Araştırma Makalesi/Derleme Makalesi

Personalized News Recommendation System

Melis ÖZKARA[†], Metin TURAN[††]
[†] İstanbul Ticaret Üniversitesi, Bilgisayar Mühendisliği, İstanbul, Türkiye
[††] İstanbul Ticaret Üniversitesi, Bilgisayar Mühendisliği, İstanbul, Türkiye
melis.ozkara@istanbulticaret.edu.tr, mturan@ticaret.edu.tr
0000-0001-9655-2886, 0000-0002-1941-6693

Atıf/Citation: ÖZKARA, M., TURAN, M., (2022). Personalized News Recommendation System, Journal of Technology and Applied Sciences 5(2), 123-133, DOI: 10.56809/icujtas.1193993

ÖZET

Öneri Sistemleri, kullanıcının daha önce yapmış olduğu tercihlere dayalı olarak, kullanıcının bir sonraki tercihlerini öngörülebilir bir şekilde öneren yöntemlerdir. Bu yöntem günümüzde daha da popüler hale gelmiştir ve eldeki verileri değerlendirerek geleceğe yönelik tahmin gerektiren herhangi bir konu veya alana uygulanabilir. Bir tür bilgi çıkarma çalışmasıdır. Ayrıca Amazon'un gelirinin yaklaşık %35'ini yönlendirme sistemlerinden elde etmesi bu yöntemin ne kadar önemli olduğunun bir göstergesidir. Ancak benzer bir uygulama alanı olan haber tavsiye sistemi de diğerleri kadar yaygın olarak kullanılmamaktadır. Bu çalışmada, kullanıcının girdiği siteler, aradığı kelimeler ve yer imleri dikkate alınarak bir haber öneri sistemi tasarlamak amaçlanmıştır. Geçmiş haber öneri sistemlerine bakıldığında bu çalışma denenmemiş özgün bir çalışmadır. Haberleri kullanıcıya ilgili olarak sunabilmek için makine öğrenmesi modeli, haber kategorilerini ve haber içeriklerini içeren bir veri seti ile eğitilmiştir. Kullanıcı ortamından gelen veriler eğitilen modele verilerek, kullanıcının bulunan ilgili kategorileri RSS (Rich Site Summary) tarafından anlık olarak işlenir. RSS'den seçilen bu haberler, günlük haber gündemine göre öncelik sırasına göre kullanıcıya gösterilir. Gerçek kullanıcı testi %89 gibi etkileyici bir doğruluk gösterdi. Bu çözüm, sorunun doğası gereği içerik tabanlı bir öneri sistemi sunar.

Anahtar Kelimeler: Kişisel Haber Tavsiye Sistemi, İçerik Tabanlı Tavsiye Sistemleri, NLP.

Personalized News Recommendation System

ABSTRACT

Recommendation Systems are the methods that suggest the next choices of the user in a predictable way, based on the preferences made by the user before. This method is become even more popular nowadays and it can be applied to any topic or field that needs future estimation evaluating the data at hand. It is a kind of information extraction study. Furthermore, the fact that Amazon receives about 35% of its revenue from referral systems is an indication of how important this method is. However, news recommendation system, which is a similar application area, is not also widely used as others. In this study, it is aimed to design a news recommendation system by considering the sites the user enters, the words that they searched for and bookmarks. Considering the previous news recommendation systems, this study is an untested original study. The machine learning model has been trained with a data set that includes news categories and news content in order to present the news to the user as interested. By giving the data from the user environment to the trained model, the found interested categories of the user is processed instantly by the RSS (Rich Site summary). This news selected from RSS are shown to the user in order of priority regarding the daily news agenda. The real user test showed impressive accuracy as 89%. This solution presents a content-based recommendation system as nature of the problem.

Keywords: Personalized News Recommendation System, Recommendation Systems, NLP, Machine Learning.

Geliş/Received	:	**24.10.2022**
Gözden Geçirme/Revised	:	**05.12.2022**
Kabul/Accepted	:	**19.12.2022**

1. Introduction

In recent years, recommendation systems have quickly entered our lives with the rise of companies such as Spotify, Netflix and Amazon. These systems, which are accessible in advertising, e-commerce, video, music, film and many other areas, will continue to effect our lives in the future. Recommendation systems try to anticipate the user's preferences and suggest solution for user action.

These systems aim to predict and present a service to the user. Companies attach importance to these systems in order to position themselves in front of other companies, since they can generate huge amounts of income in this way (Dwivedi, 2020). It offers great benefits for the user and the provider. The user can find and buy their own personalized products more easily. On the other hand, the provider obtain the opportunity to increase its sales, and also a long-term relationship is established as customer loyalty (Özkok, 2020).

Recommender system types have been categorized as follows:

- Contextual Recommendation Systems
- Collaborative Recommendation Systems
- Popularity Based Recommendation Systems
- Hybrid Recommendation Systems

In Contextual Recommendation Systems, similar products are tried to be displayed according to the products that the user has preferred in the past. In this method, we do not need any information about other users. It is very sensitive to the integrity established between the item and user (Beel vd., 2013). The term collaborative recommendation was proposed by Goldberg et al in 1992 with the idea that "information filtering can be more effective when people are involved in the filtering process". This method is a system that collects information about the products that the user prefers and makes suggestions by guessing what other users might like from similar preferences. User to user and item to item collaborative recommendation algorithms are applicable. Products selected by similar users are presented in the user, and similar products are displayed according to the product similarity in the item. In the popularity-based system, products with a trend are displayed to the user. The hybrid recommender system, on the other hand, is created by combining collaborative and contextual systems. In recent years, with the developments of information technologies, recommendation systems have also gained progress in the academic field. When we look at the scientific recommendation system studies in the literature, it is seen that the contextual recommendation type is mostly applied. Contextual recommendation type is followed by collaborative recommendation and hybrid recommendation types.

To categorize users approprimately and to provide acceptable success while creating the model, it is very important to choose and collect the right data. Consequently, the model is expected to learn about user interests and things he likes and to give a reasonable result in that direction.

There are studies on different subjects, with different methods applied. The study of N. Jonnalagedda, S. Gauch, K. Laville, and S. Alfarhood aimed to present news articles according to the interests of the user instead of presenting them in order of occurrence. Users have created profiles according to their interests and news. In the research to develop a personalized news recommendation system with the help of "Twitter", articles were ranked according to the popularity of the article identified from the Twitter's general timeline, resulting in a hybrid structured system. As a result, it was observed that the popularity-based system contributed. Generally, hybrid2 with α=0.5 outperforms hybrid1 by 7.8%, personal by 2.4%, and popularity by 6.0% (Jonnalagedda vd., 2016).

An another example of similar research is the recommendation system for text-based news executed by Seven and Alpkoçak. The algorithm created using the KNN (K-Nearest Neighbors) method was used in the entire text search architecture. The collaborative recommendation system was developed using a keyword-based system, where the users' match score was 6 out of 10 (Seven ve Alpkoçak, 2020).

At the same time, Taşçı (Taşçı, 2015) created a content-based recommendation system in his study and suggested a recommendation system that uses an object-user matrix according to the relationship of the items to be recommended to the user. Since our study is a content-based system, it has similar aspects with this study. In order to evaluate the results, the time applied for the system was kept long and users were provided with feedback. According to the results obtained, we should also consider the differences when recommending news. It is also necessary to consider the keyphrase. An object-trust relationship must be established. It is stated in the article that the source confidence value can be taken from users as 60. He concluded that identification of the right news sources and determination of the user-trust status are necessary.

Liu and Dolan applied a time-based method with a different point of view, taking into account the time people read the news and the length of the news text. 10,000 users were randomly assigned to approximately the same number of control and test groups. Users in the control group tried the collaborative referral system method, and users in the test group tried the hybrid method. Google news was used in this study. The result of research, users in the test group visited Google News 14.1% more than the control group. As we understand from the articles, we can get better results from the hybrid method (Liu vd., 2010).

Pazzani used a hybrid model in his study. According to the news read by the users in the model, the results were obtained by combining two different algorithms which are short-term and long-term. In the short-term algorithm, KNN was used and on the other hand, Naive-Bayes was used for the long-term algorithm. We used Naive Bayes in our study. Ten users trained the system for 4 to 8 days. It resulted in 300 news per user. According to the result, the accuracy score for the hybrid system was 75-80, and the F1 ratio was 55-60. By comparing these two algorithms, it was observed that users responded faster in the short term (Pazzani, 1999).

The study of Li et al. it proposes SCENE (SCalable two-stage pErsonalized News rEcommendation), a two-stages personalized news system. The system consists of three main elements. These items are news clustering, user profiling, and news sorting. They used Locality Sensitive Hashing (LSH) and hierarchical clustering for clustering. In order to test the system, news articles were collected in 9 different categories such as sports, politics and movies. The data has been preprocessed before being presented to users. After preprocessing, an average of 1,221 news is stored every day with 4,630 users. They chose Goo, ClickB, Bilinear and Bandit methods for comparison. According to the result, SCENE is the most successful method among them with 0.6930 (Li vd., 2011).

On another study execute by Saranya and G. Sudha, experiment on a collection of sports related news obtained from various news websites data was automatically obtained from news agencies. The received data are written to the database. Classification was made by news agencies. Two types of user profiles were created in the study. One is the static user profile and the other is the dynamic user profile. Static user profiles are created using information collected from the user during registration, while dynamic user profiles are used to address frequently changing user interests. In our study, there is no preliminary preparation for users. Only 3 categories of interest were requested. 500 news were used for the sports category. Under this category, the results of 25 different queries belonging to sub-categories such as cricket, tennis, football, athletics and hockey were obtained. News sites such as The Hindu, Times of India, Indian Express, India Times and the proposed system were compared. As a result, the proposed system was most successful at a rate of 0.9-0.95. From the experimental results, it has been observed that the proposed method provides good recommendation accuracy. The suggested method be able to provide content more suited to user preferences, even if the number of suggested items are small (Saranya and Sudha, 2012).

Teo and Tan asked users to enter the words they were interested in in the system they called Personalized Information Network (PIN), and the system tried to give the most appropriate advice according to the entered word. An access tool for searching and receiving news on the World Wide Web, a personal learning tool for professional learning and information filtering; and PIN, which includes a personalized news browser, consists of three main subsystems. Looking at the experimental results during the learning and estimation period of twenty-two days, training MSE (Mean squared error) is always smaller than test MSE and usually the result is zero. According to this result, the system can perform online learning effectively (Tan and Teo, 1998).

Dhruv, Kamath, Powar and Gaikwad created a model that was developed by taking into account the artists listened to by the users' friends, using a hybrid approach, using a hybrid approach that resembles the chosen artist with other artists of the same style. In this system, UBCF (User based Collaborative Filtering) algorithm, similarity matrices, content-based filtering and hybrid filtering are used. Stages such as data cleaning and filling in missing values were applied. As a result, the accuracy of the system using the UBCF method was 95%, and the accuracy of the system using the hybrid model was 69% (Dhruv vd., 2019).

In this study, data consisting of the user's browser history, bookmarks and Google searches, interests were determined by using the trained news categories model, and a news recommendation system was created according to the relevant categories by selecting the agenda contents from the RSS news. This paper brings a different perspective to the literature in terms of determining the weights of the categories that the user is interested in and the agenda news, and choosing the current and preferred news in the recommendation system beyond being a content-based recommendation system. As far as we know, it is the first attempt using the local computer information for news recommendation system. When we look at the literature studies, the hybrid system was used in general. In our study, a more successful result can be obtained if it turns to the hybrid system.

4900 news articles in 7 different categories were pre-processed and test data was obtained. Then, using the bag of words obtained from the collected resources of usage data from user computer, those that exceed the threshold value, the preference categories of the user were determined using the trained model. Finally, among the news that is the subject of the agenda (by looking at their tags), those that match the user's preference categories are presented to the user in order of priority. The proposed system showed 85% success rate for Naive Bayes learning method, which is the important statistical model for NLP (Neuro Linguistic Programming), is chosen.

In the second part of the article, algorithm, the data set and model creation are explained. In the third part, collecting data from user computer and experimental results are expressed. In the last part, the study was evaluated and the points open to improvement were emphasized.

2. Methods

In the study, certain steps were taken both on the data set, on the data received from the users and on the news pulled from RSS. Thanks to these steps, it is aimed to read the data more accurately and to obtain a better result.

The process steps performed in the study are as follows;

- Retrieve data from call history
- Extract URL (Uniform Resource Locator) path address and headers from imported data
- Removing unnecessary words (stop-words)
- Cleaning punctuation marks and numbers
- For the dataset to be used in training;
    - Cleaning the data
    - Data normalization
    - Spelling and rooting (tokenization and lemmazation)
    - Term frequency calculation
- Classification
- Determining the percentage value of news categories interested by the user according to the model output
- Extracting news suitable for interested categories from RSS
- Prioritizing news that fits the agenda and presenting it according to user preference weight

2.1. Tools and Libraries Used

In this study, PyCharm as the development environment and Python as the programming language were preferred for the following reasons.

- Python contains more comprehensive libraries for NLP and its C-based syntax makes coding easier.
- PyCharm is user-friendly development environment for the Python programming language.

Browserhistory, feedparser, numpy, pandas, nltk, sklearn, urlparse, matplotlib, seaborn, wordcloud, requests, beautifulsoup libraries were used in the study. Library descriptions are given in Table 1 briefly.

Table 1. Python Libraries Used in the Study

Browser History	It has been used to retrieve data from Chrome.
Feedparser	It is used to parse the news from RSS.
Numpy	It is used to organize the training and test data in the study.
Pandas	It is used to process data in CSV (Comma Separated Values) format.
Nltk	It is an open source library created for working on human languages.
Urlparse	It is used to parse the URL part of the data received from Chrome.
Matplotlib	It is used to visualize the data.
Seaborn	It is used to visualize statistical-based data.
Wordcloud	It is used to visually show frequently used words.

Requests	Used to call the URL page before getting the agenda tags.
BeautifulSoup	It is used to retrieve the data between the HTML (Hypertext Markup Language) tags of the page.

2.2. Training Data Set

The dataset consists of two columns. The first one expresses the category title (class information), while the second column contains the data under the text title. Table 2 shows example data in the sports category.

Table 2. Example Data in Sports Data Set

sports	oklahoma harden ın boşluğunu dolduramadı nba de geçtiğimiz sezonun finalistlerinden olan Oklahoma city thunder yaz sezonunda takımdan ayrılarak Houston rockets a takas olan James harden yokluğunu hissediyor
sports	masa tenisinde 3 etap izmir de masa tenisinde süper lig 3 etap karşılaşmalarının izmir de yapılacağı bildirildi masa tenisi federasyonu ndan yapılan açıklamaya göre müsabakalar 10 11 kasım tarihlerinde fuar celal atik spor salonu nda gerçekleştirilecek
sports	cimbom 3 te 3 peşinde Galatasaray kadın basketbol takımı fiba Avrupa ligi c grubu üçüncü maçında yarın deplasmanda polonya nın polkowice takımı ile karşılaşacak durumda bulunuyor

News texts are divided into 7 categories: sports, world, economy, culture, technology, politics and health. There are 700 pieces of data evenly distributed in each category (Yıldırım, 2017). For the data set to be trained; With pandas, each category name is matched to a numeric index and sorted by category values. The category id and categories are as follows.

0: Politics, 1: World, 2: Economy, 3: Culture, 4: Health, 5: Sport, 6: Technology

2.3. Data Preprocessing

After the data collection is completed, the first thing to do is to apply the data pre-processing steps. The data is converted into a ready-to-process form. Pre-processing steps directly affect the success of the study. Thus, the success of the pre-processing stage allows to reach the correct and precise result (Olson vd., 2008). The preprocessing steps in the study are briefly as follows:

1) Tokenization: It is the process of saving the text in sequences by breaking it down word by word as desired.
2) Removing Stop Words: It is the process of removing the words in the text which has no meaning in the language itself.
3) Removing punctuation and digits: It is the process of removing punctuation marks and numbers in the text.
4) Normalization: It is the process of removing the upper and lower case separation in the text.
5) Stemming: It is the process of recording the root of the words by removing the suffixes in the words in the text.

2.4. Term Weighting

The data is now ready to be analyzed for frequency. Next, the text document normalized is converted to a term count vector using the CountVectorize method in Python (Uslu and Akyol, 2021).

Term weighting is the process of assigning weights to indicate the importance of the term in the data. In this study, the TF-IDF (Term Frequency- Inverse Document Frequency) term weighting method, which seems suitable for the nature of the problem, was used. Special words for each category are found with TF-IDF for their categories. TF-IDF is the weight factor showing the importance of the term in the data (Kumaş, 2021).

TF calculates the number of times that term occurs in a document. IDF, on the other hand, gives us the information in how many times that term is used in the texts. By multiplying these two values, the TF-IDF value is obtained. In this study, the words in the news in the data set and the news content as a document were used as terms. The TF-IDF value was found using the tfidfvektorizer class. With TF-IDF, there are important words determined in each category. The pseudo code of this operation in Python is given below.

```
Create TfIdfVectorizer method
Set subliner_tf as True inside the method
Set min_df to 5
Set as norm 12
Set in ngrams (1,2)
Add list of stop words to method
Transform the result from the method
```

Sublinear_tf, min_df, norm, tokenizer, encoding, ngram_range and stop_words parameters are used in TF-IDF vectorizer operation. Here, sublinear tf scaling (logarithmic based frequency conversion, so that the importance of a term in a document is normalized) is applied by selecting the sublinear_tf parameter. The min_df parameter, on the other hand, allows to ignore terms that have a document frequency strictly lower than the given threshold when constructing the vocabulary. With the norm parameter, each output line has one of the forms l1 and l2. The tokenizer parameter preserves the preprocessing and n-gram generation steps. The Encoding parameter determines the detection standard for characters in the data (Deniz vd.,2021). With the ngram_range parameter, it is possible to calculate the ngram interval given by min and max. With (1,2), both 1-gram (unigram) and 2 grams (bigram) are used in the model (larger grams will not work well since the data size used for training is small). The stop_words parameter is for removing meaningless, non-valued words from the data (Çelik and Koç, 2021).

By adding the chi2 class in the Sklearn library as a library, 1-gram and 2-gram values were obtained for each news category. The pseudo code of this operation in Python is given below.

```
For each news category
Use Chi2 library
Find bigram and unigram words
Print
```

For each news category, the most common words and phrases are listed. These phrases are listed in descending order of frequency. Those with a high frequency value in the list are considered to be suitable collocations. For example, unigram and bigram outputs for world and economy categories are as follows.

#'Word':	#'Economy':
- Most relevant unigram expressions:	- Most relevant unigram expressions:
Suriye	Lira
İsrail	Yuzde
- Most relevant bigram expressions:	- Most relevant bigram expressions:
Gazze	Indeks Yuzdesi
İsrail'de	Yılın Aynı

2.5. Model Training

After 1-gram and 2-gram associations, TF-IDF transform and CountVectorizer transform were made. CountVectorizer preprocesses text data into a term/symbol count vector. Vectorized texts according to TF-IDF values were given to the Naive Bayes classifier by dividing into 25% test and 75% training data. Thus, the training is completed. The trained data were evaluated separately with Naive Bayes, Gradient Boosting, AdaBoost and Decision Tree algorithms. The results are given in Table 3.

Table 3. Classification Results

	Test Accuracy	Precision	Recall	F1
MultinomialNB()	84.84	0.85	0.85	0.85
DecisionTreeClassifier()	63.35	0.63	0.63	0.63

GradientBoostingClassifier()	82.64	0.83	0.83	0.83
AdaBoostClassifier()	57.32	0.57	0.57	0.57

3. Experiments

3.1. Pulling Data with Browser History

Browserhistory is a library used for Linux, MacOS, Windows platforms that extracts browser history from user's local computer and writes data to CSV files; In terms of browser, it is a simple Python module that supports Firefox, Google Chrome and Safari. The URL link contains the name and date of the page.

It is taken as the historical data on the computer used by the user and saved in CSV format. With BrowserHistory, the data received from the browser was estimated for each site link, bookmark and searched words and added to the category. The pseudo code of this operation in Python is given below (Pypi, Browser History).

```
Use the BrowserHistory class
Convert to Csv format
```

While analyzing the data, it is very important to decide which learning model will give the most accurate result. In this study, Multinomial Naive Bayes, GradientBoosting, AdaBoost and Decision Tree algorithms were experimented (selected for being most appropriate algorithms) separately for the same data set and classification was performed. Although GradientBoosting gave a good result in in terms of success, the best classification metrics were obtained by Multinomial Naive Bayes. Consequently, tests were executed over the Multinomial Naive Bayes. Multinomial Naive Bayes algorithm is a probability and statistics-based Bayesian classification (Kaşıkçı and Gökçen). Formula 1 shows the formula of the multi-class Bayesian model. Here, c refers to the classes (categories), and X refers to the instance whose class is desired to be found. Small x's are feature vectors and contain important words determined for each category in this problem.

$$P(c \mid x) = \frac{P(x \mid c)\,P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \text{ x } P(x_2 \mid c) \text{ x } \ldots \text{ x } P(x_n \mid c) \text{ x } P(c) \qquad (1)$$

After estimating the data obtained from a sample user environment according to the categories using this classifier, the percentage values of the categories that a user is interested in are calculated and shown in Table 4. Categories above a certain threshold value are selected as preferences. In this study, the first 3 categories with the highest percentage value were taken into account for news presentation.

Table 4. Interested category percentage values for any user

	Percentage Values
Economy	33.333
Technology	23.404
Politics	14.184
Sport	13.475
World	5.673
Culture	4.964
Health	4.964

3.2. Showing Instant News

In this study, after finding the percentage values for each category, the FeedParser library was used to extract instant data from RSS for news types that exceed the threshold value from these values. The threshold value is the value obtained by dividing the percentage number of all categories expressing the mean by seven. Feedparser library receives news from instant RSS; It is a library for capturing the title, link and description of the news (Pypi, FeedParser).

The desired HTML page was obtained with the requests module in order to determine the order of importance of these news. Requests is a module that makes get,post, put, delete requests for the web page. After making a get request for the web page, BeautifulSoup library was used. This library is used to parse the HTML tags of a page and retrieve the data between those tags (Erdinç, 2017). The pseudo code of this operation in Python is given below.

```
Get url page
Add to response variable
Return response u with content in bytes
Parse content in page variable with html parser
Just parse the block with the data we want to retrieve
Print
```

Important tags on the agenda are pulled from the agenda page on news sites, and thanks to this library, the priority order of the news is determined. According to the category result of the user, the presence of the word on the agenda in that news title is checked. The news with the agenda tag is shown to the user in the top row. For example, the simplest version of the RSS agenda tags dated 18/03/2022 is as follows.

Agenda Tags: ukraine, weather, sadık çiftpınar, putin, 1915 çanakkale

After the words on the agenda were obtained, it was checked that whether there were any words of the interested category of the user, which were above the threshold value, on the agenda in the title of the instant news from RSS. Categories that are above the threshold value were added and looked at according to the agenda tag, and presented to the user by putting them in the first place in order of priority.

3.3. Result

In the tests conducted with 15 real users, the user was first asked about the three categories he was most interested in out of 7 categories. In the next stage, the user's historical data is collected from personel computer and processed as explained above. These data are used with the permission of the user. The study was first added to personal computers. After running the application, the categories of interest and the result were compared. In Table 5, a portion from the data is presented which is gathered from User#1's browser history.

Table 5. User#1's browser history

	Browser History
User#1	x"https://www.haberturk.com/borsa-gune-yukselisle-basladi-3379626-ekonomi","Borsa güne yükselişle başladı”
User#1	"https://www.dunya.com/finans/haberler/rusya-mb-yaptirimlara-ragmen-faizi-yuzde-20de-sabit-tuttu-haberi-652356","Rusya MB, yaptırımlara rağmen faizi yüzde 20'de sabit tuttu - Dünya Gazetesi"
User#1	"https://www.webtekno.com/dunyada-yilin-otomobili-odulleri-2022-finalistler-h121852.html","Dünyada Yılın Otomobili Ödülleri Finalistleri Açıklandı"
User#1	"https://www.cnnturk.com/spor/futbol/bulent-korkmaz-kaybetsek-kirilma-macimiz-olurdu","Bülent Korkmaz: Kaybetsek kırılma maçımız olurdu - Son Dakika Futbol Haberi"
User#1	"https://www.bloomberght.com/abd-nin-kripto-adimi-bitcoin-i-vurabilir-2301609","""ABD'nin kripto adımı Bitcoin'i vurabilir"" - Bloomberg HT"

The first 3 category output of the trained model are shown in Table 6. The increasing number of matching categories gives us better model. The accuracy of the model is evaluated in terms of matching percentage of the trained model categories with the user preference categories. It is seen that the trained model is successful at a rate of 89% (the ratio of the number of correctly predicted categories to the total number of categories, 40/45).

The problem here is that it cannot create the right effect when it is both Turkish and English. For someone who only uses Turkish, this study achieves a successful result. It is clear that the result will be more effective when mixed methods are used in future studies and when there is support for more than one language.

Table 6. Categories of the trained model for 15 real user test data

	User Categories	Trained Model Output
User#1	Economy, World, Technology	Economy, Technology, World
User#2	World, Sports, Politics	Sports, Technology, Politics
User#3	Technology, Sports, Culture	Technology, Sports, Culture
User#4	Health, Culture Technology	Health, Technology, Culture
User#5	Technology, Economy, Health	Economy, Technology, Health
User#6	Economy, Politics, Health	Politics, Technology, Health
User#7	Politics, Technology, Economy	Politics, Technology, Economy
User#8	World, Economy, Culture	Economy, World, Technology
User#9	Sports, Economy, Technology	Sports, Technology, Economy
User#10	Technology, Economy, Health	Technology, Economy, Health
User#11	Technology, Health, Culture	Health, Technology, Economy
User#12	Culture, Health, Economy	Health, Culture, Economy
User#13	Sports, Economy, World	Sports, Economy, World
User#14	Technology, Sports, Culture	Technology, Sports, Economy
User#15	Sports, Economy, Technology	Sports, Economy, Technology

4. Conclusion and Discussion

In this study, a content-based recommendation system has been developed by taking into account the sites the user has entered, the words he has searched for and bookmarks. Compared to other models and the accuracy of the result, Multinomial Naive Bayes gave the best recommendation with an 85% success rate for validation. Moreover, real test data accuracy on this model was above being 89%.

While bringing the news, the agenda headlines of the news sites have been obtained so that the news that is closer to the headlines on the agenda will appear at first. Thus, the news that is both interesting and more prominent on the agenda is shown to the user. Presenting something that the user likes not only increase the reading habit, but also prevents wasting time.

Even we don't think they would be useful, perhaps, deep learning models can be experimented for comparison. Although the system has been successful in general, considering that some users use both Turkish and a different language, it has been observed that it reduces the success effect in these conditions. A more effective result can be obtained when mixed methods are used together in both Turkish and other languages (generally English) for future studies. Different classification methods can be tried. In addition to TF-IDF weighting, it is necessary to observe the performance effects of different weighting methods in the literature. In the future this work can be turned into a web application or plugin.

Considering that recommendation systems have a permanent role in our lives even now, it can be predicted that news recommendation systems will serve people who need fast daily data in the future, and more importantly, the essence of information can be presented by using summary systems.

REFERENCES

Beel J., Gipp B., Langer S., Breitinger C. Paper Recommender Systems: A Literature Survey. International Journal on Digital Libraries 2016; 17(4): 305-338.

Billsus D., Pazzani, M.J. A Hybrid User Model for News Story Classification. In: Kay, J. (eds) UM99 User Modeling. CISM International Centre for Mechanical Sciences 1999; 407:99-108 Springer, Vienna.

Çelik Ö., Koç B. C. Classification of Turkish News Texts with TF-IDF, Word2vec and Fasttext Vector Model Methods. DEÜ FMD. 2021;23(67):121-127

Deniz E., Öz V. K., Bozkurt Keser S., Okyay S. and Kartal Y. İçerik Tabanlı Bilimsel Yayın Öneri Sisteminde Benzerlik Ölçümlerinin İncelenmesi. DUMF Journal of Engineering 2021;12(2):221-228, doi:10.24012/dumf.838084

Erdinç S. 2017, Python BeautifulSoup Modülü,
https://www.sinanerdinc.com/python-beautifulsoup-modulu
(Access Date: 03.12.2021).

Dhruv A., Kamath A., Powar A., Gaikwad K. Artist Recommendation System Using Hybrid Method: A Novel Approach. In: Shetty N., Patnaik L., Nagaraj H., Hamsavath P., Nalini N. (eds) Emerging Research in Computing, Information, Communication and Applications. Advances in Intelligent Systems and Computing. 2019;(882). Springer, Singapore.

Dwivedi R. 2020, What Are Recommendation Systems in Machine Learning.
https://www.analyticssteps.com/blogs/what-are-recommendation-systems-machine-learning,
(Access Date: 03.12.2021).

Jonnalagedda N., Gauch S., Labille K., Alfarhood S. Incorporating Popularity in a Personalized News Recommender System. PeerJ Computer Science, 2016; 2: e63

Kaşıkçı T., Gökçen H. Determination of E-Commerce Sites with Text Mining.Journal of Information Technologies.2014;7(1). DOI: 10.12973/bid.2014

Kumas E. Comparison of Classifiers While Performing Sentiment Analysis from Turkish Twitter Data.Journal of ESTUDAM Information.2021;2(2):1-5.

Li, L., Wang, D., Li, T., Knox, D., Padmanabhan, B., Scene: a scalable twostage personalized news recommendation system, Proceedings of the 34th international ACM SIGIR conference on Resear. 2011;125–134

Liu, J., Dolan P., Pedersen E.R.: Personalized news recommendation based on click behavior. In: Proceedings of the 15th International Conference on Intelligent User Interfaces, IUI 2010;31–40. ACM, New York

Olson, D.L. and Delen, D., Advanced Data Mining Techniques, Springer Science & Business Media, Verlag Berlin Heidelberg, 2008.

Özkok H. 2020, Recommendation Engine (Tavsiye-Öneri Sistemleri),
https://www.datasciencearth.com/recommendation-engine-tavsiye-oneri-sistemleri/
(Access Date: 03.12.2021).

Pypi.Browser History. https://pypi.org/project/browserhistory/,
(Access Date: 01.12.2021)

Pypi.Feed Parser. https://pypi.org/project/feedparser/,
(Access Date: 01.12.2021)

Saranya KG and Sudha G Sadhasivam. Article: A Personalized Online News Recommendation System. International Journal of Computer Applications. 2012;57(18):6-14.

Seven S., Alpkoçak A. Kişisel Haber Öneri Sistemi. Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi. 2020; 22(64): 301-307.

Tan A. H., Teo C. Learning user profiles for personalized information dissemination, Neural Networks Proceedings. IEEE World Congress on Computational Intelligence. 1998;1:183-188.

Taşçı S, Content Based Media Monitoring and News Recommendation System. Master Thesis. Hacettepe University, Computer Engineering, Ankara, 2015

Uslu O. and Özmen Akyol S., Türkçe Haber Metinlerinin Makine Öğrenmesi Yöntemleri Kullanılarak Sınıflandırılması, Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi, 2021;2(1): 15-20.

Yildirim S., 2017, Text Categorization for Turkish-Multi NB,
https://www.kaggle.com/savasy/text-categorization-for-turkish-multi-nb,
(Access Date: 01.11.2021).

TEŞEKKÜR ve BEYANLAR / ACKNOWLEDGEMENT and DECLARATIONS

Bu çalışmada herhangi bir potansiyel çıkar çatışması bulunmadığını beyan ederim.

Not: *Bu makale, İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Tezli Yüksek Lisans Programı'nda, Metin TURAN danışmanlığında, Melis ÖZKARA tarafından yürütülecek olan, "Kişiselleştirilmiş Haber Tavsiye Sistemi" başlıklı yüksek lisans tezinin ön çalışmalarından yararlanılarak hazırlanmıştır.*