e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

CONTENT-BASED RECOMMENDER SYSTEM FOR AN ONLINE ADVERTISING PLATFORM

Pierre Mukisa Tshomba[*1], Md Samiul Islam[*2], Songjiang LI[*3]

[*1]Department of Computer Engineering, Mapon University, Kindu, Democratic Republic of Congo, ORCID: https://orcid.org/0009-0009-6715-9335

[*2]Department of Computer Science and Technology, Changchun University of Science and Technology, Jilin, China. ORCID: https://orcid.org/0009-0003-0572-2029

[*3]Professor, Department of Computer Science and Technology, Changchun University of Science and Technology, Jilin, China.

DOI : https://www.doi.org/10.56726/IRJMETS41650

ABSTRACT

This study proposes a content-based recommendation system utilizing Latent Dirichlet Allocation (LDA) topic modeling for a classified advertising website. The system dynamically adapts the user's profile based on implicit and explicit feedback by analyzing the user's past behavior with respect to the ads encountered. This approach ensures that the system can provide personalized recommendations that reflect the user's evolving preferences. An A/B test was conducted to evaluate the system's effectiveness, which showed a significant increase in key performance factors, including conversion rate, average session time, and the number of views. Overall, the study demonstrates the effectiveness of content-based recommendation systems using LDA topic modeling and emphasizes the importance of dynamic user profiling in personalized recommendation systems.

Keywords: Content-Based Recommendation System, LDA Topic Modeling, Classified Advertising Website, Dynamic User Profiling, A/B Test.

I. INTRODUCTION

Since its invention in the 1990s, the Internet has experienced explosive growth, revolutionizing various aspects of our societies. The Internet's profound impact is evident in how we communicate, work, and entertain ourselves. This transformative influence has been further amplified by significant technological advancements during the 2000s, leading to a remarkable explosion of data exchange on the World Wide Web.

In recent years, the Internet has become an integral part of nearly all daily activities. This significance has become even more apparent in the wake of the COVID-19 pandemic, where people increasingly rely on the Internet for remote work, online learning, and virtual socializing [1]. As technological advancements continue, the Internet's role in our daily lives will undoubtedly expand further. Prominent companies such as Google, YouTube, Facebook, and Wikipedia have become household names, dominating the online landscape and ranking among the largest internet service providers. These platforms attract billions of monthly visits, offering a vast array of services and products. As users navigate through these platforms, they are faced with an overwhelming number of choices, making it challenging to find items that align with their specific interests. Consequently, there is a growing need for filtering mechanisms to assist users in their decision-making process. Recommender systems play a pivotal role in addressing this challenge.

Recommender systems, in their various forms, refer to software applications and methodologies that offer personalized recommendations to users [2]. Leveraging user knowledge, item characteristics, and previous interactions, these systems provide relevant content and services tailored to individual preferences. In our study, we focus on developing a content-based recommender system for an online advertising platform. Our approach incorporates LDA (Latent Dirichlet Allocation) topic modeling to analyze the textual content of advertisements and make recommendations based on topic similarity. The foundation for recommendation systems can be traced back to the 1990s with the introduction of the mail filtering system, Tapestry [3]. Since then, recommendation systems have evolved into highly effective and widely used tools for information discovery on the Internet. They have revolutionized business operations by providing personalized recommendations to customers, leading to increased satisfaction and loyalty. Moreover, recommendation systems have found applications in various fields, including entertainment, education, and healthcare, with the aim of enhancing user experience and outcomes.

In the e-commerce sector, the ability of recommender systems to capture consumer interests and make targeted product or service recommendations provides companies with a sustainable competitive advantage. This advantage translates into increased views, interactions, and ultimately, revenue growth.

The advertising industry, which relies on the connection between consumers and product suppliers, has also experienced significant changes. The recent COVID-19 pandemic has further influenced consumer purchasing behaviors, leading to increased online activity and the necessity for companies and brands to adapt their marketing strategies [4]. As technology has advanced, online advertising has emerged as a crucial advertising form. Its diverse channels and formats, such as mobile advertising, search engine advertising, social media advertising, direct mail marketing, and display advertising, including web banner ads, have expanded the potential audience. Among the various types of online advertising, classified advertising has seen a steady rise in popularity, particularly in digital platforms, owing to its advantages in audience accessibility, cost-effectiveness, and post longevity. Leveraging the web experience, recommendation systems integrated into classified ad platforms allow customers to browse ads in a more personalized and convenient manner.

In this paper, we present our work on developing a content-based recommender system for an online advertising platform. Leveraging LDA topic modeling, our system analyzes the textual content of advertisements to identify underlying topics and recommend relevant ads based on topic similarity. By providing personalized recommendations, our system aims to enhance user engagement and satisfaction, ultimately creating business value for the advertising platform.

In the subsequent sections, we will discuss the methodology employed in developing our recommender system, elaborate on the LDA topic modeling approach, and present the experimental results and evaluation of our system's performance.

LDA AND TOPIC MODELING

Topic modeling is an unsupervised learning method used to identify the primary themes or topics that are present within a collection of documents based on the words that occur in them. Topic modeling is a technique that involves finding a set of words that most accurately captures the information contained in a collection of documents, through the use of unsupervised learning. Finding hidden patterns across the collection is more accessible with topic modeling. Documents are annotated based on these themes. These mentions can, afterward, be used to arrange, search, and summarize documents. Several articles have been published in the previous two decades on a variety of topic modeling methodologies, including LDA (Latent Dirichlet Allocation)

Latent Dirichlet allocation (LDA) is one of the most widely used techniques for topic modeling. It is an unsupervised clustering technique used for text analysis that can extract latent topics from a collection of documents [5]. The words used in a document can be associated with certain topics. The aim of LDA is to determine which topics a document belongs to by examining the terms it contains. The underlying assumption is that documents that cover related subjects will share the same vocabulary. LDA is based on a generative approach where words within documents are generated through a probabilistic process. The underlying topics of these documents are then identified as distributions of the words in a predefined vocabulary V using a bag-of-words representation. In this approach, the topics and words are considered as hidden and observable random variables, respectively. Once the generative process is established, the joint distribution can be defined, and statistical inference can be applied to generate probability distributions of the latent variables conditioned on the observable variables. In the case of LDA, the hidden topics in the corpus are represented as multinomial distributions of words, and the documents in the corpus are modeled as Dirichlet prior distributions over the latent topics. Let θ as a set of joint distributions for a corpus of D documents The Dirichlet distribution is given by:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta^{i-1}$$

Equation 1: Latent Dirichlet distribution

And, the joint distribution over θ and the set of N topics Z is given by:

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)$$

Equation Error! No text of specified style in document.**:** LDA joint distribution

Where α is the K-dimensional hyperparameter, and it remains constant for all the documents present in D. β is also a multinomial hyperparameter used in generating word distributions over all the vocabulary words in V.

II. METHODOLOGY

The proposed methodology in this study is founded upon the computation of similarity between the textual descriptions of advertisements in the database and the inferred profile of the active user. This method eliminates the need for manual input by relying on the system's automatic inference of the user profile. By quantifying the similarity between the advertisement and the user profile, accurate predictions can be made regarding the user's preferences for advertisements.

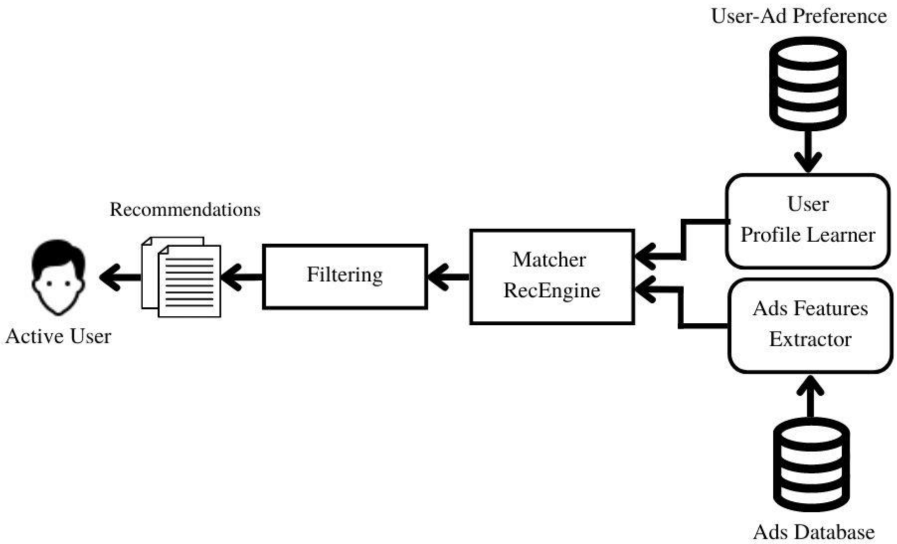

Figure 1: Overall architecture of the proposed approach

The novel aspect of the proposed methodology lies in its approach to analyzing user preferences. Rather than depending on manual input, the system generates a profile for each user based on their past activities and interactions with ads. In certain cases, recommendations are determined by identifying ads that align with the user's demonstrated interest and searching for similar ads. Alternatively, the system may explore the database for advertisements resembling the one currently being viewed by the user. This approach allows for a dynamic and personalized recommendation process that adapts to individual user behavior. By leveraging the automated inference of user profile and employing similarity computation, the proposed methodology presents an innovative approach to predicting user preferences in the realm of advertisements.

Data collection

In order to generate personalized recommendations for the active user, a certain amount of information must be gathered beforehand. This information is typically acquired through an analysis of the user's past activities and preferences. When constructing a user profile, ADrecsys takes into account both explicit and implicit feedback. Implicit feedback is derived from the user's monitoring activities, while explicit feedback is obtained from likes and alerts. Explicit feedback contributes to the development of a clear and precise user profile, while implicit feedback provides further insights into the user's habits and preferences. By combining explicit and implicit feedback, ADrecsys is able to construct an effective user profile, which can be utilized to personalize advertisements and content. As illustrated in Figure 1. of the architecture, our approach relies on two types of information.

- Ad information

In order to comprehensively depict the advertisements, a range of information is necessary. This information encompasses details such as the ad's content, title, category, author, publication date, geographic location, and other relevant criteria.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

- User-Ad interactions

This data aggregates implicit and explicit feedback on previously visited ads by the user. User explicit likes or dislikes of a particular advertisement combined with data implicitly collected of the user's past behavior and activities indicates how interested the user is in a specific advertisement.

Feature Extraction for Profile Building

ADrecsys employs natural language processing (NLP) techniques to extract structured and relevant information features from the unstructured textual data found in each advertisement's document [6]. Before training the model responsible for generating user profiles, the documents containing advertisement descriptions and other important attributes undergo a preparation step called Data cleaning and Preprocessing. This step involves lowering text and removing noise from the documents, such as punctuation marks, non-alphanumeric characters, and more importantly stop words. Other operations like tokenization and lemmatization are performed to break down the documents into their component words and to keep only the roots of the words taking into account the context. By performing this preprocessing, ADrecsys better understands document content, reduces document size, and optimizes the search algorithm. Consequently, computational costs for each profile are reduced, enabling more efficient resource utilization and faster response times when retrieving results from the model.

Profile building

Multiple approaches can be employed in the process of constructing a user profile. In the scope of this study, a particular approach has been adopted, which involves the representation of the user's profile as an LDA vector. This vector captures distinct topics inferred from the advertisements that have piqued the user's interest. The Latent Dirichlet Allocation (LDA) method leverages the user's history of liked advertisements to deduce a set of latent features that constitute the user's profile.

Table 1: LDA Vector Example

User ID	Feature 1	Feature 2	Feature 3	Feature 4	Feature...
1	0.035	0.478	0.008	0.801	....

III. MODELING AND ANALYSIS

The ADrecsys classifieds platform is constructed upon a Python-based backend, which utilizes the Flask framework to expose its functionalities through an API. Complementing this backend is a frontend application, developed in PHP with the Laravel framework, which leverages the obtained endpoints to facilitate seamless interaction with users. This architectural design ensures a cohesive system where the backend processes and provides data through the API, while the frontend, built with PHP and Laravel, efficiently communicates with the users, ensuring a smooth and interactive user experience.

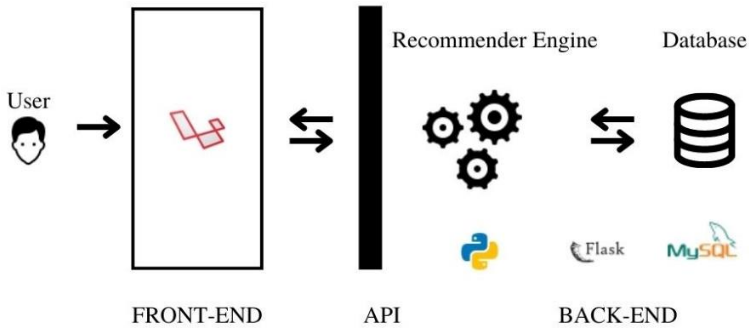

Figure 2: Technical architecture

Data Model

Figure 3. showcases the simplified data model pertaining to the entities that the recommendation engine operates with. It presents a focused representation of the entities directly utilized by the recommendation engine, omitting other components such as user management. This model offers a concise overview of the

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

essential entities involved in the recommendation process, highlighting the core elements relevant to the functioning of the recommendation engine.

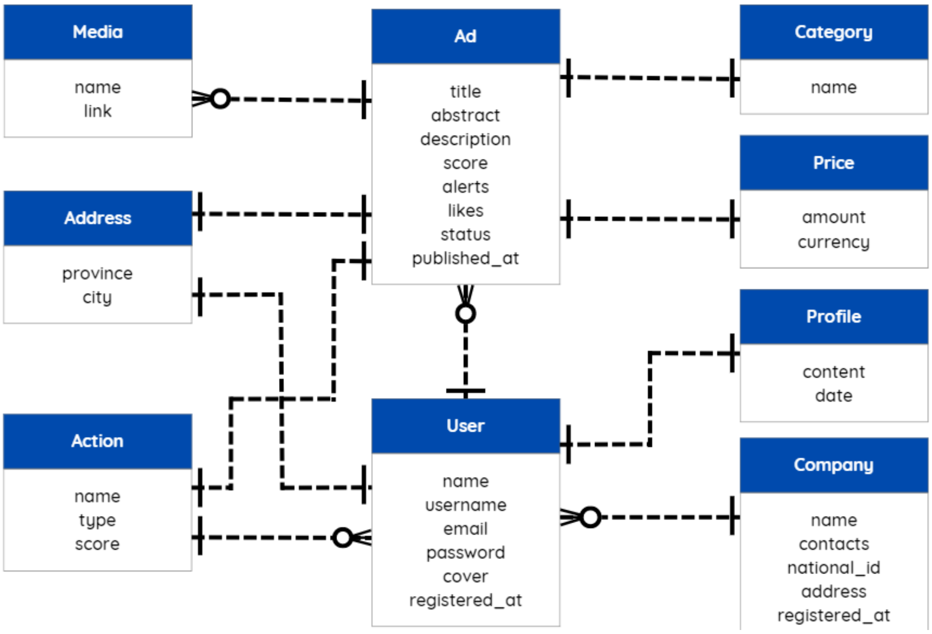

Figure 3: Data Model

Recommendation Algorithm

Figure 4. presents the key workflows encompassed within the ADrecsys Recommender System, depicted from the user's perspective. This illustration illustrates the system's functionalities available to the active user. With the necessary permissions, the user can publish new advertisements, contributing to the expansion of the platform's content. Additionally, the user has the ability to provide explicit or implicit feedback on posted advertisements. This valuable feedback plays a crucial role in refining the user's profile, enabling the system to generate personalized recommendations that align with the user's preferences and interests. By incorporating user feedback, the system continuously enhances its understanding of the user's preferences, thereby delivering tailored recommendations that cater to their specific needs.

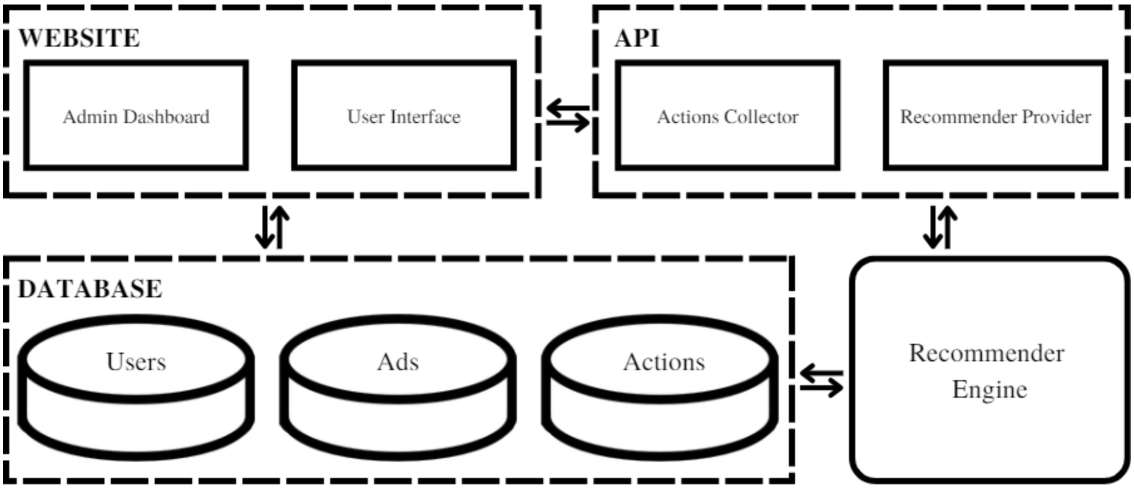

Figure 4: Workflow Architecture

The implemented ADArecsy recommendation system is based on the LDA topic modeling and cosine similarity. The recommendation engine has been implemented as a function, available on an API endpoint, which can be invoked by passing the active user's ID. The function can also be called by providing it with a user profile. The algorithm creates or updates the user's profile using a previously trained LDA model. The ads similar to the profile are collected, filtered, and finally sent to the user.

```
recommended_ads = get_recommendations_by_user_id(user_id = 1, limit= 10)
print(recommended_ads)
```
✓ 0.3s

```
[768, 45, 1819, 1857, 366, 851, 769, 180, 440]
```

Figure 5: Recommendation Function

To achieve the aforementioned outcome, several sequential steps were undertaken. The subsequent sections explain these steps in detail.

- Content analysis

The primary objective of this crucial stage is to convert the unprocessed textual data obtained for the descriptive information of the advertisement into a format that can be readily understood by machines, such as a vector. In the context of this research, solely the textual attributes of the advertisement, including the title, category, description, abstract, province, and city, are utilized in the recommendation procedure. The model's training process involves utilizing a dataset consisting of a total of 2000 entries. It is important to note that this dataset has undergone meticulous pre-processing procedures, as demonstrated in Figure 6, to ensure optimal quality and compatibility with the model's requirements.

```
def clean_text(text):
    if isinstance(text, str):
        t = text.lower()
        t = remove_tags(t)
        t = expand_contractions(t)
        t = remove_unnecessary(t)
        t = remove_stop_word(t)
        t = lemmatize_text(t)
        t = unique_words(t)
        return t
    return ""
```

Figure 6: Text cleaning function

The above function cleans and organizes the raw data into a single document. The document is initially converted to lowercase. The HTML and CSS tags are removed. Some contractions are extended for more clarity. Emojis and stop words are removed as they do not add much value to the document. The remaining words are lemmatized to keep only the roots of the words taking into account the context. The result is shown in Figure 6.

```
document = "🥳🥳🥳 Christmas promo going on right now! Our On-Going PROMO Includes: Buy 2 smartphone units and get 1 unit for free. Tecno Phantom X2 Pro is a great device to opt for if you are into new-age mobiles. This gadget provides features that you can use for both entertainment and professional work. The display of this smartphone is seamless and is teamed up with a great camera profile. Nord-kivu Steven_shop 14_mobile_phone. Follow us by clicking <a href = 'https://stevenshopgoma.cd'> on this click </a> 👍👍"

cleaned_document = clean_text(document)

print(cleaned_document)
```
✓ 0.2s Python

```
christmas promo go right ongoing include buy 2 smartphone unit get 1 free tecno phantom x2 pro great device opt newage mobile gadget provide feature use entertainment professional work display seamless team camera profile nordkivu stevenshop 14mobilephone follow us click
```

Figure 7: Result of a common ad document cleaning

After the content has been cleared of ads, The following phase is tokenization using TF-IDF (Term Frequency - Inverse Document Frequency). The tokenization procedure in this study enables us to break down documents into their component words. During the tokenization process, a TFIDF matrix is generated and used as input for subsequent steps. Several parameters are employed during this process, including the analyzer, which determines whether features are composed of word or character n-grams, as well as the min_df parameter, which sets a threshold for ignoring terms with document frequency below a certain level during vocabulary creation. The max_df parameter, on the other hand, sets a threshold for ignoring terms with document frequency above a certain level. Additionally, the ngram_range parameter specifies the range of n-values for different n-grams to be extracted, while stop_words lists the terms that will not be used when developing the vocabulary.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

- Creating user profile using LDA

In conjunction with the TF-IDF technique, the Latent Dirichlet Allocation (LDA) algorithm plays a pivotal role in constructing comprehensive user profiles that effectively capture the diverse interests and preferences of individual users. Leveraging the sparse TF-IDF matrix obtained in the previous phase, the LDA model generates a matrix that establishes associations between each advertisement and the underlying latent topics, as depicted in Figure 7. Notably, every advertisement is assigned a score for each identified topic. This model is further utilized to discern the latent topics present in the ads that the user has viewed and demonstrated interest in, enabling the system to gain deeper insights into the user's preferences. Consequently, the system becomes adept at suggesting similar items, augmenting personalized recommendations, and generating a comprehensive user profile.

```
LDA = LatentDirichletAllocation(n_components=15, random_state=42)
topic_results = LDA.fit_transform(corpus_vector)
print(type(topic_results))
✓ 3.1s

<class 'numpy.ndarray'>
```

Figure 8: LDA model learning process

Two parameters were supplied when the model LDA was instantiated: the n_components: number of topics and the random_state: for consistent results across several function calls. These parameters can be fine-tuned for even more precision. In this study, the pyLDAvis module is employed to facilitate the visualization of the outcomes obtained from the Latent Dirichlet Allocation (LDA) model, as illustrated in Figure 8.

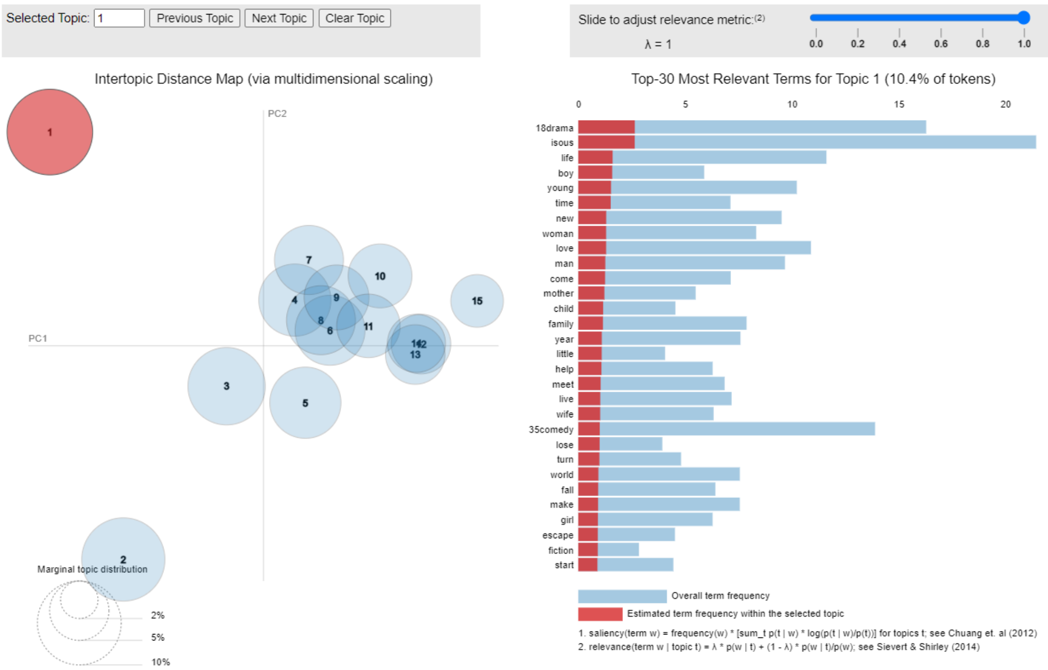

Figure 9: pyLDAvis for topics visualization

- Generating recommendations

This step takes place in two stages: The LDA model, trained earlier, first creates the user's profile, as shown in Figure 9, from the list of advertising that the user has found interesting.

```
def get_user_profile(user_id):
    user_ads = get_user_ads_list(user_id)
    if not user_ads.empty:
        document = ""
        for document in user_ads['document']:
            document += document
        profile_vector = tfidfvectorizer.transform([document])
        profile_topics_results = LDA.transform(profile_vector)
        return profile_topics_results
    return None
✓ 0.0s
```

Figure 10: User's profile generation

Once the user's profile is obtained, the system can proceed with the recommendation. The system calculates the degree of similarity between the user's profile and the published ads, then sorts them accordingly, as described in Figure 10.

```
def get_recommendations_by_user_id(user_id, limit = 20):
    user_profile =  get_user_profile(user_id)
    if len(user_profile):
        similarity_matrix = cosine_similarity(topic_results, user_profile)
        similarity_scores = list(enumerate(similarity_matrix))
        similarity_scores = sorted(similarity_scores, key=lambda x: x[1], reverse=True)
        similarity_scores = similarity_scores[1:limit]
        ad_indexes = [i[0] for i in similarity_scores]
        return ad_indexes
    return[]
✓ 0.0s
```

Figure 11: Recommendation function

- Rendering recommendations

Following the retrieval of recommendations through an API call, the client application, specifically a website in our context, leverages the acquired data to present users with advertisements that align closely with their specific requirements and match their individual profiles. The optimal approach for presenting these recommendations to users involves creating a pleasing and efficiently organized visual interface.

The primary aim is to ensure that users can easily locate the desired information without any unnecessary complexity or confusion. By prioritizing a user-friendly design and intuitive layout, the client application strives to provide a seamless and straightforward browsing experience, enabling users to swiftly access the relevant information they seek.

In this study, two ways of displaying recommendations have been explored:

On the classified website homepage of ADrecsys as shown in Figure 12, the recommendations prominently displayed are meticulously tailored to cater to the interests and preferences of individual users, taking into account their profiles. This personalized approach to recommendations not only enhances the overall user experience but also fosters a sense of loyalty among users. By showcasing items that are highly likely to captivate the user's attention, personalized recommendations contribute to a more enjoyable browsing experience, enticing users to explore the website further.

Moreover, by ensuring that the recommendations are consistently up-to-date and genuinely helpful, users gain valuable insights into the offerings of the website, thereby motivating them to engage more actively and comprehensively with the platform.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

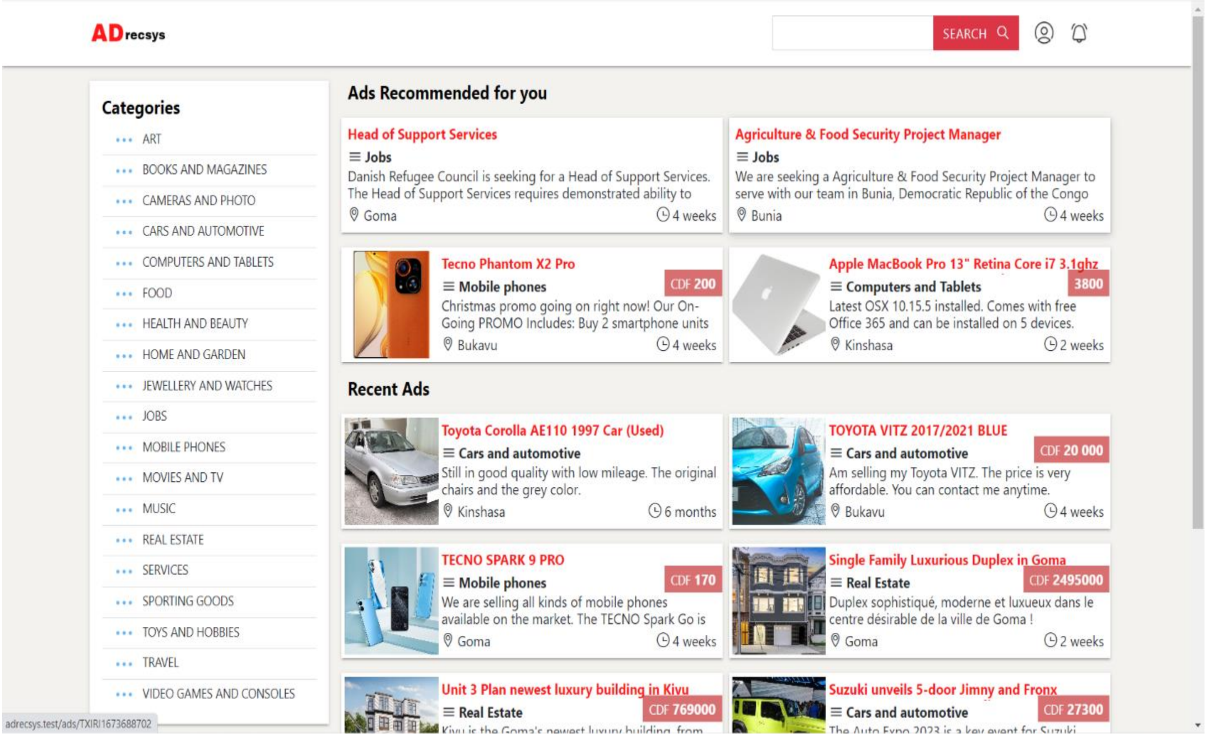

Figure 12: Homepage Recommendation

Within the scope of this study, the second approach of presenting advertisements investigated is through a dedicated single ad page. In this particular scenario, the recommendation process is not only reliant on the user's profile but also takes into consideration the specific advertisement that the user is currently viewing. By incorporating this contextual information, users are provided with a seamless browsing experience that facilitates the discovery of other products or ads that bear similarity to the one they are engaged with. The availability of related and relevant recommendations in close proximity to the user's current focus enhances their ability to navigate the platform, discover additional offerings, and prolong their engagement, thus positively impacting their browsing session's duration.

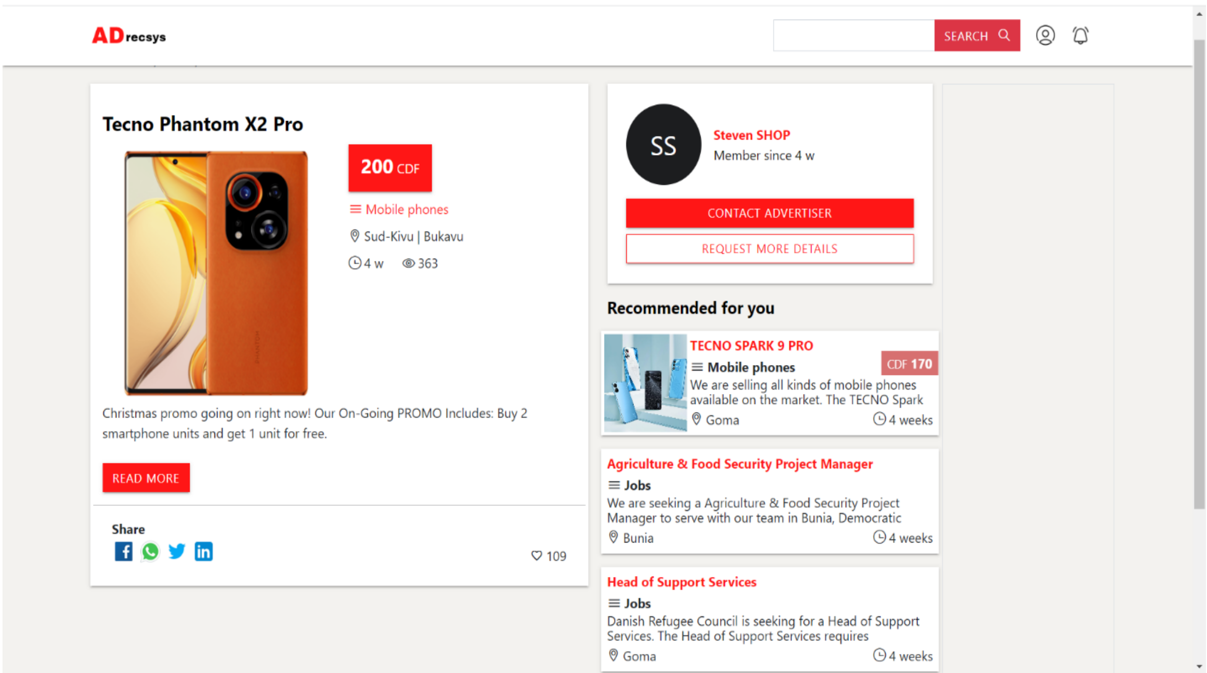

Figure 13: Single Ad Page Recommendation

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

IV. EXPERIMENTAL RESULTS

After the design and building process, ADrecsys was integrated into an existing platform named julisha.cd. This platform currently has 54 active businesses and enterprises, 1334 registered users with accounts, and 3669 published advertisements. As a result of this integration, this section focuses on evaluating this recommendation system's performance based on usage measures. The data collection for this study ran for ten consecutive weeks, from December 8, 2022, to February 9, 2023. During this time, ADrecsys gathered substantial information regarding user behavior about the recommended items.

In the context of this research, A/B testing is used to measure the relevancy and effectiveness of the suggested system. The A/B testing strategy compares the performance of a new model to that of an older model or a baseline model. Users are randomly split into two or more groups. One group will be shown recommendations from the new model, and the others will see recommendations from the old model. Afterward, the results of the two are compared and analyzed to measure the difference in user engagement or other relevant metrics. This approach is beneficial as it allows measuring a new model's impact on users and making changes accordingly.

The sections that follow highlight some of the key performance indicators (KPIs) that have witnessed significant improvements, such as conversion rate, average session length, and Bounce Rate. It is crucial to note that the test will not be performed on the entire population but on a sample estimated by Power analysis based on some factors such as the Power of the test, the alpha value, and the Effect size. The Power analysis aims to ascertain the optimal sample size required to obtain statistically significant results that are representative of the population. This approach saves time and resources while still providing reliable data.

- Conversion rate

First, a hypothesis must be formulated at the beginning of the evaluation. This will ensure the rigor and accuracy of our interpretation of the results.

H_0 = There is no difference in conversion rate between the control and experiment groups.

H_1 = The experiment group has a higher rate of conversion than the control group.

We also set the significance level $\alpha = 0.05$. System performance collected prior to the launch of the evaluation indicates an average annual conversion rate of 11%. The improvement in conversion rate could potentially lead to an increase in user engagement and satisfaction and ultimately in business revenue, making it a valuable addition to the system. However, if the new model fails to meet this target, further analysis and adjustments may be necessary to improve system performance.

	conversion_rate	std_deviation	std_error
group			
control	0.111	0.314	0.007
experimental	0.133	0.340	0.008

Figure 14: Conversion rate basic statistics: mean, standard deviation, and standard error

Previous statistics indicate an increase in conversion rate; however, additional analysis is required to evaluate whether the difference in conversion rates is statistically significant and whether the recommendation system is responsible for the higher conversion rate in the experimental group. The final part of the analysis is to test the hypothesis. We can calculate the p-value using the normal approximation given that we have a large sample size (i.e., the z-test) as demonstrated in Figure 14.

```
p-value: 0.031
ci 95% for control group: [0.097, 0.124]
ci 95% for experimental group: [0.118, 0.148]
```

Figure 15: P-value and confidence interval of the control and experimental group

Given the obtained p-value of 0.031, which is significantly lower than the predetermined level of significance alpha = 0.05, the null hypothesis is rejected. Therefore, it can be concluded that there is a statistically significant difference between the two groups in terms of the conversation rate. This means that the new recommendation system performed significantly different and better than our old recommendation model. This finding suggests

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

that implementing the new recommendation system could lead to an increase in user satisfaction and engagement.

Further research can be conducted to explore the potential impact of this change on other aspects of the advertisement business and revenues.

- Average session duration

The average session duration reflects the average time spent per session on a website. It is an important metric for understanding user engagement and behavior, as it indicates how effectively users utilize the website's content and features. A longer average session duration typically indicates that users are finding the website content engaging and valuable, while a shorter duration may suggest that users are not finding what they are looking for or are experiencing technical difficulties. It is important to analyze this metric in conjunction with other metrics, such as bounce rate and conversion rate, to gain a comprehensive understanding of user behavior on the website. The mean session duration can be determined by dividing the aggregate duration of all sessions by the total number of sessions.

Week #	Control Group ASD[sec]	Experimental Group ASD [sec]
1	150	162
2	126	184
3	166	182
4	139	143
5	98	124
6	117	167
7	145	158
8	160	144
9	156	191
10	158	189

Graph

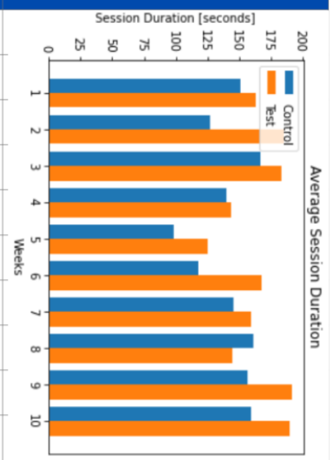

Figure 16: Evolution of the Average Session Duration during the experiment period

The average session duration grew from 141.5 seconds to 164.4 seconds towards the end of this test period, as depicted in Figure 17.

group	asd	std_deviation	std_error
control	141.5	20.61	6.87
experimental	164.4	21.37	7.12

Figure 17: Average session duration standard deviation and error

In order to ascertain the statistical significance of the observed difference, a two-sample test was conducted, which yielded a p-value of 0.033. Given that this value is below the pre-determined level of significance of 0.05, it can be concluded that there exists a statistically significant difference between the two groups with respect to the duration of the sessions, as depicted in Figure 5.7. This suggests that the two groups have different engagement levels or preferences when it comes to session duration, and the recommender systems applied, based on these experiments, explain the underlying reasons for this difference.

- Page views

The Number of Visitor Pageviews keeps count of the times a page has been viewed or refreshed by a user. The number of visitor page views tracks how often a user has viewed or refreshed a page. This KPI is essential for understanding a website's success, as it can be used to track user engagement and preferences.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:06/June-2023 Impact Factor- 7.868 www.irjmets.com

Week #	Control Group Pageviews	Test Group Pageviews	Graph
1	1596	3825	
2	2737	4542	
3	3355	5821	
4	2109	5687	
5	2142	4952	
6	6494	6907	
7	8232	8198	
8	6998	8384	
9	4428	6891	
10	5647	7289	

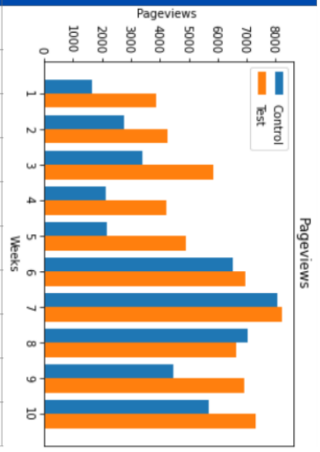

Figure 18: Evolution of the page views during the experiment period

The most notable aspect of Table 5.8 is the increased trend of page views across the ten-week experiment period. The average page views numbers went from 4353.8 to 6249.6 towards the end of this test period, as depicted in figure 5.9.

group	asd	std_deviation	std_error
control	4353.8	2189.24	729.75
experimental	6249.6	1457.89	485.96

Figure 19: Average page views standard deviation and error

To assess the statistical significance of the disparity, a two-sample test was conducted, yielding a p-value of 0.044, which is lower than the 0.05 level of significance. This result suggests that there exists a statistically significant difference between the two groups in regard to page views.

V. CONCLUSION

In recent years, the explosion of digital content has created challenges for users in navigating the vast amount of information available online. To address this issue, recommendation systems have emerged, utilizing machine learning algorithms to analyze user data and provide personalized content suggestions. Content-based recommendation systems, specifically, leverage item features to generate recommendations that align with users' previous interactions.

One effective approach to constructing a content-based recommendation system involves utilizing topic modeling algorithms like Latent Dirichlet Allocation (LDA) to infer user interests and preferences. LDA is a statistical algorithm that examines a collection of documents and identifies latent topics that best describe their content. By applying LDA to users' interaction histories, relevant topics can be extracted to generate tailored recommendations.

This paper presents a content-based recommendation system that employs the LDA algorithm to infer user profiles. The system was integrated into an existing classified advertising platform and evaluated through real-world data over a 10-week period using A/B testing. The results demonstrated the system's superiority over the existing approach with statistical significance.

The successful implementation of this system highlights the practical application and value of employing machine learning algorithms in recommendation systems. By analyzing text data and identifying latent topics, the LDA algorithm exhibits substantial potential for enhancing recommendation system accuracy and effectiveness. Furthermore, the system's integration into an established web platform underscores the flexibility and adaptability of machine learning algorithms, enabling seamless integration into existing systems to augment functionality.

As technology continues to advance, recommendation systems are poised to become more sophisticated, utilizing advanced machine learning algorithms to analyze user behavior and generate highly personalized recommendations. The LDA algorithm represents just one example among numerous techniques that can be employed to develop content-based recommendation systems. With ongoing research and development, these systems are now indispensable components of online content delivery, empowering users to efficiently navigate vast amounts of information and discover the most relevant content tailored to their needs.

VI. REFERENCES

[1] A. Feldmann et al., "Implications of the COVID-19 Pandemic on the Internet Traffic," Broadband Coverage in Germany; 15th ITG-Symposium, online, 2021, pp. 1-5

[2] Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender Systems Handbook. Springer.

[3] Goldberg, D., Nichols, D., Oki, B., & Terry, D. (1992). Using Collaborative Filtering to Weave an. Information Tapestry. ACM.

[4] Mehrbakhsh, N., Abumalloh, R. A., Alghamdi, A., Behrouz, M.-B., Abdulaziz, Mohammed, T., Sarminah, S. (2021). What is the impact of service quality on customers' satisfaction during the COVID-19 outbreak? , Telematics and Informatics.

[5] Hamed, J., Yongli, W., Chi, Y., & Xia, F. (2019) Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey, Multimedia Tools and Applications, abs/1711.04305.11: 15169.0-15211.

[6] Julia, h., & Christopher d, m. (2015) Advances In Natural Language Processing, Science, 349.6245: 261-6.