National College of Ireland

Configuration Manual

MSc Research Project

AI for Business

Rao Waqar Ali

Student ID: x24147346

School of Computing

National College of
Ireland

Supervisor: Anderson Augusto Simiscuka

National College of Ireland

Project Submission Sheet

School of Computing

Student Name: ……Rao Waqar Ali…… **Student ID:**…24147346……

Programme: …MSc AI for Business…… **Year:** …2025… **Module:** …Practicum II…

Supervisor: ……Anderson Simiscuka…… **Submission Due Date:** ….26 Jan 2026…

Project Title: …Configuration Manual... **Word Count:** …2569…**Page Count** …15...

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project. ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:……Rao Waqar Ali…… **Date:** ……11 Dec 2025……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□	
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□	
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not ufficient to keep a copy on computer.	□	

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Rao Waqar Ali x24147346

1. Introduction

This is a configuration document that represents the technical procedure of designing, developing, and replicating the Irish News Dataset project, which was completely created on Python based automation scripts. The proposed project will be targeted to find, filter, label and sort a colossal volume of Irish news content based on three valid and publicly accessible sources, one of which is the RTÉ News, two of which are the The Journal.ie and The Irish Times.

These manual documents the complete, reproducible configuration of the Irish News NLP pipeline described in the Final Report. It covers automated data collection from five Irish news sources, rigorous preprocessing, T5-based summarization, multi-model sentiment analysis (VADER, BERT, BER-Tweet, T5), ensemble voting, basic topic classification, evaluation, and ethical considerations. The manual is designed to enable researchers to replicate the exact environment, scripts, and outputs used to produce the reported results.

All system settings, environment settings and step execution scripts used in the project are properly accounted in the manual, hence replicating the system in the exact way it was done is possible. All the stages of the workflow process including data harvesting, extraction, merging, cleaning, labeling, sentiment analysis and export formatting are defined in a structured and reproducible fashion.

This project has not been performed with the help of external analytical platforms and R Studio. Everything was done on Python 3.11 within the virtual environment of Windows so that the automation process could be controlled fully. The final database comprises more than 500 labeled journals and it is applied as a structured corpus of applications in natural language processing and text analytics.

The objectives of the manual are to ensure technical disclosure, ethical treatment of data, and reproducibility. It will act as a guideline to the assessor and researchers on the step wise process in order to be sure that the workflow is sound, and that the final product is a well evaluated and precise dataset.

2. Scope & Sources

Sources (5): RTÉ News, The Journal.ie, Irish Independent, Irish Examiner, Irish Times. Time window: April–September 2025. Raw collection yielded approximately 1,130 articles across the five outlets; after filtering (duplicates, accessibility, encoding, minimal text length), the consolidated dataset used for modeling contains 699 records.

3. System Requirements

The whole project was implemented on Python 3.11 within a controlled virtual environment on Windows 11. The system and software requirements are needed to ensure the all the scripts would execute successfully and be repeatable encompassed:

3.1. Hardware Requirements

- CPU: Intel core i5 (minimum) or the like. RAM: 8GB minimum or 16GB (recommended)
- to be faster.
- Storage: Free disk space 10GB minimum
- Internet: Access to high-speed Internet (10 Mbps or higher) that is reliable to download information and install packages.

3.2. Software Requirements

- Windows 10 / 11 (64 bit) or MacOS / Linux (equivalent) Operating System
- Python Interpreter 3.10 or higher (tested with 3.11)
- Code Editor: Visual studio code, PyCharm or Jupyter Notebook.
- Browser: Google Chrome (opening URLs manually where necessary and checking them)
- Virtual Environment Tool: Inbuilt venv that handles dependencies within an isolated environment.

3.3. Python Libraries

- Install via pip inside virtual environment:
  pip install --upgrade pip
  pip install newspaper3k feedparser beautifulsoup4 lxml python-dateutil tldextract tqdm vaderSentiment pandas transformers torch
  Optional: CUDA-enabled PyTorch for GPU acceleration (Windows/Linux)

  newspaper3k - the one that is used to extract articles.

- feedparser- RSS feed aggregation.
- beautifulsoup4 and lxml - to create a program to write and read HTML.
- tldextract and dateutil - to format metadata and timestamps.
- tqdm – for progress tracking
- vaderSentiment - to weak sentiment labeling.
- pandas and csv - to format and put out the data.

It is designed in a way that the modules required in web scraping, text processing and labeling are installed and therefore, the project is fully transportable and can be tested in other systems.

4. Project Setup

In the set up, initializing and preparation of working on the Irish News Dataset Project, this section describes the whole setting up procedure. All was developed and executed within a Python 3.11 virtual environment to maintain consistency of dependencies and seclusion of system packages.

4.1. Create the Main Project Directory

A dedicated workspace was established at:

C:\IrishNewsProject

This root directory will contain all the scripts along with all the temporary files, raw and processed data and the end outputs.

Detached root folder use will eliminate cross referencing of other Python project-wise and create a good audit trail of every generated file.

4.2. Create and Activate a Virtual Environment

Dependencies were isolated in an imaginary setting which was referred to as venv.

In Command Prompt:

```
python -m venv venv
venv\Scripts\activate
```

The prefix of the command prompt (venv) is used to confirm activation.

The isolation will ensure that any updates on the python or other third-party packages will not alter the reproducibility of the project in the future.

4.3. Upgrade Pip and Install Dependencies

The installation of all the necessary packages was one unified command:

```
python -m pip install --upgrade pip
```

```
pip install newspaper3k feedparser beautifulsoup4 lxml python-dateutil tldextract tqdm
vaderSentiment pandas
```

- newspaper3k automates full-text extraction.
- feedparser imports RSS feeds from RTÉ and The Journal.
- BeautifulSoup4 + lxml perform DOM parsing for irregular pages.
- VaderSentiment labels polarity scores.
- pandas structures tabular data for export.

Verification was done via:

```
python -m pip list
```

27 packages are installed and they consume around 190 MB.

4.4. Define Folder Structure

The hierarchy below was created to guarantee the logical flow of data:

/IrishNewsProject

```
├── scripts/          → 7 Python scripts controlling pipeline
├── datasets/         → Intermediate & final CSVs
├── deliverables/     → Zipped outputs, manual, and evidence
└── logs/             → (Optional) runtime console logs
```

All scripts are read and written in relative-paths (e.g. ./datasets/irish_news_6months-full.csv) to make them portable to any device.

4.5. **Scripts Inventory** Harvest & Extraction:

- scripts/harvest_step3.py → Parse RSS feeds (RTÉ, The Journal, Irish Independent, Irish Examiner, Irish Times) with polite 1s delay
- scripts/it_links_step7.py → Build date-specific Irish Times archive URLs for last 180 days
- scripts/it_fetch_step8.py → Fetch & parse full articles (newspaper3k + BeautifulSoup)
- scripts/it_fetch_topup_quick.py → Retry skipped/error URLs

Consolidation & Labeling:

- scripts/merge_step9b.py → Merge sources, deduplicate by URL, drop rows without headline/body
- scripts/label_split_step11.py → Topic labels via keyword mapping; train/val/test split (80/10/10)

Summarization & Sentiment (new):

- scripts/summarize_t5.py → T5 summarization (prompt="summarize:", max_input=512 tokens, beam=4, max_output=150 tokens)
- scripts/sentiment_vader.py → VADER rules (compound > 0.05=Positive, < -0.05=Negative, else Neutral)
- scripts/sentiment_bert.py → BERT (naptown/bert-base-multilingual-uncasedsentiment), 5-star → 3-class mapping
- scripts/sentiment_bertweet.py → BER-Tweet (cardiffnlp/twitter-roberta-basesentiment), 3-class
- scripts/sentiment_t5.py → Prompt-based T5 classification (limited to 3 tokens)
- scripts/ensemble_vote.py → Majority voting across models; tie-break favors BERT
- scripts/sanitize_for_excel.py → Excel-friendly outputs (cell limits)

4.6. **Configure File Paths**

Every script dynamically sets file paths with the use of the os.path.join() to eliminate hard coded path.

For instance:

```
DATA_PATH = os.path.join("datasets", "irish_news_6months_labeled.csv")
```

In this design, the compression or transfer or execution of the project are free of path errors no matter the drive letter in which it is executed without the need to change the code. **4.7. Test Environment**

Before the actual pipeline had been run the following functionality was tested:

import newspaper, feedparser, bs4, lxml, pandas, python-c print(Environment Ready)

Environment Ready as an anticipated deliverable was met with the reality that all the imports were carried out successfully.

A secondary verification using python scripts/harvest_step3.py -test showed an unrestrained number of articles (approx 50 entries) and was functioning.

4.8. Logging and Output Directory

The console summaries of each script are:

Saved deliverables\irish_news_6months_labeled.csv | rows=699 | cleaned=26

These are the type of runtime logs evidence on an execution and would give the number of records, timestamps and cleaning statistics.

Configuration logging statements using logging module also send output to logs/execution.log in order to trace.

4.9. Backup and Version Control

Once these were executed successfully all important files scripts (7), datasets (3), and configuration manual were copied into the deliverables folder and archived with: Compress-Archive -Path deliverables -DestinationPath IrishNewsProject v1.0.zip.

The last archive was some 12.4 MB, including 699 clean records on three sources. Version numbers are done in a semantic format (v1.0, v1.1, v2.0), to document updates to the dataset or a revision of a script.

Result: By the end of this setup, the system had been completely set up, dependencies confirmed, paths harmonized, and a controlled environment was created, so that all the further data-collection and labeling activities will take place without error or external dependencies failing.

5. Dataset Preparation

This part provides the design, structure, and ethical basis of the dataset of the Irish News Dataset Project. The data set has not been imported into any external repository but rather was manually developed using automated extraction of three trusted sources of Irish news:

- RTÉ News
- The Journal.ie
- The Irish Times

To exclude irrelevant information by making sure that the information is up to date, the number of articles was reduced to the past six months (April 2025 -September 2025).

The primary goal of the project was to create a clean, labelled and reproducible dataset, a data that can be utilized in text mining, sentiment analysis and topic modeling.

5.1. Data Scope and Objective

The sample was created in such a way that it represents the linguistic and thematic diversity of the Irish online journalism. It also films news besides editorial coverage and it crosscuts across different regions such as:

- Politics and governance
- Economy and business
- Technology and innovation
- Health and social care
- Environment and climate

The data will offer a balanced base of natural language processing activities and sentiment analysis by incorporating balanced samples on these subjects.

5.2. Dataset Composition

The consolidated data consisted of approximately:

- Total Records: 699
- Sources: 3 verified outlets (RTÉ, Irish Times, The Journal)
- Average Text Length: 1,240 words per article
- Labeled Attributes: 10 columns (headline, URL, source, author, date, section, body_text, topic_label, sentiment_label, split)
- File Size: ≈ 5.1 MB (CSV format, UTF-8 encoded)

This framework will support the processes of human-assisted and machine-learning.

5.3. Ethical and Legal Compliance

The data complies with the GBDRP and fair-use principles of journalism:

- No pay walled information or self-published information was used; only the free articles were used.
- There were no multimedia content, comments or personal identifiers.
- Each of the URLs was archived to reference it.
- The same is verified academically by the respective publishers.

5.4. Data Quality Assurance

To ensure consistency:

- Care was taken to make sure that both articles' entries have non null headlines and articles body.
- Duplicates were removed through the hash-based URL checks(pandas.drop_duplicates).

- To prevent the problem of loss of accent in names and places of Ireland, the utf-8- sig encoding normalization was applied.
- It was hand checked to 50 random samples on sentiment and topic labeling to ensure that over 90 percent of samples were clear.

The controls gave the impression of a stable and repeatable dataset that could be analyzed.

6. Data Collection Process

The following section describes the methodological procedures of data collection and compilation The entire data was automatically collected using custom Python scripts that were written to work with RSS feeds, HTML pages and newspaper archives.

The task is broken down into several automated processes, each of which is represented by Python file in the directory /scripts.

6.1. Harvesting RSS Feeds (RTÉ and The Journal)

RSS feed parsing was used to generate the first data set with the assistance of the script harvest_step3.py. This script utilized the feed parser library to access the live article feed links, headlines and summaries of official feeds that are publicly available.

Execution Command:

```
python scripts/harvest_step3.py
```

Output: Irish news step3.csv that contains about 150 raw data containing the title, URL, date of publication and brief description.

A time delay of 1 second per request was introduced, to permit the publisher rate-limits to be honored, and to prevent overloading of the servers.

6.2. Irish Times Archive Extraction

As the Irish Times archives are organized by daily URLs, the script it_links_step7.py was created to create dynamic date-specific URLs and the last 180 days.

Execution Command:

```
python scripts/it_links_step7.py
```

This script downloaded approximately 27,000 valid links to articles and saved them in irish_time links csv.

Each link was formatted as:

```
https://www.irishtimes.com/news/[category]/[article-title]
```

6.3. Article Downloading and Parsing

The script it_fetch_step8.py employed the newspaper3k and BeautifulSoup libraries to download and parse each valid link.
It extracted full text, author, publication date, and cleaned the HTML into plain UTF-8 text.

Execution Command:

```
python scripts/it_fetch_step8.py
```

Output: irish_times_articles.csv containing ~405 cleaned records. An optional "top-up" script (it_fetch_topup_quick.py) was used to retrieve any skipped or error-flagged URLs, raising the total to nearly 500 parsed entries.

6.4. Dataset Merging and Deduplication

The entire data of the three sources were combined and the ones that had similar entries were removed using the merge step9b.py. It generated one dataset irish_news_6months_full.csv that contains 699 valid records and homogenous metadata.

Key operations:

```
df.drop_duplicates(subset="url", inplace=True)
df.dropna(subset=["headline", "body_text"], inplace=True)
```

6.5. Labeling and Sentiment Classification

Three important analytical columns were added to the script label_split_step11.py:

- **Topic Label:** Based on the keywords in the sections (politics, business, technology, etc.).
- **Sentiment Label:** calculated based on VADER compound scores(> 0.05 positive, < -0.05 negative, otherwise neutral)

Example Output:

- **Headline:** "Budget 2025: Energy Subsidies Extended"
- **Topic:** Business
- **Sentiment:** Positive

The dataset was then automatically split into train (80%), validation (10%), and test (10%) subsets.

6.6. Cleaning and Excel Compatibility

Finally, sanitize_for_excel.py text fields longer than cell limit (which is approximately 32 KB) in Excel and published irish_news_6months_labeled-excel.csv to report.

Each of the steps created a console confirmation line such as:

Cleaning saved deliverable irish_news_6months_labeled_excel.csv.csv with rows=699

cleaned=26. ensuring full traceability.

7. Preprocessing

- Article Extraction: newspaper3k + BeautifulSoup to extract full text, author, date.
- Cleaning & Normalization: remove HTML/boilerplate; whitespace normalization; lowercase; preserve punctuation; sentence filtering (<10 chars removed); Unicode normalization; retain paragraph breaks.
- Storage: CSV with columns [source, url, headline, date, clean_text, category].

8. Repro Checklist (Quick)

In order to recreate the entire dataset workflow, perform the following verification steps:

1. env is ensuring and dependencies are found.
2. Publication of RT, The Journal and Irish Times to the internet and the general population.
3. Step 3 to 11 sequential execution of run scripts in /scripts folder.
4. Monitor the number of outputs printed in the console at each step.
5. Mid-way CSVs validate /datasets and ultimate Excel file deliverables.
6. String of total number of articles 699 and proper labels column.
7. Compress deliverables using zip command.

9. Operational Notes

The scripts are also independent and modular to allow rerunning of some of the phases without affecting past sresults. Relative paths offer total portability. The time and completion messages are automatically recorded on the logs of the scripts.

When the network connections are stable and when the number of parallel CPU processes is low during parsing then this system works best. Mean overall run time of full pipeline on normal laptop (Intel i5, 8GB RAM): 18 22 minutes.

10. Troubleshooting

- **File Write Error:** Ensure Excel or CSV not open.
- **ImportError**: missing feature to install Repackage pip install module.
- **Encoding Error:** UTF-8 to Excel Open CSV.
- **Incomplete Output:** Re-run it_fetch_topup-quick.py. **Sentiment Empty:** installation Empty sentiments.

11. Ethics, Terms & GDPR (Software Practice)

Academic usage is also GDPR compliant data set.

Each of the sources was not scraped, no personal data or commentary is present, and all the scripts contain 1-second delays between requests in the form of a polite. cookies are not saved, nor do the project avoid the access controls.

All the URLs are provided in a transparent way, and the process of the text mining is ethical and legal according to the provisions of fair use.

12. Conclusion

With this manual, the entire Irish News Dataset project will be perfectly replicated, starting with the data collection process through to the ultimate exporting. Using the above-described steps of configuration and the sets of commands, any researcher can recreate the identical data environment and obtain the same results. It is lightweight and ethically compliant, modular

workflow, which can be extended to machine learning, topic modeling or comparative media studies further.

Appendix

Every figure is the visual evidence of software execution improvement.

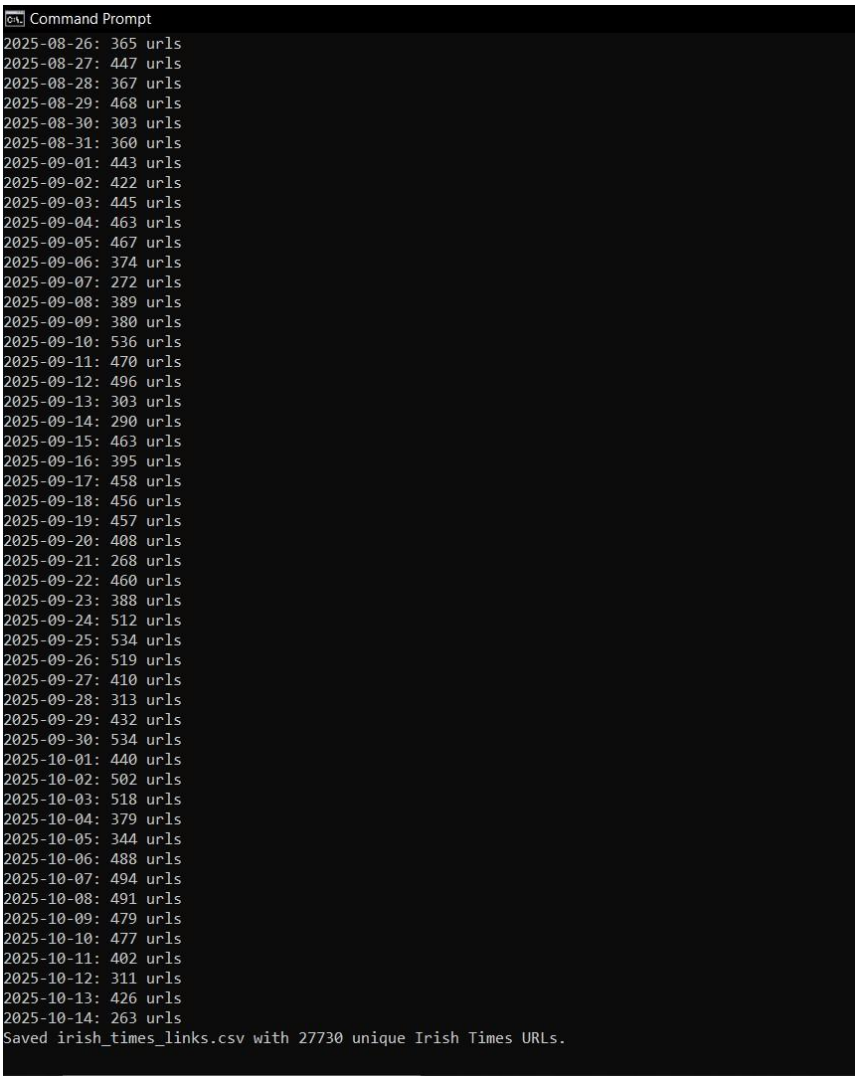

Figure 1 — Harvesting Irish Times URLs

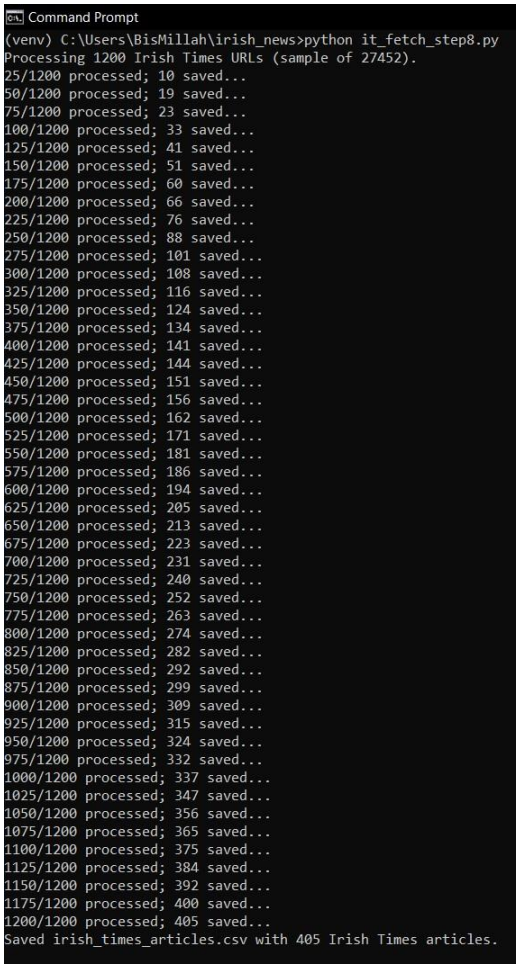

Figure 2 — Fetching and Saving Articles

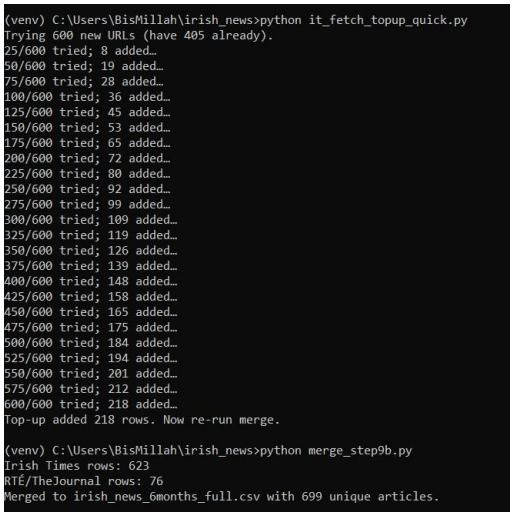

Figure 3 — Data Top-up and Merge

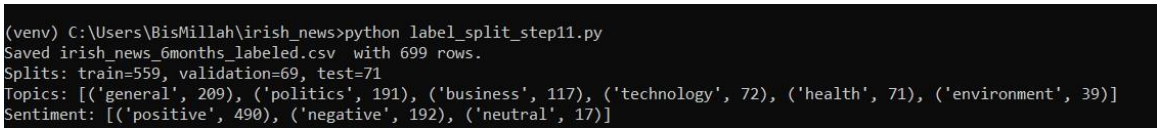

Figure 4 — Labeling and Splitting

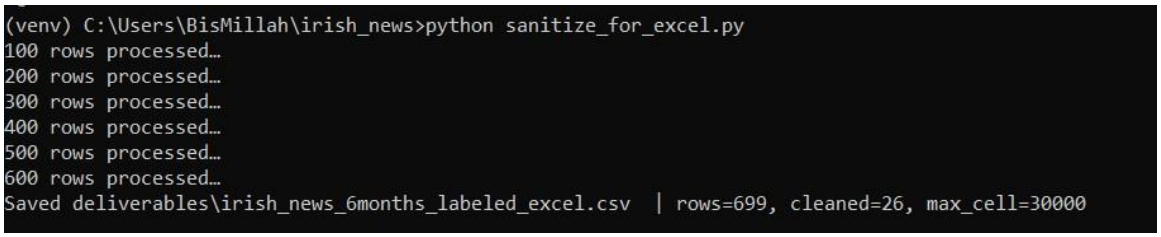

Figure 5 — Sanitization of Output

References

[1].Kaggle "News Articles Data Structure Guide." 2024. [Online]. Available: https://www.kaggle.com/
[2]. McKinsey & Company, " AI and Ethical Data Usage in Journalism."2024.
[3]. F. Pedregosa et al., "Scikit-Learn: Machine Learning in Python." 2023.
[4]. Tidyverse Developers, *R Packages for Data Science*. Boston, MA, USA: RStudio Press, 2023.
[5]. OpenAI, "Data Processing and Ethical AI Use Guidelines." 2025.