**A logo with text on it

AI-generated content may be incorrect.**

**Student name and ID: ALI JAWED DELAWARI GH1024093**

**Model Supervisor: William Morrison**

**Model code: B199 Bachelor’s Thesis**

**Word count: 10434**

**Submission date:19/12/2025**

**GITHUB link:** [**https://github.com/alijaweddelawari/B199-Dissertation**](https://github.com/alijaweddelawari/B199-Dissertation)

**Acknowledgement**

I would like to express my heartfelt appreciation to my dear Prof. Dr. William Morrison for his supervision, support, motivation and for believing in my capabilities throughout my research. A lot of thanks to GISMA university of applied sciences for their help.

Finally, a huge thanks for my family and friends for their endless love, support and

motivation especially in hard times.

An Analysis of Artificial Intelligence Representation in Afghan News Media Using Natural Language Processing

**Abstract**

Artificial Intelligence (AI) is increasingly shaping global societies, economies and media narratives. But research on how AI is perceived in conflict-affected, third world and low-resource contexts remain extremely limited. Afghanistan, despite its significant political, infrastructural and educational challenges, has slowly begun to engage with AI through media reporting, education initiatives and international cooperation. Understanding how AI is framed and viewed in Afghan media is therefore essential for assessing public awareness, optimism and concerns surrounding this emerging technology.

This thesis examines and analysis how AI is represented in Afghan news media between 2018 and 2025 using Natural Language Processing (NLP) techniques. An original dataset of AI-related news articles was manually collected from Afghan, regional and international media sources in English, Dari and Pashto. To make sure the analytical consistency and reliability, all non-English articles were manually translated and summarised into English. The dataset was then processed using a transparent NLP pipeline implemented in Python within a Jupyter Notebook environment.

The analysis combines lexicon-based sentiment analysis using VADER and topic modelling using Latent Dirichlet Allocation (LDA). Also, an external validation step was conducted using a labelled Reuters news headline dataset to contextualise the performance and limitations of the sentiment model. A lightweight and small result inspection interface was employed to support qualitative verification and interpretability of model outputs.

The findings show that Afghan media portray AI through a balanced but cautiously optimistic lens. Most articles adopt a positive tone, particularly when discussing topics like education, youth skills, innovation and international cooperation. Negative sentiment shows up primarily in coverage related to surveillance technologies, misinformation and geopolitical concerns. Topic modelling reveals three dominant themes: AI policy and international cooperation, AI tools and media coverage and AI adoption in daily life and education. When sentiment and topic analysis are combined, education-focused narratives emerge as the most positive, while media- and security-related narratives show more mixed emotional framing.

Overall, this study demonstrates that Afghan media do not present AI as a purely beneficial and positive or harmful and negative phenomenon, but rather as a technology associated with both opportunity and risk. By constructing an original dataset and applying NLP methods in a low-resource context, this thesis contributes to research on AI perception in underrepresented regions and highlights both the potential and limitations of computational media analysis in fragile environments.

**Table of Contents**

**Titles Page**

Abstract 2

Table of Contents 1

List of Figures 3

CHAPTER 1: INTRODUCTION 4

1.1 Problem and Motivation 4

1.2 Contributions 5

1.3 Research Questions 5

1.4 Significance of the Study 6

Chapter 2: Related Work (Literature Review) 7

2.1 AI Perception Studies 7

2.2 AI and Technological Development in Afghanistan 7

2.3 Socio-Cultural Barriers to AI Adoption 7

2.4 Data Science and AI Research in Low-Resource Environments 10

2.5 Media as a Mirror of AI Perception 11

2.6 Theoretical and Methodological positioning 12

2.7 Research Gap 12

CHAPTER 3: APPROACH (METHODOLOGY) 13

3.1 Overview of the Approach 13

3.2 Data Collection 15

3.2.1 Dataset Construction Process 16

3.2.2 Translation of Local Articles 17

3.2.3 Dataset Size and Composition 17

3.2.4 Why This Dataset Is Original 18

3.2.5 Rationale for focusing on Afghanistan 18

3.2.6 Limitations 18

3.3 Data Preprocessing 19

3.3.1 Loading the Dataset into the Notebook 19

3.3.2 Identifying Structural Issues 19

3.3.3 Creating the “clean\_summary” Column 20

3.3.4 Handling Missing Summaries 20

3.3.5 Handling Duplicates 20

3.3.6 Final Preprocessing Output 21

3.4 Justification of Tools and Libraries 21

3.4.1 Pandas (Data Loading and Manipulation) 22

3.4.2 Regular Expressions (Text Normalisation) 22

3.4.3 NLTK (Stop words and Tokenisation) 22

3.4.4 VADER (Sentiment Analysis) 22

3.4.5 Gensim and LDA (Topic Modelling) 23

3.5 Result Inspection Interface 23

3.5.1 Purpose and Role within the Analysis 23

3.5.2 Functional Scope 23

3.6 External Validation Method for Sentiment Analysis 24

3.7 Chapter Summary 25

CHAPTER 4: EVALUATION AND RESULTS 26

4.1 Overview 26

4.2 External Validation of the Sentiment Analysis Model 26

4.3 Sentiment Analysis Results 28

4.3.1 Sentiment Label Distribution 28

4.3.2 Examples of Positive, Negative, and Neutral Articles 29

4.3.3 Interpretation of Sentiment Trends 30

4.4 Topic Modelling Results 31

4.4.1 Topic 1: AI Policy, National Strategy & International Cooperation 33

4.4.2 Topic 2: AI Tools, Media Coverage & Public Awareness 33

4.4.3 Topic 3: AI Adoption in Afghan Daily Life & Education 34

4.4.4 Topic Coherence and Model Quality 35

4.4.5 Interpretation 35

4.5 Joint Sentiment–Topic Analysis 36

4.5.1 Sentiment Breakdown by Topic 37

4.5.2 Topic 1 – Predominantly Positive 39

4.5.3 Topic 2 – Positive but Critical 40

4.5.4 Topic 3 – Mixed with Strong Concerns 40

4.5.5 Interpretation and Implications 41

4.6 Result Inspection and Qualitative Verification 42

4.6.1 Use of the Interface for Qualitative Assessment 42

4.6.2 Illustrative Example 42

4.6.3 Analytical Value and Consistency 43

4.7 Visualisation and Interpretation of Results 43

4.7.1 Sentiment Distribution Across the Dataset 44

4.7.2 Example Sentiment Outputs: First Ten Records 45

4.7.3 Topic Model: Top Keywords for Each Topic 46

4.7.4 Word Clouds for Topics 47

4.7.5 Topic Distribution Across the Dataset 49

4.7.6 Integrated Topic–Sentiment Interpretation 50

4.7.7 Limitations of the Visualisation Approach 50

4.7.8 Summary of Visual Findings 51

4.8 Discussion of Findings 51

CHAPTER 5: DISCUSSION, CONCLUSION AND FUTURE WORK 52

5.1 Discussion 52

5.2 Conclusion 54

5.3 Future Work 55

References 56

Appendix A 59

Appendix B 62

Appendix C 64

**List of Figures Page**

FIGURE 4.1 27

FIGURE 4.2 28

FIGURE 4.3 29

FIGURE 4.4 32

FIGURE 4.5 36

FIGURE 4.6 37

FIGURE 4.7 37

FIGURE 4.8 37

FIGURE 4.9 38

FIGURE 4.10 43

FIGURE 4.11 44

FIGURE 4.12 45

FIGURE 4.13 46

FIGURE 4.14 47

FIGURE 4.15 48

FIGURE 4.16 48

FIGURE 4.17 49

**CHAPTER 1: INTRODUCTION**

The fast development and changes in Artificial Intelligence have transformed global societies, economies and communication systems. While most developed nations lead in AI development and regulation, emerging economies are just gradually exploring its applications and implications. Afghanistan despite facing infrastructural and socio-economic challenges is beginning to engage with AI through education, media, and limited policy dialogues. Understanding how AI is perceived and looked at in Afghanistan espicailly through media narratives is crucial for identifying public awareness, optimism and concerns toward this transformative technology. This introduction chapter introduces the background, the motivation and scope of this research explaining why AI perception analysis in Afghanistan is academically significant and socially relevant.

***1.1 Problem and Motivation***

AI has become one of the main forces of the 21st century effecting nearly every aspect of our modern life from things like education and healthcare to national security and communication (World Economic Forum, 2020). But global discussions and research on AI often exclude developing regions such as Afghanistan. This face unique barriers including limited technological infrastructure, low digital literacy and political instability (UNDP, 2024). This lack of representation creates a research gap for us. The perception of AI in Afghanistan remains still undocumented, even as its influence grows through international cooperation, online education and government modernization efforts.

In Afghanistan AI adoption is visible in emerging sectors such as education and youth-driven innovation projects. Yet public understanding and trust in AI is still far from what it is in developed countries. AI mostly symbolizes opportunity and a tool for progress, innovation and global connection. But it also evokes fear and mistrust often due to concerns like about surveillance, unemployment or misuse by political sectors. The local media plays a crucial role in forming this perception and effecting people’s views on AI by framing it and its development as either beneficial, neutral or risky.

Monitoring and analysing media sentiment can provide an indirect but powerful understanding of how Afghan society perceives AI. However academic exploration of this topic is minimal for now. Most studies focus on AI perception in Western or developed countries leaving emerging economies understudied. This thesis seeks to address that gap by systematically analysing Afghan news articles using Natural Language Processing (NLP) techniques to identify sentiment patterns and themes surrounding AI.

The recent global surge in AI-driven media does not focus on the need to understand how developing societies like Afghanistan view these technologies. This makes Afghanistan’s early engagement with AI a very important step toward long term digital transformation.

***1.2 Contributions***

This research contributes to both the academic and practical understanding of AI in developing nations. Its core contributions are:

**1**.Empirical Dataset Creation: A new and unique dataset was compiled containing AI related news articles from Afghan and regional media outlets focusing on Afghanistan. These articles, collected from 2018 to 2025 form one of the few structured corpora representing Afghan media discourse on AI.

**2.**Sentiment and Thematic Analysis: Using Python based NLP tools within Jupyter Notebook this study applies sentiment analysis to classify each article as positive, negative, or neutral. Topic modelling like LDA is used to find key themes such as AI in education, AI in security and AI ethics. These tools and methods provide insight that can be measured into how AI is discussed across Afghan media.

***1.3 Research Questions***

To guide our investigation the following research questions are proposed:

**RQ1:** How is Artificial Intelligence portrayed in Afghan media between 2018 and 2025?

**RQ2:** What are the sentiments (positive, negative, or neutral) expressed in AI related Afghan news articles?

**RQ3:** Which key themes and topics emerge from Afghan media coverage of AI and what do they reveal about people’s attitudes toward technological progress in the country?

These questions are aiming to capture both on how Afghanistan feels and the technical content of Afghan AI discourse forming a comprehensive picture of national perception.

***1.4 Significance of the Study***

The significance of this study is to show its ability to combine AI technology with social insight in a country where digital transformation is still emerging. Unlike most perception studies conducted in Western countries, this research focuses on an underrepresented country where AI is often introduced through NGOs and limited educational programs. By examining how AI is discussed in Afghan media, the study helps to understanding the social readiness and trust surrounding AI in fragile contexts.

From an academic perspective, the research introduces a novel, regional dataset and demonstrates how NLP techniques can be used in data scarce settings. From a practical standpoint the findings of it can inform policymakers, educators and businesses about public attitudes toward technology adoption.

So, this research aims to inspire more studies that use data science ways to explore human and societal dimensions of AI adoption in low resource settings.

**Chapter 2: Related Work (Literature Review)**

Before going into prior research, it is very important to recognize that Afghanistan’s case stands apart from most AI perception studies. Unlike technologically advanced nations, Afghanistan is a post conflict country where media, education and governance are still stabilizing. This situation makes studies like this even more important.

***2.1 AI Perception Studies***

AI has become one of the most important technologies of the 21st century, shaping societies everywhere. As AI continues to expand in education, healthcare, business and others, researchers have turned their attention toward understanding public perception of AI. How societies interpret, trust or fear these technologies. Globally, AI perception studies have revealed a complex balance between optimism about innovation and anxiety about automation, job displacement and ethical risks (World Economic Forum, 2020). While Western nations have already seen academic and industry-led investigations into AI perception, research from developing regions remains extremely limited.

In most of the global studies public attitudes toward AI are influenced by education, media exposure etc. For example, wealthy countries often display nuanced attitudes that weigh both opportunities and ethical challenges, whereas developing contexts may focus more on issues of access and employment (Oxford Insights, 2022). But when it comes to Afghanistan, there is a big research gap. The existing academic and policy literature largely overlooks how Afghans view or interact with AI technologies. This absence of research highlights a critical gap that our thesis aims to address through data science using tools like NLP to analyse how AI is discussed across Afghan media outlets.

***2.2 AI and Technological Development in Afghanistan***

Afghanistan’s engagement with digital transformation has been slow primarily due to decades of political problems, weak infrastructure and limited investment in science and technology. The period between 2002 and 2021 marked a modest expansion in connectivity and other sectors supported largely by international aid and partnerships. Universities in Kabul, Herat and Nangarhar began introducing computer science programs. Organizations like the UNDP and World Bank supported digital literacy initiatives (World Bank, 2020). However, the Taliban’s return to power in 2021 disrupted this trajectory. Many international donors and tech partners withdrew from Afghanistan that stalled educational projects and AI research efforts.

Internet access also remains limited across the country, with only around 25% of the population having reliable online connectivity (ITU, 2023). The digital divide is also along gender and rural urban lines. While urban youth show growing interest in AI tools for education and freelancing, women’s participation in STEM and digital education has declined since 2021 due to restrictions on female schooling and employment (UNESCO, 2022). This has further widened the skill and knowledge gap needed for a robust AI ecosystem.

Afghanistan lacks a formal AI policy or digital strategy. Neighbouring countries like Iran, India and Pakistan have already launched AI roadmaps and plans but Afghanistan remains without a good framework for AI development. This institutional void makes it very challenging to assess how AI could be integrated into the government, education or business sectors (OECD, 2021). Local innovation persists through small startups, university research clubs and international collaborations but these efforts are fragmented and underreported from there.

In this fragile environment, media plays crucial role in shaping the public’s understanding of AI. News coverage from outlets like TOLOnews, Pajhwok Afghan News and more often presents AI as a futuristic or external concept for Afghans who think it’s something extraordinary developed elsewhere rather than domestically. Occasional reports highlight Afghan youth using AI for freelancing, education, or other works suggesting growing curiosity and learning but there is little systematic understanding of how these narratives evolve over time.

***2.3 Socio-Cultural Barriers to AI Adoption***

Beyond all the infrastructure and policy challenges Afghanistan faces big social and cultural barriers that influence how AI and digital technologies are perceived there. The history of conflict, low literacy rates and limited access to education have fostered scepticism toward modern technologies. The Afghanistan Human Development Report (UNDP, 2021) notes that technological literacy among adults is one of lowest in South Asia. This low digital competence makes it difficult for the public to engage meaningfully with AI or to understand its implications for daily life.

Language also is a major barrier. Most AI tools and training materials are designed in English or other major global languages leaving speakers of Pashto and Dari at a disadvantage. The cultural conservatism in many parts of the country can fuel suspicion toward automation and robotics often perceived as replacements for human labour or as ethically questionable tools.

The exclusion of women from education since 2021 has created further problems. Before that several women robotics teams such as the Afghan Dreamers symbolized a new technological empowerment there. But now many such groups have been forced to relocate or shut down effecting Afghanistan’s participation in global AI innovation (BBC, 2022). The fear of surveillance and data misuse has also contributed to public anxiety, like reports of biometric data collected during international missions potentially being exploited (Reuters, 2021).

These socio-cultural constraints not only slow AI adoption there but also influence how media outlets frame AI related news Understanding these underlying perceptions is therefore important for assessing the country’s readiness for AI driven transformation.

***2.4 Data Science and AI Research in Low-Resource Environments***

Doing AI and data science research in low-resource places such as Afghanistan presents a set of technical and contextual challenges. Globally NLP and sentiment analysis have advanced rapidly due to the availability of large datasets, technical infrastructures and multilingual models (Devlin et al., 2019). These advances are largely focused in high-income and English-dominant contexts. For countries like Afghanistan for example the scarcity of labelled text data in local languages limited research infrastructure and lack of computational capacity restrict the ability to replicate such studies.

Existing regional work has shown that even in some developed South Asian places such as Pakistan and Bangladesh, sentiment analysis faces issues of data imbalance, linguistic diversity and translation inaccuracies (Rahman et al., 2022). Afghan researchers have even fewer tools to work with. Pashto and Dari, Afghanistan’s primary languages are considered low-resource languages in the field of NLP, meaning they have very limited training corpora for AI models (Joshi et al., 2020). This limitation directly affects machine learning models trained on Western datasets, which often fail to capture regional linguistic nuance or sentiment tone.

Despite these challenges there is growing academic interest in exploring localized AI and data driven approaches in post conflict contexts. For example, UNDP and UNOPS have initiated small projects using AI for development and disaster management like using satellite imagery to assess flood and earthquake damage in Herat and Baghlan provinces (UNDP, 2024). However there has been no known large academic study focused on AI perception or sentiment-based media analysis within Afghanistan itself.

This lack of digital infrastructure and available data makes Afghanistan a unique case study in low-resource AI research. It highlights the importance of approaches like the one proposed in this thesis, building a custom dataset of Afghan news articles and using NLP to extract public sentiment. From a data science standpoint such research contributes to not only understanding public attitudes toward AI but also expanding the global body of NLP work on underrepresented languages and regions.

***2.5 Media as a Mirror of AI Perception***

Media has always played an important role in shaping technological narratives. When it comes to AI, news articles often determines whether the public perceives AI as an opportunity or a threat. Media in developing countries tends to present AI as a foreign phenomenon rather than a local reality.

Afghan media fits within this pattern. Outlets such as TOLOnews, Pajhwok Afghan News and Khaama Press have reported on AI mostly through international or regional developments such as AI cooperation between Afghanistan and China. This externalized framing can influence how Afghans perceive and view AI not as a homegrown opportunity but as a foreign technology controlled by powerful states.

Now geopolitical trust becomes a key issue also because of historical and political tensions many Afghans express toward technologies originating from countries like Pakistan, Iran or even China, nations that dominate regional AI development (Ahmad, 2023). This distrust extends to digital systems and tools that are seen as foreign made, reflecting broader anxieties about trust, surveillance, and dependency. For example, studies on digital governance in post conflict societies note that the provenance of technology can shape public trust more than its technical quality (Kellerman, 2021). In Afghanistan’s case, perceptions of external control can dampen enthusiasm for AI even among younger more digitally active demographics.

Media then becomes both a reflection and amplifier of these perceptions here. Local coverage often points to foreign ownership, outside funding or regional competition, inadvertently reinforcing the fact that AI is not ours. At the same time, Afghan journalists themselves face institutional constraints, political censorship, limited technical experts and restricted internet access that limit how critically and deeply they can report on AI related topics. This dynamic makes Afghan media an especially rich but complex data source for sentiment analysis.

***2.6 Theoretical and Methodological Positioning***

While much of the existing literature on AI perception is grounded in formal theoretical frameworks like technology acceptance models, innovation diffusion theory, or media framing theory. This study adopts a deliberately applied and data-driven perspective. This methodological choice reflects on both the research context and the nature of the available data. Afghanistan represents a low-resource, post-conflict environment where formal public opinion datasets, surveys and theory-driven perception studies are largely absent (Kellerman, 2021). As a result, conventional theory-testing approaches are difficult to operationalise in a rigorous and meaningful way.

Rather than positioning itself as a theory-confirming or theory-building study, this research is exploratory in its nature and focuses on empirical observation of media narratives. Media discourse is treated as a proxy for broader societal perceptions, allowing sentiment and thematic patterns to be examined through computer analysis. This approach aligns with applied NLP research, where inductive methods are often used to surface patterns from unstructured text before formal theorisation is possible (Eisenstein, 2019).

Concepts such as trust, risk, optimism and technological opportunity are interpreted through observed sentiment and topic distributions rather than being mapped directly onto a pre-defined theoretical model. This decision also avoids overfitting Western-centric frameworks to a context with distinct political, cultural and infrastructural characteristics (Oxford Insights, 2022).

***2.7 Research Gap***

The review of existing literature clearly reveals several interlinked gaps that this thesis wants to address.

There is a lack of research examining how Afghans perceive AI through local media narratives. While sentiment analysis has been widely applied to global news, social media and policy documents in Western and some Asian contexts, there has been no equivalent study in Afghanistan or even in much of Central Asia.

Most global AI perception studies rely on big scale open datasets such as Twitter feeds or online reviews that are not available or accessible in Afghanistan due to low digital participation and data restrictions (ITU, 2023). This absence necessitates a manual collection approach, which this study implements by building a dataset of over 100 Afghan and regional news articles.

Prior literature has not addressed the intersection of media trust, geopolitics, and AI perception. Afghanistan’s case is distinct in that public attitudes toward technology are not only shaped by literacy and access but also by historical suspicion of neighbouring powers and foreign interventions. As noted before scepticism toward AI technologies developed by countries like Iran or Pakistan reflects big political perceptions rather than purely technical reasoning (Ahmad, 2023).

From a data science standpoint also there has been minimal exploration of NLP techniques in Afghan contexts or other conflict affected low-resource regions. The lack of localized resources and technical groups which do the sentiment analysis as also been a cause. By addressing these voids, this thesis contributes to both the social and technical dimensions of AI research.

**CHAPTER 3: APPROACH (METHODOLOGY)**

***3.1 Overview of the Approach***

This research bellow follows a structured, end-to-end Natural Language Processing pipeline designed to analyse how Artificial Intelligence is represented in Afghan news media between 2018 and 2025 which is the main lifetime of AI. The methodology combines manual dataset construction, systematic preprocessing, sentiment classification and topic exploration. The goal in here is to transform raw news articles that we collected into interpretable sentiment and thematic insights while ensuring transparency, reproducibility, and methodological clarity (Jurafsky & Martin, 2023).

The pipeline was implemented in Python coding using Jupyter Notebook which allows full visibility of code execution, preprocessing steps, intermediate outputs and results.

The approach begins with manual data collection. Since no public dataset was found by the researcher that met the required criteria on Afghan AI-related news. All articles were gathered manually by the researcher. This produces an original dataset of 112 articles covering English, Dari and Pashto sources. This manual collection approach shows the limited digital infrastructure of Afghanistan and ensures cultural and contextual accuracy (Newman et al., 2016).

Because headlines alone do not contain enough detail for reliable sentiment classification, the dataset also includes summaries written or translated by the researcher. Several articles originally published in Dari or Pashto were manually translated into English to maintain language consistency for the NLP pipeline. This translation process is explicitly acknowledged to avoid ethical issues around data manipulation and to align with good practice in cross-linguistic research (Temple & Young, 2004).

After data collection is done, the next stage involved data cleaning and preprocessing, which prepared the summaries for later analysis. These steps included lowercasing, punctuation removal, whitespace normalization, the detection of empty or malformed entries etc (Manning et al., 2008). Preprocessing was done minimal at this stage on purpose as deeper NLP transformations like tokenisation for topic modelling are performed later in Chapter 4.

The cleaned dataset is then prepared for two main analytical components:

1.lexicon-based sentiment analysis using VADER

2.Latent Dirichlet Allocation (LDA) topic exploration (Hutto & Gilbert, 2014) (Blei et al., 2003)

Sentiment classification for which VADER is used is described later in Chapter 4 as it forms part of the modelling and evaluation stage rather than preprocessing. Topic modelling specific tokenisation is also done in Chapter 4 to preserve clarity and avoid mixing preprocessing with modelling logic.

Together these steps form a pipeline that transforms unstructured Afghan news text into meaningful sentiments and thematic insights. The separation between preprocessing Chapter 3 and modelling Chapter 4 is intentional addressing a big key weakness found in other dissertations where methods, tools and results were mixed in a way that obscured and missed with the research logic. This structure ensures that each component of the methodology is transparent and aligned with the research objectives (Bird et al., 2009).

***3.2 Data Collection***

Because no publicly available dataset was found for analysing AI-related news coverage in Afghanistan this dessertation required the construction of a fully original dataset by the researcher. All articles were collected, translated and summarised manually by the researcher. This makes this dataset one of the very first structured dataset of Afghan AI news addressing a documented gap in digital research resources for the region.

The dataset includes articles from local, regional, and international news outlets, including:

- Pajhwok Afghan News

- Tolo News

- 8am Daily

- AFP Fact Check

- BBC Persian

- The Diplomat

- Al Jazeera

- DW

- Context News

- fact-checking organisations

- niche research outlets like GNET, policy institutes, etc.

The sources were selected based on:

**1**.Relevance= The article must involve artificial intelligence, digital technology, misinformation, surveillance, innovation, education or AI related policy in Afghanistan.

**2.**Credibility= Only registered media outlets, academic publications or recognised bodies were included.

**3.**Accessibility= Some Afghan news sites most of the times go offline due to censorship, restrictions or instability. Then possible archives or cached versions were used.

Part of materials came from Dari and Pashto outlets. The researcher himself translated the relevant content into English while keeping meaning.

This multilingual process also shows the reality that AI related reporting often is deferent by language English media tends to frame AI in global terms, where local sources emphasise practical impacts such as education, security, misinformation and public services (Ahmad, 2023).

**3.2.1 Dataset Construction Process**

The dataset was assembled in several stages:

**1.** Article Identification

Articles were located using:

1.keyword searches like AI Afghanistan, artificial intelligence, machine learning, deepfake, misinformation, technology, digital education and more.

2.manual browsing of Afghan outlets

3.fact check archives

4.research blogs and policy institutions

Every entry was verified and added manually.

**2.** Recording Structured Fields

Each row in the dataset includes:

* ID
* Source
* Publication Date
* Title
* URL
* Summary (manually written by the researcher)

This summary field is the most important component for NLP analysis. Headlines alone were not enough and didn’t have enough data for sentiment detection because they are usually short (Fast & Horvitz, 2017).

**3.2.2 Translation of Local Articles**

For the articles that were in Dari or Pashto:

* the researcher translated the content manually
* checked terminology consistency like AI”, automation, robotics, digital systems etc.
* preserved factual meaning

**3.2.3 Dataset Size and Composition**

After initial collection:

* 112 total articles were compiled
* 7 entries initially contained empty or malformed summaries due to incomplete archived pages, inaccessible paywalls, or metadata errors

These were reviewed during preprocessing in Section 3.3. The final dataset used for analysis is 112 rows.

**3.2.4 Why This Dataset Is Original**

This dataset is original because:

* articles were collected manually
* summaries were written or translated by the researcher
* multilingual Afghan sources were included
* it introduces an 8-year span (2018–2025)
* pre-Taliban era
* collapse period
* post 2021 restrictions
* global AI boom like ChatGPT era

**3.2.5 Rationale for focusing on Afghanistan**

Afghanistan is a very underrepresented region in AI perception studies because of:

* limited digital infrastructure
* censorship and content restrictions
* collapse of educational institutions after 2021
* reduced international media presence
* withdrawal of tech investments and NGOs
* low public trust in technology coming from outside countries due to political tensions
* high exposure to misinformation, particularly via social media
* multilanguage media landscape

These factors above make Afghanistan an ideal case study for understanding how places with limited digital knowledge and unstable political conditions perceive AI technologies (Kellerman, 2021).

**3.2.6 Limitations:**

- Accessing some local news archives was difficult because of domain blocks.

- Manual translation may introduce minor sometimes causes issue though neutrality was prioritised

- Fact checking all articles form big portion of the dataset due to rising deepfake misinformation in the region (Blodgett et al., 2020).

- The dataset does not include videos or social media posts because it’s focusing strictly on news text.

Despite all these limitations above, the dataset remains a strong and reliable foundation for sentiment and topic analysis.

***3.3 Data Preprocessing***

Once the dataset was fully compiled the next step is to prepare the text for analysis. Since the summaries were manually written and translated by the researcher himself, the dataset had less structural errors compared to scraped data. But several cleaning and standardisation steps were still necessary to make the text suitable for NLP, sentiment analysis and topic modelling (Bird et al., 2009).

**3.3.1 Loading the Dataset into the Notebook**

The dataset was imported into Python using tools like pandas. This step also served as a verification step to confirm:

* total row count (112)
* correct import of all columns
* detection of null values or missing entries

**3.3.2 Identifying Structural Issues**

A review of the imported dataset showed:

* 7 rows with empty or partially missing summaries
* several rows containing unwanted whitespace
* inconsistent text casing
* punctuation differences across translated and English summaries
* duplicated publisher text in a few entries

Cleaning focused on repairing the text rather than deleting data (Eisenstein, 2019).

**3.3.3 Creating the clean\_summary Column**

To prepare the textual content for analysis, a new field called clean\_summary was created while keeping the original column too.

The cleaning process included:

- converting all text to lowercase

- removing numbers, extra spaces, and punctuation

- removing english stop words

- tokenising and rejoining text

These steps follow the standard NLP preprocessing used in sentiment analysis and topic modelling (Manning et al., 2008; Jurafsky & Martin, 2023).

They were implemented by using the NLTK library (Bird et al., 2009).

**3.3.4 Handling Missing Summaries**

For the few entries where the original summary field was empty:

* records were flagged
* the clean\_summary column represented them as an empty string

Rather than deleting all these rows, they were left in the dataset to keep the temporal and thematic coverage of the corpus (Eisenstein, 2019).

This preserves dataset completeness while also maintaining academic transparency about limitations.

**3.3.5 Handling Duplicates**

A duplicate check was performed on:

* URLs
* Titles
* Summaries

This approach ensured:

* no two rows represented the same article
* no metadata disruption would bias sentiment distributions
* each article remained unique for analysis (McKinney, 2010)

**3.3.6 Final Preprocessing Output**

After all preprocessing steps:

* the dataset contained 112 cleaned summaries
* all articles were preserved
* the clean\_summary column became the main input for the modelling stage in Chapter 4
* the original Summary column remained untouched for transparency

The final structure of the dataset included:

* ID
* Source
* Date
* Title
* URL
* Summary (original)
* clean\_summary (processed)

The cleaned summaries support NLP algorithms, while the original summaries maintain the integrity and reproducibility of the dataset (Jurafsky & Martin, 2023).

**3.*4 Justification of Tools and Libraries***

The methodological choices in this project were shaped by two key constraints:

**1.** the low-resource nature of Afghanistan-related AI news

**2.** the need to build a reliable sentiment analysis and topic modelling pipeline (Joshi et al., 2020).

**3.4.1 Pandas (Data Loading and Manipulation)**

It was used many times including import the Excel dataset, inspect missing and duplicated values, create new columns like clean\_summary and prepare structured data for modelling (McKinney, 2010).

**3.4.2 Regular Expressions (Text Normalisation)**

The Python re library was used to remove punctuation, numbers, formatting artefacts and others from the text. Given the small and lightweight nature of the dataset, a simple regex-based approach provided transparent behaviour with minimal dependencies supporting reproducibility and interpretability (Jurafsky & Martin, 2023).

**3.4.3 NLTK (Stop words and Tokenisation)**

For this thesis the Natural Language Toolkit (NLTK) was selected for its built-in English stop word lists, straightforward tokenisation utilities and integration with the VADER sentiment lexicon used later in the analysis (Hutto & Gilbert, 2014).

**3.4.4 VADER (Sentiment Analysis)**

VADER is a lexicon and rule-based sentiment analysis tool designed for short texts such as headlines and summaries (Hutto & Gilbert, 2014). It produces both a compound sentiment score and categorical labels (Positive, Neutral, Negative) which align directly with the research questions of this thesis.

**3.4.5 Gensim and LDA (Topic Modelling)**

Topic modelling was performed in here using Gensim’s implementation of LDA a probabilistic model for identifying latent themes within document collections (Blei et al., 2003).

LDA was selected by the researcher because it does not require labelled data. It performs well on small corpora and produces interpretable topic word distributions. Gensim’s clear documentation and its lightweight design furthermore support reproducibility and examiner verification.

***3.5 Result Inspection Interface***

**3.5.1 Purpose and Role within the Analysis**

In addition to the core analytical pipeline, this study includes a lightweight and small result inspection interface. It is designed to support structured exploration of the annotated dataset. It serves as an auxiliary analytical tool that improves transparency and interpretability of model’s outputs. It is a key consideration in applied natural language processing research (Eisenstein, 2019).

**3.5.2 Functional Scope**

The result inspection interface provides a small set of controlled interaction options in it that allow the researcher to explore the dataset in a structured manner. Specifically, it enables:

* inspection of individual articles by unique identifier
* filtering of articles by topic
* filtering of articles by sentiment
* combined topic–sentiment filtering
* display of summary distributions across topics and sentiment classes

The interface does not alter analytical results at all but offers an additional layer of interpretability and traceability (McKinney, 2010).

***3.6 External Validation Method for Sentiment Analysis***

Although the VADER sentiment model is primarily applied to the manually constructed Afghan news dataset, it is very important to evaluate how well the pipeline performs on text with established ground truth sentiment labels. This external validation step increases the methodological reliability of the study.

The Reuters sentiment headline dataset contains news headlines labelled on a five-point sentiment scale (1 = very negative, 5 = very positive). To make these labels compatible with the three-class scheme used in this thesis and on our dataset, the following widely accepted mapping was applied:

* 1–2 → Negative
* 3 → Neutral
* 4–5 → Positive

This mapping then allowed the external dataset to follow the same structure used for Afghan dataset, supporting a direct comparison of predicted and reference sentiments.

The exact same cleaning of text described earlier in Chapter 3 was also applied to the Reuters sample. VADER was then used for this sample to generate sentiment predictions for each cleaned headline. These predictions were compared to the mapped ground truth labels using standard evaluation metrics including accuracy, precision, recall, and F1-score.

The results of this evaluation are shown in Chapter 4.

***3.7 Chapter Summary***

This chapter outlined the methodological framework used to analyse how artificial intelligence is discussed, framed and understood within Afghan media between 2018 and 2025. It detailed the full analytical pipeline from dataset construction and preprocessing to sentiment analysis and topic modelling. It layed the foundation for the evaluation and results presented in Chapter 4 (Eisenstein, 2019).

A main contribution of this study is the construction of an original dataset, manually complied and sorted from publicly available Afghan and international news sources excluding the available data that is behind paid sources like GEDLT. This manual approach was chosen on purpose to preserve contextual accuracy, avoid automated scraping biases, and capture Afghanistan specific AI narratives often absent from global datasets.

Preprocessing steps were then applied to convert the dataset into an analysis ready format (Bird et al., 2009). Such steps were needed to ensure reliable operation of downstream methods such as VADER sentiment analysis and LDA topic modelling.

The chapter also uses English summaries as the primary analytical text given the limitations of current NLP tools in low resource languages like Dari and Pashto and the political sensitivity of Afghan media discourse, manually written and translated English summaries were considered the most reliable input for computational analysis (Joshi et al., 2020).

The chapter introduced a lightweight result inspection interface integrated within the notebook environment.

Ethical and cultural considerations were discussed too including translation risks, political neutrality, algorithmic bias and researcher positionality. Chapter 4 applies this framework to empirical analysis and interpretation of the results.

**CHAPTER 4: EVALUATION AND RESULTS**

***4.1 Overview***

This chapter presents the empirical results of the NLP pipeline described in Chapter 3 of this thesis. The analysis focuses on three main components:

* Sentiment analysis of Afghan AI news summaries using VADER
* Topic modelling using Latent Dirichlet Allocation (LDA) to uncover dominant themes
* Combined interpretation of sentiment and topics, supported by an interactive prototype

All results are based on the 112 manually collected and pre-processed summaries described earlier with an external dataset used as validation of our models. Unless otherwise stated, the variable df refers to the final cleaned dataset containing both the original Summary and the processed clean summary columns as well as the derived sentiment and topic labels (Hutto & Gilbert, 2014; Blei et al., 2003).

***4.2 External Validation of the Sentiment Analysis Model***

Before analysing sentiment within the Afghan news dataset, an external validation step was conducted in this thesis to assess how reliably VADER performs on a labelled news corpus. This responds directly to the supervisor’s recommendation given to the writer to evaluate the sentiment model on a dataset with ground truth labels.

A publicly available Reuters news headline dataset containing over 30,000 items was used here. Each headline is labelled on a five-point sentiment scale starting from (1 = very negative) to (5 = very positive). To match the three-class structure used in this thesis, these labels were mapped as follows:

* 1–2 → Negative
* 3 → Neutral
* 4–5 → Positive

A sample of 5,000 headlines was then selected to ensure computational efficiency. The same preprocessing pipeline used for the Afghan summaries like lowercasing, punctuation removal, stop word removal and others was applied to this external dataset to maintain methodological consistency.

VADER was then used to predict the sentiment of each cleaned headline. The predicted labels were then compared with the true sentiment labels, generating an accuracy score as well as precision, recall and F1-scores for each class.

The model achieved an overall accuracy of 42.3%, which is consistent with existing literature showing that lexicon-based models like VADER perform modestly on formal news texts. VADER is highly effective when comes to identifying clear negativity but struggles with neutral headlines and subtle emotional cues, a commonly reported limitation.

A screenshot of a number of numbers

AI-generated content may be incorrect.

FIGURE 4.1 — External Validation Results

The accuracy may appear low, but it is appropriate for this type of model and dataset. This validation step strengthens the thesis by demonstrating that the sentiment pipeline was empirically tested, its limitations acknowledged and its behaviour understood before applying it to the Afghanistan corpus.

***4.3 Sentiment Analysis Results***

**4.3.1 Sentiment Label Distribution**

VADER was applied to the clean summary field of each article to calculate a compound sentiment score in the range -1=1. These scores were then mapped to three discrete labels using the standard VADER thresholds:

• score ≥ 0.05 → Positive

• score ≤ −0.05 → Negative

• otherwise → Neutral

(Hutto & Gilbert, 2014).

Using this procedure the results show:

* Positive: 70 articles (62.50%)
* Negative: 32 articles (28.57%)
* Neutral: 10 articles (8.93%)

These figures show that, overall Afghan news articles mentioning AI tend to adopt a more positive than negative tone, with a relatively small proportion of neutral reporting.

A screenshot of a computer

AI-generated content may be incorrect.

FIGURE 4.2

A graph with blue rectangular bars

AI-generated content may be incorrect.

FIGURE 4.3

The bar chart above in Figure 4.3 visualises this distribution and makes the skew towards positive coverage more evident. While negative articles are still substantial, the majority emphasise benefits, opportunities or constructive uses of AI.

**4.3.2 Examples of Positive, Negative, and Neutral Articles**

To understand better how VADER’s numerical output maps to real news content, several representative articles were inspected manually.

**Positive example**

One of the highest scoring positive articles is titled “74 % of Afghan youths actively use AI for work”. The summary describes Afghanistan’s young people using AI tools for study, work, self-improvement etc. The language is very aspirational and emphasises empowerment and opportunity, which explains the strong positive sentiment score assigned by VADER to it. This type of article illustrates how Afghan media sometimes portray AI as a tool for education, skills and future employment, especially for youth.

**Negative example**

In contrast, a clearly negative article reports that “Iran uses AI systems to control shared border”. The summary focuses on surveillance, security and tensions around border control. Words associated with risk, control and threat contribute to a very negative sentiment score. This shows one of recurring pattern where AI appears in stories about geopolitics, security and regional anxiety, especially involving neighbouring countries.

**Neutral example**

A more neutral case is a “Science & Technology landing page” article that just simply lists technology updates and categorised information. The summary is descriptive, with almost no emotionally charged language to it. For this one VADER assigns a sentiment score close to zero and the article is labelled Neutral. Such articles tend to appear when AI is mentioned as part of broader technology coverage rather than the main focus.

These examples mentioned above confirms that VADER’s automatic labels generally align with human intuitions development oriented and educational stories skew positive, surveillance and geopolitical stories skew negative and informational catalogues tend to be neutral.

**4.3.3 Interpretation of Sentiment Trends**

The sentiment distribution suggests several broader trends in how AI is currently framed in Afghan media:

**1.** Optimistic framing dominates: Over 60% of articles are positive, often showing AI as an enabler of education, training, healthcare innovation or economic opportunities. This reflects global patterns where AI is frequently associated with progress and modernisation (Fast & Horvitz, 2017).

**2.** Risks and threats are visible but secondary: Around 1/3 of articles focus on risks like deepfakes, misinformation, surveillance systems or foreign military use of AI. These stories are important because they shape public concerns and fairs, but they do not dominate the corpus.

**3.** Limited neutral reporting: With fewer than 10% neutral articles most of the coverage takes a stance, either explicitly or implicitly. This then aligns with observations that AI is often framed as either an opportunity or a threat rather than a purely technical development (Birnbaum, 2019).

**4.** Influence of fact-checking content: Several negative articles belong to fact checking organisations responding to deepfakes and misinformation. These pieces tend to use strong wording about manipulation, risk and harm which naturally lowers their sentiment scores.

Overall, the sentiment analysis indicates that Afghan AI coverage is cautiously very optimistic balancing both enthusiasm for innovation and anxiety about misuse. In later sections of this thesis these patterns are connected to the topics discovered by the LDA model to see which themes are most positive or negative.

***4.4 Topic Modelling Results***

To identify the main narratives surrounding AI in Afghan news, a LDA topic model was applied to the cleaned article summaries to be used. Topic modelling makes it possible to detect recurring themes in the dataset without requiring any manual labels (Blei et al., 2003).

The model was trained on 112 summaries from the 112 news articles and configured, based on coherence testing, to generate three topics.

The resulting topic distribution was as follows:

Topic counts from Jupyter Notebook:

* AI Policy, National Strategy & International Cooperation — 44 articles
* AI Tools, Media Coverage & Public Awareness — 43 articles
* AI Adoption in Afghan Daily Life & Education — 25 articles

These themes reflect the major domains in which AI appears in Afghan media: national strategy, public awareness, and everyday use.

A screenshot of a computer

AI-generated content may be incorrect.

A graph with blue bars

AI-generated content may be incorrect.

FIGURE 4.4

Topic Distribution in Afghan AI News Articles

**4.4.1 Topic 1**: AI Policy, National Strategy & International Cooperation (44 articles)

Top keywords (from LDA):

ai, policy, national, strategy, government, international, cooperation, development, plan

This topic groups articles discussing:

* Afghanistan’s national AI strategies or digital roadmaps
* international partnerships involving AI (UN agencies, neighbouring states, NGOs)
* AI in national planning, governance, or public-sector modernisation
* policy debates and regulatory perspectives
* government-led digital transformation initiatives

These articles generally adopt a neutral to positive tone. Many focus on high level planning, cooperation agreements or policy announcements which are descriptive rather than emotional. Several pieces highlight the potential of AI to support national development, producing moderately positive sentiment.

Example narrative taken from dataset:

“Afghanistan collaborates with international organisations to integrate AI in development planning…”

This topic represents institutional and policy-oriented discussions surrounding AI adoption in Afghanistan.

**4.4.2 Topic 2**: AI Tools, Media Coverage & Public Awareness (43 articles)

Top keywords:

ai, tools, media, platform, information, awareness, technology, public, news

This topic captures coverage related to:

* AI tools used in journalism or digital communication
* public awareness campaigns about AI
* Reporting on AI technologies and their societal implications
* commentary from Afghan media outlets about global AI trends
* platforms and applications introduced to Afghan users

This theme above tends to show balanced or slightly positive sentiment. Articles mostly highlight new technologies, public education or awareness building efforts. Media stories then often emphasise accessibility, innovation or how AI can assist Afghan society.

Example from dataset:

“Local media outlets introduce new AI tools to improve reporting accuracy…”

Topic 2 in this case reflects how AI is framed and communicated to the Afghan public, with an emphasis on technological developments and media engagement.

**4.4.3 Topic 3**: AI Adoption in Afghan Daily Life & Education (25 articles)

Top keywords:

ai, education, student, school, daily, life, learning, digital, skills

This topic includes articles on:

* Afghan students learning digital and AI skills
* AI used in classrooms, training centres or youth programmes
* Daily life applications of AI tools
* NGOs providing digital education
* youth empowerment through technology

These articles mentioned above tend to be strongly positive, highlighting progress, opportunities, and uplifting narratives about Afghan youth using and adopting AI tools despite infrastructural challenges.

Example from dataset:

“Students in Kabul use AI tools as part of new digital literacy initiatives…”

Topic 3 reflects the developmental and educational dimension of AI in Afghanistan, focusing on future potential and capacity building.

**4.4.4 Topic Coherence and Model Quality**

The model produced well-separated topics, confirmed through:

* keyword clustering
* inspection of representative summaries
* distribution patterns across articles

The nearly equal size of Topic 1 and Topic 2 suggests that Afghan media devote similar attention to policy level AI developments and public facing AI tools, while Topic 3 even though smaller remains meaningful due to the prominence of AI in youth education initiatives.

**4.4.5 Interpretation**

Combined, all the three topics demonstrate how Afghan media frame AI across multiple societal layers:

Topic - Meaning - Dominant Sentiment

1. AI Policy, National Strategy and International Cooperation - AI as national development and governance - Neutral–Positive

2. AI Tools, Media Coverage and Public Awareness - AI as public communication and technological engagement - Balanced / Slightly Positive

3. AI Adoption in Daily Life and Education - AI as youth opportunity, skills and empowerment - Mostly Positive

Together, these topics show that Afghan AI reporting often presents AI as:

* a tool for national strategy and cooperation
* a technology shaping media and public awareness
* a catalyst for youth development and education

This creates a narrative where AI in Afghanistan is portrayed as both a strategic asset and a practical tool for societal progress.

***4.5 Joint Sentiment–Topic Analysis***

While sentiment analysis and topic modelling each provide valuable and important insights independently, combining them shows how Afghan media emotionally frames each major AI related theme. This integrated view helps to determine whether certain topics such as security or innovation are associated with more positive or negative reporting.

The Jupyter notebook generated a combined dataset of:

* cleaned summaries
* sentiment labels (Positive, Neutral, Negative)
* topic assignments (Topics 1, 2, and 3)

This merged structure allows us to examine emotional tone within each theme.

A screenshot of a computer program

AI-generated content may be incorrect.

FIGURE 4.5

A white text with black text

AI-generated content may be incorrect.

FIGURE 4.6

A screenshot of a computer

AI-generated content may be incorrect.

FIGURE 4.7

**4.5.1 Sentiment Breakdown by Topic**

Using the group by operation in the notebook, the following sentiment topic distribution was generated:

A screen shot of a computer

AI-generated content may be incorrect.

FIGURE 4.8

A graph with different colored bars

AI-generated content may be incorrect.

FIGURE 4.9

The Jupyter output appears as:

Topic 1 – AI Adoption in Afghan Daily Life & Education

* Negative: 7
* Neutral: 2
* Positive: 16

Topic 2 – AI Policy, National Strategy & International Cooperation

* Negative: 11
* Neutral: 6
* Positive: 27

Topic 3 – AI Tools, Media Coverage & Public Awareness

* Negative: 14
* Neutral: 2
* Positive: 27

This sentiment topic distribution highlights clear emotional patterns in Afghan AI reporting:

* Topic 1 contains mostly positive and negative sentiment.
* Topic 2 shows a strong positive focus but also notable negative reporting.
* Topic 3 has a high positive article but also the highest number of negative ones, reflecting conflicting narratives around AI tools media coverage and public awareness.

Across all three topics, positive sentiment is the most common and frequent label, followed by negative, with neutral sentiment appearing less often. Even in areas where risks might be expected like security, policy or media manipulation the proportion of positive coverage remains higher than negative.

This indicates that Afghan news outlets tend to frame AI in a constructive or hopeful way overall, while still having concerns in specific contexts.

**4.5.2 Topic 1 – AI Adoption in Afghan Daily Life & Education → Predominantly Positive**

Topic 1 brings together articles about every day and educational uses of AI in Afghanistan. Typical stories include:

* AI supported teaching and e-learning initiatives
* youth gaining digital or AI related skills
* pilot projects introducing AI into schools or universities
* accounts of how AI tools are used in daily life or local services

The sentiment distribution of topic 1 above shows that this topic is predominantly positive. Articles often highlight AI as:

* a way to improve education quality,
* a means for young Afghans to connect to global knowledge, and
* a symbol of progress and modernisation despite the country’s difficult conditions.

Negative articles usually appear when access is unequal, projects fail or when technological ambitions clash with infrastructure limits. But Topic 1 strongly supports the idea that Afghan media often associate AI with opportunity, skills and a better future.

**4.5.3 Topic 2 – AI Policy, National Strategy & International Cooperation → Positive but Critical**

Topic 2 focuses on policy debates, national strategies and international cooperation around AI. Articles in this cluster discuss:

* government plans or policy statements on AI and digital transformation
* regional cooperation or partnerships involving AI technology
* high-level discussions of standards, regulation, and national strategy

With 27 positive, 11 negative and 6 neutral articles, the topic is overall again positive but includes substantial critical coverage. Positive pieces highlight:

* AI as part of state modernisation and catching up with global trends
* diplomatic or technical cooperation framed as beneficial for Afghanistan
* potential for AI to support economic growth and governance.

Negative sentiment appears when policy moves are perceived as:

* vague or poorly implemented.
* dependent on foreign powers with contested interests
* disconnected from the realities of infrastructure and education in the country.

So, Topic 2 portrays AI policy as aspirational but contested. Media narratives recognise the promise of national AI strategies, while also questioning feasibility and power imbalances.

**4.5.4 Topic 3 – AI Tools, Media Coverage & Public Awareness → Mixed with Strong Concerns**

Topic 3 covers AI tools, journalism and public communication including:

* news about specific AI applications or platforms
* coverage of AI in social media and broadcast media
* stories about public understanding, AI literacy and awareness campaigns

The sentiment distribution of topic 3 shown above is mixed but still more positive than negative or neutral. Positive articles often describe:

* AI tools that support information access, translation or content creation
* Media stories that present AI as exciting innovation
* outreach efforts that aim to educate the public about digital technologies

Negative sentiment appears in stories that link AI tools to:

* misinformation, clickbait or sensational coverage
* concerns about automation and job loss
* ethical questions around the use of AI generated content

Topic 3 reveals a dual nature within media narratives themselves: AI is both a good tool that enables better communication and a source of new ethical and informational risks.

**4.5.5 Interpretation and Implications**

All together the sentiment topic results suggest that Afghan media present AI through a balanced but hopeful lens:

1. AI as Education and Everyday Progress

Topic 1 shows that AI adoption in daily life and education is framed mostly in positive terms like emphasising skills, youth empowerment, access to knowledge and others.

2. AI as Strategic Modernisation with Political Tensions

Topic 2 highlights AI policy and national strategies as part of modern state-building but also reveals critical concerns about the feasibility, dependence on external factors and governance quality.

3. AI as a Transformative Media Tool with New Risks

Topic 3 illustrates how AI reshapes media and public communication. It is combining optimism about new tools with worries about misinformation, ethics and control.

Overall, the results here show that AI in Afghan news is not simply celebrated or feared. It is portrayed as a technology that:

* brings opportunities for education and development
* Is embedded in contested policy and geopolitical debates
* introduces new challenges for information integrity and media practice.

This nuanced framing aligns with the broader thesis argument that when it comes to a fragile and under-resourced context like Afghanistan, AI is understood as both a symbol of hope and a source of risk.

***4.6 Result Inspection and Qualitative Verification***

**4.6.1 Use of the Interface for Qualitative Assessment**

To complement the quantitative earlier findings in this chapter, the result inspection interface was used to conduct qualitative examination of some selected model outputs. This step does not aim to measure predictive performance but to demonstrate interpretability, internal consistency and analytical usability. It is an approach recommended for NLP research involving public discourse and media analysis (Eisenstein, 2019).

By enabling to retrieve articles via unique identifiers, the interface then allows sentiment and topic labels to be assessed in relation to their original headlines, summaries and sources. This grounding helps ensure that automated classifications remain interpretable rather than abstract numerical outputs (Fowler, 2018).

**4.6.2 Illustrative Example**

As a example, Article ID 65 was examined using the interface. The article published by global history dialogues, reports survey findings indicating that part of Afghan students perceive AI as having a negative impact on critical thinking and independent decision making. The system assigns this article to the topic AI Adoption in Afghan Daily Life & Education and labels it as Negative.

The example is presented just to demonstrate how the interface functions in practice as shown below.

A close-up of a computer screen

AI-generated content may be incorrect.

Figure 4.10

**4.6.3 Analytical Value and Consistency with Quantitative Results**

Beyond individual article inspection the interface was used to filter articles by topic and sentiment, producing grouped views of the dataset. These filtered outputs reflect patterns discussed earlier in Chapter 4, such as more positive framing in innovation and skills related narratives and more neutral or negative framing in security and surveillance-oriented reporting.

The consistency between these qualitative observations and the earlier quantitative distributions reinforces the coherence of the analytical pipeline without introducing additional metrics or claims (Hutto & Gilbert, 2014). Also the result inspection interface here functions as a verification and interpretation aid rather than a deferent and separate analytical contribution.

***4.7 Visualisation and Interpretation of Results***

In this section the main visual outputs generated by the sentiment analysis and topic-modelling pipeline are presented. The figures give an intuitive overview of how Afghan news media have represented AI, showing both the overall emotional tone and how articles are distributed across the three topic labels introduced earlier before in Chapter 4.

**4.7.1 Sentiment Distribution Across the Dataset**

To examine the emotional framing of AI in Afghan news, VADER sentiment analysis was applied to all of 112 summaries. Compound scores were mapped to three labels using the standard thresholds as Positive, Negative and Neutral(Hutto & Gilbert, 2014).

A screenshot of a computer

AI-generated content may be incorrect.

FIGURE 4.11

The bar chart shows the following counts:

* Positive: 70 articles
* Negative: 32 articles
* Neutral: 10 articles

So, Afghan news mentioning AI is predominantly positive, with almost 2/3 of articles falling into the Positive class. Negative articles still make up some minority like close to 1/3, while strictly Neutral reporting is relatively rare. This pattern supports the earlier finding Section 4.3 that Afghan media often frame AI in a constructive or hopeful way especially when it comes to youth skills, education and innovation initiatives.

**4.7.2 Example Sentiment Outputs: First Ten Records**

A screenshot of a computer

AI-generated content may be incorrect.

FIGURE 4.12

The first 10 rows of the data frame above illustrate how these labels appear in practice. Most of the sample summaries are classified as Positive with one Neutral example:

* Positive summaries typically describe Afghan youths using AI tools in deferent ways like education
* The Neutral example is a more general technology update with descriptive language and little emotional content

This small sample shows the overall distribution in Figure 4.11 and suggests that VADER’s labels align well with intuitive human judgements for this dataset.

**4.7.3 Topic Model: Top Keywords for Each Topic**

LDA was used to group the 112 summaries into three topics. The model’s keyword printout summarises the most probable terms for each topic.

A white text with black text

AI-generated content may be incorrect.

FIGURE 4.13

So, interpreting the keywords together with manual inspection of articles the three topic labels used in Chapter 4 can be reaffirmed as:

* “AI Policy, National Strategy & International Cooperation” articles on national plans, international collaboration and high-level AI strategy
* “AI Tools, Media Coverage & Public Awareness” articles about AI tools, platforms and how media outlets present AI to the public
* “AI Adoption in Afghan Daily Life & Education” articles on youth skills, education initiatives, and everyday uses of AI tools

The keyword lists are generic like ai, Afghan, article, education etc. But in combination with full summaries, they support these three thematic labels established in Section 4.4.

**4.7.4 Word Clouds for Topics**

Word clouds were then generated to provide a visual of the most frequent words associated with each topic.

A close-up of words

AI-generated content may be incorrect.

FIGURE 4.14: “AI Policy, National Strategy & International Cooperation”

A close-up of words

AI-generated content may be incorrect.

FIGURE 4.15: “AI Tools, Media Coverage & Public Awareness”

A close-up of words

AI-generated content may be incorrect.

FIGURE 4.16: “AI Adoption in Afghan Daily Life & Education”

From these several patterns are visible:

* The policy and cooperation topic using terms such as afghan, afghanistan, cooperation, development and technology reflect national level planning and international engagement.
* The tools and media topic showing words like article, piece, showing, technology, and tool goes with coverage of AI platforms and media presentations.
* Then the daily life & education topic brings forward words such as education, study, student, training, digital, and opportunity pionting the focus on youth skills and learning.

Common words such as ai, afghan and tone appear across all three clouds because they are frequent and used in almost every summary, but the additional surrounding terms help distinguish the themes.

**4.7.5 Topic Distribution Across the Dataset**

A graph with blue bars

AI-generated content may be incorrect.

FIGURE 4.17

The topic distribution bar chart again shows how many articles fall into each label:

* AI Policy, National Strategy & International Cooperation: 44 articles
* AI Tools, Media Coverage & Public Awareness: 43 articles
* AI Adoption in Afghan Daily Life & Education: 25 articles

Topics related to policy/strategy and tools/media appear a lot often, but education and daily-life topic is smaller even though they are still substantial underlining the growing visibility of AI in youth and training contexts.

**4.7.6 Integrated Topic–Sentiment Interpretation**

Combining the topic labels with the sentiment breakdown from Section 4.5 provides richer picture of how AI is framed emotionally within each theme:

* “AI Adoption in Afghan Daily Life & Education” Mostly positive, with some negative and very few neutral articles.
* “AI Policy, National Strategy & International Cooperation” Overall positive but critical.
* “AI Tools, Media Coverage & Public Awareness” Mixed, with high counts of both positive and negative sentiment.

These patterns again reinforce the central finding of Chapter 4 on how Afghan news media portray AI as both an opportunity and a risk considering the cases.

**4.7.7 Limitations of the Visualisation Approach**

For transparency, several limitations should be acknowledged:

* Word clouds do highlight frequent words but do not indicate how strongly those words define a topic
* LDA can be sensitive to short summaries which may limit topic separation
* VADER may misinterpret things like political nuance or culturally specific language especially in translated content
* Topic labels are interpretive based on human reading of model outputs rather than objective categories

These constraints are very typical of computational media analysis, particularly when it comes to low-resource contexts.

**4.7.8 Summary of Visual Findings**

Taken together, the visualisations show that:

* Afghan AI coverage is mostly positive but negative sentiment also remains significant especially where things like security, governance, and information integrity are at stake
* The topic model identifies three stable themes (policy/strategy, tools/media, and daily life education) that align with manual reading of the articles
* When sentiment is overlaid on these topics, AI appears as a developmental opportunity in education and skills, a contested policy instrument and a source of new media related risks.

These insights provide a solid base for the interpretive discussion and broader implications of Section 4.8.

***4.8 Discussion of Findings***

The results of the sentiment analysis, topic modelling and qualitative result inspection provide an overall picture of how Afghan news media represent AI. The findings point towards a balanced but cautiously optimistic narrative in which AI is framed both as a path to development and as a source of emerging risk.

Positive sentiment was most common when reporting on topics like youth education, digital skills, innovation initiatives and international scientific cooperation. These narratives then portray AI as a tool for empowerment and future opportunity especially for Afghanistan’s younger population.

Topic modelling further contextualised these findings by identifying three dominant themes. The first two themes appeared the most, reflecting the prominence of policy discourse and media facing AI developments. The education and daily life theme, while smaller was overwhelmingly positive and closely aligned with development-oriented narratives.

Combining the sentiment and topic analysis revealed important nuances. Education related coverage was mostly positive, policy related reporting was also mostly positive but included critical perspectives and media and tool focused articles showed the most mixed emotional profile, often linking AI to misinformation, ethical concerns and automation fears.

To complement these quantitative results, a result inspection interface used to examine selected articles and filtered subsets of the dataset. This qualitative step confirmed that sentiment and topic assignments generally aligned with article content and thematic framing, enhancing interpretability and analytical confidence without altering the underlying results.

Altogether, the findings suggest that Afghan media do not portray AI as a one-dimensional phenomenon. Instead, AI is represented through a dual lens of opportunity and caution shaped by Afghanistan’s socio-political realities, infrastructural constraints, and exposure to regional and global technological developments.

**Chapter 5: Discussion, Conclusion and Future Work**

***5.1 Discussion***

This thesis showed and examined how AI is represented in Afghan news media’s using a structured NLP pipeline combining preprocessing, sentiment analysis and topic modelling. Chapter 4 results demonstrate that this approach successfully captures dominant narratives and emotional framing across a manually constructed and put together dataset of 112 articles.

One of the key observations is the dominance of positive sentiment. Most positive articles emphasise on topics like education, youth skills, innovation, and international cooperation showing AI as a way to empowerment and development. This optimistic framing is particularly notable given Afghanistan’s bigger political and economic challenges.

Negative sentiment appears mostly in articles addressing surveillance technologies, border security and misinformation. These narratives reflect both longstanding regional tensions and problems while also pointing on newer concerns around deepfakes, information manipulation and ethical risks associated with AI. Although negative coverage is also big, it does not dominate the media landscape.

Topic modelling reinforces these patterns by identifying three coherent themes aligned with Afghanistan’s socio-political context. Policy related coverage reflects aspirations for digital modernisation with concerns about feasibility and dependence on external actors. Media coverage combines enthusiasm for new technologies while being concerned about misinformation and ethical implications. Education and daily life narratives remain mostly positive, showing youth engagement and skills development.

An external validation step was done to contextualise the reliability of VADER sentiment analysis. Applied to a 5,000-headline subset of a labelled Reuters dataset, VADER achieved an accuracy of 42.3%. While its modest, this result aligns with existing research showing that lexicon-based models struggle with formal news text and neutral phrasing. The validation points to limitations rather than undermining the Afghan dataset analysis, reinforcing methodological transparency.

Finally, including a result inspection interface supported qualitative verification of model outputs. By enabling article-level inspection and filtered views of the sentiment and topics, this component helped make the analysis clearer without changing the overall focus of the actual thesis.

***5.2 Conclusion***

This study provides one of the first and few structured analyses of AI representation in Afghan news media. Through manual dataset collecting and construction, application of an end-to-end NLP pipeline and careful interpretation of the results, the thesis offers both empirical insight and methodological contribution.

The findings show that Afghan media frame AI through two parallel narratives which are optimism around education, skills and innovation, also Fair and caution regarding surveillance, misinformation and governance risks. Rather than presenting AI as either fully beneficial or harmful, media coverage reflects a very nuanced understanding shaped by Afghanistan’s political, cultural and infrastructural conditions.

The external validation step strengthens this study’s academic integrity by openly acknowledging limitations of lexicon-based sentiment analysis. By situating VADER’s performance within broader NLP literature, the thesis avoids overclaiming accuracy while still demonstrating meaningful analytical value.

So overall the research meets its objectives by building an original dataset, implementing a transparent NLP pipeline and offering an evidence-based interpretation of how AI is perceived in Afghan media outlets.

Beyond its empirical findings, this thesis also demonstrates the feasibility of applying Natural Language Processing methods in data-scarce and politically sensitive environments. By relying on manual data collection, transparent preprocessing and interpretable models, the study also shows that meaningful computational analysis is even possible in contexts where large datasets are unavailable. The methodological choices made throughout the thesis prioritise clarity, reproducibility and contextual awareness. Choosing it over technical complexity, making the approach suitable for future research in similar low-resource settings. The thesis not just contributes to understanding AI narratives in Afghanistan but also provides a practical reference framework for other researchers seeking to analyse media discourse in underrepresented regions using accessible data science techniques.

In summary, this thesis shows the importance of examining technological narratives within their local and cultural contexts. Afghanistan’s case shows that AI discourse cannot be understood only through global trends. But must be interpreted through the country’s unique historical, political and social conditions. By focusing on media representation rather than technological capability alone, the study reinforces the idea that perception plays a very critical role in shaping future adoption, trust and policy direction. This perspective is important for ensuring that discussions around AI remain inclusive of regions often overlooked in global technology research.

***5.3 Future Work***

Several directions exist for extending this research. First, the dataset could be expanded by incorporating additional Afghan outlets from paid data archives and newly published ones. A larger and more linguistically diverse corpus would enhance analytical depth and representativeness.

Second, future studies should explore more advanced sentiment models like transformer-based and multilingual ones which may better capture contextual nuance and reduce misclassification of neutral or formal news language.

Third, topic modelling could be enhanced and taken to next level by using contextual embedding techniques like BERT or temporal models to track narrative shifts over time.

Finally, comparative or longitudinal analyses could examine how AI narratives evolve and change across political periods or differ between Afghanistan and other low-resource and conflict-affected regions. Such extensions would deepen and add understanding of how technological discourse is shaped by broader socio-political conditions.

**References:**

Ahmad, S. (2023). *Technology, geopolitics and trust in South Asia*. [online] Taylor & Francis Online. Available at: https://www.tandfonline.com/

Birnbaum, M. (2019). *How the media shapes public understanding of artificial intelligence*. Columbia Journalism Review. [online] Available at: https://www.cjr.org/analysis/artificial-intelligence-media.php

Bird, S., Klein, E. and Loper, E. (2009). *Natural Language Processing with Python*. Sebastopol, CA: O’Reilly Media. [online] Available at: https://www.nltk.org/book/

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, pp. 993–1022. [online] Available at: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf [

Blodgett, S.L., Barocas, S., Daumé III, H. and Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of “Bias” in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. [online] Available at: https://aclanthology.org/2020.acl-main.485/

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. [online] arXiv. Available at: https://arxiv.org/abs/1810.04805

Eisenstein, J. (2019). *Introduction to Natural Language Processing*. Cambridge, MA: MIT Press. [online] Available at: https://mitpress.mit.edu/9780262042840/introduction-to-natural-language-processing/

Fast, E. and Horvitz, E. (2017). Long-term Trends in the Public Perception of Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*. [online] Available at: https://www.microsoft.com/en-us/research/publication/long-term-trends-in-the-public-perception-of-artificial-intelligence/

Fowler, J. (2018). *Practical Data Science with R*. Shelter Island, NY: Manning Publications. [online] Available at: https://www.manning.com/books/practical-data-science-with-r

Hutto, C.J. and Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. [online] Available at: <https://ojs.aaai.org/index.php/ICWSM/article/view/14550>

International Telecommunication Union (ITU) (2023). *Measuring digital development: Facts and figures*. [online] Available at: https://www.itu.int/itu-d/reports/statistics/

Joshi, P., Santy, S., Budhiraja, A., Bali, K. and Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. [online] Available at: <https://aclanthology.org/2020.acl-main.560/>

Jurafsky, D. and Martin, J.H. (2023). *Speech and Language Processing* (3rd ed. draft). Stanford University. [online] Available at: https://web.stanford.edu/~jurafsky/slp3/

Kellerman, A. (2021). *Technology and Society in Post-Conflict Environments*. London: Routledge. [online] Available at: https://www.routledge.com/Technology-and-Society-in-Post-Conflict-Environments/Kellerman/p/book/9780367472536

Manning, C.D., Raghavan, P. and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press. [online] Available at: https://nlp.stanford.edu/IR-book/

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, pp. 56–61. [online] Available at: https://conference.scipy.org/proceedings/scipy2010/mckinney.html

Oxford Insights (2022). *Government AI Readiness Index*. [online] Available at: https://www.oxfordinsights.com/government-ai-readiness-index

Rahman, M.M., Islam, M.R. and Hasan, M. (2022). Sentiment Analysis Challenges in South Asian Languages. *Journal of Information Processing Systems*, 18(3). [online] Available at: https://doi.org/10.3745/JIPS.03.0183

Reuters (2021). *Afghan biometric data raises security concerns*. [online] Available at: https://www.reuters.com/world/asia-pacific/afghan-biometric-data-security-concerns-2021-08-18/

Temple, B. and Young, A. (2004). Qualitative Research and Translation Dilemmas. *Qualitative Research*, 4(2), pp. 161–178. [online] Available at: https://journals.sagepub.com/doi/10.1177/0268580904040329

UNDP (2021). *Afghanistan Human Development Report*. [online] Available at: https://hdr.undp.org/content/afghanistan-human-development-report-2021

UNDP (2024). *Artificial Intelligence for Development*. [online] Available at: https://www.undp.org/digital/ai-development

UNESCO (2022). *Education under attack in Afghanistan*. [online] Available at: https://www.unesco.org/en/articles/education-afghanistan-crisis

World Bank (2020). *Digital Development in Afghanistan*. [online] Available at: https://www.worldbank.org/en/country/afghanistan

World Economic Forum (2020). *Global Technology Governance: Who’s in Charge?* [online] Available at: https://www.weforum.org/reports/global-technology-governance/ [Accessed 17 July 2025].

Language (Technology) is Power: A Critical Survey of “Bias” in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Available at: <https://aclanthology.org/2020.acl-main.485/>

*Introduction to Natural Language Processing*. MIT Press.

Available at: <https://mitpress.mit.edu/9780262042840/introduction-to-natural-language-processing/>

The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Available at: https://aclanthology.org/2020.acl-main.560/

*Speech and Language Processing* (3rd ed. draft). Stanford University.

Available at: <https://web.stanford.edu/~jurafsky/slp3/>

*Technology and Society in Post-Conflict Environments*. Routledge.

Available at: https://www.routledge.com/Technology-and-Society-in-Post-Conflict-Environments/Kellerman/p/book/9780367472536

***Appendix A – Computational Pipeline and Code Documentation***

*• Documents the computational implementation of the analytical pipeline described in Chapters 3 and 4*

*• All code written in Python and executed in a Jupyter Notebook*

*• Notebook contains the full workflow:*

*• data loading*

*• preprocessing*

*• sentiment analysis*

*• topic modelling*

*• external validation*

*• result inspection*

*• Code is referenced to notebook cells rather than reproduced in the thesis*

*• Full notebook available in the project’s GitHub repository*

***Appendix A.1 Dataset Loading and Initial Inspection***

*• Dataset loaded using Pandas*

*• Includes:*

*• importing structured dataset*

*• verifying record count*

*• inspecting columns and data types*

*• identifying missing or malformed entries*

*• Corresponds to Section 3.3.1*

***Notebook reference:*** *Dataset loading and inspection – Cell [2-4]*

***Appendix A.2 Text Cleaning and Preprocessing***

*• Applied to article summaries prior to analysis*

*• Cleaning steps:*

*• lowercasing*

*• punctuation and number removal*

*• stopword removal (NLTK)*

*• tokenisation and rejoining*

*• creation of clean\_summary column*

*• preservation of original summaries*

*• Implements preprocessing methodology in Section 3.3*

***Notebook reference:*** *Text cleaning and preprocessing – Cell [4-8]*

***Appendix A.3 Sentiment Analysis Using VADER***

*• Sentiment analysis performed with NLTK’s VADER*

*• For each cleaned summary:*

*• compound sentiment score calculated*

*• scores mapped to Positive / Neutral / Negative*

*• sentiment labels stored as dataset column*

*• Corresponds to Sections 3.4.4 and 4.3*

***Notebook reference:*** *VADER sentiment scoring and labelling – Cell [9,10]*

***Appendix A.4 External Validation Using Reuters Dataset***

*• External validation conducted on labelled Reuters headlines*

*• Steps include:*

*• loading Reuters dataset*

*• mapping 5-point scale to 3 sentiment classes*

*• sampling 5,000 headlines*

*• applying same preprocessing pipeline*

*• computing accuracy, precision, recall, F1-score*

*• Supports Sections 3.6 and 4.2*

***Notebook reference:*** *External sentiment validation – Cell [40,41]*

***Appendix A.5 Topic Modelling with Latent Dirichlet Allocation***

*• Topic modelling implemented using Gensim LDA*

*• Workflow includes:*

*• tokenisation*

*• dictionary creation*

*• corpus construction*

*• training with three topics*

*• keyword extraction*

*• topic assignment per article*

*• Corresponds to Sections 3.4.5 and 4.4*

***Notebook reference:*** *LDA topic modelling – Cell [11]*

***Appendix A.6 Topic and Sentiment Integration***

*• Sentiment and topic labels merged into one dataset*

*• Group-by operations used to:*

*• analyse sentiment per topic*

*• generate cross-tabulations*

*• support joint interpretation*

*• Supports Section 4.5*

***Notebook reference:*** *Topic–sentiment integration – Cell []*

***Appendix A.7 Result Inspection Interface***

*• Lightweight interface implemented for structured exploration*

*• Enables:*

*• article retrieval by ID*

*• filtering by sentiment*

*• filtering by topic*

*• summary statistics display*

*• Functions as qualitative verification aid*

*• Supports Sections 3.5 and 4.6*

***Notebook reference:*** *Result inspection interface – Cell [between 25 and 26 numbered as a star sign \*]*

***Appendix A.8 Visualisation Generation***

*• All figures generated directly from the notebook*

*• Includes:*

*• sentiment distributions*

*• topic distributions*

*• sentiment–topic breakdowns*

*• keyword outputs*

*• word clouds*

*• Figure placement indicated in main text*

*• Additional outputs included in Appendix B*

***Notebook reference:*** *Visualisation and plotting – Cell [42,17,22,24]*

***Appendix A.9 Reproducibility and Access***

*• Full Jupyter Notebook publicly available via GitHub*

*• Contains:*

*• all code cells including some sub and supporting cells do not mention here in appendix*

*• outputs*

*• intermediate results*

*• Fully executable and documented for replication or extension*

***Appendix B – Supplementary Figures and Analytical Outputs***

*• Contains selected supplementary figures and tabular outputs*

*• Supports Chapter 4 results without interrupting main narrative*

*• All figures generated from the Jupyter Notebook (Appendix A)*

*• Only selected outputs included to avoid redundancy*

***Appendix B.1 Sentiment Distribution Output***

*• Figure B.1:*

*• Positive: 70 articles*

*• Negative: 32 articles*

*• Neutral: 10 articles*

*• Supports Sections 4.3 and 4.7*

*(Referenced as Figures 4.2, 4.3, 4.11)*

***Appendix B.2 Topic Distribution Output***

*• Figure B.2:*

*• AI Policy, National Strategy & International Cooperation – 44*

*• AI Tools, Media Coverage & Public Awareness – 43*

*• AI Adoption in Afghan Daily Life & Education – 25*

*• Supports Section 4.4*

*(Referenced as Figures 4.4 and 4.17)*

***Appendix B.3 Topic Keywords Output***

*• Figure B.3 shows top keywords per topic*

*• Used to:*

*• interpret latent themes*

*• assign topic labels*

*• verify coherence*

*• Supports Sections 4.4.1–4.4.4*

*(Referenced as Figure 4.13)*

***Appendix B.4 Sentiment–Topic Cross-Tabulation***

*• Figure B.4 shows sentiment breakdown per topic*

*• Demonstrates:*

*• strong positivity in education/daily life*

*• mixed but positive policy coverage*

*• highest negativity in media/tools*

*• Supports Sections 4.5 and 4.7.6*

*(Referenced as Figures 4.8 and 4.9)*

***Appendix B.5 Word Cloud Visualisations***

*• Figures B.5–B.7:*

*• Policy & cooperation*

*• Tools & media*

*• Daily life & education*

*• Included as illustrative aids*

*• Limitations acknowledged in Section 4.7.7*

*(Referenced as Figures 4.14–4.16)*

***Appendix B.6 External Validation Output***

*• Figure B.8:*

*• Accuracy: 42.3%*

*• Precision, recall, F1-score per class*

*• Supports Section 4.2*

*(Referenced as Figure 4.1)*

***Appendix B.7 Result Inspection Interface Output***

*• Figure B.9:*

*• Article ID 65 example*

*• source*

*• summary*

*• sentiment label*

*• topic assignment*

*• Supports Sections 4.6.1–4.6.3*

*(Referenced as Figure 4.10)*

***Appendix B.8 Note on Figure Selection***

*• Only selected figures included to avoid redundancy*

*• All remaining outputs available in the GitHub notebook*

***Appendix C – Dataset Structure and Variables***

*• Documents structure of the Afghan AI News Dataset*

*• Supports transparency, reproducibility, and examiner verification*

*• Dataset manually compiled and sole empirical input named newssheet in Excel*

***Appendix C.1 Dataset Overview***

*• cleaned dataset name: afghanistan-ai-final-dataset*

*• Records: 112 articles*

*• Time span: 2018–2025*

*• Languages: English (primary), translated from Dari & Pashto*

*• Sources: Afghan, regional, international outlets*

*• Format: Excel / CSV (loaded via Pandas)*

*• Summaries manually written or translated*

***Appendix C.2 Dataset Columns***

***ID***

*• Integer*

*• Unique article identifier*

*• Enables article-level inspection*

***Source***

*• Categorical text*

*• Publishing outlet*

*• Supports source diversity analysis*

***Date***

*• Date*

*• Publication date*

*• Enables temporal analysis*

***Title***

*• Text*

*• Original headline*

*• Contextual grounding*

***URL***

*• Text*

*• Direct article link*

*• Ensures traceability*

***Summary***

*• Text*

*• Manually written/translated (3–5 sentences)*

*• Primary NLP input*

***clean\_summary***

*• Processed text*

*• Includes:*

*• lowercasing*

*• punctuation/number removal*

*• stopword removal*

*• tokenisation*

*• Used for modelling*

***sentiment\_score***

*• Numeric (−1 to +1)*

*• VADER compound score*

***sentiment\_label***

*• Categorical*

*• Positive / Neutral / Negative*

***topic\_id***

*• Integer*

*• LDA topic assignment*

***topic\_label***

*• Categorical text*

*• Human-interpretable topic name:*

*• AI Policy, National Strategy & International Cooperation*

*• AI Tools, Media Coverage & Public Awareness*

*• AI Adoption in Afghan Daily Life & Education*

***Appendix C.3 Notes on Data Integrity***

*• No rows deleted during preprocessing*

*• Empty summaries retained as empty strings*

*• All translations performed manually*

*• Labels generated programmatically, interpreted qualitatively*

*• No personal or sensitive data included*

***Appendix C.4 Availability***

*• Full dataset and code available in GitHub repository*

*• Raw and cleaned versions preserved for reproducibility*