NewsK

Your News, Your Way.

Group 5

Osama

Khalil

Ahmed

Ayadh

Aysha

Yousif

Abdulaziz

Table of Contents

Executive Summary

The 2016 World Press Trends report by the World Association of Newspapers and News Publishers (WAN-IFRA) revealed a pivotal milestone: digital news readership surpassed **1.3 billion individuals**, accounting for approximately **40% of global internet users** who regularly access news online. With digital news consumption projected to continue its upward trajectory. Statista estimates the number of digital news readers will approach **2.0 billion by 2029**. In response to this demand, NewsK emerges as an innovative News-as-a-Service (NaaS) platform, specifically designed to address the information overload faced by bilingual individuals in the GCC region. Leveraging cutting-edge machine learning (ML) and natural language processing (NLP) technologies, NewsK delivers personalized, multilingual news recommendations tailored to the unique preferences of individual users and the strategic needs of corporate clients. By bridging the gap between overwhelming information and user-centric access, NewsK redefines how audiences interact with digital news, fostering efficiency and personalization in a globalized, multilingual environment.

At the heart of NewsK are two primary datasets

The Microsoft News Dataset (MIND) and **Saudi Newsnet**. MIND provides a vast collection of English news articles, capturing user interactions and metadata, while Saudi Newsnet offers localized Arabic news content from various Saudi Arabian online newspapers. This blend ensures that NewsK delivers high-quality, contextually relevant news in both languages, making it a versatile tool for its users.

NewsK's platform engine uses a hybrid approach, combining content based and collaborative filtering, which analyze user interaction data to identify preferences, with content-based, which employs NLP to match similar topics and themes within articles. This dual strategy ensures that recommendations are both accurate and relevant, even for new users with limited interaction data. Additionally, the platform leverages transformer-based models like llama3, distilbert and classify news by AraBERT to generate high-quality contextual embeddings, and mT5 for summarizing Arabic News, with Gen AI assistant ensuring that users receive concise and pertinent information.

The platform of NewsK, built with flask framework to be responsive and engaging, allowing users to easily navigate, view personalized recommendations, and read articles in English and Arabic. This commitment to inclusivity and multilingual support significantly enhances user engagement and satisfaction.

To maintain the quality of its recommendations, NewsK incorporates MLOps principles, which support continuous deployment and monitoring of machine learning models with the admin panel developed. Kubernetes and Docker facilitate to be included in our future scalable and efficient deployments, monitoring tools like MLflow and Prometheus track model performance and detect instances of data and model drift. Automated retraining processes ensure that the recommendation engine adapts to changes in user behavior and content trends, keeping the platform up-to-date and relevant.

For businesses, NewsK offers substantial value by aligning news content with strategic objectives, helping companies stay informed about industry trends, risks, and opportunities. The platform supports proactive decision-making, public image management, and investor confidence. Its scalable architecture, robust data processing capabilities, and adherence to quality standards and best practices make NewsK a reliable and future-proof solution for both individual users and corporate clients.

In summary, NewsK is an advanced news recommendation platform designed to streamline information consumption, reduce cognitive overload, and aid informed decision-making. Utilizing transformer-based models, it delivers high-quality contextual embeddings and concise content summaries with Gen AI assistant. Positioned NewsK as a leading solution to make a significant impact in the GCC region and beyond.

Section 1

Objective

Problem Statement

In today's fast-paced digital world, people are constantly bombarded with information from all directions. This can be especially tough for bilingual individuals in the Gulf Cooperation Council (GCC) countries who juggle both Arabic and English. The sheer volume of information in two languages can overwhelm the brain, leading to stress and lower productivity.

With news coming from various sources in different languages, it becomes hard to figure out what's true and what's not, causing confusion and misinformation. This constant need to sift through and analyze information can also lead to decision fatigue, making it harder to make choices.

Organizations, particularly publicly listed companies, face similar challenges. They need to carefully manage their public image because news can greatly affect their reputation and stock prices. It's crucial for them to monitor and control the narrative in both Arabic and English media to keep their image positive. Effective reputation management means quickly addressing any negative news to prevent damage and maintaining investor confidence by managing the flow of information that could impact stock prices.

However, current solutions often fall short in effectively aggregating and personalizing bilingual content, making it hard for both individuals and organizations to navigate this complex information landscape. There's a clear need for advanced strategies and tools that can seamlessly integrate and tailor bilingual information to reduce the negative effects of information overload.

Scope of the work

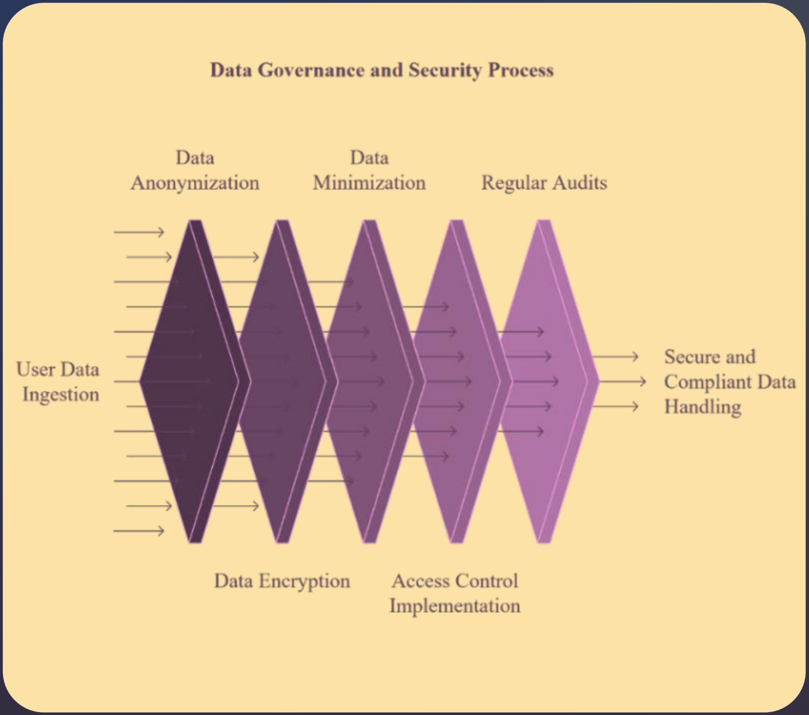
NewsK is developed to address the challenges of information overload, particularly for bilingual individuals in the GCC region who navigate both Arabic and English content. By offering a unified and personalized news recommendation platform, NewsK integrates Generative AI and Natural Language Processing (NLP) technologies to deliver tailored, multilingual content. This platform meets the diverse needs of individual users and corporate entities by providing AI-driven recommendations, contextualized summaries, and actionable insights. NewsK enhances information consumption and supports informed decision-making by reducing cognitive overload, filtering out misinformation, and mitigating decision fatigue. For businesses, NewsK aligns news content with strategic objectives, ensuring companies remain competitive by staying informed of industry trends, risks, and opportunities. This comprehensive approach helps organizations manage their public image, respond swiftly to negative news, and maintain investor confidence.

Section 1

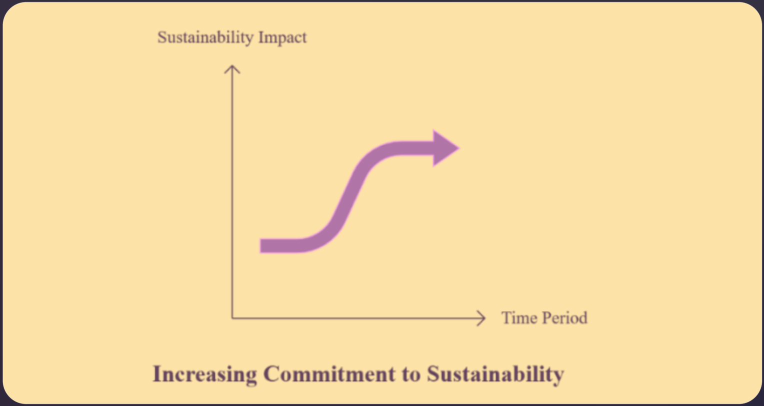
Goal

The primary goal of NewsK is to seamlessly integrate and tailor bilingual information to enhance the user's ability to efficiently process and utilize information from multiple sources in both Arabic and English.

Additionally, NewsK aims to achieve robust data governance by ensuring data privacy and compliance while maintaining high standards of data security and access controls. Designed with a comprehensive data governance framework, NewsK aligns with global regulations such as the General Data Protection Regulation[1] (GDPR) and the Saudi Personal Data Protection Law[2] (PDPL). This framework mandates user consent, data minimization, and strict access controls to protect user privacy at every step of data handling and processing. Advanced anonymization and encryption techniques are employed to safeguard personally identifiable information (PII) from unauthorized access. By adhering to data minimization principles, NewsK collects only the necessary data to deliver targeted, relevant news recommendations, reflecting its commitment to responsible data use. Future NewsK implements role-based access controls (RBAC) and multi-factor authentication (MFA) to ensure that only authorized personnel have access to sensitive data, thereby limiting exposure and reducing the risk of data breaches. Regular audits and assessments of NewsK's data infrastructure ensure compliance with internal policies and external regulatory requirements, fostering user trust and establishing NewsK as a secure platform for information dissemination.


Data Governance and Security Process
Data Anonymization
Data Minimization
Regular Audits
User Data Ingestion
Secure and Compliant Data Handling
Data Encryption
Access Control Implementation

From a sustainability perspective, NewsK aims to promote informed, ethical, and sustainable data usage practices that align with broader environmental and social responsibility objectives. NewsK supports environmental sustainability by optimizing energy efficiency through advanced data processing techniques and streamlined server architectures, thereby minimizing its digital footprint. On the social front, NewsK enhances user awareness and engagement with critical societal topics such as environmental challenges, diversity, and corporate governance[3] by integrating sustainability-focused content into its recommendation algorithms. This approach fosters a culture of informed decision-making and public engagement, contributing to a more aware and proactive society. NewsK's commitment to ethical AI usage and robust data governance ensures that technology is deployed responsibly, protecting user privacy and promoting trust. By addressing both environmental and social imperatives, NewsK exemplifies responsible digital innovation, supporting the creation of a sustainable and equitable society.



Team, Responsibilities, and Duration

We are a team of seven students in a master's course in Business Analytics and Big Data, each bringing diverse backgrounds in Engineering, IT, and Management. This project marks our first venture into developing a tool like NewsK, and we are excited to undertake this challenge under the supervision of our lecturer. The project is scheduled to run from September to December 2024, giving us a total of four months to bring NewsK to life.

Our responsibilities are divided to leverage our diverse backgrounds effectively. Students with business and management backgrounds focus on developing the business case, ensuring that NewsK aligns with market needs and corporate goals. Meanwhile, students with engineering and IT expertise concentrate on the technical development of the tool, addressing challenges related to AI, NLP, and data security. This balanced approach allows us to cover all aspects of the project comprehensively. We frequently discuss and communicate to ensure that the entire team is aware of each member's progress and contributions. This collaborative environment helps us stay aligned and work efficiently towards our common goal.

Planned work

The planned work for the development of NewsK follows a structured Machine Learning Operations (MLOps). In the scoping phase, we defined the project objectives, timelines, and success metrics to ensure alignment with user needs and corporate goals. During the data preparation stage, we established a robust pipeline for collecting and organizing bilingual news content, ensuring data accessibility, quality, and accurate labeling. This phase also prioritized validating data to maintain high standards for personalized recommendations and summaries.

In the modeling, deployment, and monitoring stages, we fine-tuned advanced NLP models to provide relevant, multilingual recommendations, contextual summaries, and misinformation filtering. The platform emphasized scalability, reproducibility, and auditing during development to ensure consistent performance. Post-deployment, the system continuously monitors for data drift, user feedback, and fairness, with regular updates to enhance content relevance and platform reliability.

Scoping	Define Project			Wrong conceptual definition Business Case Timelines
Data	Define data and establish baseline	Label and organize Data	Data Accessibility Data Format	Is Data labeled correctly? Data Validation
Modeling	Select and train model	Perform error analysis	Reproducibility Scalability	Reproducibility Model versioning Auditing
Deployment	Deploy in production	Monitor& maintain system	Reliability Scalability	Observability Auditing Data Drift Fairness Quality Explicability

Limitation:

There are several potential limitations for the NewsK project. Firstly, the limited timeframe of four months may constrain the depth and breadth of development, potentially impacting the robustness and scalability of the final product.

Additionally, the dive0rse backgrounds of the team, while beneficial for a balanced approach, might also pose challenges in terms of communication and coordination, especially when integrating business and technical aspects. Ensuring seamless collaboration and maintaining a unified vision throughout the project could be demanding.

Another limitation could be the availability and quality of bilingual data, which is crucial for training the AI and NLP models. Inadequate or biased data could affect the accuracy and reliability of the recommendations and summaries provided by NewsK. Lastly, ensuring compliance with data privacy regulations requires meticulous attention to detail and robust security measures.

Section 2

Data Sources

Overview of Datasets

NewsK leverages two primary datasets for delivering personalized news recommendations: Microsoft News Dataset (MIND) and SaudiNewsNet. Each dataset offers unique attributes that enable NewsK to provide diverse, high-quality news content in both English and Arabic, tailored to user preferences.

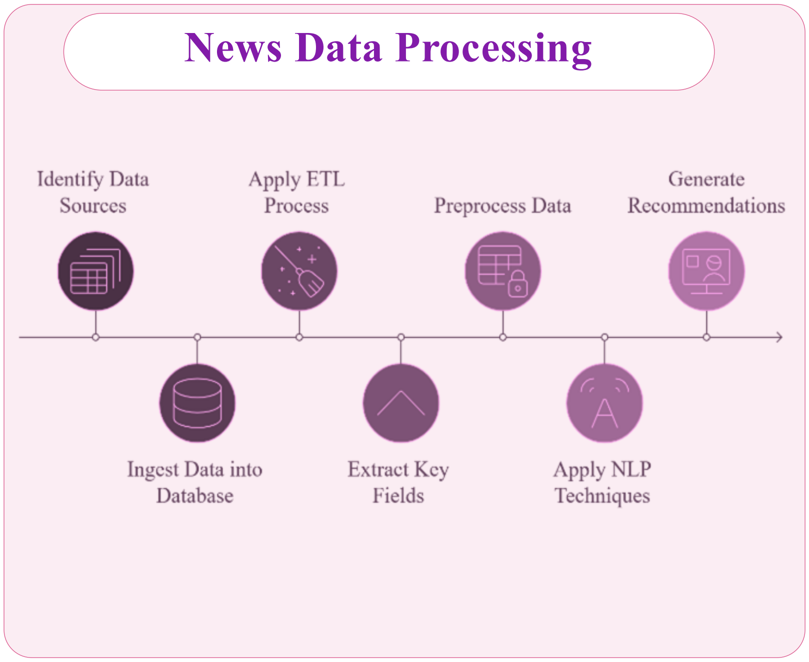
Microsoft News Dataset[4] (MIND): MIND is a large-scale English news dataset designed for news recommendation research. Comprising over 50K impression logs, it captures user interactions and metadata related to news articles, including categories, subcategories, and publication dates. This dataset provides a foundation for understanding English-language news consumption patterns.

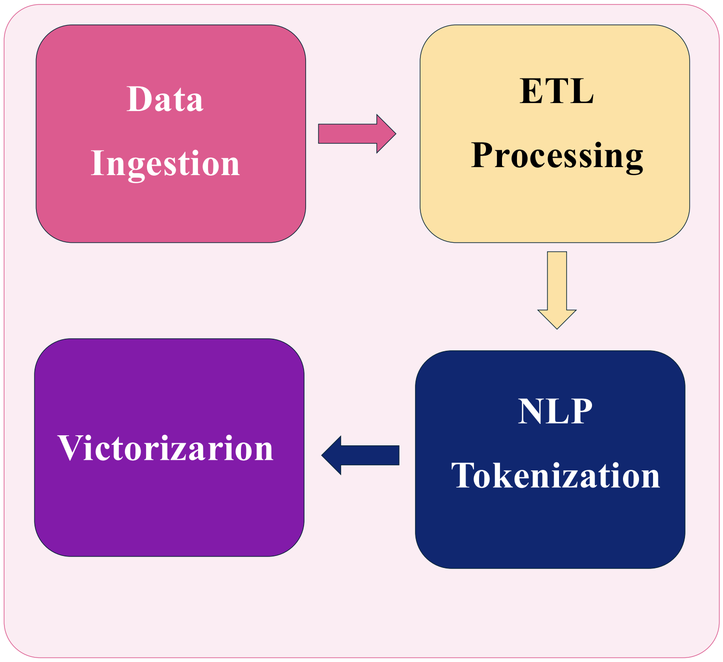
SaudiNewsNet[5]: The SaudiNewsNet dataset has 31,030 entries in both metadata and full-text articles, initially released as several JSON files. Quite heavy preprocessing was necessary to unify them into one Excel sheet that would be easier to work with using pandas. Key columns are "Source" for the originating news outlet, "Title" for headline, "Content" as the body of the news article, "Date_extracted" which gives the date and time of extraction, and "Author," for the writer of the article.

News Datasets

Data preparation for NewsK involves a multi-step process to ensure the highest level of data quality, relevance, and security. Both datasets are ingested into a centralized database, where an Extract, Transform, Load (ETL) pipeline systematically cleans and structures the data. Key fields, such as title, category, and publication date, are extracted and preprocessed to remove duplicates, normalize formats, and enhance overall quality.

For personalization, NewsK applies Natural Language Processing (NLP) techniques to tokenize and vectorize news content. Tokenization divides text into manageable units, while vectorization (via LLMs embeddings) transforms the text into numerical representations, enabling efficient content-based filtering and recommendation generation.





Section 3

Methodology

The development of NewsK leverages a range of advanced tools and technologies to address the challenges of information overload, particularly for bilingual individuals in the GCC region. For backend management, flask is used to develop the core functionalities, including data processing and model implementation, while Flask serves as the backend framework for managing API requests and handling server-side logic. MongoDB, a NoSQL database, is employed for efficient storage and retrieval of large datasets, supporting real-time data access.

On the frontend, Bootstrap.js[6] a responsive JavaScript framework, is used to build the user interface, ensuring a dynamic and engaging user experience across both mobile and desktop platforms. The machine learning and natural language processing (NLP) models at the heart of NewsK include distilbert[7] for generating high-quality contextual embeddings for English news articles and AraBERT[8] for Arabic text. mT5 is utilized for generating summaries of Arabic content, providing concise and relevant information, which ensuring multilingual support.

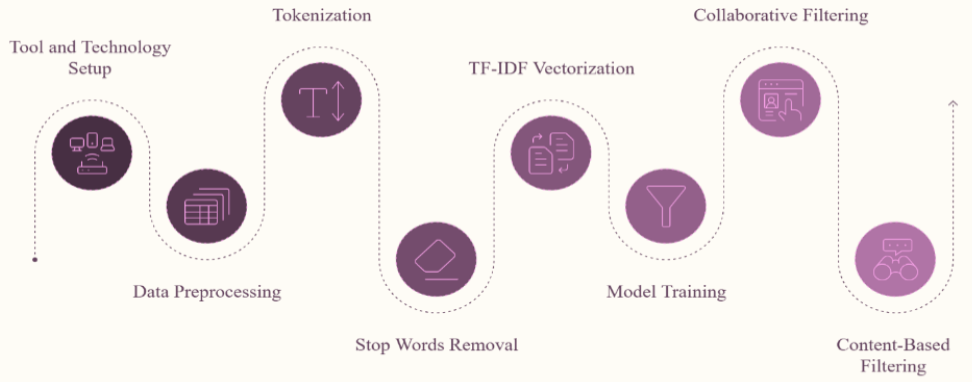
To maintain the effectiveness and accuracy of the recommendation system, NewsK integrates an MLOps pipeline that automates model retraining, deployment, and monitoring. Models are retrained monthly using the latest data, incorporating user feedback to adapt to changing preferences and news trends. Continuous Integration/Continuous Deployment (CI/CD) pipelines manage code integration and model updates, ensuring seamless deployments and minimal disruption. Kubernetes and Docker enable scalable deployments, while caching and load balancing techniques optimize

The deployment and maintenance of future NewsK are facilitated by Kubernete[9], which is employed for container orchestration, enabling scalable and efficient deployment of the platform. Docker is used to containerize applications, ensuring consistency across different environments. MLflow[10] is integrated for tracking model performance and managing the lifecycle of machine learning models, and Prometheus is used for monitoring system performance and detecting model drift.

The data preparation and processing involve several steps to ensure high-quality input for the recommendation engine. Data from the MIND and SaudiNewsNet datasets are ingested into a centralized database. An Extract, Transform, Load (ETL) pipeline cleans and structures the data, extracting key fields such as titles, categories, and publication dates. Text data is tokenized into manageable units and vectorized using techniques like TF-IDF, transforming it into numerical representations suitable for machine learning algorithms.

User interaction and feedback are integral to the continuous improvement of NewsK. The frontend interface, built with Bootstrap.js, allows users to interact with personalized news feeds, view summaries in Arabic, and read full articles. User feedback on article relevance is collected and integrated into the model retraining process, continuously refining the recommendation algorithms.

Newsk Recommendation System Process



Section 4

Solution Overview

System Architecture and Components

The NewsK platform is designed as a pioneering News-as-a-Service (NaaS) solution, built upon a robust architecture to deliver real-time, personalized, and multilingual news recommendations. This architecture empowers organizations to streamline information consumption and align insights with strategic priorities. The data ingestion pipeline forms the foundation of NewsK, efficiently capturing and integrating data from diverse sources, primarily the MIND (Microsoft News Dataset) and SaudiNewsNet datasets. These datasets provide multilingual news content in English and Arabic, which is then stored in MongoDB, a highly scalable NoSQL database. MongoDB supports efficient querying and retrieval, ensuring low latency and high availability, which are critical for a real-time news recommendation system.

Before the data can be utilized by the recommendation engine, it undergoes extensive preprocessing. This preprocessing module standardizes and structures the data for optimal performance. Text data is tokenized into smaller, manageable units, and vectorized using techniques like Term Frequency-Inverse Document Frequency (TF-IDF). Language-specific NLP models such as BERT and AraBERT are employed to handle semantic nuances in English and Arabic, respectively. This preprocessing ensures that all text data is accurate and ready for efficient input into the recommendation model.

At the core of NewsK's functionality is its recommendation engine, which combines collaborative and content-based filtering to generate highly personalized news recommendations. Collaborative filtering leverages user interaction data, such as article clicks and time spent on articles, to identify user preferences and recommend articles based on similarities with other users' behaviors. Content-based filtering, on the other hand, utilizes NLP to examine article content, matching similar topics or themes to generate recommendations. This hybrid approach ensures that NewsK can provide accurate and relevant recommendations even for new users with limited interaction data.

The frontend of NewsK is built using Bootstrap.js, a popular JavaScript framework known for its reactivity and flexibility. This allows NewsK to provide a responsive, engaging interface that enhances user interaction and accessibility. The user interface is designed to enable easy navigation, allowing users to select topics, view article summaries, and read full articles in both English and Arabic. This commitment to inclusivity and multilingual support ensures high levels of engagement and satisfaction among users.

Alternative Approaches and Rationale for Selection

In developing NewsK's recommendation system, several alternative approaches were evaluated to ensure an optimal balance between accuracy, scalability, and user experience. Initially, collaborative filtering alone was considered for news recommendations, leveraging user interaction data to identify trends and suggest articles. However, this approach faced limitations in providing recommendations to new users or for niche topics with limited user engagement data. To address these gaps, a hybrid approach integrating content-based filtering was selected, allowing for broader, contextually relevant recommendations regardless of user history.

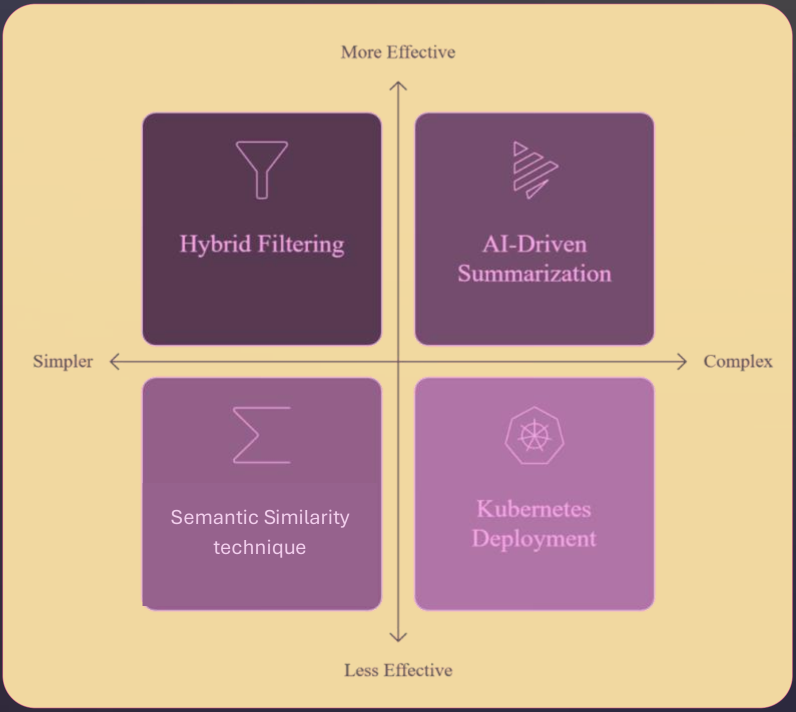
Another alternative considered was semantic similarity evaluation technique for summarization, which rely on keyword extraction and sentence scoring to generate concise news summaries. While computationally simpler, this approach lacked the sophistication needed to produce coherent evaluation summaries that captured the essence of each article. AI-driven summarization, utilizing models mT5 for Arabic, was chosen to deliver more contextually rich and accurate summaries.

As an alternative solution, NewsK employs the FastText model for news classification, offering a lightweight and efficient approach to categorizing news topics. FastText excels in speed and scalability, making it ideal for handling large datasets with minimal computational resources. Its ability to generate word embeddings and support multilingual content ensures accurate classification across diverse news sources. This approach complements transformer-based models, providing a cost-effective solution for real-time applications while maintaining reliable performance

Solution Pathway Evaluation

Finally, manual scaling was considered for handling variable traffic associated with news platforms. However, this required significant manual intervention and was less efficient in managing high traffic volumes. Kubernetes, with its container orchestration and automated scaling capabilities, was ultimately selected to ensure that the platform could handle high traffic volumes efficiently without manual intervention.

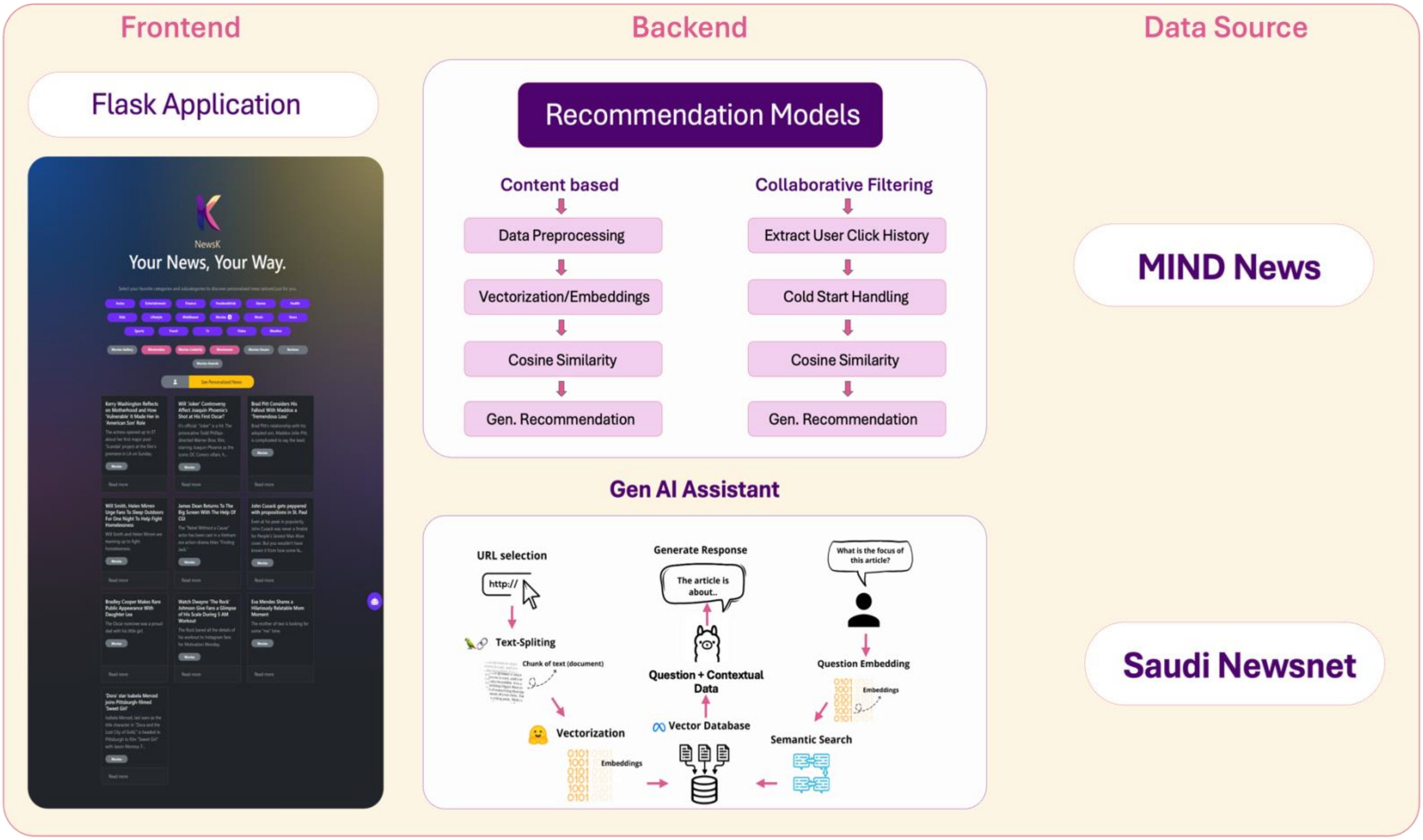
By evaluating these alternatives and selecting the most effective approaches, NewsK has developed a scalable, personalized, and user-friendly platform that meets its goals of delivering real-time, multilingual news recommendations. This comprehensive solution architecture ensures that NewsK remains a robust and adaptable platform, capable of meeting the diverse needs of its growing user base.



Section 5

Implementation and Development

The implementation and development of NewsK follow a modular and structured approach, ensuring that each component of the system is developed, tested, and deployed efficiently. The codebase is organized into distinct modules, each responsible for a specific aspect of the platform, which facilitates ease of maintenance and scalability.



Frontend

Flask Application

Backend

Recommendation Models

Content based: Data Preprocessing → Vectorization/Embeddings → Cosine Similarity → Gen. Recommendation

Collaborative Filtering: Extract User Click History → Cold Start Handling → Cosine Similarity → Gen. Recommendation

Gen AI Assistant

Data Source

MIND News

Saudi Newsnet

The frontend module is developed using flask ensuring a responsive and engaging user interface. The code is structured to allow rapid updates and feature enhancements, providing a seamless user experience across both mobile and desktop platforms. The framework is chosen for its reactivity and flexibility, which are essential for delivering a dynamic user interface.

To ensure continuous performance feedback, the monitoring and logging module integrates tools like MLflow and Prometheus. MLflow tracks model performance metrics, while Prometheus monitors system performance and detects model drift. Automated logging captures user interactions, model performance, and error rates, aiding in timely issue resolution and model refinement.

The development process adheres to established quality standards and best practices. Python code follows PEP 8 guidelines, ensuring readability and consistency. Comprehensive unit tests are written using frameworks like PyTest to validate the functionality of individual components. Version control is managed using Git, with a branching strategy that supports feature development, bug fixes, and releases. This approach ensures that code changes are tracked, reviewed, and integrated systematically.

Security and data privacy are paramount. Sensitive data is encrypted both in transit and at rest, ensuring data privacy and security. Access controls are implemented to restrict data access to authorized personnel only. The platform adheres to relevant data protection regulations, such as GDPR, ensuring that user data is handled responsibly and ethically.

Structured Report: Insights and Efforts on Building a Personalized Recommendation System for the MIND Dataset

1. Introduction

The News Dataset is a cornerstone of this recommendation system, providing a wide array of content-related features. It consists of over 51,000 training examples and 42,000 development examples, each meticulously annotated with fields such as unique news IDs, categories, subcategories, headlines, summaries, URLs, and metadata like tags for people and organizations. With 17 unique categories, including dominant themes like "News," "Sports," and "Finance," the dataset offers diverse yet structured data. Minimal missing data—5% in the "abstract" field for the training set and 4.8% for the development set—ensures high reliability for modeling purposes. Additionally, the presence of named entities enhances the dataset's richness, although some entries lack detected entities, posing a challenge for semantic modeling.

Behaviors Dataset

The Behaviors Dataset captures user interactions with news articles, offering insights into user preferences and engagement patterns. It comprises 156,965 rows in the training set and 73,152 rows in the development set. Each row contains key information, including unique impression IDs, user IDs, timestamps, click histories, and impressions of recommended articles. The dataset spans 50,000 unique users, reflecting a diverse range of behaviors and interaction frequencies. Notably, the dataset includes minimal missing values in critical fields such as `click_history`, with a missing rate of only 2.06% in training and 3.03% in development. This comprehensive behavioral data makes it highly suitable for building personalized recommendation systems that adapt to varied user preferences.

2. Observations

Modelling Challenges and Opportunities

The MIND dataset is well-structured and demonstrates high-quality data that is crucial for effective modeling. The News Dataset's diversity, with 17 unique categories and detailed metadata, provides a robust basis for feature engineering. Notably, categories like "News" and "Sports" dominate, indicating a potential skew that may require normalization or stratification during modeling. The Behaviors Dataset, on the other hand, reveals significant variation in user engagement levels. While some users display high activity, the majority exhibit moderate interaction patterns. This uneven distribution underscores the need for techniques that can effectively model both frequent and infrequent user behaviors. The minimal missing data across both datasets ensures a stable foundation for preprocessing and analysis, while the presence of named entities in the News Dataset presents opportunities for semantic enrichment.

3. Methodology

Preprocessing

The preprocessing phase focused on ensuring data consistency and enhancing feature quality. Missing values in textual fields such as `title` and `abstract` were replaced with empty strings to maintain uniformity. Text normalization involved removing special characters and numbers, converting all text to lowercase, and applying stopword removal to retain only meaningful words. Lemmatization was employed to standardize words to their base forms, thereby reducing redundancy in the feature space. To consolidate content-related features, the `title` and `abstract` fields were combined into a unified "content" field, providing a richer input for subsequent modeling tasks. The cleaned datasets were saved as `train_news.csv` and `dev_news.csv` for streamlined use in feature extraction and model training.

Modeling

Feature extraction played a pivotal role in the methodology, leveraging both traditional and modern approaches. TF-IDF vectors were generated with a maximum feature limit of 10,000 and support for bigrams, capturing term relevance while managing dimensionality. Pre-trained embeddings using the SentenceTransformer model (`distilbert-base-nli-stsb-mean-tokens`) offered deep semantic understanding, enabling the representation of textual content in a high-dimensional, context-aware space. Similarity computations, critical for both content-based and personalized recommendations, were performed using cosine similarity matrices. To handle large datasets efficiently, batch processing was employed, ensuring scalability without compromising data integrity. The recommendation framework incorporated content-based algorithms for article similarity and personalized models that leveraged user click history to rank unseen articles by relevance.

4. Results

The evaluation of the recommendation system was grounded in widely accepted metrics, including AUC (Area Under the Curve), MRR (Mean Reciprocal Rank), and nDCG (Normalized Discounted Cumulative Gain) at cutoffs 5 and 10. These metrics provided a comprehensive view of the system's effectiveness in ranking relevant articles. Results indicated that TF-IDF-based recommendations slightly outperformed embedding-based models in AUC and nDCG scores, likely due to their precise representation of term-level relevance. However, embedding-based models demonstrated superior semantic understanding, making them more adaptable to nuanced contexts. Statistical comparisons revealed robust performance across both methods, with mean and median metrics consistently favoring TF-IDF for direct relevance ranking.

Visualizations, including box and bar plots, highlighted the distribution of performance metrics and offered insights into areas for improvement. While TF-IDF excelled in ranking accuracy, embedding models showed potential for capturing complex relationships between articles, suggesting a possible advantage in hybrid approaches that combine the strengths of both methods.

5. Challenges Addressed

Several challenges were encountered and systematically addressed during the project. The issue of missing data in textual fields was mitigated through imputation strategies, ensuring no loss of critical information. Feature engineering posed a significant challenge due to the high dimensionality and complexity of the dataset; this was addressed by consolidating fields and employing dimensionality reduction techniques. Scalability concerns, particularly in similarity computations, were tackled through batch processing, enabling efficient handling of large datasets. Finally, evaluating personalized recommendations for users with limited interaction history required tailored strategies, such as incorporating fallback mechanisms and exploring collaborative filtering alternatives.

Insights and Efforts on Building a Classification and Summarization system on **Saudi Newsnet-Arabic**

Introduction

The need for Arabic automatic classification and summarization systems is growing because of the increase in Arabic digital information. Arabic, being a structurally complex language with several dialects and script variants, is one of the most challenging languages for natural language processing. The project aims to develop and apply machine learning models capable of content summarization and article classification into predefined groups on Saudi Newsnet dataset. The aims include exploring and understanding the features and structure of the dataset, preprocessing Arabic text to improve the quality of the data, classifying and representing texts using both supervised and unsupervised methods, utilizing text summarizing to convey information succinctly, and evaluating summarization, deep learning, and clustering models for their effectiveness.

Observations

Python scripts were created in order to scrape together individual news articles in JSON format, combining them into an Excel sheet for smoother management via pandas. There is also about 0.4% in the column "content" and about 15% regarding the "author" field without values. Besides that, from the Arabic language specifics also non-standard characters usage in one text, mixed usages of Arabic scripts and dialectical marks require additional special preprocessing in order for assurance of quality and suitability.

Methodology

1. Preprocessing

The dataset, coming from several JSON files, was combined into one Excel sheet using Python, hence guaranteeing that there was consistency in formatting and missing values handling. For text preparation, content cleaning, tokenization, and stopword removal were performed after exploration showed trends in missing values and the variation in the length of articles.

2. Clustering Features

Text was converted into numerical data using TF-IDF vectorization, which, besides other things, was also optimized for KMeans. Key word tagging improved the accuracy in identifying clusters and let the articles fall into politics, sports, and other financial categories.

3. Supervised Learning

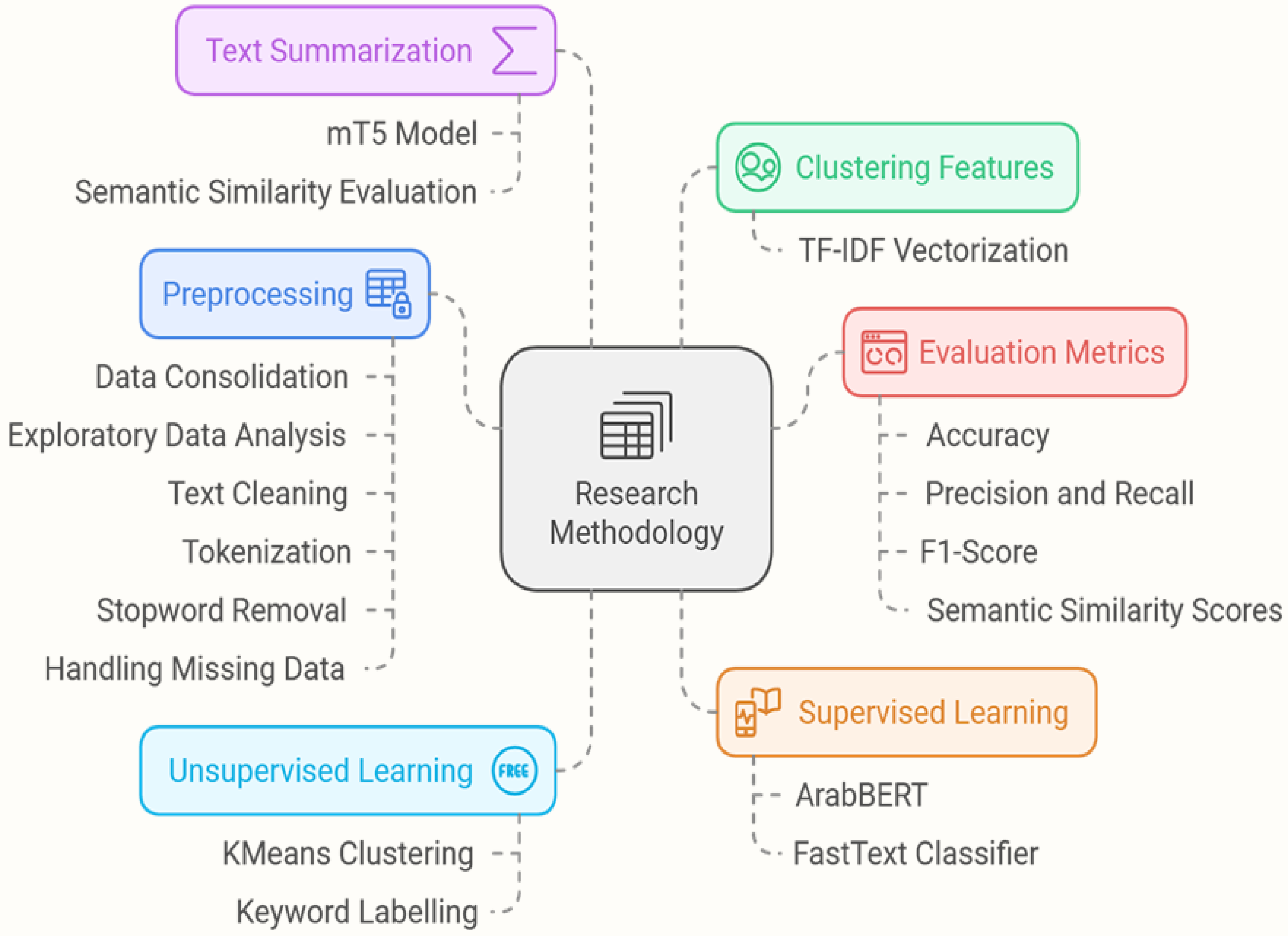
Categorization was performed using FastText and ArabBERT. FastText had the added value of being an efficient, portable alternative in low-resource scenarios, while ArabBERT served to further refine the tagged data for more in-depth study.

4. Text Summarization

The mT5 model has generated the summary of Arabic news stories in brief. Semantic similarity evaluations were performed to ensure the accuracy and relevance of the summary to the source material.

5. Evaluation

Accuracy, precision, recall, and F1-scores were used for the performance evaluation of the model, and semantic similarity scores were used for the quality of the summary; hence, it justifies the reliability of the method.



Results

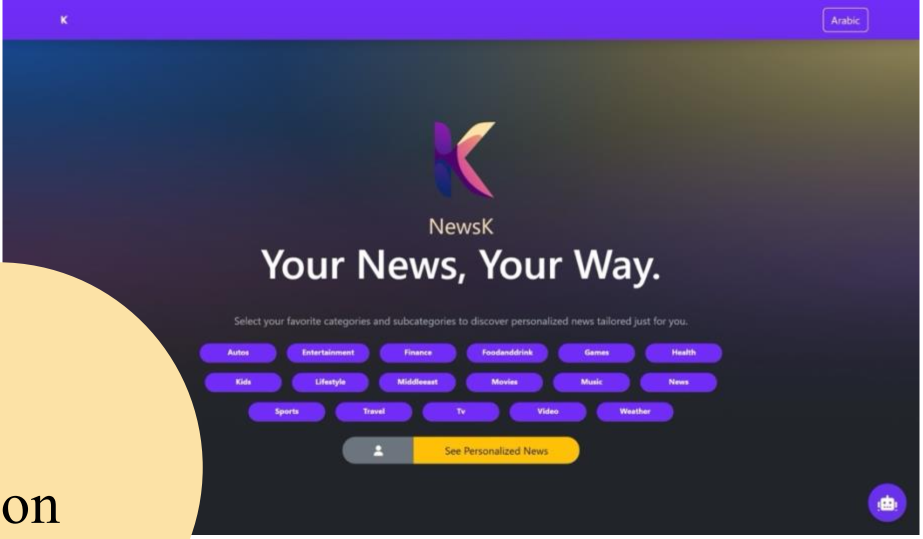
ArabBERT performed the best in the fine-grained category prediction, which handled all the intricacies of the Arabic text. The mT5 model has given very short and clear summaries with high semantic similarity scores; thus, it retained the key information. FastText provided an efficient and lightweight solution, which provided competitive accuracy in resource-constrained environments, succeeding in finding unique patterns within data. KMeans clustering provided relevant groupings but required refinement through keyword labeling for better accuracy.

Challenges Addressed

Data quality was enhanced by minimizing noise through preprocessing procedures to ensure a clear and consistent dataset. Class imbalance was partially addressed using sample balancing strategies and optimizing cluster sizes. For long-text summarization, articles were condensed into insightful summaries using the mT5 model, and the effectiveness of the generated summaries was validated through an assessment of semantic similarity.

Deployment

of the News Recommendation and Summarization System



NewsK

Your News, Your Way.

Select your favorite categories and subcategories to discover personalized news tailored just for you.

High-Level Understanding of the Structure

The system’s architecture ensures clear separation of responsibilities:

Data Handling

News data is stored in structured formats (TSV and CSV files) and processed using pandas Data Frames. The News Category class provides methods for grouping and filtering data, ensuring efficient management of news articles

Model Integration

The English Recommender uses pre-trained embeddings for content similarity, while the Arabic Summarization relies on a multilingual summarization model to handle Arabic content. Pre-trained models are loaded via joblib for optimized performance.

API Design

A Flask application acts as the interface for the system, exposing endpoints for user interaction. These endpoints include category retrieval, article recommendations, and summarization features, allowing seamless integration with user-facing applications.

Methodology

The system implements a modular and scalable architecture for news recommendation and summarization. It is structured into three primary components:

01 Arabic News Category Class

This component focuses on organizing and processing news data. It provides functionality to load datasets, group articles by categories and subcategories, and retrieve recent news articles based on user-selected criteria.

02 English Recommender Class

This module leverages pre-trained models, such as TF-IDF vectors and LLM embeddings, to recommend similar articles. It calculates content similarity using cosine similarity, enabling personalized recommendations.

03 Arabic Summarization Class

Focused on Arabic news, this component integrates a fine-tuned multilingual mT5 model for summarization. It also supports retrieval of articles by category or ID and generates concise summaries for news content.

The entire system is unified through a Flask-based web API, which exposes endpoints for retrieving categories, recommending articles, managing users, and summarizing content.

Integration of Models

The system effectively integrates pre-trained models to enhance its functionalities:

English Recommendations

TF-IDF vectors and LLM embeddings enable the English Recommender to provide personalized article suggestions based on cosine similarity metrics. This ensures that users receive highly relevant recommendations.

Arabic Summarization

The Arabic Recommender employs a multilingual mT5 model to summarize Arabic news articles. This capability improves content accessibility by providing concise summaries.

Data Management

Both models rely on structured datasets loaded into pandas Data Frames, ensuring consistency and ease of processing across different modules.

Challenges Addressed

The system addresses several key challenges effectively:

1 Efficient News Retrieval

Methods like get_recent_news and get_latest_arabic_news_by_category allow users to fetch relevant articles quickly based on categories, subcategories, or IDs.

2 Cross-Language Support

Separate recommenders for English and Arabic ensure adaptability to different languages and datasets, enabling the system to cater to diverse user bases.

3 Content Discovery and Summarization

By leveraging cosine similarity for English articles and a multilingual summarization model for Arabic content, the system enhances both discovery and accessibility.

4 API-Driven Interaction

The Flask application simplifies user interaction by exposing well-defined endpoints for various functionalities, including recommendations, summarization, and user management.

5 Scalability

The modular design of the system allows for easy addition of new datasets, features, or languages. The clear separation of data handling, model integration, and API functionalities ensures maintainability and scalability.

This comprehensive system demonstrates an effective approach to news recommendation and summarization. By combining modular design, robust model integration, and user-centric APIs, it delivers a powerful tool for personalized news discovery and cross-language content accessibility.

Section 6

Marketing

The marketing strategy for NewsK aims to position it as the premier News-as-a-Service (NaaS) solution, catering to the unique needs of bilingual individuals and corporate clients in the GCC region. This strategy focuses on highlighting the platform's advanced technological capabilities, personalized user experience, and significant business benefits.

The primary target audience for NewsK includes bilingual individuals who consume news in both Arabic and English, as well as corporate entities that require timely and relevant news to inform strategic decisions. This includes professionals in industries such as finance, technology, healthcare, and media, who need to stay updated on industry trends, risks, and opportunities.

NewsK's value proposition centers on its ability to reduce cognitive overload and decision fatigue by providing personalized, multilingual news recommendations. The platform's use of advanced machine learning and natural language processing technologies ensures high precision and relevance in news delivery. For businesses, NewsK offers the added advantage of aligning news content with strategic objectives, helping companies maintain a competitive edge and manage their public image effectively.

To reach a broad audience and engage with potential users, digital marketing will be a key component. Utilizing social media platforms such as LinkedIn, Twitter, and Facebook will help increase visibility through targeted ads, informative posts, and interactive content. Search engine optimization (SEO) and search engine marketing (SEM) will further enhance online presence and drive traffic to the NewsK website.

Content marketing will play a crucial role in educating the target audience about the benefits of NewsK. This will include creating blog posts, whitepapers, case studies, and webinars that showcase success stories and provide insights into industry trends and best practices. Email marketing campaigns will nurture leads and keep existing users informed about new features, updates, and industry news, with personalized content to build strong relationships and encourage engagement.

Forming strategic partnerships with industry influencers, news organizations, and professional associations will expand reach and credibility. Collaborations can include co-hosted webinars, guest blog posts, and joint marketing campaigns. Participating in industry events, conferences, and trade shows will provide opportunities to showcase NewsK and network with potential clients. Hosting workshops and speaking engagements will position NewsK as a thought leader in the news recommendation space.

Tracking key performance indicators (KPIs) such as website traffic, user engagement, conversion rates, and customer retention will measure the effectiveness of marketing efforts. Analytics tools will provide insights into user behavior and preferences, allowing for continuous optimization of marketing strategies.

Highlighting positive feedback and success stories from early adopters of NewsK will build trust and demonstrate the platform's value in real-world applications. Customer testimonials and detailed case studies will be essential in showcasing the impact of NewsK.

By implementing this comprehensive marketing strategy, NewsK can effectively reach its target audience, communicate its unique value proposition, and establish itself as the leading news recommendation platform in the GCC region. This approach will drive user acquisition, enhance brand awareness, and ultimately contribute to the platform's long-term success.

Section 7

Conclusion

The development and implementation of NewsK have demonstrated significant potential in addressing the challenges of information overload, particularly for bilingual individuals in the GCC region. By leveraging advanced machine learning and natural language processing technologies, NewsK provides personalized, multilingual news recommendations that enhance information consumption and support informed decision-making. The platform's hybrid recommendation engine, which combines collaborative and content-based filtering, has shown high precision in delivering relevant news articles tailored to user preferences. Metrics such as precision, recall, and user engagement rates indicate that NewsK effectively reduces cognitive overload and filters out misinformation, thereby mitigating decision fatigue.

The potential impact of NewsK on businesses is substantial. By aligning news content with strategic objectives, companies can stay informed of industry trends, risks, and opportunities, ensuring they remain competitive in a rapidly changing market. The platform's ability to manage public image and respond swiftly to negative news helps maintain investor confidence and supports proactive decision-making. Furthermore, the implementation viability of NewsK is reinforced by its scalable architecture, robust data processing capabilities, and adherence to quality standards and best practices. The integration of MLOps principles ensures continuous improvement and adaptability, making NewsK a reliable and future-proof solution for both individual users and corporate clients.

In summary, NewsK not only enhances the user experience by providing timely and relevant news recommendations but also offers significant business value by supporting strategic decision-making and maintaining a competitive edge. The platform's comprehensive approach to data processing, recommendation generation, and user interaction positions it as a state-of-the-art solution in the domain of news recommendation systems.

Ethics Statement

NewsK is committed to maintaining the highest standards of ethical conduct in data handling, AI implementation, and user privacy. In line with regulatory frameworks like GDPR and PDPL, NewsK emphasizes transparency, data security, and user consent in all its operations. The platform ensures user privacy by anonymizing data, minimizing data collection, and implementing secure access controls, guaranteeing responsible data handling.

The recommendation algorithms at NewsK follow guidelines for fairness and accuracy to prevent biases in content delivery. By integrating diverse news sources and conducting regular audits, NewsK aims to provide a fair and ethical platform for all users. Additionally, the platform's dedication to responsible AI supports sustainable and ethical technology use, benefiting both corporate and societal well-being.

References

We acknowledge the use of Generative AI. The prompts used include paraphrasing content and organizing the report. The outputs from these prompts were utilized to enhance the clarity, structure, and coherence of the technical report.

01 **GDPR 2018** General Data Protection Regulation (GDPR). (2018). Official Journal of the European Union. Retrieved from https://gdpr.eu

02 **PDLP 2021** Personal Data Protection Law (PDPL) – Saudi Arabia. (2021). https://sdaia.gov.sa/en/SDAIA/about/Documents/Personal%20Da

03 **Saudi Aramco** 2023 Sustainability Summary Report. Retrieved from https://www.aramco.com/en/sustainability/sustainability-report.

04 **MIND** Microsoft News Dataset for Recommendation Research. (n.d.). Retrieved from https://msnews.github.io/

05 **Saudi Newsnet** Saudi News Network for Arabic Content. (n.d.). Retrieved from https://github.com/inparallel/SaudiNewsNet

06 **n.d** Vue.js Documentation. (n.d.). Retrieved from https://vuejs.org/

07 **BERT 2018** Bidirectional Encoder Representations from Transformers. (2018). Retrieved from https://arxiv.org/abs/1810.04805

08 **ARA BERT 2020** Pre-trained Arabic Language Model. (2020). Retrieved from https://arxiv.org/abs/2003.00104

09 **n.d** Kubernetes Documentation. (n.d.). Retrieved from https://kubernetes.io/docs/

10 **n.d** MLflow Documentation. (n.d.). Retrieved from https://mlflow.org/