Recommender of relevant stories
Project Proposal for NLP Course, Winter 2023

Ángela H. Jiménez and **Iñigo Barandiaran**
WUT
01190845@pw.edu.pl and 01190800@pw.edu.pl

Martin Goicoechea
WUT
01190817@pw.edu.pl

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

In the digital age, the deluge of narrative content poses a challenge for users seeking stories that resonate with their interests. This project examines the state-of-the-art in natural language processing (NLP) as applied to the development of a recommender system for relevant story selection. We explore the transition from traditional recommendation algorithms to sophisticated NLP techniques that leverage deep learning models such as word embeddings and transformer architectures. Our focus is on how these methods can discern and match the semantic content of stories with user preferences to enhance the personalization of recommendations. We address the inherent challenges in story recommendation, including contextual nuance, diversity of user preferences, and ethical considerations such as bias mitigation and privacy. The project anticipates the integration of advanced NLP techniques like transfer learning and the exploitation of large-scale datasets to improve recommendation accuracy and user satisfaction. The outcome will be a prototype model that encapsulates the cutting-edge in NLP-driven recommendation systems, offering a tailored story discovery experience.

1 Introduction

The surge of digital content has transformed the web into a vast ocean of stories, where users often find themselves adrift in information overload. Identifying content that resonates on a personal level can be a formidable task, which calls for more than just conventional filtering methods. This project presents a solution in the form of a sophisticated news recommender system, designed to navigate the complexities of individual preferences and the subtleties of narrative content.

The aim is to go beyond the surface, allowing the system to not just look at what's popular, but to understand and align with the user's unique interests. It leverages advanced NLP techniques to sift through the narrative nuances and deliver news that is not just newsworthy, but also personally relevant.

Acknowledging the complexities of language and diversity in user tastes, the project confronts these challenges head-on. It addresses concerns of fairness and privacy with an ethical framework built into its core. By harnessing robust datasets and innovative NLP strategies, the project seeks to refine the accuracy of recommendations and enhance user engagement.

The envisioned outcome is a tool that filters out the noise and zeroes in on stories that matter to the individual, redefining the experience of discovering and enjoying news in the digital age.

2 Literature Review

2.1 Statement of the art

In today's digital ecosystem, personalized content delivery is not just a luxury—it's essential. The transformation of recommendation systems is pivotal, shifting from basic algorithms to complex entities that leverage Natural Language Processing (NLP) to meet individual user demands. NLP's ability to parse and understand human language is crucial, enabling systems to recommend stories that resonate more deeply with users by analyzing content beyond mere keywords. This paper delves into how NLP enriches recommendation systems, enhancing the connection between users and the stories that captivate them, setting the stage for the next leap in content personalization.

2.1.1 Historical context

The genesis of recommendation systems can be traced back to the implementation of basic algorithms designed to mimic the personal touch of a human curator. Early systems employed collaborative filtering, harnessing the power of collective user behaviors and preferences to make suggestions. Users were recommended content based on the likes and dislikes of similar profiles, which, while effective in certain contexts, often missed the mark in understanding the individual's specific tastes.

Content-based filtering emerged to address this, recommending items by comparing the content of the products with a user profile. The focus shifted to the attributes of the items themselves, promoting stories with themes or genres that the user had previously enjoyed. However, this approach struggled with the nuance and depth of human language, often overlooking the contextual and sentimental subtleties inherent in storytelling.

These limitations paved the way for the integration of NLP. By understanding language at a granular level, NLP allows systems to capture the essence of content, bridging the gap between user preferences and content relevance. The synergy of collaborative and content-based methods with NLP has revolutionized recommendation systems, leading to a more sophisticated, personalized experience.

2.1.2 Advancements in NLP for Recommendation Systems

The landscape of NLP has undergone a remarkable transformation, particularly in its application to recommendation systems. Initial forays into NLP were dominated by rule-based systems that relied on manually crafted linguistic rules. However, these systems were rigid, struggled with the nuances of language, and failed to scale with the ever-expanding corpus of data.

The advent of machine learning ushered in a new era for NLP. Instead of relying on static rules, machine learning models learn patterns from data, enabling a more dynamic understanding of language. This shift was accelerated with the introduction of word embeddings, like Word2Vec and GloVe, which represented words in a continuous vector space, capturing semantic and syntactic meanings based on the context of their use.

Deep learning further revolutionized NLP by utilizing neural networks to process language data in complex ways. The development of transformer architectures, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), marked a significant milestone. These models use self-attention mechanisms to process words in relation to all other words in a sentence, drastically improving the ability to understand the intent and sentiment of text.

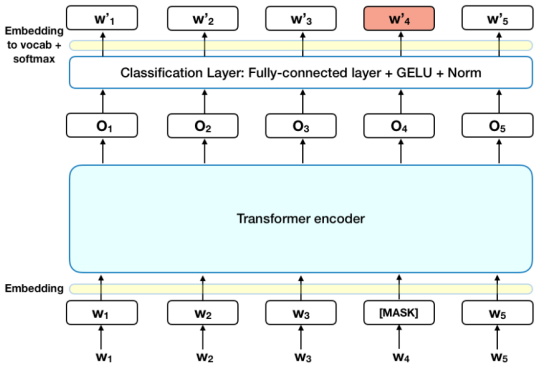

Figure 1: BERT structure.

In recommendation systems, these advancements mean algorithms can now consider a story's writing style, plot intricacy, character development, and thematic depth, providing recommendations that are not just topic-appropriate but emotionally and intellectually in tune with the reader's preferences. The integration of NLP into recommendation systems has therefore not only improved accuracy but also enriched the user experience by offering content that is contextually and emotionally resonant.

2.1.3 Personalization Techniques

Personalisation sits at the heart of modern story recommendation, where NLP plays a vital role. Current methods analyze user interactions—like browsing history and reviews—to fine-tune story suggestions. User embeddings, inspired by word embeddings, are key to this process. They convert complex user data into a vector space model, capturing the essence of individual preferences. This allows for nuanced recommendations that align with a user's unique taste, going beyond genre to match writing styles and thematic preferences. Sentiment analysis further enriches this personalisation, enabling a system to cater to both explicit and implicit user needs. Together, these techniques elevate the story-finding experience, making it as personalized as it is intuitive.

2.1.4 Semantic Analysis and Relevance Matching

Sophisticated NLP techniques are now integral to recommendation systems, enabling a nuanced understanding of story content. These systems go beyond mere keywords, analyzing context, sentiment, and deeper meanings. Tools like Latent Semantic Analysis and neural networks discern themes and emotional tones, which inform how stories are ranked and recommended. This allows for a more personalized selection, where stories are matched to a user's preferences not just in topic but in tone and depth, creating a resonant and contextually rich reading experience.

2.1.5 Challenges and Considerations

Despite remarkable progress, state-of-the-art recommendation systems face notable challenges. The nuanced nature of language means that detecting sarcasm or irony remains a complex task for NLP, potentially skewing content relevance. Diversity in the user base introduces variability that systems must accommodate to avoid homogenized recommendations. Data sparsity, especially with new users or less common stories, can hinder the system's ability to make accurate predictions, leading to the 'cold start' problem. Ethically, the potential for bias in algorithmic decisions is a concern, as is ensuring user privacy in an age where personal data is a valuable commodity. Addressing these challenges is crucial for advancing recommendation systems to be both sophisticated in their technology and responsible in their application.

2.1.6 Future Directions

The trajectory of NLP in story recommendation systems is set towards leveraging emerging technologies that promise to overcome current limitations and open new possibilities. Transfer learning, for example, stands at the forefront of these technologies, with its ability to apply knowledge from one domain to another, potentially ameliorating issues like the cold-start problem by using pre-trained models that require less user-specific data to make accurate recommendations.

Growing datasets and their diversity are expected to fuel the enhancement of recommendation systems, providing a richer soil for algorithms to learn from a broader spectrum of user interactions and story content. This expansion, coupled with more powerful computational resources, will enable the processing of this data at unprecedented speeds and complexity, making real-time, contextually aware, and emotionally resonant recommendations a tangible reality.

Furthermore, these advancements hold the promise of refining algorithms to better handle linguistic subtleties such as sarcasm, thus improving the semantic understanding of content. Ethical challenges, particularly bias and privacy, are likely to be addressed through more transparent algorithms and regulations that ensure the fair and respectful treatment of user data.

As we venture into this future, we anticipate recommendation systems that are not only more accurate but also fairer and more aligned with user expectations, heralding a new chapter in personalized storytelling.

2.1.7 Conclusions

The integration of NLP into story recommendation systems represents a confluence of linguistic acuity and technological advancement. While current systems adeptly navigate the complexities of language to deliver personalized content, they are not without challenges. The future, however, is bright with potential. Advancements in transfer learning, richer datasets, and more robust computational power are poised to address issues such as the cold-start problem, data sparsity, and linguistic nuances including sarcasm. Ethical considerations, particularly concerning bias and privacy, will remain at the forefront of technological evolution. As we advance, we envisage recommendation systems that are not only more sophisticated and responsive to the subtleties of human language but also ethically sound and transparent. This progression promises a new era of personalized storytelling, finely attuned to the preferences and emotional contours of its audience.

2.2 Relevant Datasets

In the burgeoning field of Natural Language Processing (NLP), the quest to develop sophisticated story recommendation systems is an endeavor that combines art with the precision of data science. An efficacious recommender system hinges on the quality and relevance of the datasets it is trained on. This are some expamples of the most used datasets:

The MovieLens 25M Dataset, while ostensibly designed for cinematic recommendations, offers a compelling starting point. This dataset provides a

multifaceted view of user preferences and behaviors through a vast array of ratings. The transposition of this data to the domain of story recommendation is predicated on the notion that narratives, irrespective of medium, share core attributes that resonate with audiences. The granular ratings encompassing myriad genres, themes, and storytelling techniques can be instrumental in discerning patterns and preferences that transcend the boundary between film and literature.

In a similar vein, the Book-Crossings Dataset emerges as an indispensable resource. Its explicit ratings of books embody a treasure trove of information regarding reader preferences. Although the dataset's density is relatively low, this characteristic engenders a dataset that is manageable and ripe for deep analysis. By leveraging this dataset, an NLP-based recommender system can be attuned to the nuances of reader engagement and satisfaction, facilitating the recommendation of stories that not only align with users' tastes but also have the potential to broaden their literary horizons.

Kaggle's Book Recommendation Dataset provides a more contemporary arena for model training. The dataset's modern compilation of book recommendations is a reflection of current reading trends and user interactions. This data is invaluable for NLP applications, as it can be used to train algorithms on both the content and context of user preferences, enabling the delivery of story recommendations with remarkable relevance and personalization.

Lastly, the Microsoft News Dataset (MIND) stands as a paradigmatic example of how digital interaction data can be harnessed for the recommendation of textual content. This dataset is particularly salient for story recommender systems owing to its origin from a news platform, offering a real-world application scenario with rich behavioral logs. The dataset facilitates a deep understanding of content consumption patterns, which is essential for any NLP-driven recommendation engine that aims to provide users with compelling and contextually relevant narrative content.

In conclusion, the utility of these datasets in crafting an NLP-based story recommendation system is multifaceted. Each dataset serves as a unique lens, offering insights into user preferences and behaviors across various narrative forms. When synergistically combined, they empower recommendation systems to not only comprehend the explicit preferences of users but also to infer latent needs and desires, thereby elevating the standard for personalized storytelling experiences in the digital age.

3 Solution Concept

For the solution concept of this project, we have searched different ways that we could choose for solving this task of recommender systems. On the one hand we have content based recommendation engine. Content-based recommender systems are a subset of recommendation systems, fine-tune suggestions for users by evaluating the inherent qualities and attributes of items. They specialize in comprehending the content of items and aligning it with user preferences. By scrutinizing features such as genre, keywords, metadata, and other descriptive elements, content-based recommender systems establish profiles for both users and items. This empowers the system to propose items with content traits akin to user preferences. Unlike collaborative filtering techniques that depend on users' past interactions, content-based systems function independently, rendering them especially valuable in situations where user history is limited or unavailable. Through this personalized strategy, content-based recommender systems assume a critical role in enhancing user experiences across various domains, encompassing the recommendation of movies, articles, and the guidance of users in product or destination selection.

3.1 Types of Recommender Systems

3.1.1 Content-Based

A content-based recommender system operates with data supplied by the user, either through explicit actions like providing ratings or implicit actions such as clicking on links. Using this data, the system generates a user profile, which serves as the foundation for making personalized suggestions. As the user continues to contribute input or engage with the recommendations, the system's accuracy steadily improves over time.

The concepts that we use when talking about content-based recommender systems are Term Frequency (TF) and Inverse Document Frequency (IDF), which are used in information retrieval systems and also content based filtering mechanisms (such as a content based recommender). They are used to determine the relative importance of an articleTF is simply the frequency of a word in a

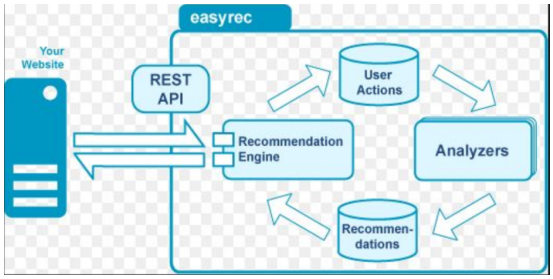

Figure 2: Content based structure.

document. IDF is the inverse of the document frequency among the whole corpus of documents. TF-IDF weighting negates the effect of high frequency words in determining the importance of an item (document).

3.1.2 Collaborative Filtering

Collaborative methods rely on measuring the similarity between users. This technique initiates by identifying a group of users, denoted as "X," whose preferences closely align with those of user A, forming A's neighborhood. Subsequently, items that are favored by the majority within X are suggested to user A. The effectiveness of a collaborative algorithm hinges on its ability to accurately pinpoint the neighborhood of the target user. Historically, collaborative filtering-based systems grapple with issues like the cold-start problem and privacy concerns due to the necessity of sharing user data. However, they excel in not requiring knowledge of item features for generating recommendations and can introduce users to new interests by unveiling novel items. Collaborative methods are further categorized into two types: memory-based and model-based approaches. Memory-based collaborative approaches offer recommendations by factoring in the preferences of a user's neighborhood. They directly utilize the utility matrix for prediction, commencing with model construction. The model is essentially a function that takes the utility matrix as its input.

Then recommendations are made based on a function that takes the model and user profile as input. Here we can make recommendations only to users whose user profile belongs to the utility matrix. Therefore, to make recommendations for a new user, the user profile must be added to the utility matrix, and the similarity matrix should be recomputed, which makes this technique computation heavy.

Memory-based collaborative approaches are again sub-divided into two types: user-based collaborative filtering and item-based collaborative filtering. In the user-based approach, the user rating of a new item is calculated by finding other users from the user neighbourhood who has previously rated that same item. If a new item receives positive ratings from the user neighbourhood, the new item is recommended to the user.

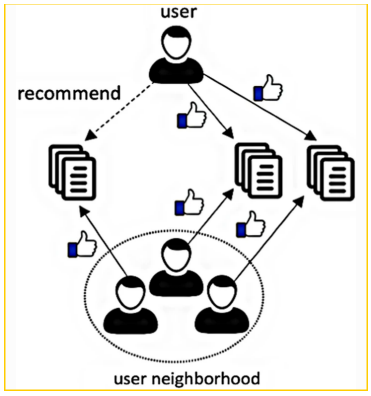

Figure 3: User-based collaborative filtering.

In the item-based approach, an item-neighbourhood is built consisting of all similar items which the user has rated previously. Then that user's rating for a different new item is predicted by calculating the weighted average of all ratings present in a similar item-neighbourhood.

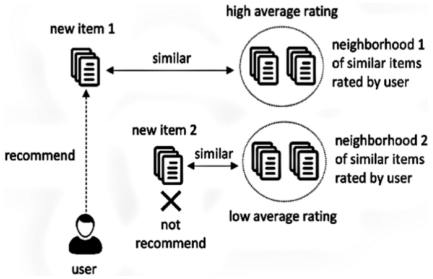

Figure 4: Item-neighbourhood filtering.

Lastly, we have the hybrid technique: This one is about an aggregation of two or more techniques employed together for addressing the limitations of individual recommender techniques. The incorporation of different techniques can be performed in various ways. A hybrid algorithm may incorporate the results achieved from separate techniques, or it can use content-based filtering in a collaborative method or use a collaborative filtering technique in a content-based method. This hybrid incorporation of different

techniques generally results in increased performance and increased accuracy in many recommender applications. Some of the hybridization approaches are meta-level, feature-augmentation, feature-combination, mixed hybridization, cascade hybridization, switching hybridization and weighted hybridization

3.2 Recommender System Challenges

3.2.1 Cold Start Problem

The cold start problem arises when the recommender system encounters insufficient data to draw meaningful inferences. This issue occurs when the system is unable to generate efficient recommendations for new users who have not provided ratings for any items or have only rated a limited number of items. Typically, the cold start problem occurs when a new user joins the system or when new items or products are added to the database.

Several solutions have been proposed to address this problem. Firstly, new users can be requested to explicitly specify their item preferences. Secondly, new users can be prompted to rate a selection of items at the outset. Lastly, demographic information or metadata can be collected from users to inform item recommendations.

3.2.2 Synonym Problem

This issue occurs when there are inconsistencies in the entries or names of similar or related items within a system, or when a single item is represented by multiple names. For instance, the terms ”babywear” and ”baby cloth” may refer to the same product. Unfortunately, many recommender systems struggle to differentiate between these variations, leading to a decrease in the accuracy of their recommendations.

To mitigate this problem, several methods are employed, including demographic filtering, automatic term expansion, and Singular Value Decomposition . These techniques aim to address the issue of item naming inconsistencies and improve the precision of the recommender system’s recommendations.

3.2.3 Shilling Attack Problem

This issue arises when an unscrupulous user deliberately assumes a false identity and infiltrates the system to manipulate item ratings dishonestly. This scenario occurs when the malicious user aims to artificially boost or diminish the popularity of certain items by introducing biases towards specific target items. Shilling attacks significantly undermine the trustworthiness and dependability of the system.

One potential solution to address this problem is to promptly detect and identify the attackers, and subsequently eliminate the fraudulent ratings and fabricated user profiles from the system.

3.2.4 Grey Sheep Problem

The grey sheep problem is a unique challenge observed in pure collaborative filtering methodologies, where the feedback provided by a user does not align with any existing user neighbourhood. Consequently, the recommender system encounters difficulty in accurately predicting suitable items for that particular user.

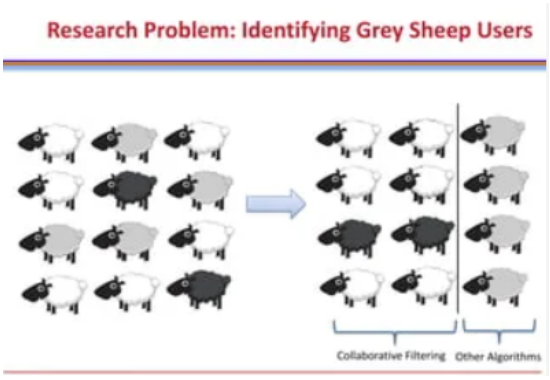

Figure 5: Grey Sheep problem.

To address this issue, a potential solution lies in employing pure content-based approaches, wherein predictions are made by leveraging the user’s profile and the inherent properties of the items themselves. By relying on these content-based techniques, the recommender system can mitigate the grey sheep problem and enhance its ability to generate relevant recommendations for users in such scenarios.

4 Project Proposal

4.1 Objectives

- Develop a robust recommendation engine that analyses user preferences and behaviour to suggest relevant stories.

- Build a scalable and efficient system capable of handling a large volume of users and stories.

- Provide a personalised user experience by tailoring recommendations to individual preferences and interests.

- Evaluate the performance of the recommender system using appropriate metrics and user feedback.

4.2 Methodology

The proposed methodology for building the relevant stories recommender system consists of the following steps:

4.2.1 Data Collection

For the proposed project, the data acquisition will be conducted through the Slovenian Press Agency (STA). This reputable source will provide a comprehensive dataset encompassing various elements such as publication times, categories, keywords, and authorship. Such detailed information will facilitate a robust analysis of the news content disseminated by this esteemed agency.

It is imperative to underscore the importance of ethical considerations and adherence to data protection regulations during the collection and utilization of this data. Meticulous attention must be paid to the data's integrity, ensuring that it is systematically categorized and sanitized for analysis purposes. This approach will not only ensure compliance with legal standards but also enhance the reliability and validity of any consequent research findings derived from this dataset. In order to do so an keep the data protection a NDA has been signed with the STA agency.

4.2.2 Data Preprocessing

Clean and preprocess the collected data to ensure data quality and consistency. This includes tasks such as text normalisation, removing duplicates, handling missing values, and addressing data quality issues.

4.2.3 Feature Extraction

Extract relevant features from the story data to represent them effectively for recommendation purposes. This may involve techniques such as text embeddings, metadata attributes and user behaviour signals.

4.2.4 User Modelling

Develop user profiles by analysing their historical interactions, preferences, and demographics. This involves creating user representations that capture their interests, preferences, and behaviour patterns. User modelling techniques may include collaborative filtering, matrix factorization, or deep learning models.

4.2.5 Recommendation Algorithms

Implement and evaluate various recommendation algorithms to generate personalised recommendations. This may include collaborative filtering, content-based filtering, and hybrid approaches that combine multiple techniques. The selection of algorithms will be based on their performance and suitability for the given problem.

4.2.6 Model Training and Validation

In the development of the recommendation engine, specific machine learning and optimization techniques tailored to the nuances of story data will be employed. The approach will include the use of Singular Value Decomposition (SVD) for matrix factorization, which is well-suited for decomposing large datasets into lower-dimensional representations while preserving the structure of user-story interactions.

Furthermore, we will experiment with deep learning architectures, particularly Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) cells, which are adept at processing sequential data and capturing temporal dynamics in user interaction patterns.

For optimization, we will implement stochastic gradient descent with momentum, an algorithm that is efficient for large-scale data and helps in avoiding local minima by considering the history of gradients in the optimization process.

To ensure the robustness of our models, we will train them on a comprehensive dataset comprising historical user interactions, enriched with metadata such as reading time and engagement metrics. Validation will be performed using a stratified cross-validation approach to maintain the distribution of stories across different folds, and a holdout set will be reserved for the final evaluation to mimic real-world performance.

4.3 System Components

The relevant stories recommender system will consist of the following key components:

4.3.1 Data Pipeline

Set up a robust and automated data pipeline for collecting, cleaning, and preprocessing the story

data. The pipeline will ensure regular updates to incorporate new stories and maintain data freshness.

4.3.2 Recommendation Engine

Implement the recommendation algorithms and integrate them into the system to generate personalised recommendations. The recommendation engine will process user profiles, story features, and historical interactions to generate relevant story suggestions.

4.3.3 User Interface

Design and develop a user-friendly interface where users can view and interact with recommended stories. The user interface will provide a seamless and intuitive user experience, allowing users to provide feedback, bookmark stories, and explore additional content.

4.3.4 Feedback loop

Incorporate user feedback mechanisms to gather explicit and implicit feedback on the recommended stories. This feedback loop will enable continuous improvement of the recommender system by incorporating user preferences and adapting to changing user interests.

4.4 Evaluation

To assess the performance and effectiveness of the relevant stories recommender system, the following evaluation methods will be employed:

4.4.1 Performance Metrics

Evaluate the recommender system using standard metrics such as precision, recall, mean average precision, and area under the curve (AUC). These metrics will measure the relevance and accuracy of the recommendations generated by the system.

4.4.2 A/B Testing

Conduct A/B tests to compare the performance of different recommendation algorithms or system variations. This will help identify the most effective algorithms or system configurations to maximise user engagement and satisfaction.

4.4.3 User Feedback

Gather user feedback through surveys, interviews, or feedback forms to assess user satisfaction and the relevance of the recommended stories. User feedback will provide insights into user preferences, identify areas of improvement, and guide future enhancements of the recommender system.

References

2019. *Machine Learning with PySpark*.

2022. *Recommender Systems Handbook*.

Natural language processing-based e-news recommender system using information extraction and domain clustering.