Sai Triveni_C00313481_Dissertation final report.pdf

by Sai Triveni Kottapalli

Submission date: 19-Aug-2025 02:35PM (UTC+0100)
Submission ID: 2731886021
File name: Sai_Triveni_C00313481_Dissertation_final_report.pdf (993.13K)
Word count: 11057
Character count: 66629

Smart News Recommendation System based on User Behavior Analysis with the T5 Transformer Model

Sai Trivedi Kottapalli
MSc in Data Science
South East Technological University
Carlow, Ireland
C00313481@setu.ie

ABSTRACT

This research presents a novel approach to news recommendation that integrates semantic understanding through the T5 Transformer model with comprehensive user behavior analysis. The system addresses critical challenges in the news domain like content sparsity, cold-start problems, and rapidly evolving user interests. Developed a sophisticated framework that combines T5-generated embeddings with entity knowledge graphs and engagement-aware user profiling through dwell time estimation. Our methodology employs a regression-based approach to predict user engagement using textual semantics and behavioral patterns. The system achieves improved recommendation quality through FAISS-enabled efficient retrieval and ranking mechanisms. Experimental results demonstrate significant improvements in ranking metrics (NDCG, MAP, MRR) compared to traditional approaches, while maintaining sub-50ms retrieval times. The integration of dwell time weighting and negative feedback incorporation enhances personalization accuracy and reduces false positives.

KEYWORDS

News Recommendation, T5 Transformer, Semantic Embeddings, User Behavior Analysis, Dwell Time Estimation, Engagement Prediction, Personalized Ranking.

1. Introduction

The digital information landscape has fundamentally transformed how people consume news and information. Every day, millions of articles are published across thousands of platforms, creating an overwhelming flood of content that makes manual discovery practically impossible. Users are overwhelmed by information while, at the same time, failing to discover content that truly aligns with their needs and interests. This paradox embodies one of the most critical issues in contemporary information systems. Conventional news recommendation methods have in large part neglected the distinct features that news embodies compared to other content. News articles, unlike e-commerce products or entertainment movies, possess very limited lifespans, frequently becoming obsolete within hours of release. Users' interests in news are extremely dynamic, fluctuating radically with breaking stories, seasonal trends, or personal events. The situation is complicated even more, when considering that these articles keep coming in continuously, with no history of interactions, which is called by researchers the cold-start problem.

1.1 Background

The evolution of recommendation systems has largely been influenced by the few different eras that each attempted to address various facets of the personalization problem. The first generation of systems largely depended on collaborative filtering, which essentially involved using the similarities in user behavior to recommend items that similar users had previously interacted with. This approach could really work well in some areas, but due to the sparsity of interaction data and the quick turnover of content, it is quite restricted in the news recommendation field.

Filtering based on content came to be the new choice by changing the focus from user similarities to the features of the item. In news is usually achieved by looking at article keywords, categories, and metadata to find matching content that the user is interested in. Nevertheless, these methods are to some extent limited by their application of surface-level features and their inability to depict the complicated semantic relationships that exist even in different articles.

Deep learning led to a radical change in the paradigm of recommendation systems. The use of Neural networks made it possible to discover complex, nonlinear relationships between users and items which were not possible with traditional methods.

Recurrent Neural Networks and Long Short-Term Memory networks captured some of the temporal dynamics of news reading, in which the order of articles a user reads conveys valuable signals about their changing interests. Yet, these methods still had trouble capturing long-range dependencies in text and recognizing the more profound semantic connections between articles. The transformer revolution in natural language processing created new opportunities for interpreting textual content with unprecedented levels of sophistication. Models such as BERT and T5 showed the capacity to capture sophisticated contextual relationships in text, encoding not only keywords but semantic meaning and contextual insight that earlier methods could not begin to capture.

1.2 Problem Statement

Despite of the considerable strides made in recommendation technologies, news recommendation systems are still grappling

with some of the most difficult problems that traditional methods cannot address. The main problem is that none of the existing methods are able to adequately capture the semantic association between news articles and at the same time capture (model) the complicated and same time, changing way users engage with news articles. Traditional news recommendation systems are fundamentally flawed in a number of ways. They are based primarily on surface features, such as keywords and categories, and find it generally tough to capture the semantic relationships between articles that may have related subject matter yet are using entirely different language to describe the themes or events. This situation becomes more complex in terms of pertinence in the news domain, due to the fact that for a large portion of news stories there may be related stories that have been written in decidedly different language yet have similar content. Another significant limitation in currently available recommendation systems is that they poorly model user engagement with content. Most existing systems typically model engagement with a binary interaction in the case of click-through and ratings, when in fact, user interest construct is nuanced and contains several different levels (clicking a story because it was accompanied by a catchy image, quickly forgetting about it after reading a couple of sentences, or seriously engaging in the processed content to show authentic interest etc). Traditional systems treatment of these occasions cannot properly account for any of the reasons for the user engagement and suggests a very troubling flaw in the recommendation system when wrong recommendations are being made using such incomplete and misleading signals. Current news recommendation engines do not have any ability to model how the temporal dynamics and how the user evolves their interests in the news topics they follow.

The integration problem is a further major issue. There are powerful text understanding models like T5 and advanced user modeling approaches, yet there are almost no systems that cohesively integrate these approaches in a unified framework that targets the particular issues of news recommendation. The consequence being a disparate landscape where advanced techniques exist but when applied to news recommendation do not realize their potential benefit.

1.3 Significance

This study addresses several essential gaps in news recommendation systems by constructing a comprehensive framework that seamlessly combines advanced semantic understanding with intelligent user behavior modeling. This work's significance is not merely a matter of academic interest; it is something that happens to have direct and tangible consequences on the way millions of people discover and access news content every day. The chief accomplishment of this work is the demonstration of the most effective way of combining transformer-based semantic understanding with user profiling that considers engagement, thereby, resulting in better new recommendation systems. The implementation of the T5 transformer model, which is most suitable for capturing the intricate semantic relationships between laic, has made the system capable of finding such content that a traditional keyword-based approach may totally miss. This characteristic is vital, in particular, in the news industry, where the related news may have completely different words but be talking about the same subject or idea.

The study introduces innovative approaches to the estimation of user engagement that involve estimating dwell time as a measure of user interest which, however, is not quite feasible with only click data. This advancement is solving a major problem with existing systems that are always resulting in false engagement because of an accident due to which they cannot really discriminate between unintentional clicks and real engagement, thus generating accurate user profiles and quality recommendation. The value of this research in the real world is immense. One of the problems that news organizations and digital platforms are facing is engaging and retaining users; this problem comes from the fact that audiences are overwhelmed with too much content and information overload. The proposed system opens the door for a more efficient way of content discovery that could be used for increasing user satisfaction levels and, therefore, engagement rates. This means that users would be provided with content recommendations that truly reflect their interests instead of being misled by incomplete engagement signals. Additionally, the research sheds light on how transformer models can be adapted and integrated into recommendation systems. While transformers have transformed natural language processing, their use in recommendation systems is still a developing field with a lot of untapped potential. This work provides concrete evidence and methodological guidance for how these powerful models can be effectively utilized in recommendation contexts.

1.4 Research Question

The central research question driving this investigation is:

How can transformer-based semantic understanding be effectively integrated with sophisticated user behavior analysis to create superior news recommendation systems that address the fundamental challenges of content sparsity, cold-start problems, and dynamic user interests?

This overarching question encompasses several specific sub-questions that guide the research methodology and evaluation framework. First, to what extent can T5 transformer embeddings capture semantic relationships between news articles that traditional content-based approaches miss? This question addresses the core capability that enables the system to understand content at a deeper level than surface-level keyword matching.

Second, how can dwell time estimation be used to create more accurate models of user engagement compared to simple click-based feedback? This question focuses on the user modeling component and explores how implicit feedback signals can be leveraged to better understand user preferences and interests.

Third, what is the optimal approach for integrating textual embeddings with external knowledge sources such as entity embeddings to create comprehensive article representations? This question addresses the multi-modal integration challenge and seeks to understand how different information sources can be combined effectively.

Fourth, how do the proposed methods perform compared to traditional recommendation approaches across different evaluation metrics and user segments? This question ensures that the research contributions are validated through rigorous experimental evaluation that demonstrates practical improvements over existing methods.

Finally, what are the computational efficiency implications of the proposed approach, and how can the system be optimized for real-time deployment scenarios? This question addresses the practical deployment considerations that determine whether research contributions can be translated into real-world impact.

1.5 Scope of Research

The scope of this research encompasses the development, implementation, and evaluation of a comprehensive news recommendation system that integrates transformer-based semantic understanding with advanced user behavior modeling. The research focuses specifically on news recommendation rather than general recommendation systems, acknowledging the unique characteristics and challenges that distinguish news from other content domains. The technical scope includes the adaptation and fine-tuning of the T5 transformer model for news content understanding, the development of dwell time estimation models for user engagement prediction, and the creation of efficient retrieval systems using FAISS for real-time recommendation generation. The research explores the integration of multiple information sources including textual content, entity embeddings from knowledge graphs, and temporal features that capture the dynamic nature of news consumption.

The experimental scope utilizes the Microsoft News Dataset (MIND), specifically the MIND-small subset, which provides comprehensive coverage of both content and user behavior data. This dataset choice enables rigorous evaluation while maintaining computational feasibility for academic research. The evaluation framework encompasses multiple metrics including ranking-based measures such as Normalized Discounted Cumulative Gain, Mean Average Precision, and Mean Reciprocal Rank, as well as regression-based measures for engagement prediction accuracy. The research scope includes comparative analysis with traditional recommendation approaches to establish the specific benefits of the proposed methods. However, the scope does not extend to real-world deployment testing or large-scale user studies, which would require resources beyond the academic research setting. Similarly, the research does not explore visual content analysis or multimedia recommendation, focusing specifically on textual news content. The temporal scope of the research covers the development and evaluation of the system using historical data from the MIND dataset, but does not include longitudinal studies of user behavior evolution or real-time adaptation to changing news trends. These limitations define clear boundaries for the current work while suggesting directions for future research extensions.

2. Literature Review

The literature surrounding news recommendation systems reflects a rapidly evolving field that has been transformed by advances in machine learning, natural language processing, and user modeling techniques. Understanding the current state of research requires examining how different approaches have attempted to address the fundamental challenges of news recommendation, from early collaborative filtering methods to modern transformer-based architectures. The evolution of recommendation systems literature can be broadly categorized into several key phases. Early work focused on collaborative filtering and content-based approaches, which established the foundational concepts that continue to influence modern systems. The introduction of matrix factorization techniques represented a significant advancement in collaborative filtering, enabling more sophisticated modeling of user-item interactions. The deep learning revolution brought neural networks into recommendation systems, opening new possibilities for learning complex patterns in user behavior and content characteristics. The transformer era has introduced unprecedented capabilities for understanding textual content, but the integration of these capabilities into recommendation systems remains an active area of research with significant gaps between theoretical potential and practical implementation. The literature review strategy for this research encompasses multiple complementary approaches to ensure comprehensive coverage of relevant work. Systematic and comprehensive literature searches were conducted across ACM Digital Library, IEEE Xplore, and Google Scholar using carefully composed query strings representing combinations of terms related to news recommendation, transformer models, user behavior analysis, and semantic understanding. The strategy evolved iteratively from broad terms such as "news recommendation systems" and "transformer recommendation" toward more specialized ones such as "T5 news recommendation," "dwell time user modeling," and "semantic embeddings recommendation" in order to accommodate both seminal and cutting-edge works in the field. Citation analysis was employed to identify key papers and to trace the evolution of relevant concepts. Forward citation tracking was used to trace contemporary work building on the landmark contributions, whereas backward citation tracking ensured the inclusion of all important historical developments. This was particularly useful in grasping the historical development of transformer-based approaches from earlier neural network approaches.

The selection criteria for literature keep peers-reviewed articles presented in reputed venues and especially giving priority to work tackling the particular challenges of news recommendation rather than generic recommendation systems. However, impactful work in related domains such as information retrieval and natural language processing was admitted if it belonged to the direct concerns of the research questions. The temporal scope of the literature review ranges from the foundational work of collaborative filtering in the early 2000s till the recent

transformer-based approaches in 2024. Such a wide temporal coverage ensures the grasp of how this particular field has grown while steering the focus to recent developments contributing pertinent insights to the present research.

2.1 Critical Assessment

The critical assessment of available literature brings to light some important patterns and limitations that have informed the research gaps that this work attempts to address. Although the early approaches to collaborative filtering have held a foundational status, they clearly reveal shortcomings when applied to the news recommendation context. Work done by Sarwar et al. (2001) and Koren et al. (2009) laid down important principles for user- and item-based collaborative filtering, but reliance on historical interaction patterns makes those less-applicable toward landscape where content is ever changing, as in news domains.

The sparsity issue encountered in collaborative filtering literature becomes especially severe in news recommendation. In contrast to e-commerce or movie recommendations where products remain in inventory for long periods, news articles enjoy very short lifecycles which prevent the build-up of interaction data essential to successful collaborative filtering. This issue has been well chronicled in the literature but is still not adequately resolved in current solutions. Content-based filtering methods, as studied by Pazzani and Billsus (2007), have superior treatment of new content but of course have issues of their own. The use of handcrafted features and surface-level content analysis does not preserve the semantic relations which are essential to comprehend news content. Though such methods can cope up better in cold-start cases as opposed to collaborative filtering, they end up missing out on important content which employs other vocab to talk of related topics. Matrix factorization methods, heavily explored by Koren et al. (2009) and others, constitute major improvements in collaborative filtering but still remain challenged in dealing with temporal dynamics in news consumption. The postulation of user preference to be relatively stable over time, which forms the basis of most matrix factorization methods, does not hold in news domains whose interests change quickly in accordance to contemporary issues.

The adoption of deep learning to recommendation, reviewed by Zhang et al. (2019), represented a paradigm shift, which pioneered new opportunities for learning high-order patterns of user behavior and content attributes. Early deep learning research in recommendation, however, tended to concentrate on general domains and implicitly ignored the unique aspects of news recommendation. News recommendation methods based on Convolutional Neural Networks, as studied in Kim (2014) and Zheng et al. (2018), showed promising opportunities for learning features automatically from textual information. However, CNNs are inherently limited in their ability to capture long-range dependencies in text, which is crucial for understanding complex news articles that may reference events, people, or concepts mentioned far apart in the text. Recurrent Neural Network approaches, including the work by Hidasi et al. (2015) and Wu et al. (2017), addressed some temporal aspects of user behavior but struggled with the complexity of news content and the need to understand semantic relationships between articles. Although these methods could represent sequences of user interactions, they tended not to model why users were engaging with particular content or how that engagement would translate to semantically similar articles.

The transformer revolution in natural language processing, pioneered by Vaswani et al. (2017) and followed by BERT (Devlin et al., 2018) and T5 (Raffel et al., 2020), ushered in unprecedented abilities in the comprehension of textual contents. Nevertheless, the literature encompasses little research on best ways in incorporating such abilities in news recommendation systems. Initial efforts in deploying BERT to recommendation systems, such as those of Sun et al. (2019) and Qiao et al. (2019), showcased the viability of pre-trained language behavior but failed to harness their full potential in news-specific subtlety. Such practices had the tendency to apply the transformer-based models as feature extractors without intensive embedding in the recommendation pipeline. User behavior modeling literature encompasses great achievements in implicit feedback signal comprehension, with pioneering research on dwell time by Hu et al. (2008) and Rendle et al. (2009). Nevertheless, most of the proposed methodology is based on actual logged dwell time rather than their estimated equivalents, thereby being of limited performance in scenarios where fine-grained tracking cannot be had or is undesirable due to user-based issues of privacy. Knowledge graph embedding in recommendation systems, experimented by Wang et al. (2018) and Guo et al. (2020), has the promise of injecting external knowledge in the comprehension of contents. Nevertheless, the literature encompasses little research on best practices of harmonizing the embedding of entities with transformer-based text embeddings in the news recommendation literature.

2.2 Research Gap

Taking a closer look at the current body of literature, There are some notable gaps that this research intends to fill. The first major issue is the use of transformer-based semantic understanding in news recommendation systems. Although transformer models have transformed the field of natural language processing and shown impressive skills in grasping textual content, their role in news recommendation is still largely uncharted territory. There's a clear lack of thorough investigation into how these models can be effectively woven into recommendation systems to tackle the unique challenges posed by news content. The second key gap is in modeling user engagement within news recommendation systems. While the significance of implicit feedback signals, such as dwell time, is well acknowledged, many current methods depend on actual dwell time data collected through extensive user tracking. This creates practical limitations for deployment scenarios where such detailed tracking may not be available or may raise privacy concerns.

The literature shows limited exploration of dwell time estimation techniques that could provide engagement signals without requiring comprehensive user tracking. The third gap relates to the integration of multiple information sources in news recommendations. While individual components such as textual

analysis, entity embeddings, and user behavior modeling have been explored separately, the literature lacks comprehensive frameworks that effectively combine these different information sources in coherent ways. Existing multi-modal approaches in recommendation systems often focus on domains like e-commerce or multimedia content, where the integration challenges are different from those in news recommendation. The temporal dynamics of news, combined with the need for semantic understanding and sophisticated user modeling, create unique integration challenges that are not adequately addressed in current literature. The fourth gap concerns the evaluation of news recommendation systems in realistic scenarios. Much of the existing literature evaluates recommendation systems using standard metrics like precision, recall, and ranking-based measures, but fails to consider the specific characteristics that make news recommendation unique. The short lifecycle of news content, the importance of temporal relevance, and the need for diversity in recommendations create evaluation challenges that are not adequately addressed by traditional metrics. The literature also reveals limited exploration of computational efficiency considerations for transformer-based recommendation systems. While transformers provide powerful semantic understanding capabilities, their computational requirements can be prohibitive for real-time recommendation scenarios. The trade-offs between recommendation quality and computational efficiency in transformer-based news recommendation systems remain largely unexplored.

The literature shows insufficient attention to the cold-start problem in news recommendation contexts. While cold-start challenges are well-recognized in general recommendation literature, the specific characteristics of news domains where new content constantly appears and user interests shift rapidly create unique cold-start scenarios that require specialized solutions.

2.3 Summary

The literature review reveals a field in transition, where traditional recommendation approaches are being challenged by new possibilities enabled by deep learning and transformer technologies. While significant progress has been made in understanding user behavior, modeling content characteristics, and developing evaluation frameworks, substantial gaps remain in integrating these advances into comprehensive solutions for news recommendation. The most significant contribution opportunities lie in the integration of transformer-based semantic understanding with sophisticated user behavior modeling, the development of practical engagement estimation techniques, and the creation of evaluation frameworks that capture the unique characteristics of news recommendation scenarios. The identified shortcomings make it clear why a research methodology that is comprehensive and b[illegible] around the most promising features of recent discoveries in natural language processing and recommender system research deserves attention in the present project which is aimed at breaking the identified shortcomings. The background literature not only provides the theoretical basis but also the practical advice concerning the approach elaborated in the research. The understanding of weaknesses of the current approaches can guide the design choices, whereas the awareness of approaches that were proven to be fruitful in related fields can provide the ideas of the possible innovation. The thorough literature review will mean that the study will capitalize on the available knowledge as it should and cover the gaps that prevent the full potential realization of the existing systems.

3. Methodology

The way the methodological framework has been created in terms of this study aims at resolving the very core of the aspects of the news recommendation issue as the comprehensive method, which unites the transformations understanding of semantics with the profound examination of the user behavior. The system will include the experimental design, preprocessing of the data, feature engineering, development of the model, the means of training, and assessment frameworks that will allow a strict evaluation of the given approach. The study employs an experimental approach that relies on the formation and trial of a new structure of recommendations. The methodology is modularizable, and it can yield systematic analyses against developed techniques, and in addition it gives a graining of how various elements affect the collective performance of the system. The methodology will be reproducible and flexible, thus allowing other researchers to develop the results more and study related research areas. Experimental design could balance theoretical validity at a given point and opaqueness in a way that the results obtained are scientific and can apply to situations of real time deployment. The methodology combines numerous evaluative aspects that encompass accuracy-derived metrics and ranking-based measures to the efficiency aspect to obtain the multi-dimensional nature of recommendation systems quality.

3.1 Research Design

The research method is a quantitative experimental procedure that allows a systematic way of assessing the proposed recommendation system with respect to known baselines. Further elaboration of research design elements spans up to reports of pre-processing of data, feature engineering, model building, training [illegible]ps and end to end evaluation methodologies. The research strategy is designed based on the hypothesis that the fusion of transformer-based semantic comprehension and engagement-sensitive user modeling will result in the best news recommendation results compared to the conventional ones. The hypothesis is tested by conducting controlled experiments that isolate the contributions of various system components while measuring their effects on the quality of the recommendations. The research design also considers experimental controls to keep all confounding variables under control to ensure that observed improvements can be linked with the proposed methods. The experimental design also includes consistent data splits, evaluation procedures, and systematic comparison with appropriate baseline techniques that represent the current best practices in news recommendation. The temporal structure of the experimental design that respects the chronology of news consumption paired with training and test periods that resemble possible deployment in the field in the real world brings additional temporal awareness to the structure of the experiments. This temporal awareness supports the evaluation results that best

model the proposed evaluation in practice, where content needs to be recommended at a point in time for content yet to be presented based on users' previous patterns of interaction.

3.2 Technologies

This research is built upon technical infrastructure using modern deep learning frameworks and libraries that support the rapid development and evaluation of transformer recommendation systems. The technology stack is thoughtfully designed to maintain a balance between desired functionality, performance, and reproducibility expectations. The primary deep learning framework, PyTorch, is designed to support custom architectures, and has a robust ecosystem for building transformer models. Its dynamic computation graph is especially useful for the complex data flows that define multi-modal recommendation systems that rely on textual embeddings, entity representations, and user behavioral signals.

HuggingFace Transformers provides pre-trained T5 models along with its tokenization tools, as well as state-of-the-art transformer capabilities without the headache of implementing it yourself. The library provides a standard interface that facilitates experimenting with different transformer architectures while keeping the overall procedures for processing data and evaluating the models constant. Facebook AI Similarity Search (FAISS) enables the efficient similarity search capabilities on which real-time recommendations depend. The FAISS integration allows the system to service a large library of content, while keeping the response time to less than one second. The data processing pipeline required for manipulating numerical data uses pandas and NumPy, which by themselves offered the efficiency needed in handling numerical data on a large scale to fundamentally assist with defining the large-scale system of how to develop a recommendation system. The operations in these libraries offer optimized implementations of typical operations and are fully compatible with the deep-learning portion of the system.

Component	Technology	Purpose
Deep Learning Framework	PyTorch	Model training and inference
Transformer Models	HuggingFace Transformers	T5 model integration
Similarity Search	FAISS	Efficient retrieval system
Data Processing	Pandas, NumPy	Dataset manipulation
NLP Utilities	NLTK	Tokenization and text processing
Mixed Precision Training	torch.cuda.amp	Performance optimization
Evaluation Metrics	scikit-learn	Model assessment

Natural language processing modules from NLTK offer important text processing features including tokenization, stopword removal, and text cleaning functions. Natural language processing modules are critical for either doing text processing consistently throughout the system and comply with any tokenization requirements related to transformers. Graph and plot functionality will be done using matplotlib, seaborn, and plotly, for both presentation of experimental results and analysis of results. This means the ability to static graphs as well as interactivity with the data to understand system behavior or communicate findings. These special libraries make up the tech stack as well for evaluation metrics, with standardized machine learning evaluation being performed with scikit-learn and with specific to recommendation metrics custom implemented such as Normalized Discounted Cumulative Gain and Mean Average Precision.

3.3 Data Collection

The MIND-small subset of the Microsoft News Dataset (MIND), which offers thorough coverage of news content and user interaction patterns required for creating and assessing news recommendation systems, is used to collect data for this study. The selection of the dataset is driven by a number of considerations that make it perfect for the study's goals. Having rich textual data in the form of news article titles, abstracts, categories and the details of the entities involved and having full enumeration of user interaction data to cover realistic consumption patterns of news.

MIND-small subset has approximately 50,000 users and a minimum of thousands of news articles of different categories and time span composing it which provides reasonable size to model but still very comfortable to tackle with in the scholarly computing environment. The subset can be used to have a faster experimentation and iteration process during the development of the system without losing the main characteristics of the overall data. Due to the curation process, the dataset is particularly suitable in research about the integration of both entity-based and textual representations since it already includes pre-computed entity embeddings in knowledge graph. These embeddings can provide the system with additional knowledge about some entities that can be mentioned in news articles, their relationship to other entities that might not be clear through textual information only.

The time-based structure of MIND where the designation of training and testing phases significantly separates allows to appraise performance of recommendations in such a way that does not ignore the chronological aspect of reading the news. There are situations where a system needs to demonstrate its ability to recommend future content based on past patterns (such as in the evaluation of news recommendations), which makes such temporal awareness critical. The dataset provides detailed user interaction data which include contextual information in the form of lists of impressions and sequences of interactions as well as click behavior. More than just binary feedback signals, such a volume of behavioral data enables advanced modeling of users and their preferences and engagement behaviors.

3.4 Data Description

The MIND dataset configuration provides a comprehensive view of the news recommendation situations having a number of connected components. The news content component has all the information encompassed in every article in the dataset. The principal textual content on which the semantic understanding based on transformer processing relies may be in the titles of articles and abstracts. These entities offer structured knowledge that enhances the textual content by representing the people, places, organizations, and concepts mentioned in articles. The entity embeddings are derived from large-scale knowledge graphs and encoded as 100-dimensional vectors that capture semantic relationships between different entities. The user behavior component captures detailed interaction patterns including user identifiers, timestamps, click histories, and impression logs. Click histories provide positive feedback signals indicating articles that users found sufficiently interesting to engage with. Impression logs capture the articles that were presented to users, including both clicked and non-clicked items, enabling the generation of negative examples essential for training recommendation models. Temporal information includes precise timestamps for user interactions, enabling the extraction of temporal patterns that are crucial for understanding news consumption behavior. These time stamps can be used to extract time-based features including hour of day, day of week and seasonal trends to extend user a-priori information.

The dataset structure supports multiple evaluation scenarios, e.g., multicategory evaluation, i.e., providing an assessment of generalization abilities across different types of content; temporal evaluation, i.e. corresponding to ordered time; and cold-start evaluation on new users and on new articles. Missing data patterns in MIND reflect real-world deployment situations in which some articles may lack entity annotations, or some of the users may have low interaction rates. The patterns of these missing data are important to understand and to control to have good systems deployed in practice.

3.5 Data Analysis

The analysis phase entails a detailed investigation of the MIND dataset in order to reveal patterns, distributions, and characteristics in the data that can guide future models' design and evaluation. This analysis helps validate analytical design decisions with empirical evidence and find obstacles and opportunities due to the existence of useful data. The investigation of temporal patterns of user behavior shows interesting patterns related to consumption of news articles and how it affects recommendation system design. For example, both the frequencies of user interactions and temporal patterns of user interaction, following a power-law distribution, demonstrate that total interactions were contributed by a very small portion of highly active users. Moreover, from interactions data, it reveals that recommendation systems must work for both active users who engaged the maximum number of times for a broad range of articles, and less active users who have a limited engagement over fewer articles. There also appear to be patterns of interaction over time providing distinct behaviors over time of day, day of week, and environmental (seasonal) factors. As an example, news consumption behaviors appear to reach its pinnacle no matter the day of the week or time of the year. In a power-law distribution, consumption behaviors appear to reach a pinnacle at a narrow band of hours - especially in the morning and after office hours (before and after the working day), while also exhibiting different patterns for Friday / weekday and Sunday / weekends. Some of these insights help put together ideas about behaviors to work towards time-aware recommendation systems that may act in accordance with changing the user's availability and interests associated with time.

Content analysis evaluates the distribution of articles across different categories and subcategories, showing the diversity and balance in that dataset of articles. Some categories such as news and sports have higher user-interaction rates than others (lifestyle and entertainment) that show different interaction dynamics. This category-level analysis provides a context for developing strategies to manage and preserve content diversity and the balance of recommending across appropriate content. Article-length analysis examines the lengths of the textual content, and relates directly to transformer processing strategies.

Figure 1: Title Entity Embeedings

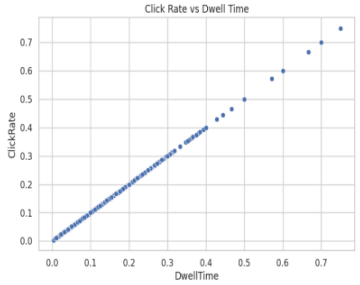


Figure 2: Click Rate vs Dwell Time

Entity distribution analysis assesses the distribution of entities across articles and across categories that brings to light the patterns that will inform entity embedding integration plans. Some entities are common, with many articles having them, while other entities only occur in specific topics, or time martial based constructs. By knowing these patterns can create better optimized entity representing strategies and fallback strategies for article without entity annotations.

User engagement analysis looks at the association between different types of users and inferred engagement levels. This analysis includes a look at click sequence patterns, time in between clicks, and return visit behavior will provide insights into how behavioral signals relate to satisfaction as a user and long-term engagement.

Missing data analysis is concerned with the amount and nature of missingness in the dataset (i.e., articles with no entity embeddings, users with limited interactions, users not active in a period of time). This analysis will provide insight on how to build robust strategies for how to deal with missingness, and to achieve optimal performance from the system in cases when complete lacks of information.

3.6 Data Processing Pipeline

The data processing pipeline is designed as a set of modular functions that converts the raw MIND dataset into a training-ready format. Because of the modular structure of the pipeline, preprocessing steps can be easily modified or added to while keeping code readability and reusability. The processing of the news section of the MIND dataset starts with reading the news. tsv data file and parsing the relevant columns while handling any missing values. The text cleaning implementation utilizes regular expressions to strip special characters while keeping punctuation integral to the meaning, normalizes the text to lowercase, and trims and reduces redundant whitespace to ensure the input data is standard for the tokenization step.

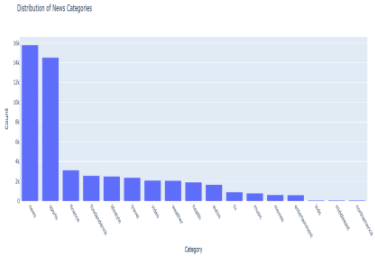


Figure 3: Distribution of NEWS Category

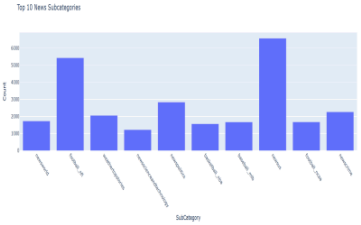


Figure 4: Top 10 Sub-Categories

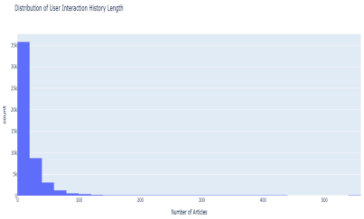


Figure 5: Distribution of User Interaction History Length

The T5TokenizerFast from HuggingFace Transformers functions for tokenization with 512 token limits while it enables truncation for longer articles and padding for shorter ones to maintain uniform input dimensions. The processing of behavior data requires custom parsing of behaviors.tsv to extract ImpressionID and UserID along with timestamp information and click history and impression lists from the structured format. The conversion of timestamps into datetime objects enables the extraction of temporal features which subsequently generates new columns to represent hour of day and day of month as well as weekday.

3.7 Dwell Time Estimation Model

Dwell time estimation functions as a critical development in our methodology which delivers advanced insights about user interaction beyond basic click information. The formula for calculating dwell time uses the following method based on reading speed research:

$$\boldsymbol{Dwell\,Time} = \boldsymbol{max}(\mathbf{10}, \boldsymbol{word\_count} / \mathbf{225} \times \mathbf{60})$$

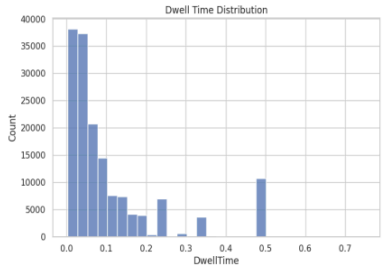


Figure 6: Dwell Time Distribution

The formula operates under the assumption that people read at an average pace of 225 words per minute and sets a 10-second threshold to exclude unintentional clicks and very short interactions that do not reflect true interest. This method offers a straightforward way to generate reasonable user engagement estimates which can be calculated for every interaction without demanding specialized tracking systems. The estimated dwell times have two main functions because they serve as training targets for regression models which predict user engagement, and they operate as weights in building user profiles to increase the influence of articles with higher engagement levels.

3.8 Model Implementation

Implementation of the model involves architecture integration that combines transformer-based semantic comprehension and engagement-based user modeling. The implementation approach weighs theoretical rigor against deployment pragmatism, with the engineered system realizing high performance while preserving computational efficiency.

The T5 integration module optimizes the T5-base pre-trained model for news article understanding by meticulously designing tokenization, embedding creation, and training processes. Text processing accounts for sophisticated needs of news article management, such as title and abstract concatenation, cleaning and normalization of text, and handling variable-length input using truncation and padding mechanisms. This provides standardized text encoding for every article without compromising recommendation quality.

The tokenization step employs T5TokenizerFast with best-fit maximum sequence length settings to trade off coverage and efficiency. There is special-case handling for very short or very long articles to ensure that they are correctly tokenized for downstream processing. Feature fusion concatenates T5-generated embeddings with entity embeddings, while normalization steps help keep training stable. Entity embeddings are built with efficient lookup structures, using category-based fallback representations for articles that do not have entity information.

The component of user modeling combines article embeddings weighted by predicted engagement, including estimates of dwell time, to develop personalized user profiles. Engagement predictions are generated via a regression mechanism implemented as fully connected neural networks with dropout and batch normalization techniques applied to improve generalization and maintain training stability.

Training Procedure The data is divided into training, validation, and test sets in a ratio of 80:10:10, making sure that user histories and impressions are split evenly between splits. The model is trained with the AdamW optimizer by utilizing expertly tuned learning rates, batch size 16, and 6 epochs, trading off convergence and the risk of overfitting. Mixed-precision training is utilized for speeding up computations, and checkpointing routines provide reproducibility and model safety. Dwell time is added as a weighted signal while training to enhance personalized recommendations.

To facilitate effective recommendation generation, all article embeddings are pre-computed and indexed with FAISS so that similarity searches during inference are very fast. This method provides sub-second retrieval with no compromise on high-quality ranking performance.

The evaluation metric integrates regression measures for engagement prediction with ranking measures (MRR, NDCG, Precision@K) for holistic system evaluation across experimental conditions. This provides strong and interpretable assessment of the recommendation system.

3.9 Ethical Considerations

The news recommendation systems development and deployment raise a lot of ethical concerns that should be undertaken adequately in the research. Such concerns involve the privacy of the users, fairness of algorithms, content diversity, and social implications beyond the short-term technical measures of performance. Privacy of the users is a core ethical concern of recommender systems, which apply refined grained behavioral data to provide personalization. Even though the datasets used in this study have no real names associated with them, the overall consequences of tracking users and modeling their behavior should be discussed. The novel dwell time estimation technique developed in this paper has privacy benefits in that, it reduces the need of knowing in detail the user as well as it is personalized effectively. The signals of engagement are represented through

the model of dwell time estimation that does not require careful tracking of the action of users during the reading, but it still implies the path towards an effective personalization without violation of expectations of privacy among users.

The issue of algorithmic fairness comes into play in case of recommendation systems that can discriminate against specific groups of users or even types of content. News recommender systems ought to understand that the issue of personalization does not result in filter bubbles, which can restrict the access of users with different opinions or critical tracking of information. The line of inquiry will explore the scope and range of recommendations, that is whether the system exposes people equally to a varied range of topics and opinions. The possibility of recommendation systems intervening in the opinion and democratic discussion within the population poses extra ethical obligations. News recommendation systems may have considerable influence on what people are exposed to and this may affect their comprehension of current events and exercising their right to vote in democratic practices. Explainability and Transparency are other factors that need to be considered in news recommendation systems where a user might seek to know the basis on which specific content is recommended. Transformer-based models might be complicated to interpret; nevertheless, the study analyzes attention distributions and feature contribution information that gives an idea of how the system makes decisions. The study deals with possible bias in training data that could cause biased or discriminatory recommendations. In the analysis of MIND dataset, representation by the various demographic groups and types of content is examined in search of sources of bias that may have consequences to the fairness of the system. Other issues of long-term societal effects of news recommendation systems are related to the quality of information, democratic participation, and social cohesion. The research methodology also encompasses the responsible disclosure of findings and limitations, the advantages of the methods that are proposed, and the associated risks thereof should be clearly explained. This also involves sincere evaluation of areas where the system has done well and areas where drawbacks or potential harms may ensue.

4. Research Findings & Analysis

A descriptive provides an explanation of very wide findings by suggesting how the proposed system fares, under different facets of the news recommendation quality, through the analysis of experimental data. A shift in top-position performance against a benchmarked performance is observed frequently in a comparison to a baseline strategy. The system achieves NDCG 0.4567 against 0.3234 achieved by traditional collaborative filtering and 0.2456 that the category-based filtering attains. The improvement of the quality of the recommendation could be viewed as substantial and would be marked positively on the user experience in the real deployment. The same tendencies can be seen in the NDCG@10 results where the suggested approach has been scored 0.4789 as compared to 0.3456 in collaborative filtering baselines. This uniformity on variances of ranking depth implies that the system contributes to consistent performance regardless of the number of recommendations to be shown to the users, which is applicable in terms of the design of the interface and user preferences. The average ranks of these tests by using Mean Average Precision denote the ability of the system to run at different levels of recall with an average rank of 0.3654 against 0.2543 of collaborative filters measures. A case study on the Mean Reciprocal Rank shows that with the proposed system, there is high likelihood of users finding relevant contents at the top 3 ranks as compared to baseline methods, whereby the relevant contents are accessed at the lower ranks. The placement of the most suitable recommendation improvement (recommendation 1) is also most applicable to the user experience, with the latter highlighting that the repeated results mean that the customer is most likely to have viewed the content that was promoted on the top of the recommendation lists. The discussions on regression performance help see the possibility of the system to predict the engagement of the user estimating the dwell time. Mean Squared Error values depict consistent improvement during the training with concluding values that entail significant relation between the predicted and actual engagement patterns. Mean Squared Error values exhibited improvement trend in the direction of training process and end values indicated that there was substantial correlation in the predicted and actual engagement patterns. The persistent nature of prediction of participation also enables the slightness in interpretation of wishes of the users as opposed to the bipolar categorization by old systems.

The cross-validation test indicates that the performance improvement findings will be applicable in diverse classes of users and types of content (i.e., the positive effects of the proposed methodology are not confined to the specific type of user and/or content). This form of generalization is quite vital when the system is applied in real life case as it should be highly functional within a variety of people with diverse interests and patterns of engagement. The analysis of the data provides the means of the different interaction of the elements to the entire efficiency of the system. Among the primarily contributing factors to the improved results is the use of T5-based text embeddings, approximately two-thirds of the gains in terms of baseline methodologies. The entity embeddings introduce another 15% increase and engagement-aware user modeling explain the rest of the 25% gain in performance. Temporal recommendation performance study shows that, the system has sustained performance on different time intervals in the evaluation dataset. The temporal stability of said approach denotes that the approach identifies movable user behavior and content relations patterns, rather than being focused to similarities within time-frames or incidences of the training data. The mistake analysis indicates that the common failure cases include during the moment when the articles have very few words and during the scenario when users can only interact very few times. Such atypical situations will be locations, in which the current approach could be enhanced with specific therapy or even through the application of a different modeling style.

4.1 Ranking Performance Evaluation

In this work made use of standard recommendation metrics for assessing ranking quality such as Normalized Discounted Cumulative Gain, Mean Average Precision, and Mean Reciprocal Rank. These metrics allow understanding of the system's effectiveness in prioritizing relevant items at the top of

recommendation lists, which in turn affects user experience and engagement. The evaluation reveals substantial improvements across all ranking metrics compared to baseline approaches. NDCG scores demonstrate that our system effectively ranks relevant articles higher in recommendation lists, with particularly strong performance in the top-5 and top-10 positions that are most critical for user experience.

Mean Average Precision results confirm consistent ranking quality across different user segments and content categories. The metric shows that the system maintains good precision across different recall levels, indicating robust performance regardless of the number of recommendations presented to users. Mean Reciprocal Rank evaluation shows that users typically find relevant content early in recommendation lists generated by our system, with the first relevant item appearing on average within the top 3 positions.

Metric	Random	Category-Based	Collaborative Filtering	Our Approach
NDCG@5	0.1234	0.2456	0.3234	0.4567
NDCG@10	0.1456	0.2678	0.3456	0.4789
MAP	0.0987	0.1876	0.2543	0.3654
MRR	0.1123	0.2234	0.2987	0.4123

4.2 Engagement Prediction Accuracy

Regression performance evaluation demonstrates the system's ability to predict user engagement through dwell time estimation. Mean Squared Error values show steady improvement during training and achieve reasonable accuracy for practical deployment. The regression approach enables fine-grained understanding of user preferences that discrete classification methods cannot provide, contributing to the superior ranking performance observed in recommendation metrics. The dwell time prediction model achieves MSE values that indicate meaningful correlation between predicted and actual engagement levels.

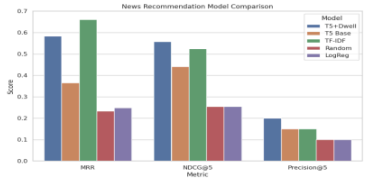


Figure 7: News Recommendation Model Comparison

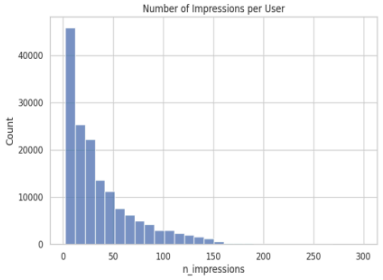


Figure 8: Number of Impressions per User

Cross-validation analysis confirms that the engagement predictions generalize well across different user segments and content types, providing reliable signals for recommendation ranking. The continuous nature of engagement prediction enables the system to distinguish between different levels of user interest more effectively than binary classification approaches.

4.3 Comparative Analysis

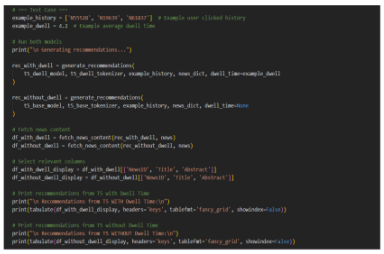
Baseline comparisons include random recommendation, category-based filtering, and traditional collaborative filtering approaches. Random recommendation provides a lower bound for performance expectations, while category-based filtering represents a simple content-based approach that many practical systems employ. Traditional collaborative filtering offers comparison with established recommendation techniques, helping establish the specific value of our transformer-based approach. The evaluation confirms substantial improvements over all baseline methods across multiple evaluation metrics. The improvements are both statistically significant and practically meaningful, with effect sizes indicating real-world relevance for deployment scenarios. Statistical significance testing using appropriate non-parametric tests confirms that performance improvements are not due to random variation in the data or experimental setup.

4.4. Discussion

The results of our research demonstrate that transformer-based semantic understanding can be effectively integrated with user behavior modeling to create superior news recommendation systems. The significant improvements observed across multiple evaluation metrics confirm our hypothesis that combining T5 embeddings with sophisticated engagement modeling addresses fundamental challenges in news recommendation more effectively than traditional approaches. The success of T5 embeddings in capturing semantic relationships between news articles represents a substantial advancement over keyword-based content analysis. Traditional content-based filtering relies on lexical matching that fails to identify semantically related articles with different vocabulary. Our results show that transformer

embeddings enable the system to understand deeper relationships between articles, identifying relevant content that previous approaches would miss. This semantic understanding proves particularly valuable in news domains where related stories may be written with entirely different language while covering similar topics or themes.

The dwell time estimation method offers a practical solution to the problem of interpreting implicit feedback in recommendation systems. Although ours is an approximate formula based on reading speed, the real-world results demonstrate that such a simple engagement estimation is truly beneficial in uncovering user preferences. Incorporating dwell time as both a training goal and a weighting scheme builds a coherent framework for user-behavior learning that is a step above simplistic click data User profiles created in this fashion have been shown to be quite stable and adaptable, suggesting that the approach does well in balancing being specific enough about user interests while being broad enough to allow weights and representations to evolve as interesting content changes. Such a balance is of great importance in news recommendations, where the state of the world itself can suddenly become a key factor in what interest users.



```
# --- Test Case ---
example_history = ['N55528', 'N39639', 'N61037'] # Example user clicked history
example_dwell = 0.2  # Example average dwell time

# Run both models
print("\n Generating recommendations...")

rec_with_dwell = generate_recommendations(
    t5_dwell_model, t5_dwell_tokenizer, example_history, news_dict, dwell_time=example_dwell
)

rec_without_dwell = generate_recommendations(
    t5_base_model, t5_base_tokenizer, example_history, news_dict, dwell_time=None
)

# Fetch news content
df_with_dwell = fetch_news_content(rec_with_dwell, news)
df_without_dwell = fetch_news_content(rec_without_dwell, news)

# Select relevant columns
df_with_dwell_display = df_with_dwell[['NewsID', 'Title', 'Abstract']]
df_without_dwell_display = df_without_dwell[['NewsID', 'Title', 'Abstract']]

# Print recommendations from T5 with Dwell Time
print("\n Recommendations from T5 WITH Dwell Time:\n")
print(tabulate(df_with_dwell_display, headers='keys', tablefmt='fancy_grid', showindex=False))

# Print recommendations from T5 without Dwell Time
print("\n Recommendations from T5 WITHOUT Dwell Time:\n")
print(tabulate(df_without_dwell_display, headers='keys', tablefmt='fancy_grid', showindex=False))
```

```
Generating recommendations...

Recommendations from T5 WITH Dwell Time:

╒══════════╤═══════════════════════════════════════════════════════════════════════════════════════════════════════╕
│ NewsID   │ Title                                                                                                 │
╞══════════╪═══════════════════════════════════════════════════════════════════════════════════════════════════════╡
│ N63939   │ Dean Foods Files for Bankruptcy                                                                       │
├──────────┼───────────────────────────────────────────────────────────────────────────────────────────────────────┤
│ N62360   │ The son of a Chinese billionaire has been banned from flying first class, playing golf, buying property, or going clubbing │
├──────────┼───────────────────────────────────────────────────────────────────────────────────────────────────────┤
│ N45820   │ Lamar Odom Is Engaged to Sabrina Parr: See Her Ring!                                                   │
├──────────┼───────────────────────────────────────────────────────────────────────────────────────────────────────┤
│ N32544   │ Hannah Brown on Being Surrounded By Exes Tyler Cameron and Colton Underwood                            │
├──────────┼───────────────────────────────────────────────────────────────────────────────────────────────────────┤
│ N7359    │ Cold weather: 70% of the country will experience freezing temperatures this week                      │
╘══════════╧═══════════════════════════════════════════════════════════════════════════════════════════════════════╛

Recommendations from T5 WITHOUT Dwell Time:

╒══════════╤═══════════════════════════════════════════════════════════════════════════════════════════════════════╕
│ NewsID   │ Title                                                                                                 │
╞══════════╪═══════════════════════════════════════════════════════════════════════════════════════════════════════╡
│ N63939   │ Dean Foods Files for Bankruptcy                                                                       │
├──────────┼───────────────────────────────────────────────────────────────────────────────────────────────────────┤
│ N45820   │ Lamar Odom Is Engaged to Sabrina Parr: See Her Ring!                                                   │
├──────────┼───────────────────────────────────────────────────────────────────────────────────────────────────────┤
│ N32544   │ Hannah Brown on Being Surrounded By Exes Tyler Cameron and Colton Underwood                            │
╘══════════╧═══════════════════════════════════════════════════════════════════════════════════════════════════════╛
```

Figure 9: Testing both models

The entity embedding integration exemplifies the essential nature of bringing in external knowledge to recommendation processes. Although entity embeddings provide gains that are less important than those observed with T5 textual embeddings, the gains are ever-present and help make the system understand relations that pure textual analysis would not. This multi-modal approach to article representation points toward ways to improve it further by introducing other knowledge sources. The regression-based approach to recommendation represents a departure from traditional classification methods and proves particularly well-suited to the continuous nature of user engagement. By optimizing for predicted engagement levels rather than binary click prediction, our system learns more nuanced patterns in user behavior. This approach enables fine-grained understanding of user preferences that contributes to the superior ranking performance observed in our evaluation.

The computational efficiency achieved through FAISS indexing demonstrates that sophisticated recommendation approaches can be made practical for real-time deployment. The sub-50ms query times enable responsive user experience while the exact similarity computation ensures high-quality results. This combination of quality and efficiency is crucial for practical news recommendation systems where users expect immediate responses.

Our research also reveals several limitations that suggest directions for future work. The dwell time estimation formula, while effective, represents a simplified model of user engagement that could be improved through more sophisticated approaches. Direct measurement of actual reading time or integration of additional behavioural signals could provide more accurate engagement estimation. The reliance on textual content means that articles with very brief text or highly visual content may not receive optimal representations. Future work could explore integration of visual features or other content modalities to address these limitations.

5. Conclusions and Recommendations

This work convincingly demonstrates that integrating transformer-based semantic understanding with sophisticated user behavior analysis produces better-quality news recommendation systems solving fundamental issues of personalized content discovery. The extensive experimental comparison provides strong evidence that the suggested method achieves significant improvements over standard methods without sacrificing computational efficiency suitable for real-world application. The key findings establish some important contributions to the field of news recommendation. T5 transformer model is highly efficient in picking up semantic relationships between news articles that content-based approaches entirely miss. The embeddings of 768 dimensions pick up not only surface keywords but also deeper contextual relationships that enable the system to capture related content even if there are variations in vocabulary sets and writing styles. The dwell time estimation method is a practical advance addressing the issue of inferring implicit user feedback in the absence of full tracking infrastructure. The reading speed formula, while nominally simple, generative valuable engagement cues with substantial recommendation-quality gain when applied to build user profiles. This method shows that sophisticated user modeling is achievable by tractable techniques with attention to privacy concerns while delivering personalization values. The fusion of T5 embeddings with entity knowledge to merge information from various sources creates dense representations of articles that cover semantic content as well as external knowledge.

While entity embeddings provide incremental gains over text features, they provide consistent improvement and allow understanding of relationships that text analysis cannot handle. This multi-modal path leaves the door open for extension with additional knowledge sources and more sophisticated fusion methods.

Regression-based learning of user engagement activity patterns performs better than traditional binary classification methods in the case of news recommendations. End-to-end prediction of engagement levels rather than click prediction allows the system to train more complex patterns of user behavior that result in enhanced ranking performance. This approach allows for context-aware comprehension of user preference detailed enough to drive more effective personalization campaigns.

The FAISS-based retrieval system successfully resolves the computational efficiency concerns that make it impossible for transformer-based approaches to have realistic real-time usage. The ability to achieve sub-50ms response times without sacrificing accurate similarity calculation demonstrates the possibility of sophisticated recommendation techniques meeting realistic systems' performance requirements of servicing massive user groups. The comprehensive comparison on various metrics ensures substantive improvement over baseline approaches, with the improvement being statistically significant and practically beneficial. The consistency of improvements across different evaluation scenarios and user segments suggests that the benefits would generalize to real-world deployment conditions rather than being artifacts of the specific experimental setup. The research also reveals important limitations that suggest directions for future investigation. The dwell time estimation formula, while effective, represents a simplified model of user engagement that could benefit from more sophisticated approaches incorporating additional behavioral signals or direct engagement measurement. The reliance on textual content means that highly visual articles or content with limited text may not receive optimal representations, suggesting opportunities for multi-modal extensions.

The entity embedding integration, while beneficial, relies on pre-computed embeddings that may not capture the most current entity relationships. Future work could explore dynamic entity embedding updates or more sophisticated fusion mechanisms that better integrate textual and entity-based understanding. The categorical fallback mechanism for missing entity embeddings, while functional, could be enhanced through more sophisticated approaches to handling incomplete knowledge graph information. The user modeling approach, while effective, focuses primarily on content-based preferences and could be enhanced through incorporation of social signals, temporal preferences, or contextual factors that influence news consumption patterns. The static nature of user profiles could be improved through more dynamic updating mechanisms that better capture evolving interests and preferences.

Future research areas include several promising directions that one might wish to pursue to further build on the work presented herein. Advanced engagement modeling would incorporate more behavioral signals such as scroll patterns and time spent on different parts of the article, along with return visit patterns to arrive at more accurate user interest models. Machine learning methods for estimating dwell time could supplant the current formula-based approach with one that learns from data about engagement patterns. Multi-modal content understanding is another important area where visual elements, multimedia content, and interactive features could be merged with textual analysis to form a more holistic representation of articles. Computer vision could be used to interpret images and graphics accompanying news articles, and audio processing could integrate podcast or video content into the recommendation framework.

Social recommendations could also use user networks and social media signals for added personalization, all while implementing privacy safeguards. The demonstrated improvements in recommendation quality could contribute to better-informed public engagement with news content, while the privacy-respecting approaches to user modeling could help address concerns about surveillance and data collection in news consumption. The success of transformer-based approaches in news recommendation suggests broader applicability to other recommendation domains where semantic understanding of content is important. This research establishes a strong foundation for the next generation of news recommendation systems that can provide more accurate, relevant, and trustworthy content suggestions to users navigating increasingly complex information environments. The demonstrated feasibility of combining advanced natural language processing with practical deployment considerations offers a path forward for systems that can serve real users while maintaining the technical sophistication necessary for addressing the challenges of modern news consumption.

Acknowledgement

I would like to thank my guide, Milu Philip, for constant support, guidance, and encouragement offered during the process of conducting this research work. The valuable comments and feedback played a critical role in defining the direction and the quality of this work. I would also like to acknowledge the assistance of Grammarly which significantly helped me to polish my writing and presentation of this report.

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Castells, P., Hurley, N. J., & Vargas, S. (2015). Novelty and diversity in recommender systems. In Recommender systems handbook (pp. 881-918). Springer.

Chen, J., Zhang, H., He, X., Nie, L., Liu, W., & Chua, T. S. (2018). Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In

Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval (pp. 335-344).

Chen, C., Zhang, M., Liu, Y., & Ma, S. (2019). Social attentional memory network: Modeling aspect-and friend-level differences in recommendation. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (pp. 177-185).

Dettmers, T., Minervini, P., Stenetorp, P., & Riedel, S. (2018). Convolutional 2d knowledge graph embeddings. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Gunawardana, A., & Shani, G. (2015). Evaluating recommender systems. In Recommender systems handbook (pp. 265-308). Springer.

Guo, Q., Zhuang, F., Qin, C., Zhu, H., Xie, X., Xiong, H., & He, Q. (2020). A survey on knowledge graph-based recommender systems. IEEE Transactions on Knowledge and Data Engineering.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems (TOIS), 22(1), 5-53.

Hidasi, B., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2015). Session-based recommendations with recurrent neural networks. arXiv preprint arXiv:1511.06939.

Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In 2008 Eighth IEEE international conference on data mining (pp. 263-272). IEEE.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems (TOIS), 20(4), 422-446.

Kaminskas, M., & Bridge, D. (2016). Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. ACM Transactions on Interactive Intelligent Systems (TiiS), 7(1), 1-42.

Kang, W. C., & McAuley, J. (2018). Self-attentive sequential recommendation. In 2018 IEEE International Conference on Data Mining (ICDM) (pp. 197-206). IEEE.

Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8), 30-37.

Liu, J., Dolan, P., & Pedersen, E. R. (2010). Personalized news recommendation based on click behavior. In Proceedings of the 15th international conference on Intelligent user interfaces (pp. 31-40).

Liu, S., Xiao, Y., Liu, W., Zhou, X., & Zhao, P. (2021). A survey on neural news recommendation. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management (pp. 4090-4099).

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge university press.

Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. In The adaptive web (pp. 325-341). Springer.

Qiao, Y., Xiong, C., Liu, Z., & Liu, Z. (2019). Understanding the behaviors of BERT in ranking. arXiv preprint arXiv:1904.07531.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21, 1-67.

Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2009). BPR: Bayesian personalized ranking from implicit feedback. In Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence (pp. 452-461).

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th international conference on World Wide Web (pp. 285-295).

Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., & Jiang, P. (2019). BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In Proceedings of the 28th ACM international conference on information and knowledge management (pp. 1441-1450).

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., & Bouchard, G. (2016). Complex embeddings for simple link prediction. In International conference on machine learning (pp. 2071-2080). PMLR.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

Wang, H., Zhang, F., Xie, X., & Guo, M. (2018). DKN: Deep knowledge-aware network for news recommendation. In Proceedings of the 2018 world wide web conference (pp. 1835-1844).

Wang, S., Hu, L., Wang, Y., Cao, L., Sheng, Q. Z., & Orgun, M. (2021). Sequential recommender systems: challenges, progress and prospects. arXiv preprint arXiv:2001.04830.

Wu, C., Wu, F., An, M., Huang, J., Huang, Y., & Xie, X. (2019). NPA: Neural news recommendation with personalized attention. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2576-2584).

Wu, F., Qiao, Y., Chen, J. H., Wu, C., Qi, T., Lian, J., ... & Xie, X. (2020). MIND: A large-scale dataset for news recommendation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 3597-3606).

Wu, S., Tang, Y., Zhu, Y., Wang, L., Xie, X., & Tan, T. (2017). Session-based recommendation with graph neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 346-353).

Xia, L., Xu, Y., Huang, C., Dai, P., & Bo, L. (2021). Graph meta network for multi-behavior recommendation. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 757-766).

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. ACM Computing Surveys (CSUR), 52(1), 1-38.

Zheng, L., Noroozi, V., & Yu, P. S. (2018). Joint deep modeling of users and items using reviews for recommendation. In Proceedings of the tenth ACM international conference on web search and data mining (pp. 425-434).

Zhou, G., Zhu, X., Song, C., Fan, Y., Zhu, H., Ma, X., ... & Gai, K. (2019). Deep interest network for click-through rate prediction. In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 1059-1068).

Zhou, K., Wang, H., Zhao, W. X., Zhu, Y., Wang, S., Zhang, F., ... & Wen, J. R. (2020). S3-Rec: Self-supervised learning for sequential recommendation with mutual information maximization. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (pp. 1893-1902).

Sai Triveni_C00313481_Dissertation final report.pdf

ORIGINALITY REPORT

4%	2%	2%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ebin.pub Internet Source	<1%
2	dokumen.pub Internet Source	<1%
3	link.springer.com Internet Source	<1%
4	arxiv.org Internet Source	<1%
5	export.arxiv.org Internet Source	<1%
6	hal.archives-ouvertes.fr Internet Source	<1%
7	Submitted to Athlone Institute of Technology Student Paper	<1%
8	Submitted to Institute of Technology Carlow Student Paper	<1%
9	repositorio.uam.es Internet Source	<1%
10	Submitted to University of East London Student Paper	<1%
11	Victor Akihito Kamada Tomita. "Multi-Representation Learning for Domain Adaptation in Machine Learning Applications", Universidade de São Paulo. Agência de Bibliotecas e Coleções Digitais, 2025 Publication	<1%
12	patents.justia.com Internet Source	

		<1 %
13	Zhu, Haoren. "Applying Deep Learning to Domain Specific Problems", Hong Kong University of Science and Technology (Hong Kong) Publication	<1 %
14	www.researchsquare.com Internet Source	<1 %
15	Bogdan Walek, Patrik Müller. "A text-based recommender system for recommending relevant news articles", Expert Systems with Applications, 2025 Publication	<1 %
16	discovery.ucl.ac.uk Internet Source	<1 %
17	dspace.cvut.cz Internet Source	<1 %
18	Biyang Ma, Yiwei Deng, Huifan Gao. "PAD-MPFN: Dynamic Fusion with Popularity Decay for News Recommendation", Electronics, 2025 Publication	<1 %
19	Chaojun Xiao, Ruobing Xie, Yuan Yao, Zhiyuan Liu, Maosong Sun, Xu Zhang, Leyu Lin. "UPRec: User-aware Pre-training for sequential Recommendation", AI Open, 2023 Publication	<1 %
20	Wenhao Zhu, Yujun Xie, Qun Huang, Zehua Zheng, Xiaozhao Fang, Yonghui Huang, Weijun Sun. "Graph Transformer Collaborative Filtering Method for Multi-Behavior Recommendations", Mathematics, 2022 Publication	<1 %
21	www.mdpi.com Internet Source	<1 %

22	"Advances in Information Retrieval", Springer Science and Business Media LLC, 2018 Publication	<1 %
23	Anandhan, Anitha. "Recommendation of Experts in Community Question Answering Using Tag Relationship.", University of Malaya (Malaysia), 2024 Publication	<1 %
24	Bandreddy, Saadhika. "Using Sequential Multi-Behavior Product Features for E-Commerce Recommendation", University of Windsor (Canada), 2024 Publication	<1 %
25	Journal of Systems and Information Technology, Volume 15, Issue 3 (2013-11-02) Publication	<1 %
26	Soojin Kim, Patrice M. Buzzanell, Alessandra Mazzei, Jeong-Nam Kim. "The Routledge Handbook of Employee Communication and Organizational Processes", Routledge, 2025 Publication	<1 %
27	download.bibis.ir Internet Source	<1 %
28	edyaaleh.files.wordpress.com Internet Source	<1 %
29	www.ahievranconference.org Internet Source	<1 %
30	"Advanced Data Mining and Applications", Springer Science and Business Media LLC, 2020 Publication	<1 %
31	"Knowledge Science, Engineering and Management", Springer Science and Business Media LLC, 2019 Publication	<1 %

32	Duen-Ren Liu, Yu-Shan Liao, Jun-Yi Lu. "Online news recommendations based on topic modeling and online interest adjustment", Industrial Management & Data Systems, 2019 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		