fine-tuning process involves feeding the model pairs of input-output sequences to minimize the difference between the predicted and expected outputs.

3.2.7 Live Dashboard

After passing through all modules the mentions will be viewed in a dashboard with all statistical information obtained from the results.

3.3 System Scope

- To provide an Arabic Brand Listening tool with acceptable results.
- To provide a variety of options of platforms to search in.
- To provide a text summary for the text and audio retrieved using NLP techniques.
- To provide sentiment analysis to display the number of positive and negative mentions.
- To provide links for the mentions.
- To provide daily and monthly reports with the results visualized.

3.4 System Context

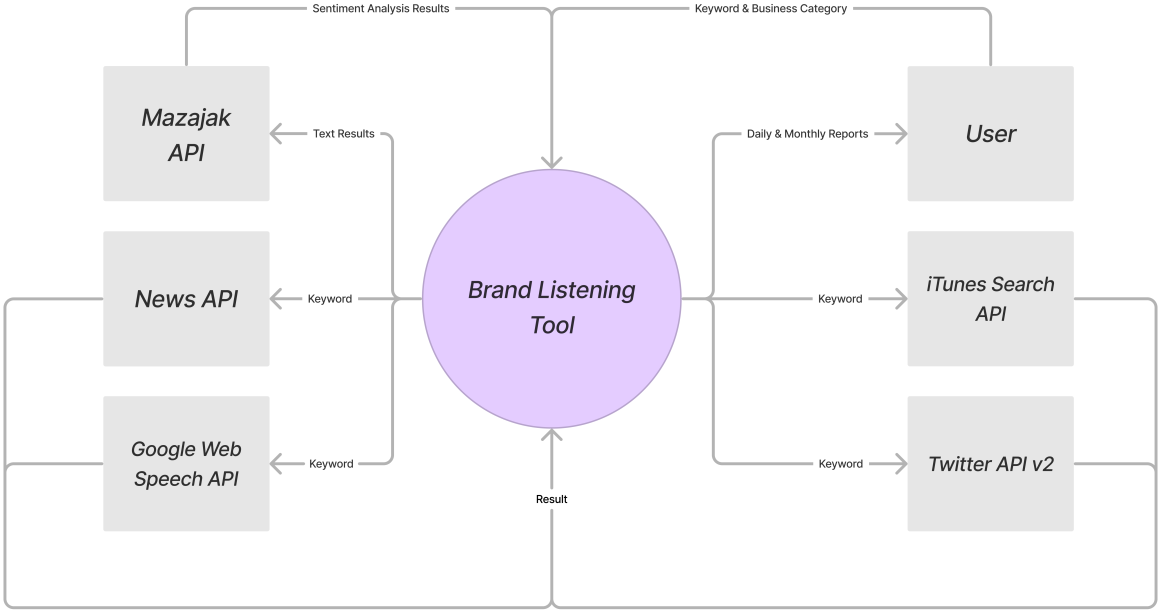


Figure 6: Context Diagram

3.5 Objectives

The main goal is to develop a user-friendly web application that will allow all business owners to make more informed decisions. The client will be asked to enter a keyword and their business category; if the system finds content that matches the specific keywords, the following information will be presented to the user: a text summary and sentiment analysis results of the mentions, analytical information, and hyperlinks to the mentions.

3.6 User Characteristics

- The user can be an individual, company, organization, firm, or any type of business.
- The user can be an adult or elderly.
- The user can be a student or a graduate.
- The user must have basic computer skills.
- The user can use the system for various reasons, such as monitoring their brand's mentions or exploring the industry.

4 Functional

4.1 System Functions

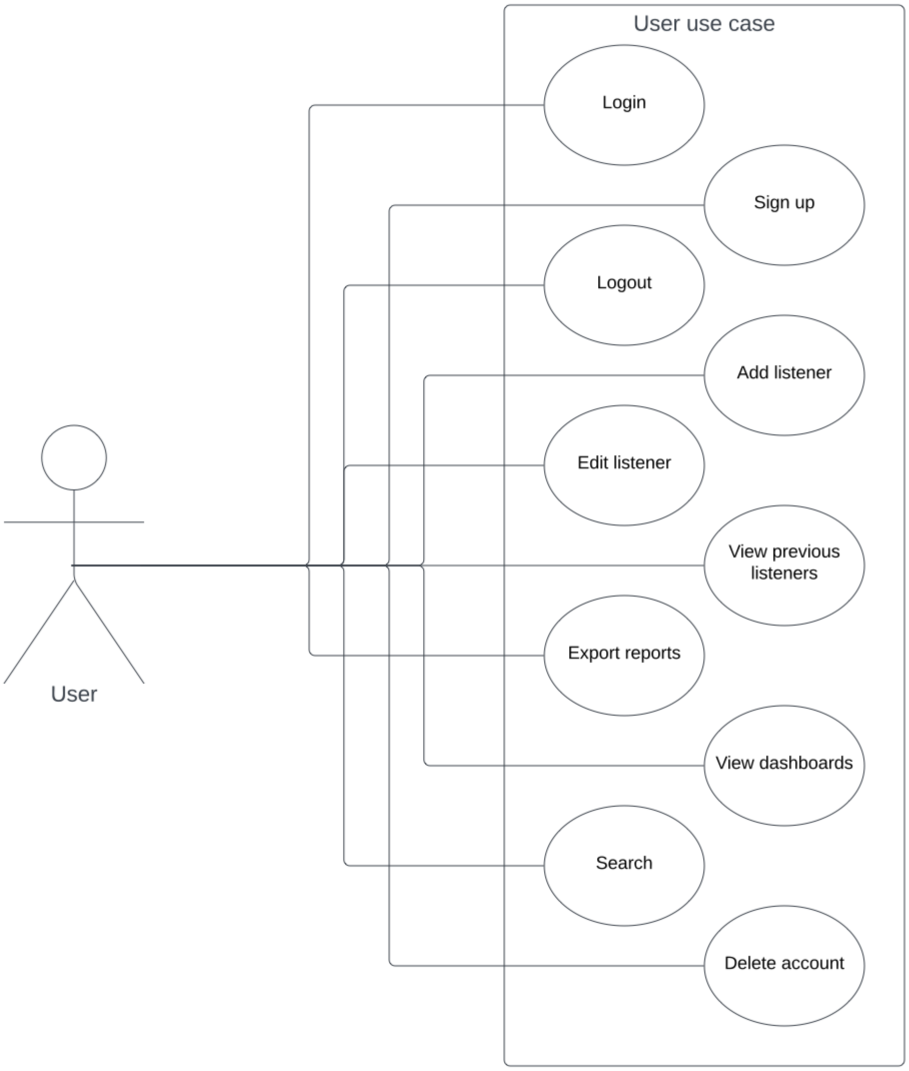


Figure 7: User Use Case

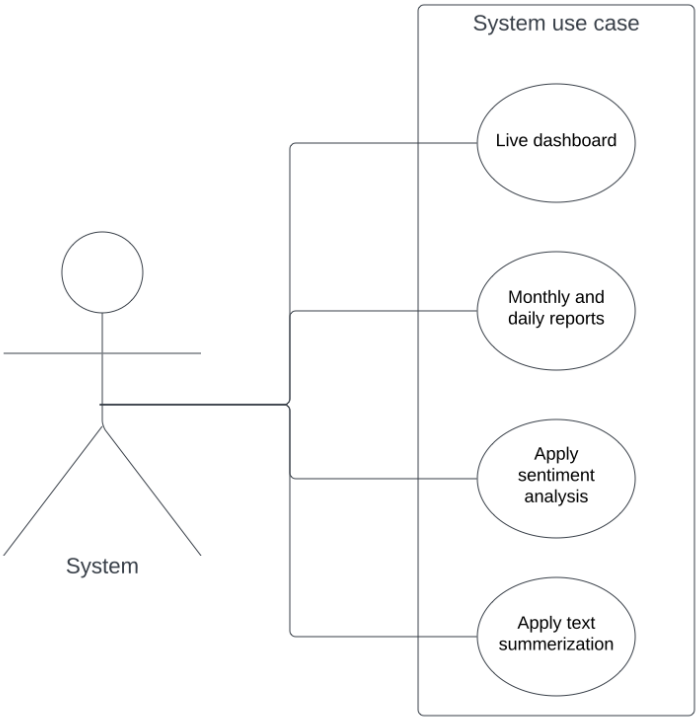


Figure 8: System Use Case

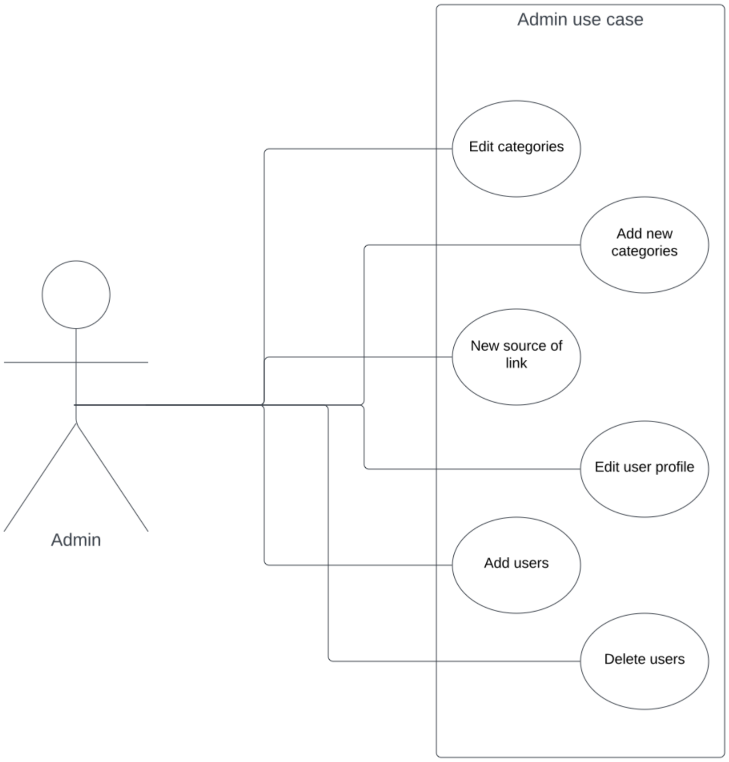


Figure 9: Admin Use Case

FR01	The User shall login.
FR02	The User shall sign up.
FR03	The User shall logout.
FR53	The User shall edit profile.
FR04	The User shall give the proposed system a keyword and the category related to the keyword.
FR05	The User shall add the listener to the query of the keyword.
FR06	The User shall edit the listeners.
FR07	The User shall check previous keyword listeners and its results.
FR08	The User shall view the dashboards.
FR09	The User shall export reports
FR10	The User shall delete account

MR01	The System shall provide the user a dashboard with all statistics related to the keyword given.
MR02	The System shall provide the user with monthly or daily reports.

CR01	The Admin shall edit categories.
CR02	The Admin shall add new categories.
CR03	The Admin shall add new source links for each category.
CR04	The Admin shall add new users.
CR05	The Admin shall edit users.
CR06	The Admin shall delete users.

4.2 Detailed Functional Specification

Table 1: Listener

Name	Listener.
Code	AR01.
Priority	High.
Critical	The user must enter a keyword to view analytical information.
Description	After signing in, he/she will be able to enter the keyword into our system.
Input	Enter the name of the brand.
Output	Group of analytical information.
Pre-condition	Must be logged in.
post-condition	Re-direct to home page.
Dependency	FR01
Risk	User enter invalid keyword or keyword that does not exists.

Table 2: Sentiment analysis

Name	Sentiment analysis.
Code	AR02.
Priority	Medium.
Critical	The user must enter a keyword to view the sentiment analysis.
Description	After the user sign in, he/she will be able to search for any brand then, Sentiment analysis will be displayed.
Input	Enter the name of the brand.
Output	Sentiment(Positive, negative, neutral)
Pre-condition	Must be logged in and, keyword detected.
post-condition	Sentiment will be displayed.
Dependency	AR01.
Risk	Low accuracy information will be displayed.

Table 3: Summarization

Name	Summarization.
Code	AR03.
Priority	Medium.
Critical	The user must enter a keyword to view the summary.
Description	After signing in, a detailed summary shall be displayed to the user from the mentions of the keyword given.
Input	Text from media sources.
Output	Summary.
Pre-condition	Must be logged in and, keyword detected.
post-condition	Summary will be displayed.
Dependency	AR01.
Risk	Low accuracy information will be displayed.

Table 4: Dashboard

Name	Live Dashboard.
Code	AR04.
Priority	High.
Critical	The user must enter a keyword to view the live dashboard.
Description	After entering the keyword into our system, live dashboard will be displayed.
Input	Information from media sources.
Output	live dashboard.
Pre-condition	Must be logged in and, keyword detected.
post-condition	live dashboard will be displayed.
Dependency	AR01, AR02, AR03
Risk	Low accuracy information will be displayed.

Table 5: Reports

Name	Reports.
Code	AR05.
Priority	Low.
Critical	The user can export the information as daily or monthly reports.
Description	After viewing the live dashboard the user will be able to export the information gathered into forms of PDF as daily or monthly reports.
Input	Information from media sources.
Output	Daily and Monthly reports.
Pre-condition	Must be logged in and, keyword detected.
post-condition	Daily and Monthly reports.
Dependency	AR04.
Risk	Low accuracy information will be exported.

5 Design Constraints

5.1 Standards Compliance

The web application shall be hosted on a web server for final deployment. Internet connection and web browser are required in order to display or create listener.

6 Non-functional Requirements

6.1 Usability

The website will provide a user-friendly interface with a professional UX design.

6.2 Availability

The system shall be available at all times so that the user can access the dashboard at any time.

6.3 Portability

The system will be available on the internet, so it is accessible from any internet-connected device regardless of the operating system.

6.4 Scalability

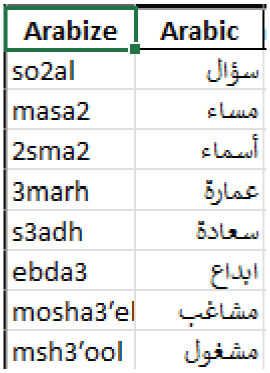
The system should be able to handle an increasing volume of data while maintaining acceptable performance.

7 Data Design

There are 3 datasets that are used for the proposed system. The models that utilise the datasets are the Franco-Arabic translation, topic modeling and spam detection models. The datasets included are 'Arabic-Franco Dataset'[24], 'OSIAN: Open Source International Arabic News Corpus'[25], and 'Arabic Text Summarization 30_000'[26].

7.1 Arabic-Franco Dataset

The dataset[24] used for translating Franco-Arabic to Arabic language in order to ease the process of analyzing the text. The dataset contains nearly 20000 instance of Franco-Arabic text and its corresponding Arabic text.

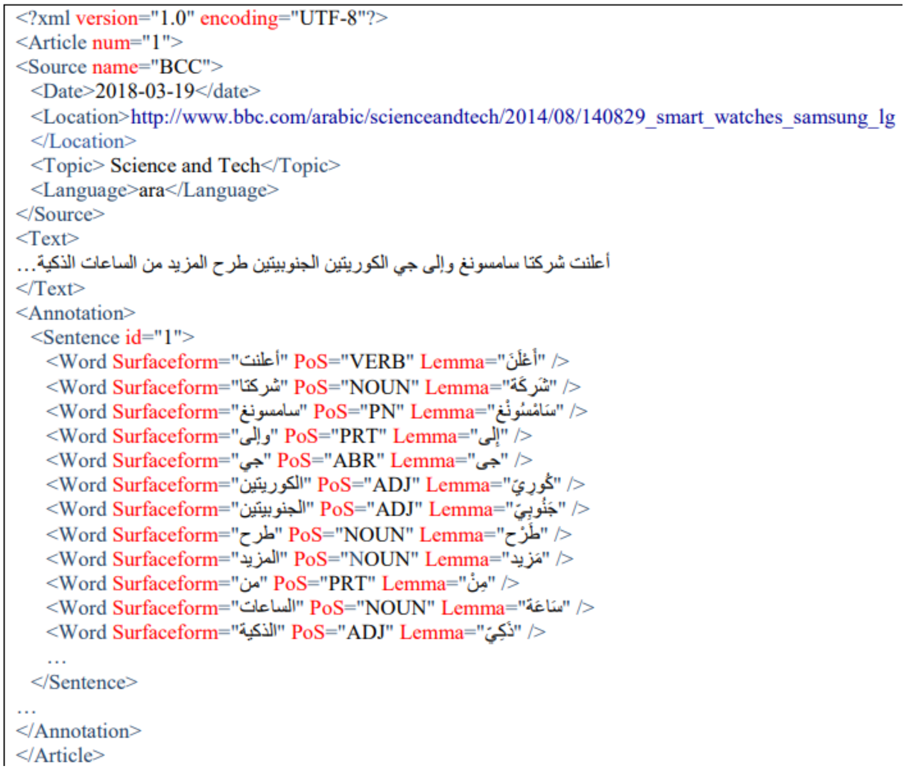


Arabize	Arabic
so2al	سؤال
masa2	مساء
2sma2	أسماء
3marh	عمارة
s3adh	سعادة
ebda3	ابداع
mosha3'el	مشاغب
msh3'ool	مشغول

Figure 10: Arabic-Franco Dataset Dataset Sample

7.2 OSIAN: Open Source International Arabic News Corpus

The OSIAN (Open Source International Arabic News) Corpus[25] was used along with the Arabertv2 algorithm[8] in the topic modeling stage. The corpus contains over 3 million articles collected from various international news sources, spanning over 10 years. The articles cover a wide range of topics such as politics, sports, and entertainment, and are written in Modern Standard Arabic. The corpus was preprocessed and integrated into the CLARIN infrastructure, allowing for easy access and searchability. By utilizing this corpus, the proposed system was able to improve the BERTopic technique[7] to analyze and classify the mentions based on their topic and context.

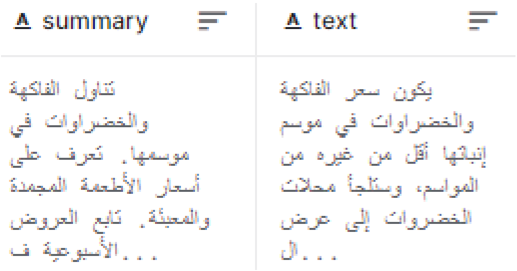


```
<?xml version="1.0" encoding="UTF-8"?>
<Article num="1">
<Source name="BCC">
  <Date>2018-03-19</date>
  <Location>http://www.bbc.com/arabic/scienceandtech/2014/08/140829_smart_watches_samsung_lg
  </Location>
  <Topic> Science and Tech</Topic>
  <Language>ara</Language>
</Source>
<Text>
أعلنت شركتا سامسونغ وإلى جي الكوريتين الجنوبيتين طرح المزيد من الساعات الذكية...
</Text>
<Annotation>
  <Sentence id="1">
    <Word Surfaceform="أعلنت" PoS="VERB" Lemma="أَعْلَنَ" />
    <Word Surfaceform="شركتا" PoS="NOUN" Lemma="شَرِكَة" />
    <Word Surfaceform="سامسونغ" PoS="PN" Lemma="سَامْسُونْغ" />
    <Word Surfaceform="وإلى" PoS="PRT" Lemma="إِلَى" />
    <Word Surfaceform="جي" PoS="ABR" Lemma="جى" />
    <Word Surfaceform="الكوريتين" PoS="ADJ" Lemma="كُورِيّ" />
    <Word Surfaceform="الجنوبيتين" PoS="ADJ" Lemma="جَنُوبِيّ" />
    <Word Surfaceform="طرح" PoS="NOUN" Lemma="طَرْح" />
    <Word Surfaceform="المزيد" PoS="NOUN" Lemma="مَزِيد" />
    <Word Surfaceform="من" PoS="PRT" Lemma="مِنْ" />
    <Word Surfaceform="الساعات" PoS="NOUN" Lemma="سَاعَة" />
    <Word Surfaceform="الذكية" PoS="ADJ" Lemma="ذَكِيّ" />
    ...
  </Sentence>
...
</Annotation>
</Article>
```

Figure 11: OSIAN corpus sample in XML format

7.3 Arabic Text Summarization 30_000

Finally, for the text summarization model, the dataset used is "Arabic Text Summarization 30_000[26]". This dataset contains about 30000 instance of the text and its corresponding summarization.



summary	text
تناول الفاكهة والخضراوات في موسمها. تعرف على أسعار الأطعمة المجمدة والمعبئة. تابع العروض الأسبوعية ف...	يكون سعر الفاكهة والخضراوات في موسم إنباتها أقل من غيره من المواسم، وستلجأ محلات الخضروات إلى عرض ال...

Figure 12: Dataset Sample

7.4 API Response Description

The APIs used in the proposed system for the data retrieval process are YouTube, Twitter, and iTunes. The response of YouTube API is a list of links, titles, publish dates, views counts, descriptions, channel details. Twitter API returns a response that contains a list of the tweet IDs and the text of each tweet. Finally, iTunes Search API returns list of links of the resulted podcasts.

7.5 Database Description

The database is based on MongoDB. There will be 7 tables in the schema. The data stored are in the string format.

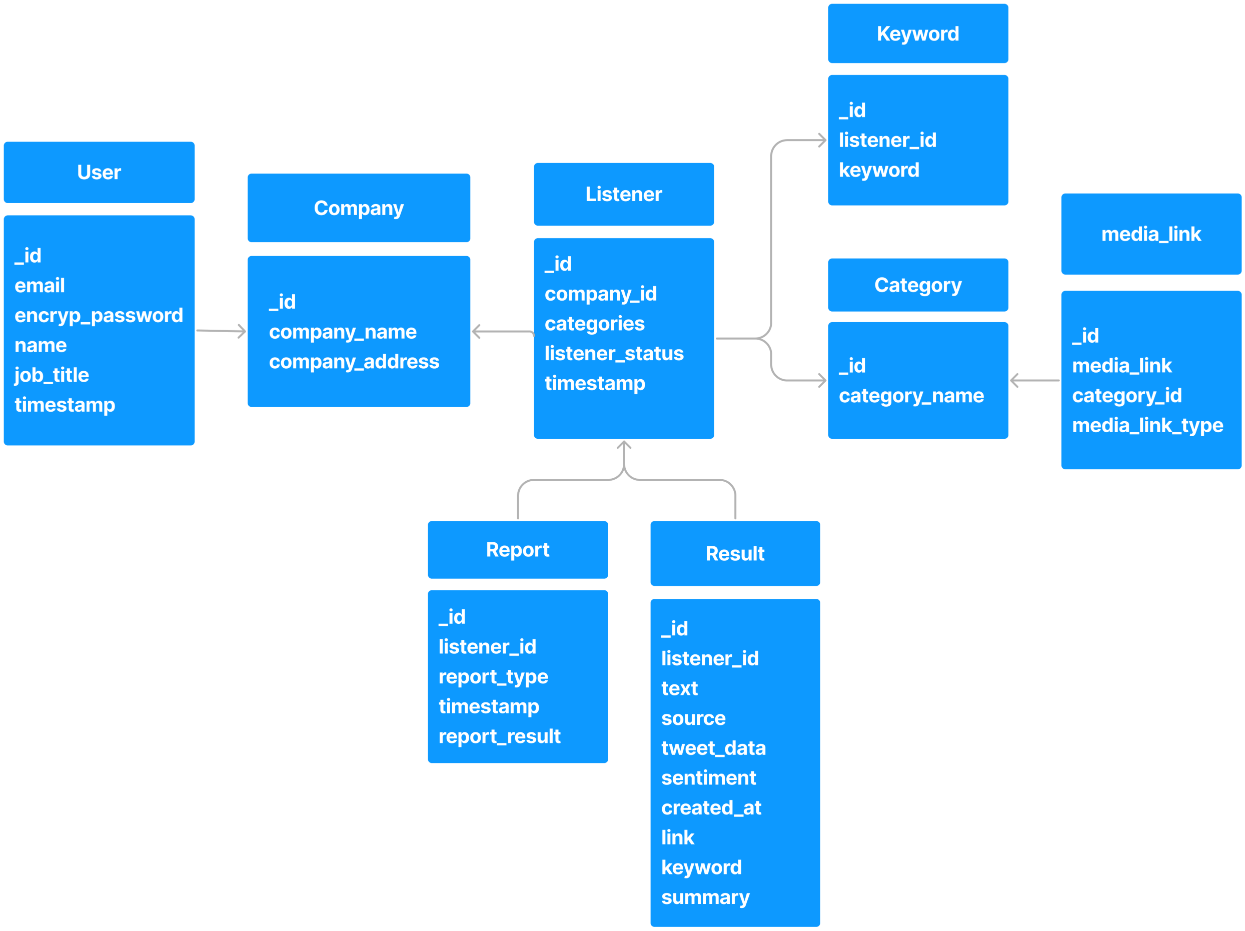


Figure 13: Database Description

8 Preliminary Object-Oriented Domain Analysis

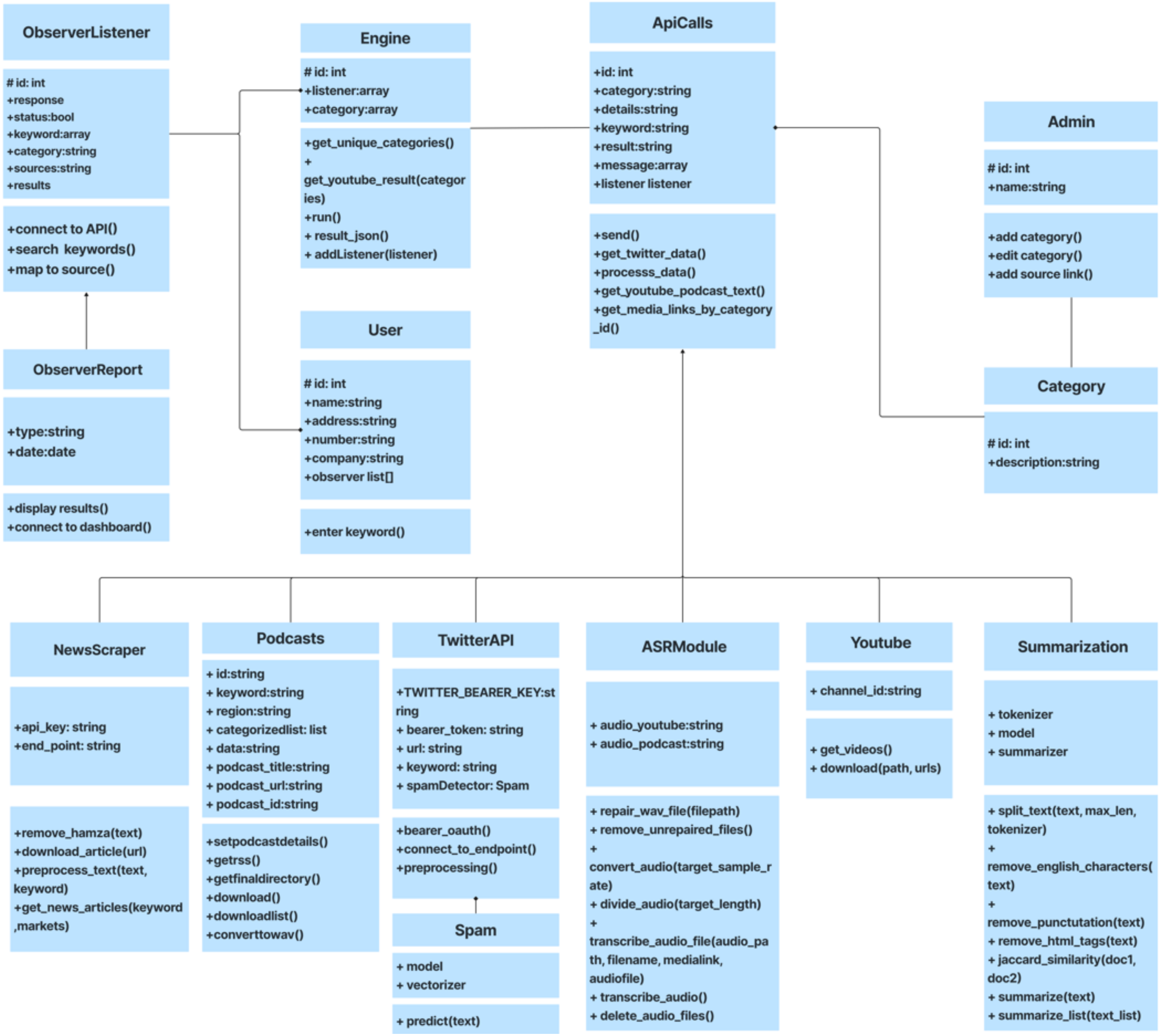


Figure 14: Class Diagram

9 Operational Scenarios

9.1 Scenario 1

The user is able to access the system by registering or logging in. While registering, the user selects their business category and the keywords to be monitored. The system starts listening on these platforms for the keywords and retrieving data with the keywords' mentions.

9.2 Scenario 2

The user logs into the system and if their brand was mentioned, the user is able to view the text with the mention, along with the link to the mention if applicable in a dashboard. The system will process the audio or text retrieved from the media source then sentiment analysis and text summarization will be implemented. The user can find the sentiment and summary of each mention on the dashboard, along with related statistics such as the volume of the keyword's mentions.

9.3 Scenario 3

The user starts by logging into the system and clicking view report. The user is then able to view either daily reports or monthly reports with the user's request results visualized and summarized.

10 Project Plan

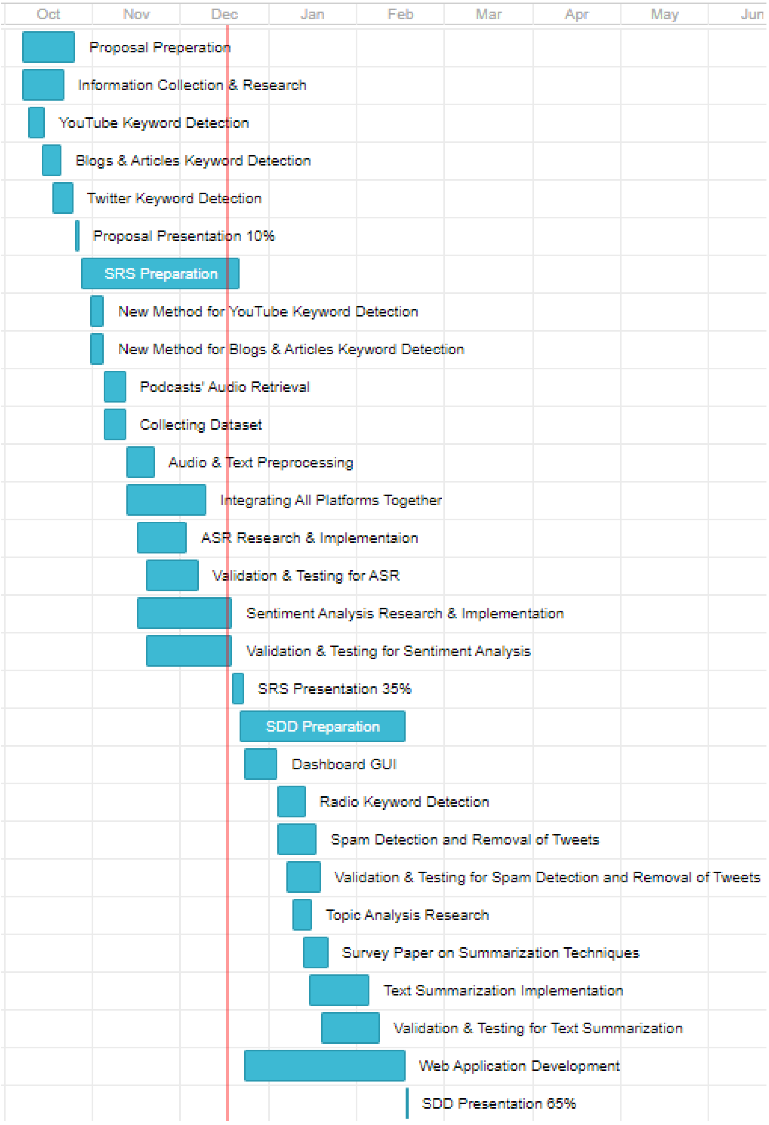


Figure 15: Mass Media Brand Listening Tool Gantt Chart

Team members are assigned numbers to be used in the time plan table

1. Lara Ghoniem
2. Mohammed Yasser
3. Salma Helmy
4. Omar Wael

ID	Task	Start Date	End Date	Assigned To
1	Proposal Preperation	10/7/22	10/25/22	1,2,3,4
2	Information Collection & Research	10/7/22	10/21/22	1,2,3,4
3	YouTube Keyword Detection	10/9/22	10/14/22	3
4	Testing ASR Tools	10/9/22	10/14/22	4
5	Blogs & Articles Keyword Detection	10/14/22	10/20/22	1
6	Twitter Keyword Detection	10/18/22	10/24/22	2
7	Proposal Presentaion	10/26/22	10/26/22	1,2,3,4
8	SRS Preparation	10/28/22	12/21/22	1,2,3,4
9	Collecting Datasets	11/24/22	12/1/22	1,2,3,4
10	New Method for Blogs and Articles Keyword Detection	12/1/22	12/8/22	3
11	Update On Twitter Keyword Detection	12/1/22	12/8/22	2
12	Podcasts' Audio Retrieval	12/1/22	12/8/22	1
13	Audio & Text Preprocessing	12/1/22	12/8/22	1,2,3,4
14	Conducting a Survey	12/1/22	12/8/22	4
15	New Method for YouTube Keyword Detection	12/8/22	12/15/22	4
16	Local API implementation	12/8/22	12/15/22	3
17	ASR Implementation & Validation	12/8/22	12/15/22	2
18	Sentiment Analysis Implementation and Validation	12/8/22	12/15/22	1,2
19	SRS Presentation	12/22/22	12/22/22	1,2,3,4
20	SDD Preparation	12/22/22	2/15/23	1,2,3,4
21	Web Application Development	12/23/22	2/15/23	1,2,3,4
22	Dashboard GUI	12/23/22	1/3/23	1,2,3,4
24	Radio Keyword Detection	1/4/23	1/13/23	3,4
25	Spam Detection and Removal of Tweets	1/4/23	1/17/23	1,2
26	Topic Analysis Research	1/9/23	1/15/23	3,4
27	Text Summarization Implementation and Validation	1/15/23	2/4/23	1,2,3,4
28	SDD Presentation	2/16/23	2/16/23	1,2,3,4

11 Appendices

11.1 Definitions, Acronyms, Abbreviations

CSV	comma-separated values
API	Application Programming Interface
WAV	Waveform Audio File Format
Hz	Hertz
ASR	Automated Speech Recognition
FFMPEG	Fast Forward Moving Picture Experts Group
RSS	Really Simple Syndication
JSON	JavaScript Object Notation
NLP	Natural Language Processing
ArSaS	Arabic Speech Act and Sentiment corpus
TF-IDF	Term Frequency-Inverse Document Frequency
DMNB	Discriminative Multinomial Naïve Bayes

11.2 Supportive Documents

A survey was conducted to confirm which media sources would have the most brand mentions for the different business categories. The survey also provides insights on which Egyptian and Arab media companies, creators, and influencers consumers rely on for reviews, recommendations, and general brand mentions. The results are shown in the figures below.

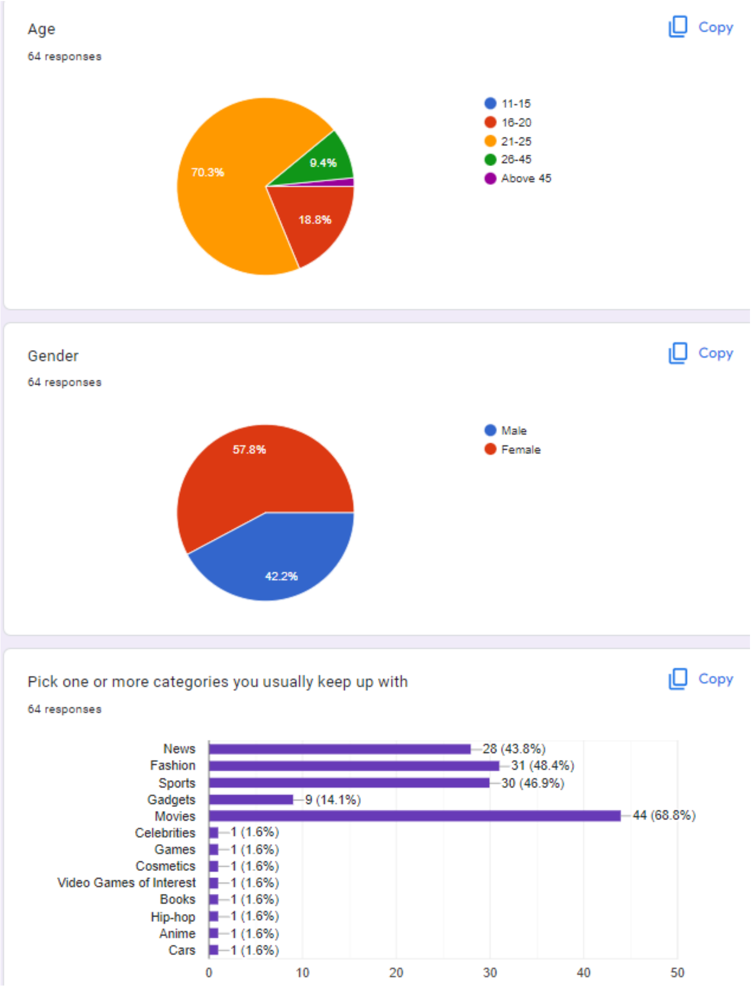
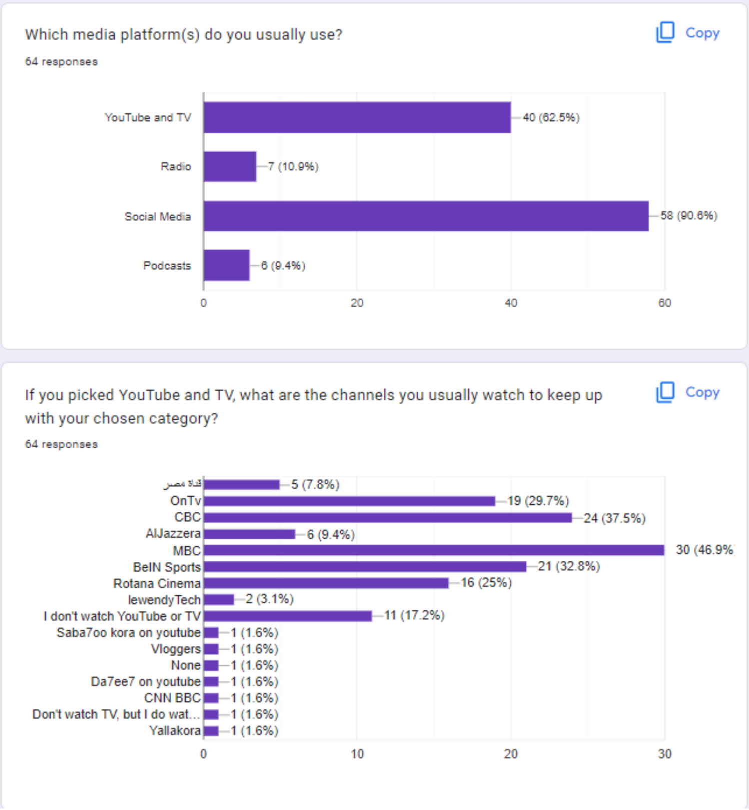


Figure 16: Statistics 1



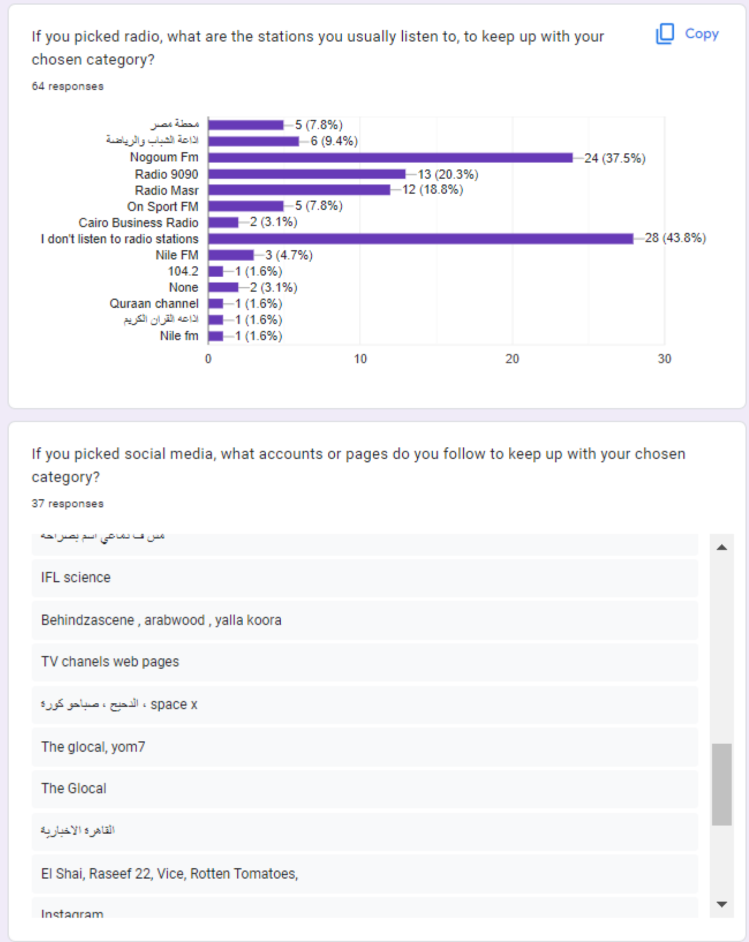
Which media platform(s) do you usually use?

64 responses

If you picked YouTube and TV, what are the channels you usually watch to keep up with your chosen category?

64 responses

Figure 17: Statistics 2



Copy

If you picked radio, what are the stations you usually listen to, to keep up with your chosen category?

64 responses

If you picked social media, what accounts or pages do you follow to keep up with your chosen category?

37 responses

مش فـ دماغي اسم بصراحه

IFL science

Behindzascene , arabwood , yalla koora

TV chanels web pages

الدحيح ، صباحو كوره ، space x

The glocal, yom7

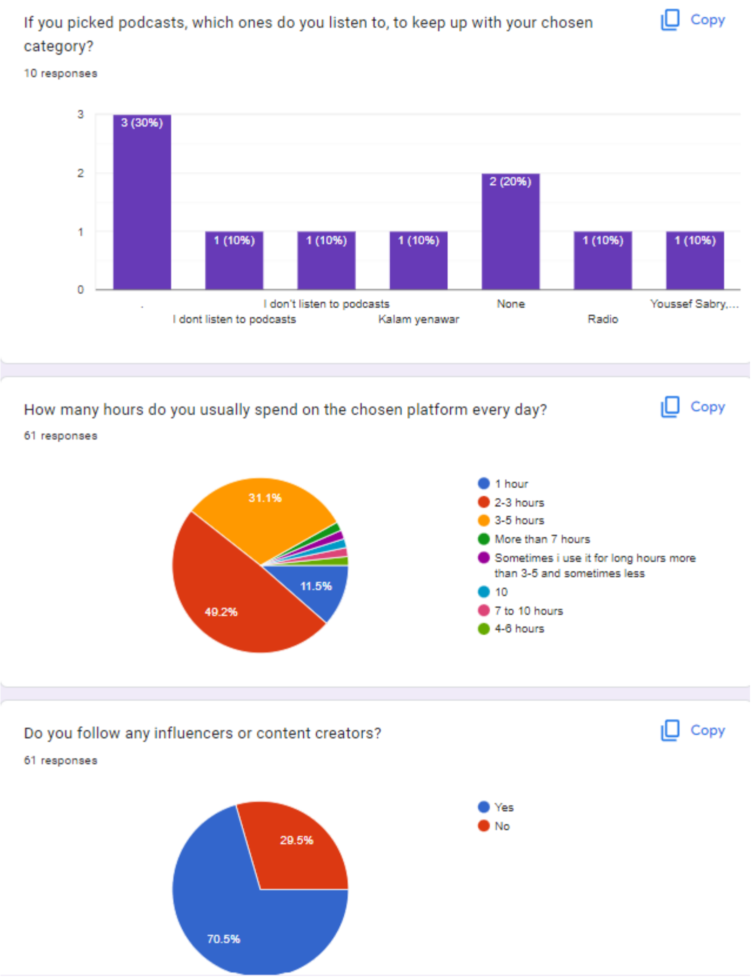
The Glocal

القاهره الاخباريه

El Shai, Raseef 22, Vice, Rotten Tomatoes,

Instagram

Figure 18: Statistics 3



If you picked podcasts, which ones do you listen to, to keep up with your chosen category?

10 responses

How many hours do you usually spend on the chosen platform every day?

61 responses

Do you follow any influencers or content creators?

61 responses

Figure 19: Statistics 4

References

[1] Hussain AlSalman. “An Improved Approach for Sentiment Analysis of Arabic Tweets in Twitter Social Media”. In: *2020 3rd International Conference on Computer Applications Information Security (ICCAIS)*. 2020, pp. 1–4. DOI: 10.1109/ICCAIS48893.2020.9096850.

[2] Mythilisharan Pala, Laxminarayana Parayitam, and Venkataramana Appala. “Real-Time Transcription, Keyword Spotting, Archival and Retrieval for Telugu TV News Using ASR”. In: *Int. J. Speech Technol.* 22.2 (June 2019), pp. 433–439. ISSN: 1381-2416. DOI: 10.1007/s10772-019-09598-6. URL: https://doi.org/10.1007/s10772-019-09598-6.

[3] Yue Gao, Fanglin Wang, Huanbo Luan, et al. “Brand Data Gathering From Live Social Media Streams”. In: *Proceedings of International Conference on Multimedia Retrieval*. ICMR ’14. Glasgow, United Kingdom: Association for Computing Machinery, 2014, pp. 169–176. ISBN: 9781450327824. DOI: 10.1145/2578726.2578748. URL: https://doi.org/10.1145/2578726.2578748.

[4] *Trust in Influencer marketing*. Mar. 2022. URL: https://izea.com/resources/insights/2022-influencer-marketing-trust/.

[5] Ibrahim Abu Farha and Walid Magdy. “Mazajak: An Online Arabic Sentiment Analyser”. In: *Proceedings of the Fourth Arabic Natural Language Processing Workshop*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 192–198. DOI: 10.18653/v1/W19-4621. URL: https://aclanthology.org/W19-4621.

[6] *BERT2BERT*. URL: https://huggingface.co/malmarjeh/bert2bert.

[7] Maarten Grootendorst. *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. 2022. arXiv: 2203.05794 [cs.CL].

[8] Wissam Antoun, Fady Baly, and Hazem Hajj. *AraBERT: Transformer-based Model for Arabic Language Understanding*. 2021. arXiv: 2003.00104 [cs.CL].

[9] *iTunes Search API*. Sept. 2017. URL: https://developer.apple.com/library/archive/documentation/AudioVideo/Conceptual/iTuneSearchAPI/index.html#//apple_ref/doc/uid/TP40017632-CH3-SW1.

[10] Fredrik Blank. *Poddl*. 2022. URL: https://www.fredrikblank.com/poddl/.

[11] Hitesh Kumar Saini. *YouTube-Search-Python*. 2021. URL: https://pypi.org/project/youtube-search-python/.

[12] *Pytube*. 2021. URL: https://pytube.io/en/latest/index.html.

[13] *Twitter API documentation | docs | twitter developer platform*. URL: https://developer.twitter.com/en/docs/twitter-api.

[14] *Bing Search API*. URL: https://www.microsoft.com/en-us/bing/apis/bing-web-search-api.

[15] *Fulltext*. URL: https://pypi.org/project/fulltext/.

[16] *FFmpeg*. URL: https://ffmpeg.org/documentation.html.

[17] Brian McFee, Colin Raffel, Dawen Liang, et al. “Librosa: Audio and Music Signal Analysis in python”. In: *Proceedings of the 14th Python in Science Conference* (2015). DOI: 10.25080/majora-7b98e3ed-003.

[18] André Natal, Glen Shires, Philip Jägenstedt, et al. *Web Speech API*. Aug. 2020. URL: https://wicg.github.io/speech-api/.

[19] Rachel Wicks and Matt Post. “A unified approach to sentence segmentation of punctuated text in many languages”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 3995–4007. DOI: 10.18653/v1/2021.acl-long.309. URL: https://aclanthology.org/2021.acl-long.309.

[20] Dhara J. Ladani and Nikita P. Desai. “Stopword Identification and Removal Techniques on TC and IR applications: A Survey”. In: *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*. 2020, pp. 466–472. DOI: 10.1109/ICACCS48705.2020.9074166.

[21] Mariem Bounabi, Karim El Moutaouakil, and Khalid Satori. "A comparison of text classification methods using different stemming techniques". In: *International Journal of Computer Applications in Technology* 60.4 (2019), pp. 298–306. DOI: 10.1504/IJCAT.2019.101171. eprint: https://www.inderscienceonline.com/doi/pdf/10.1504/IJCAT.2019.101171. URL: https://www.inderscienceonline.com/doi/abs/10.1504/IJCAT.2019.101171.

[22] Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. "Improving Lemmatization of Non-Standard Languages with Joint Learning". In: *CoRR* abs/1903.06939 (2019). arXiv: 1903.06939. URL: http://arxiv.org/abs/1903.06939.

[23] *transformers*. URL: https://huggingface.co/docs/transformers/index.

[24] Bashar Talafha, Analle Abu Ammar, and Mahmoud Al-Ayyoub. "Atar: Attention-based LSTM for Arabizi Transliteration". In: *International Journal of Electrical and Computer Engineering (IJECE)* 11.3 (June 2021), pp. 2362–2370.

[25] Imad Zeroual, Dirk Goldhahn, Thomas Eckart, et al. "OSIAN: Open Source International Arabic News Corpus - Preparation and Integration into the CLARIN-infrastructure". In: *Proceedings of the Fourth Arabic Natural Language Processing Workshop*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 175–182. DOI: 10.18653/v1/W19-4619. URL: https://aclanthology.org/W19-4619.

[26] *Arabic text summarization 30*$_0$00. URL: https://www.kaggle.com/datasets/fadyelkbeer/arabic-text-summarization-30-000.