ie University Microsoft SOKONEWS MEDIA

Ai-Driven News Recommenders

Completed by:

***Hugo** A. Borjórquez Gallardo, **Sofia** Depoortere, **Filippo** Lisanti,*

***Fernando** Moreno Borrego, **Christopher** Stephan*

Master in Business Analytics & Big Data **March 2025**

ETHICS STATEMENT

This capstone project adheres to strict ethical guidelines to ensure confidentiality, transparency, integrity, and fairness throughout all phases of our research and collaboration. We recognize the importance of ethical responsibility in developing AI-driven personalization systems, particularly in the context of news recommendations, where trust, neutrality, and user privacy are critical. Our ethical commitments include:

- Confidentiality: We will safeguard any proprietary, sensitive, or personally identifiable information provided by the company or collected during the project. No confidential data will be shared, stored, or used beyond the agreed-upon scope.
- Compliance: We will strictly adhere to all relevant laws, regulations, and institutional policies, as well as best practices in data protection. We will ensure that all data handling, model training, and deployment practices align with legal and ethical requirements.
- Integrity: All research, analysis, and reporting will be conducted honestly, transparently, and without bias. We are committed to ensuring that our recommendations and insights are evidence-based, free from manipulation, and ethically sound.
- Respect for Stakeholders: We will treat all stakeholders—including users, business partners, and regulatory bodies—with fairness, professionalism, and respect. Our goal is to produce outcomes that benefit both the company and the broader community, while minimizing potential harms such as algorithmic bias, misinformation, or filter bubbles.

By embedding these ethical principles into the core of our research and system design, we establish a clear framework for responsible AI development. This explicit commitment not only guides our decision-making at every stage but also assures stakeholders of the integrity, fairness, and accountability of our project.

LINK TO GITHUB REPOSITORY

The following is the link to the github repository where you can find supplementary information, including:

- Technical notebook implementations
- Report Documents, such as the annex, supporting documents, and diagrams.

https://github.com/Christopher-Stephan/Microsoft_Capstone_Project

The datasets used for model implementations can be found in the following link:
https://msnews.github.io/

Processed datasets stored in .parquet can be found in the following google drive link:
https://drive.google.com/drive/folders/1JU3C_SzxmmjvE2jJ2Hw_Ab0-nR-HaE6k?usp=sharing

If you have any questions or inquiries regarding this, do not hesitate to contact us at:

- christopher.stephan@student.ie.edu

TABLE OF CONTENTS

PROBLEM STATEMENT **1**
INTRODUCTION **1**
HYPOTHESIS FORMULATION **2**
DATA SOURCES **3**
METHODOLOGY **4**
EXPLORATORY DATA ANALYSIS **5**
DATA SCIENCE IMPLEMENTATION **6**
Baseline and Target Models 6
Feature Engineering and Selection 7
Model Training, Validation, and Tuning 7
Performance Metrics and Improvement 7
REVENUE IMPACT **8**
Advertisement Revenue 8
Subscription Revenue 9
INVESTMENT **10**
Cloud Infrastructure 10
AI Model Training 10
Personnel 11
Compliance and Security 11
ROI **12**
BUSINESS KPIs & STREAMLIT **12**
PRODUCTION LEVEL CONSIDERATIONS **13**
End-to-End Solution Architecture for Production Readiness (Low-Level Architecture) 13
Scalability, Deployment, Monitoring, and Maintenance 14
Legal, Privacy, and Intellectual Property (IP) Considerations 14
Pathways for Building an In-House Dataset for Long-Term Sustainability 14
Multi-Pathways to Serve Different Business Units 14
Governance, Bias, and Fairness 15
Security and Optimization 15
Future Costs of Entire Cloud Infrastructure 15
CONCLUSION **15**
BIBLIOGRAPHY **17**

PROBLEM STATEMENT

Personalizing news content to increase user engagement is a critical challenge for SokoNews as it competes in a rapidly evolving digital news landscape. As a growing news agency, SokoNews is looking to implement a data-driven recommendation system that can dynamically adapt to user preferences and optimize content delivery. However, due to internal constraints, there is currently no available in-house dataset to support the development of such a system. Therefore, our goal is to explore external datasets, design and evaluate different recommendation methodologies, and build a working proof-of-concept system that demonstrates clear value. This system should align with our mission to increase user engagement through personalized content, while also considering ease of deployment, scalability, and maintenance within our organization.

Furthermore, personalizing news content presents unique challenges distinguishing it from other recommendation domains such as movies or e-commerce. One of the most significant issues is the cold-start problem, which affects both users and articles (Wu et al., 2020). New users without prior interaction history pose difficulties in tailoring relevant suggestions, while newly published news—often the majority of platform content—cannot be recommended based on historical behavior due to their recent nature. Additionally, news articles have extremely short life cycles, with 80% achieving clicks within the first 24 hours, making traditional collaborative filtering or ID-based approaches ineffective (Neary, 2023). Personalization efforts must rely on deep content understanding, leveraging textual elements like titles, abstracts, and entities, and sophisticated modeling of user interest that adapts to rapidly evolving trends. Addressing these issues is crucial for SokoNews to remain competitive in delivering timely and relevant news content to a broad user base.

Personalized news recommendation differs significantly from other domains due to a combination of rapid content turnover, cold-start issues, implicit feedback mechanisms, and dynamic user interests. Unlike movies or products, news articles have fleeting relevance, making it difficult to build stable user-item matrices. The reliance on implicit feedback, such as clicks, introduces ambiguity—a click may reflect curiosity, disagreement, or mere headline attractiveness, complicating the interpretation of user preferences. Furthermore, the diverse nature of user interests, often spanning multiple topics and shifting daily, requires recommendation models to be highly adaptive and context-aware. Addressing these challenges necessitates a combination of real-time content understanding, advanced user modeling, and robust privacy and fairness frameworks, all of which are central to SokoNews's future strategy for scalable and ethical news personalization.

INTRODUCTION

Though the primary objective of our work at SokoNews is to build a technologically sound system capable of delivering personalized news recommendations, it is equally important to assess its business impact. The success of such a system would ultimately be measured not only by its ability to enhance user engagement but also by its ability to generate tangible financial benefits and drive long-term sustainable bottom-line growth.

At the heart of our analysis was the question of *how a news recommender system could translate into bottom-line impact*. Extensive research into similar implementations by major online newspapers, such as *The New York Times* and *The Financial Times*, provided valuable insights. The introduction of personalized recommendation systems in these organizations had led to significant improvements in key performance indicators. These included an increase in conversion rates from casual readers to paying subscribers, a rise in click-through

rates (CTR) that directly boosted advertisement revenue, an overall increase in platform traffic, and improvements in retention and user engagement. For example, *The Financial Times*' 'MyFT' feature, which allows subscribers to follow topics of interest and receive tailored content, led to an average 86% increase in engagement levels (FT Strategies, 2023). Furthermore, Scandinavian local newspaper, iTromsø, witnessed a 23% increase in unique visitors within 3 months of implementing news recommender systems (Gulla, 2017).

In order to quantify the financial impact of our own implementation at SokoNews, it was essential to break down the components of our analysis. Two primary factors were considered. The first was the increase in revenue, which was expected to be driven by higher advertising earnings and a rise in subscription revenue. The second was the cost of investment associated with developing and maintaining the recommender system. By focusing on these two elements, we calculated the return on investment (ROI) using an incremental approach, meaning that we isolated the financial impact of the recommender system without taking into account pre-existing revenues and costs. The formula for this calculation involved subtracting the new costs from the additional revenues generated and dividing this result by the total investment.

With this framework in place, we began analyzing the impact of the recommender system on revenue generation, focusing on two key sources: ad revenue and subscription revenue, which will be discussed further after understanding the technical implementations of our news recommender.

HYPOTHESIS FORMULATION

To address the challenge of personalizing news recommendations, SokoNews has identified several analytical approaches that represent viable paths for exploration. One hypothesis is that article titles are the most influential component in driving user engagement, which suggests that content-based models focusing on title similarity could generate effective recommendations. Another approach considers that users are more likely to engage with articles containing specific named entities, implying that entity-level similarity and embeddings could enhance relevance. Additionally, we hypothesize that users prefer content aligned with their historical reading preferences, supporting collaborative filtering or hybrid methods that combine both user history and article features. A further direction is the exploration of recency-based preferences, recognizing that users may favor more recently published news articles over older ones. Finally, category-based filtering, where users are recommended articles from their most frequently read categories, offers a simple, interpretable baseline that may provide quick value. These hypotheses form the foundation for SokoNews's technical exploration of personalized news recommendation systems.

Our analysis will primarily focus on a subset of these five hypotheses, tackled through the combined use of BERT-based content similarity models and CTR prediction models. The BERT-based recommender directly addresses the hypothesis that "*article titles are the most influential component in driving user engagement*", as it builds user profiles based on previously read content and recommends new articles primarily using title and abstract similarity. By leveraging textual information, including categories and subcategories, the BERT model also partially touches on the idea that users prefer content aligned with their historical reading preferences, albeit from a content (rather than behavioral) perspective.

In parallel, the CTR prediction model is designed to address multiple hypotheses. First, it contributes to testing "*users are more likely to engage with articles containing specific named*

entities", by incorporating entity similarity between users' reading history and new articles. Second, it directly targets "*users prefer content aligned with their historical reading preferences*", by including features derived from user interaction history such as categories, subcategories, and click patterns. Third, the CTR model also allows us to explore *recency-based preferences*, as it includes temporal features that assess how the timing of article publication may influence user clicks.

Our primary objective will be operationalized through the BERT and CTR models. This focused approach enables us to explore both content-based and interaction-based personalization strategies, and we will provide further analysis and validation of these models in the following sections.

DATA SOURCES

To explore these hypotheses, SokoNews is leveraging the Microsoft MIND dataset (Microsoft News Dataset), a large-scale dataset designed for news recommendation research. This dataset includes behavioral logs of over 1 million users, more than 160,000 news articles, and approximately 15 million user interactions collected over a six-week period. It is structured around two key files: behaviors.tsv, which captures detailed user interactions and impressions, and news.tsv, which provides rich metadata about each article, including titles, abstracts, categories, subcategories, and entity embeddings. The MIND dataset offers significant technical value as it contains a wide range of content and user interaction features necessary for training and testing different recommendation models, including entity-based, content-based, and collaborative filtering approaches.

However, the dataset presents some limitations. Its fixed time window of six weeks restricts longitudinal analysis, and as a dataset collected from Microsoft's ecosystem, it may reflect biases specific to that audience, which may not fully align with SokoNews's target demographic. The lack of in-house data also raises challenges in tailoring the models to SokoNews's unique user base. Nonetheless, MIND serves as an essential starting point to develop and validate proof-of-concept models.

Currently, there are no alternative datasets integrated into our analysis, but future work will likely involve developing an internal SokoNews dataset to better capture the specific behaviors, preferences, and linguistic patterns of our own user base. Such a dataset would allow us to fine-tune and validate models in a way that reflects SokoNews's unique audience, potentially improving both recommendation quality and business relevance. This process would require designing robust data collection pipelines, including mechanisms to log user interactions, preferences, and feedback in a privacy-compliant manner.

A critical challenge inherent to the MIND dataset—and to news personalization at large—is the cold-start issue for both users and news articles. The dataset itself demonstrates that most news items have a lifespan of less than two days, meaning new content continuously flows into the system without sufficient interaction data to inform recommendations. This underlines the need for models that can leverage textual and semantic content immediately upon article release, rather than waiting for behavioral signals to accumulate. Additionally, as news items are composed of various textual elements such as titles, abstracts, bodies, categories, and entities, effective recommendation requires multi-view representation learning to integrate these components holistically. While MIND provides rich metadata and entity embeddings, the absence of images, videos, and broader user signals (such as

time-on-page or scrolling behavior) represents a limitation, emphasizing the future need for multi-modal and behavioral data capture in SokoNews's internal datasets (Wu et al., 2020).

In terms of legal and intellectual property (IP) considerations, the MIND dataset is publicly released for research purposes under Microsoft's terms and conditions, and we are careful to ensure that our usage aligns with these guidelines. Any models or insights derived from this dataset remain within the boundaries of acceptable research use. When SokoNews begins to collect its own user data, strict adherence to data protection regulations, such as GDPR or local Sokovian privacy laws, will be essential. This includes ensuring that users provide explicit consent for data collection, maintaining transparent communication about how their data is used, and implementing technical safeguards to protect user information. Additionally, internally generated datasets and resulting models would become part of SokoNews's proprietary assets, giving the organization full control over their use and commercialization while carrying responsibility for their ethical management.

METHODOLOGY

To develop the SokoNews personalized news recommender, we first implemented preprocessing pipelines and models locally using Python, Jupyter notebooks, and data science libraries like Pandas, NumPy, and Scikit-learn. Our initial modeling efforts focused on Click-Through Rate (CTR) prediction using LightGBM and Logistic Regression, along with a content-based recommender leveraging TF-IDF vectorization over article texts. Natural Language Processing (NLP) tasks—tokenization, stopword removal, and lemmatization—were handled via NLTK, while NumPy was used for entity and relation embeddings. This local prototype facilitated rapid iteration on data understanding and initial model design.

Once validated, we plan to transition the pipeline to Azure for scalability and production-readiness. Our cloud-based setup will utilize Azure Blob Storage, Databricks, and Azure ML. Blob Storage will serve as the central data repository for raw and processed datasets (in Parquet format), while Databricks will handle both preprocessing and model training.

Preprocessing included handling missing values, deduplication, and time-based dataset splitting for realistic modeling. Data transformations involved timestamp conversions, session segmentation, and user-ID mapping for efficient processing.

For feature engineering, we designed user-level statistics (e.g., history length, average clicks per user, recency of interaction) and content-specific features (e.g., word count, category frequency, entity similarity scores). We also incorporated pre-trained Wikidata embeddings to capture semantic relationships between articles and user preferences.

The pipeline was built to be modular and reproducible. Raw MIND data was cleaned, behavior and news datasets were aligned, and enriched features were stored in Parquet format for downstream use. Key assumptions included treating missing user history as "No_History" and using Wikidata IDs for entity alignment.

Locally, we developed a full end-to-end pipeline using .ipynb & .py notebooks, from data ingestion to serving via Streamlit. On Azure, training will run on Databricks, with models stored in Azure ML Model Registry for traceability and versioning. The final system will be deployed as a Dockerized Streamlit app, dynamically querying Parquet datasets and trained models. This ensures a scalable, modular, and maintainable architecture.

Given the short lifespan of news content, our methodology emphasizes content-driven personalization via BERT-like approaches and CTR modeling. By leveraging article titles, abstracts, and entities, we mitigate the item cold-start problem, allowing recommendations even for newly published articles. Our CTR models balance long-term preferences with short-term session-based interests, capturing the dynamic nature of news consumption.

To address the user cold-start problem, we integrate an LLM-powered questionnaire in Streamlit, enabling preference elicitation for new users. This approach ensures personalized recommendations from the first interaction, enhancing user engagement.

EXPLORATORY DATA ANALYSIS

The exploratory data analysis paints a rich picture of how users interact with news content on the platform, revealing important dynamics around content design, user behavior, and engagement patterns. News titles and abstracts are crafted to be short, impactful, and easy to consume, with the majority of abstracts under 100 words and titles dominated by high-frequency, attention-grabbing words such as "Trump," "win," "game," "week," and "new." These frequent mentions of current events, trending figures, and sports terms suggest a strong emphasis on timeliness and relevance in content creation. Interestingly, titles vary across categories, with words like "team," "game," and "season" dominating in sports, while finance and travel categories focus on terms like "year," "home," and "city." Even lifestyle and news categories feature their own signature keywords, pointing to a targeted approach in language that resonates with specific audiences.

When looking at user engagement patterns, there's a clear concentration of activity around specific times of day, with sessions peaking between 8 AM and 1 PM, and a notable drop in the evenings and late nights. This indicates that news consumption is likely tied to daily routines, perhaps during commutes or work breaks. Additionally, weekdays, especially Monday through Wednesday, see significantly higher user activity compared to weekends, underscoring that news reading may be more habitual during the workweek, or simply it is a byproduct of limitation in the dataset.

User behavior within sessions also reveals a preference for focused, quick interactions. The vast majority of sessions involve just one or two clicks, and click history lengths are generally very short, indicating that users are not spending long periods browsing but are instead coming with specific intentions—perhaps seeking updates on topics of immediate interest. Supporting this, the distribution of number of clicks per session is heavily skewed, with few users engaging in longer browsing.

In terms of content complexity, abstracts across categories show a balanced structure, where word counts, character counts, and punctuation counts remain within moderate ranges. Unique word counts and stop word counts suggest that while abstracts are concise, they still maintain variety and natural language flow, avoiding overly repetitive phrasing. Furthermore, sentiment analysis of abstracts reveals that most content is neutral to mildly positive, but some categories like food, entertainment, and lifestyle exhibit higher positivity, whereas others like news and North America-related content tend to be more neutral or serious in tone.

The analysis of entity mentions within titles provides another layer of insight. "Donald Trump," "United States," and "National Football League" are among the top entities, indicating that politics and sports are major drivers of attention. Notably, seasonal events like

"Halloween" also appear frequently, reflecting the role of timely, culturally relevant events in drawing user interest. At the same time, categories such as North America and Middle East contain a higher average number of entities per article, signaling that some topics require richer context and are more information-dense.

Overall, the data reflects a platform optimized for quick, routine-based interactions, with content designed to be highly relevant, concise, and emotionally attuned to the needs of each category's audience. Whether through short, engaging titles, neutral or positive abstracts, or content shaped around top entities and trends, the platform seems to align well with user preferences for accessible, targeted, and timely news consumption.

For more insights and details on all the EDA results, refer to Figures 1-23 in the annex attached.

DATA SCIENCE IMPLEMENTATION

In this section, we outline the modeling approaches and system design behind our personalized news recommender. We describe how we developed and optimized three complementary models — CTR prediction, content-based (BERT-like), and LLM-driven recommendations — each addressing different user needs. We also explain how these models were trained, validated, and integrated into our interactive Streamlit application, with a focus on improving recommendation quality, user engagement, and business impact.

Baseline and Target Models

To address the task of personalized news recommendations, we developed three complementary solutions: a CTR prediction model, a content-based recommender (BERT-like model), and a lightweight LLM-based recommender embedded within our Streamlit app. Each approach serves a distinct role in fulfilling diverse user needs. The CTR model is our primary target model, designed to predict the likelihood that a user will click on a given news article based on a wide set of behavioral and content-based features. As a starting point, we used Logistic Regression as a baseline, given its simplicity and interpretability, but its linear nature quickly proved insufficient for capturing the complex, non-linear interactions inherent in user preferences and article content.

To overcome this, we shifted to LightGBM, a gradient boosting framework well-known for its efficiency and effectiveness on large, structured datasets. LightGBM was chosen for its ability to model sophisticated feature interactions, handle categorical variables natively, and deliver fast training and prediction speeds—critical for real-time recommendation scenarios. Alongside CTR prediction, we implemented a content-based recommender, referred to as BERT. This content-based approach generates personalized suggestions based on the similarity of article content to a user's reading history, making it a valuable fallback when explicit click data is sparse. Finally, to cover another dimension of user interaction, we introduced a third approach: an LLM-powered recommender in the Streamlit application. Unlike CTR and BERT models that rely on user history, this LLM-based system engages users through a conversational questionnaire—allowing them to express their interests directly, and generating recommendations based on their responses. This allows for cold-start personalization where neither behavior nor history is available, completing a well-rounded recommendation ecosystem.

Feature Engineering and Selection

Building an effective CTR model required substantial feature engineering and thoughtful selection of inputs that capture both user preferences and article characteristics. We engineered a suite of features including entity vector similarity (to assess topical alignment between user and article), category and subcategory matches, article content word count, and user behavioral patterns such as history length, number of past clicks, recency of activity, and user-specific average clicks. These features enable the CTR model to personalize recommendations at the individual level, reflecting both what the user typically engages with and how they consume content over time.

Additionally, contextual signals like hour of the day were included to capture temporal reading patterns, which can influence click likelihood. Categorical fields, such as combined category-subcategory labels, were encoded for modeling in LightGBM. While CTR relies heavily on these explicit, structured features, the BERT recommender focuses on article text embeddings, emphasizing semantic similarity over behavioral patterns, making it well-suited for users with limited interaction history. Meanwhile, our LLM recommender requires no feature engineering from structured data—instead, it transforms user preferences collected via natural language questions into tailored article selections, thus filling the gap for first-time or low-activity users.

Model Training, Validation, and Tuning

Considering the scale and complexity of the dataset, we strategically limited training to a 100,000-row sample to ensure efficient experimentation while still preserving rich interaction patterns. We began by training a Logistic Regression baseline, primarily to establish a reference point. However, the model struggled to handle non-linearities inherent in user-item interactions. Consequently, we focused on LightGBM, leveraging its ability to capture complex patterns and interactions efficiently.

We performed grid search hyperparameter tuning over parameters such as num_leaves, max_depth, learning_rate, n_estimators, and min_child_samples, optimizing for AUC performance. This meticulous tuning led to an optimized LightGBM model with balanced complexity and strong generalization, supporting its use for real-time inference in production. In parallel, the BERT recommender required no supervised training—it simply computes TF-IDF-based content similarity on demand, making it an inherently fast and lightweight model for content-driven suggestions. The LLM recommender, designed for conversational interaction, similarly avoids heavy pre-training, instead relying on dynamically collected user inputs to query relevant articles, offering an immediate and flexible user experience.

Performance Metrics and Improvement

To evaluate our CTR model rigorously, we used a comprehensive suite of classification and ranking metrics, focusing on AUC, precision, recall, MAP, and nDCG. These metrics were carefully chosen to reflect both the accuracy of click predictions and the quality of ranked article lists—a critical concern in news recommendation where showing the right articles at the top is key to engagement. AUC served as a measure of the model's overall ability to distinguish between clicked and non-clicked articles, while precision and recall evaluated the trade-offs between relevance and coverage of the recommendations. Most importantly, MAP and nDCG (including nDCG@5 and nDCG@10) were used to assess ranking quality, directly corresponding to user satisfaction since users often interact with only the top few articles recommended.

The LightGBM model outperformed Logistic Regression across all these metrics, with AUC improving from 0.6185 (Logistic) to 0.7574 (LightGBM), and nDCG@10 rising to 0.6348, indicating better-ranked, more relevant recommendations. Refer to Figures 24-26 in the annex for overall metric analysis on these models. These improvements highlight LightGBM's superiority in capturing the complex factors that influence clicks. The BERT recommender, while not evaluated on these metrics due to its unsupervised nature, is optimized for textual relevance, offering strong semantic matches when behavioral signals are absent. Lastly, the LLM-based recommender focuses on personalization through direct user input, ensuring that even in cases where no historical data exists, users receive targeted and relevant article suggestions based on their stated interests—a key element for new user acquisition and satisfaction.

REVENUE IMPACT

The next step in our analysis is to understand how these models impact our revenue model. The first major component of revenue growth was advertising. Online newspapers typically earn ad revenue through two main models. The first is the Cost Per Mille (CPM) model, in which revenue is generated based on the number of times an ad is viewed by users. The second is the Cost Per Click (CPC) model, where revenue is earned when users actively click on advertisements. To determine the extent to which the recommender system would contribute to these ad revenues, we first needed to estimate the total number of impressions on the platform.

Given the constraints of our dataset, which only covered a limited timeframe from November 9 to November 22, 2019, we had to extrapolate the available data to project an annualized figure. The dataset included a total of 4,979,946 recorded impressions over a span of 13 days. Since this period represented approximately 3.56% of the full calendar year, we scaled up the numbers accordingly and estimated that the total annual impressions would amount to approximately 139,821,561.

Beyond simply understanding the total number of impressions, we also needed to assess how the introduction of a recommender system, derived from the models we built, would enhance this figure. Based on benchmark studies from major media organizations, we found that personalized recommendation systems could drive an increase of 50% in total impressions (Mediahaus, 2021). Applying this projection to our dataset, we estimated that the additional impressions generated purely as a result of the recommender system would amount to approximately 69,910,780 (Table 1), bringing the total number of impressions to nearly 210 million.

Advertisement Revenue

With this increase in impressions, we then calculated the corresponding rise in ad revenue. For the CPM model, one of the critical factors in determining revenue was the click-through rate (CTR), which represents the percentage of impressions that result in users clicking on an article. Before implementing the recommender system, the CTR in our dataset stood at 5.6% (impressions equal to one divided by total impressions). However, based on performance benchmarks from other news platforms, we expected this figure to rise to 20% following the introduction of personalization (Mediahaus, 2021). This resulted in a CTR increase of approximately 14 percentage points. Additionally, we assumed that each article contained an average of five ads and that the estimated effective CPM (eCPM), which represents the revenue earned per thousand impressions, would be approximately €5 (Top Draw, 2025).

After accounting for these factors, we calculated that the additional revenue generated from the CPM model alone would amount to approximately €755,036 in year 1 (Table 2).

The CPC model, on the other hand, required a slightly different approach to revenue estimation. In this model, revenue is based on the number of ads clicked rather than simply viewed. The key factors influencing this calculation were the increase in the article click-through rate, which we had already determined to be 14 percentage points, the number of ads per page, which remained at five, and the difference in ad click-through rate (Ad CTR), which we set at 1% based on industry benchmarks (Chaffey, 2024). Additionally, we determined that the average cost per click (CPC) paid by advertisers, which directly contributes to our revenue, is €0.60 (Dogtiev, 2025). Applying these values, we calculated that the increase in CPC-based ad revenue would be approximately €906,043 in year 1 (Table 3).

When combining the results of both ad revenue models, we determined that the recommender system would lead to a total increase in advertising revenue of €1,661,080 in its first year, with further growth expected in subsequent years.

Subscription Revenue

In addition to advertising revenue, the second major component of revenue growth is subscription revenue. A well-implemented recommendation system can drive higher subscription revenue in two primary ways. First, it can increase the conversion rate of casual readers into paying subscribers. Second, it can help reduce churn by improving engagement and retention, thereby ensuring that a higher percentage of existing subscribers remain active over time.

To estimate the impact of the recommender system on subscriptions, we first needed to determine the total number of users on the platform. Our dataset indicated that over the 13-day period, there were 876,956 unique users. Since approximately 40% of these users remained active from the first day recorded, to the last, we extrapolated these figures to an annual timeframe. This led us to estimate that the platform has 15,124,118 unique users per year (Table 4).

Based on industry research, the introduction of a recommender system is expected to double the conversion rate from 2% to 4%. Applying this increase to the estimated total user base, we determined that an additional 302,482 users would convert into paying subscribers as a direct result of the system. Given that the annual subscription fee was set at €120 per user, this translated into an additional subscription revenue of €36,297,840.

Another factor influencing subscription revenue was the reduction in churn rate. Prior to implementing the recommender system, the industry standard churn rate stood at approximately 10%. However, research suggested that the system could reduce this figure by half, bringing it down to 5%. This improvement meant that an additional 15,124 subscribers would be retained. With each retained subscriber generating €120 in annual revenue, this resulted in an additional revenue gain of €1,814,894.

When combining the revenue gains from new subscriptions and reduced churn, we estimated that the recommender system would lead to an overall increase in subscription revenue of approximately €38,112,734 in its first year (Table 5).

In total, after accounting for both advertising and subscription revenue, the recommender system is projected to generate an additional €39,773,814 in revenue in its first year. These figures highlight the financial potential of implementing a personalized news recommender system, demonstrating a substantial boost in both advertising and subscription revenue (Table 6).

However, while these revenue projections are promising, it is equally crucial to evaluate the investment required to develop, deploy, and maintain such a system to determine its overall profitability and long-term viability.

INVESTMENT

The implementation of the personalized news recommender system at SokoNews requires a carefully structured investment in cloud infrastructure, model training, personnel, and compliance. These expenditures are essential to ensure operational efficiency, scalability, and regulatory compliance, while delivering the expected increase in user engagement and revenue generation outlined in the business case.

The total cost for the initial implementation of the system amounts to €3,462,944, covering all the necessary technological and operational expenses. This number is based on the current volume of impressions that SokoNews receives (139,821,561 yearly impressions). Below, we outline the reasoning behind each investment category and its impact on the successful deployment of the system.

Cloud Infrastructure

Cloud infrastructure forms the backbone of the recommender system, providing compute power, storage, and networking capabilities necessary for real-time inference and content personalization. Given the high frequency of user interactions and real-time nature of news recommendations, a cloud-based solution ensures that the system can scale efficiently without requiring large upfront capital expenditures on hardware. Serverless computing, estimated at €559 per year, ensures that resources are allocated only when needed, reducing idle costs while maintaining high availability for recommendation requests (Simsek, 2025). Dedicated virtual machines, costing €1,398 per year, provide additional computational power for handling sustained workloads, particularly for batch processing and data transformation (Azure). GPU instances, with an annual cost of €26,806, are essential for deep learning models, enabling fast and efficient inference for generating personalized recommendations. Storage costs of €470 per year account for the large volumes of interaction data, user preferences, and recommendation logs that need to be maintained for historical data tracking and model retraining. Networking and API costs, estimated at €1,150 per year, ensure seamless communication between data sources, machine learning models, and the application interface, covering expenses for data transfers, API calls, and security protocols. Given the high frequency of user interactions and the real-time nature of news recommendations, a cloud-based solution ensures that the system can scale efficiently without requiring large upfront capital expenditures on hardware. This approach also minimizes maintenance costs, as cloud providers handle software updates, security, and operational uptime. The total estimate amounts to €30,383/year. (Table 7)

AI Model Training

AI model training plays a critical role in optimizing recommendation accuracy and adapting to user preferences, accounting for a significant investment of €1,805,560 (Antaris et al., 2020). GPU training costs, amounting to €25,560 per year, support continuous fine-tuning

and optimization of models, allowing the system to adapt to evolving user preferences (Aherne, 2023). Prototype development, budgeted at €60,000, enables the testing of different recommendation algorithms and the evaluation of key performance indicators such as click-through rates, engagement metrics, and response times. The minimum viable product (MVP) deployment phase, requiring €120,000, allows for validation of real-world performance and the collection of early user feedback (Zealousys, 2024). Full-scale deployment, which costs €400,000, ensures the successful integration of the recommender system into the SokoNews platform, optimizing robustness, fault tolerance, and performance (Verma, 2024). Continuous model retraining, an essential part of maintaining recommendation quality, accounts for €1,200,000 to ensure that entity embeddings, CTR models, and content similarity rankings remain up to date as user behaviors and news trends evolve (Artificial Intelligence Cost Estimation: Key Factors & Examples). The total amounts to €1,805,560. (Table 8)

Personnel

Personnel costs represent another major expenditure, with a total annual allocation of €1,450,000 to support the team responsible for developing, deploying, and maintaining the recommender system (Glassdoor). Data scientists, with a combined salary cost of €500,000 per year, focus on feature engineering, model selection, and evaluation to ensure accurate user preference predictions. Machine learning engineers, whose salaries total €440,000 annually, oversee model deployment, optimization, and scalability, ensuring smooth system performance in production. Data engineers, with an annual expenditure of €270,000, are responsible for data ingestion, storage, and transformation, ensuring structured and accessible datasets for machine learning pipelines (Indeed). Cloud and DevOps engineers, with a total cost of €240,000 per year, ensure that cloud infrastructure is reliable, automate model deployment, and monitor system performance to prevent failures and inefficiencies. Developing an AI-powered recommendation engine requires continuous monitoring and optimization, and this specialized team ensures that the system remains operational, scalable, and responsive to business needs. The total amounts to €1,450,000/year. (Table 9)

Compliance and Security

Compliance and security investments are crucial to ensuring regulatory adherence and protecting user data. With a total annual cost of €177,000, these measures include GDPR initial compliance, which costs €30,000 and ensures that data collection meets European privacy regulations while giving users control over their personal information. Annual GDPR maintenance, requiring €20,000, supports audits and updates to maintain ongoing regulatory compliance. AI Act compliance, budgeted at €52,000, ensures that the system meets emerging AI regulations, prioritizing fairness, transparency, and explainability in recommendations. Cybersecurity implementation, which accounts for €75,000, covers encryption, monitoring, and risk assessment protocols to prevent unauthorized access, cyber threats, and data breaches. Failing to comply with data privacy and AI regulations could result in legal risks, reputational damage, and financial penalties, making this investment essential for safeguarding the system while reinforcing user trust in personalized recommendations. The total amounts to €177,000. (Table 10)

As SokoNews anticipates a 50% increase in impressions (209 million), the expansion of the recommender system will require additional resources to support higher traffic volumes and computational workloads. The projected post-expansion cost of €4,325,382 reflects an increase in cloud infrastructure costs, which will rise from €30,383 to approximately €32,041 due to higher GPU usage and storage capability (approx. 19.46 terabytes). AI model training

costs will scale from €1,805,560 to €2,298,340, accounting for increased retraining frequency and larger datasets. Personnel expenses will increase proportionally, reaching €1,750,000 to accommodate additional engineering and data science support for system scalability. Compliance and security costs will also rise from €177,000 to €245,000, ensuring continued adherence to GDPR, AI Act requirements, and cybersecurity standards as data volumes grow (Irwin, 2024).(Tables 11-14)

ROI

Following the development of both revenue and investment models, we conducted an assessment of the Return on Investment (ROI) for our proposed recommender system. The ROI was calculated using the following equation:

$$\text{ROI} = \frac{\text{Revenue from Recommender System} - \text{Investment Cost}}{\text{Investment Cost}}$$

In the first year, the system is projected to generate a remarkable ROI of 8.2, highlighting its strong financial viability. To ensure a more comprehensive analysis, we extended our ROI projections into the second and third years, factoring in anticipated changes in cost structure and revenue growth.

Assuming, based on industry trends, a 10% revenue growth rate in the second year, we projected an ROI of 8.64, a slight increase despite the increased costs associated with scaling and new technological enhancements. In the third year, with a 20% revenue growth rate, the ROI is expected to climb to 9.51, reinforcing the system's long-term profitability (Table 15), which when benchmarked against industry standards, these figures indicate that the recommender system delivers a market-leading ROI, proving not only its financial feasibility but also its strategic necessity for sustaining and enhancing SokoNews' competitiveness.

BUSINESS KPIs & STREAMLIT

With strong financial projections and a compelling ROI, the next step is ensuring that our technical implementations directly support business goals. The recommender system is designed to maximize user engagement, increase CTR, and extend session durations—all of which directly influence ad revenue, subscription retention, and overall platform profitability.

Our CTR prediction model, powered by LightGBM, enhances the likelihood of users interacting with recommended articles, sustaining engagement over time. Meanwhile, the BERT-based content model ensures that even users with little to no history receive relevant recommendations, tackling the common cold-start problem. The LLM-powered recommendation system further enriches personalization by allowing users to explicitly articulate their preferences through an interactive, conversational interface in Streamlit.

All three models are seamlessly integrated into the Streamlit application, providing users with flexible recommendation options—CTR-driven, content-based, or conversational. The system also offers transparency, explaining why specific articles are recommended based on user preferences (e.g., categories, entities). This builds trust and engagement, helping users feel in control of their news feed while aligning with ethical AI principles to mitigate bias, misinformation, and filter bubbles.

By combining behavioral learning (CTR), content-based filtering (BERT), and conversational LLM recommendations, we ensure a robust and adaptable personalization system. The Streamlit interface not only makes these models accessible and interactive but also aligns them with real-world user workflows, translating directly into higher engagement, improved retention, and increased revenue. This scalable foundation positions SokoNews for continuous growth, real-time updates, and evolving user expectations in the dynamic news industry.

PRODUCTION LEVEL CONSIDERATIONS

Our current simple architecture provides a clear pathway from data ingestion to serving personalized news recommendations through Streamlit. In this pipeline, semi-structured files such as .vec and .tsv from Google Drive and GitHub are first ingested into Azure Blob Storage. From there, data is processed in batches using Azure Databricks, where all feature engineering, model training, and evaluation are performed. After processing, these enriched datasets flow directly into Streamlit, where they power an interactive recommendation engine for users. This setup provides a functional, lightweight, and agile framework to rapidly build, test, and deploy recommender systems (Figure 27). However, while effective for small-scale testing, this architecture limits our capacity to scale for large production environments, handle real-time data, or serve diverse business units. This brings us to envision a more advanced low-level Azure architecture, designed for enterprise-level production readiness and long-term sustainability.

End-to-End Solution Architecture for Production Readiness (Low-Level Architecture)

The low-level architecture, as seen in Figure 28, represents a comprehensive and future-proof ecosystem, addressing the complexity of a real-world news recommendation platform capable of serving millions of users and multiple stakeholders. It starts with data ingestion from multiple sources, including streaming data (e.g., live news feeds, clickstreams, IoT devices) and batch data (e.g., images, videos, relational databases). This allows us to combine user behavior in real time with historical datasets, enriching our model and continuously learning from user interactions. These sources are ingested using Azure Event Hubs, IoT Hub, API Management, and Data Factory, enabling both real-time and batch ingestion pipelines to operate simultaneously.

Once ingested, data flows into centralized storage hubs such as Azure Data Lake and Azure Cosmos DB, separated into hot, cold, and archive layers for optimized cost and access. This enables us to store high-frequency clickstream data in hot storage for immediate access, while archiving less frequently used historical datasets. Azure SQL and Azure Synapse provide structured storage and allow powerful querying across vast datasets. From here, batch processing and stream analytics — powered by Azure Databricks, Synapse, and Stream Analytics — allow us to process click behavior, generate embeddings, compute recommendations, and update models on the fly.

Critically, enrichment services like Azure Machine Learning and Azure Cognitive Services introduce advanced AI capabilities such as personalization, NLP, and sentiment analysis. These can enable our models to not only recommend content but also generate dynamic summaries, flag fake news, or adjust for user mood in real-time, offering a sophisticated personalized experience.

Scalability, Deployment, Monitoring, and Maintenance

This envisioned pipeline addresses scalability by enabling horizontal scaling at each layer. Azure Event Hubs can handle millions of events per second, and Azure Databricks and Synapse can dynamically adjust compute capacity depending on processing needs. Deployment pipelines using Azure DevOps, Logic Apps, and Functions ensure automated CI/CD for fast updates and consistent deployments across environments (development, staging, production).

Monitoring and maintenance are ensured through Azure Monitor, Sentinel, and Purview, providing end-to-end observability, security auditing, and data governance. For instance, Purview allows us to track data lineage and ensure compliance with data privacy laws like GDPR and CCPA. Real-time health monitoring with Azure Health Models would allow for automatic detection of failures in the ingestion or recommendation pipeline.

Legal, Privacy, and Intellectual Property (IP) Considerations

Given the sensitive nature of user behavioral data, privacy is paramount. Our production-level architecture would fully leverage Azure's compliance suite, including Key Vault for secure secret management, Azure Active Directory for identity and access, and private links to avoid public exposure of sensitive data. Data masking and anonymization techniques can be applied before storing or processing user data to ensure that personal identifiable information (PII) is protected. Additionally, as we generate proprietary embeddings and models (IP), we would secure them within Azure Container Apps or Private Repos to prevent leakage of intellectual property.

Pathways for Building an In-House Dataset for Long-Term Sustainability

A notable limitation of relying on external datasets is data ownership and relevance. Therefore, an essential long-term goal for our news agency would be building a first-party, in-house dataset. By deploying custom API endpoints via Azure API Management, we could start collecting direct user interaction data, such as click logs, reading times, scrolling behavior, and feedback loops. Integrating Azure AI Personalizer would allow us to gather context-aware interaction data, greatly enhancing the personalization engine while keeping the data proprietary.

Moreover, Azure Data Share could facilitate collaborative datasets with trusted partners like other news agencies, universities, or research firms, giving us access to diverse and high-quality datasets. Over time, these in-house datasets would feed into custom model training in Azure ML, providing us with models that are unique to our user base and content, giving a competitive advantage in personalized news delivery.

Multi-Pathways to Serve Different Business Units

The architecture envisions flexible serving mechanisms for different types of business users. For instance, news consumers can access personalized news streams through web and mobile apps, while internal editors and analysts can use Power BI dashboards to monitor trending topics, reader engagement, and recommendation quality. Executives might use automated decision systems connected to Azure AI to adjust marketing campaigns or push specific content based on user behavior trends. Furthermore, Azure Cognitive Search can provide semantic search capabilities across the news archive, enhancing both user experience and editorial workflows.

Governance, Bias, and Fairness

Another crucial production-level consideration is the need for bias mitigation and fairness in the recommendation pipeline. Without careful monitoring, news recommenders can exacerbate filter bubbles or political echo chambers, continuously showing users content aligned with their previous clicks while neglecting diversity. Future integration of services such as Azure AI Personalizer and Azure Machine Learning fairness modules could help address this by rebalancing recommendations to include diverse perspectives, while Azure Monitor and Sentinel can track system behavior and detect patterns indicating bias or drift. Additionally, as SokoNews begins collecting in-house datasets, governance and compliance will be enforced via Azure Purview, ensuring all data usage adheres to legal and ethical standards, including GDPR and Sokovian regulations on data privacy.

Security and Optimization

Finally, production readiness mandates robust security, governance, and cost optimization. Using Azure Policy and Compliance Center, we ensure that all deployed services adhere to organizational and legal policies. Autoscale and Spot VMs allow us to keep compute costs manageable, while Azure Advisor continuously suggests cost-saving opportunities. Sustainability and serverless computing principles would be employed through Azure Container Apps and Functions, ensuring that we minimize our carbon footprint and resource wastage.

Future Costs of Entire Cloud Infrastructure

The full-scale production implementation of the recommender system requires an estimated €1,480,000 annually to support horizontal scaling and enterprise-grade infrastructure. This cost is separate from the initial and expansion investments, reflecting the additional requirements for serving millions of users in a highly scalable and resilient environment. Data ingestion costs will rise significantly, with Azure Event Hubs, IoT Hub, and Data Factory accounting for €158,000, ensuring real-time streaming capabilities and batch data processing at scale. Storage expenses for Azure Data Lake, Cosmos DB, SQL, and Synapse will increase to €345,000, accommodating high-frequency clickstream data and historical archives. AI processing and model training will require €300,000, enabling advanced machine learning models to personalize content in real-time. Deployment and monitoring solutions, including Azure DevOps, Logic Apps, and Sentinel, will amount to €230,000, supporting automated CI/CD pipelines and security monitoring across development, staging, and production environments. Ensuring compliance with regulatory frameworks, data privacy, and secure user authentication will contribute €175,000, reinforcing adherence to industry standards. These investments are necessary to support millions of users and maintain the reliability, scalability, and efficiency of SokoNews' recommendation system in a fully operational production environment, ensuring long-term sustainability beyond the initial implementation and scaling phase (Azure).

CONCLUSION

The development of a personalized news recommender system for SokoNews presents a transformative opportunity to enhance user engagement, drive ad revenue, and increase subscription growth. By leveraging CTR prediction models, BERT-based content similarity, and LLM-powered conversational recommendations, we have built a robust, scalable, and modular framework that addresses key industry challenges such as the cold-start problem, rapid content turnover, and implicit feedback ambiguity. The system's impact extends beyond personalization—it fosters user trust through transparency, optimizes content discovery, and ensures that even new users receive relevant recommendations from their first interaction.

With a projected first-year revenue increase of €39.8 million and a strong ROI of 8.2, our findings demonstrate that implementing AI-driven recommendation strategies is not only viable but also essential for SokoNews' long-term competitiveness.

Beyond financial gains, this recommender system provides a technologically sound foundation for future innovation. The integration of Azure cloud infrastructure, scalable ML pipelines, and modular deployment architectures ensures that the system can handle increased user traffic, real-time personalization, and continuous model retraining. Furthermore, transitioning from the MIND dataset to an internal first-party dataset will enhance recommendation quality and data ownership, further strengthening SokoNews's market position. Additionally, ethical considerations—such as bias mitigation, privacy protection, and regulatory compliance (eg. GDPR, Sokovian Regulations)—are central to our approach, ensuring that recommendations are fair, transparent, and aligned with user expectations.

Looking ahead, SokoNews is well-positioned to evolve its recommender system into a fully production-ready, enterprise-grade solution. Future iterations will focus on multi-modal recommendations (incorporating images, videos, and sentiment analysis), real-time clickstream processing, and adaptive learning models to improve responsiveness to breaking news trends. By continuously refining its AI-driven approach, SokoNews will not only enhance user engagement and retention but also maintain a competitive edge in an increasingly AI-powered media landscape.

BIBLIOGRAPHY

Aherne, N. (2023, November 7). Cost of training AI models. https://www.linkedin.com/pulse/cost-training-ai-models-nathan-aherne-8ojtc/

Antaris, S., Rafailidis, D., & Aliannejadi, M. (2020, November 10). On estimating the training cost of conversational recommendation systems. arXiv.org. https://arxiv.org/abs/2011.05302

Artificial intelligence cost estimation: Key factors & Examples. (n.d.). https://www.run.ai/guides/machine-learning-engineering/ai-cost-estimation

Chaffey, D. (2024, January 12). 2024 average ad click through rates (ctrs) for paid search, display and social media. Smart Insights. https://www.smartinsights.com/internet-advertising/internet-advertising-analytics/display-advertising-clickthrough-rates/

Dogtiev, A. (2025, January 21). CPC rates. Business of Apps. https://www.businessofapps.com/ads/cpc/research/cpc-rates/

Gulla, J. A., Svendsen, R. D., Zhang, L., Stenbom, A., & Frøland, J. (2021). Recommending news in traditional media companies. AI Magazine, 42(3), 55–69. https://doi.org/10.1609/aimag.v42i3.18146

Glassdoor. "Average Salaries for AI and Machine Learning Roles." Glassdoor, 2025, www.glassdoor.com.

Indeed. "Data Engineer Salaries in Europe - 2025." Indeed, 2025, www.indeed.com.

Irwin, L. (2023, May 10). How much does GDPR compliance cost in 2023? IT Governance Blog En. https://www.itgovernance.eu/blog/en/how-much-does-gdpr-compliance-cost-in-2020#:~:text=Still%20getting%20to%20grips%20with,according%20to%20a%20PwC%20report.

Mediahaus transforms the readers' experience with Personalization. (2021). https://www.froomle.ai/reports/mediahuis

Mehta, P. (2024, November 26). Recommendation System development cost: the complete guide. Zealous System. https://www.zealousys.com/blog/recommendation-system-development-cost/

Neary, J. (2023, June 14). What's the lifespan of an article? | Chartbeat Blog. Chartbeat Blog. https://blog.chartbeat.com/2023/06/14/whats-the-lifespan-of-an-article/

Pricing – Azure Dedicated Host | Microsoft Azure. (n.d.). Microsoft Azure. https://azure.microsoft.com/en-us/pricing/details/virtual-machines/dedicated-host/

Şimşek, H. (2025, March 12). Top 10 Serverless GPU Clouds & 14 Cost-Effective GPUs. AIMultiple. https://research.aimultiple.com/serverless-gpu/

Strategies, F. (2025, January 30). Personalisation has a huge impact on customer engagement | FT Strategies - Media consultancy from the Financial Times. FT Strategies. https://www.ftstrategies.com/en-gb/insights/personalisation-has-a-huge-impact-on-customer-engagement-how-mature-are-your-capabilities

Top Draw. (2025, March 4). Online advertising costs in 2025: Top draw. Top Draw Inc. https://www.topdraw.com/insights/is-online-advertising-expensive/

Verma, S., & Verma, S. (2024, October 9). How much does it cost to build an MVP for AI applications? Biz4Group. https://www.biz4group.com/blog/how-much-does-it-cost-to-build-an-mvp-for-ai-applications

Wu, F., Qiao, Y., Chen, J.-H., Wu, C., Qi, T., Lian, J., Liu, D., Xie, X., Gao, J., Microsoft Research, Microsoft, & Tsinghua University. (2020). MIND: a large-scale dataset for news recommendation. Microsoft News. https://msnews.github.io/assets/doc/ACL2020_MIND.pdf