Machine Learning Model for Personalizing Online Arabic Journalism

Nehad Omar[1], Yasser M. K. Omar[2], Fahima A. Maghraby[3]
College of Computing and Information Technology
Arab Academy for Science, Technology and Maritime Transport, Cairo, Egypt

***Abstract*—The paper discusses a model of generating dynamic profile for Arabic News Users, capturing user preference by analyzing his review of historical news, and recommend him news as soon as he creates account on News Mobile App, Preference is calculated based on article main keywords score, which is extracted from article headline & body as NLP (Natural Language Processing), when user reads an article, its keywords are calculated with rate of interest to his profile. Machine Learning techniques are used in the proposed model to recommend user the relevant news to his preferences and provide him personalization. The model used hybrid filtering techniques to recommend user suitable articles to his preferences, as Content-Based, Collaborative, and Demographic filtering techniques with KNN (K-nearest neighborhood). The model combined between those techniques to enhance the recommendation process, after recommendation happened, that the model tracks User behavior with the recommended articles, whether he reviewed it or not, and the actions he did on the article page to calculate his rate of interest, then dynamically updates his profile in real time with interested keywords score , thus By having User profile and defined preference, the model can help Arabic news publisher to classify users into segments, and track changes in their opinion and inclination, using observation method of read news from different user segments, and which articles attract them, thus it leads publishers to visualize their data and raise their profitability, and to follow the international trend in e-journalism industry to be a data driven organization.**

Keywords*—Personalization; e-journalism; KNN (K-nearest Neighborhood); dynamic user profile; NLP (Natural Language Processing); data driven organization

I. Introduction

In the last decade, smart phones technologies impose a great transformation in different fields, especially for media and journalism, many electronic newspapers and smart phone applications have been published in the whole world, and most of international news providers work on personalizing news for their customers on time, based on user reading history and predefined preference on site or application.

This technology revolution extends to be implemented in many Arab countries, a lot of Arabic online newspapers were launched via smart phone applications, its target was publishing news regardless which articles attract him.

Most Arabic newspaper websites & applications display news lists to its audience either as "The Most Read" or "The Latest Published", that may extend to the type of news that user prefers, and he defined in his favorite topics, i.e., International, Political, Sports, Science, Technology, etc., while as many international newspapers start classifying their readers, and build their dynamic profiles based on news they circulate through its website, using different techniques and technologies, such as the BBC News Agency.

Text classification and categorization researches had been increased, many researchers are interested in investigating Arabic Natural Language Processing, to find how to automate categorization & classification for Arabic Text [1]. Arabic language consists of 28 letters and are written from right to left, it's the native language of more than 400 million people in Middle East countries [2-21], which means that 23 country are speaking Arabic language and use it as an official language in their Media, Journalism, and public speech.

The proposed model discusses how to generate a dynamic profile for Arabic News User based on tracking news article s/he is interested in, it always updates his profile based on different factors (e.g., New topics he adds or deletes from his favorite topics, topics he's fully/partially interested in, and the preferable time to read during day, etc.).

The proposed model works on enhancing the extracting Arabic News' user preferences based on the historical news and current articles he reviews, and enhancing the mechanism of information retrieval for similar articles that match his preference, then recommending the related article in real time to him, capturing preference is mainly works on the extracting major 'keywords' are used in the article's headline and body and save the highly rated keywords from user to his preferences, for example 'NASA' is a keyword are mostly come under scientific topics, but if we use "Rocket" as keyword, we find it's used in both "Politics Topics" and "Scientific Topics", then if user is mostly interested in politics topics more than Scientific topic, then he'll read the full article of politics that contains those keywords, and the model calculated the keywords with highly rate in user preference, while if he is not interested in reading article in scientific topics, and read part of the article and spent less time on it, in this case the model calculates different score for the same keywords, and counts the rate of user interest in those keywords, and based on the *current profile* of the User, the model will nominate him the most relevant articles to his interests.

Arabic Language has special characteristic, especially in its morphology & orthography principles [1], because it's so rich Language in its grammar, which need special handling for morphological analysis [13-15].

Moreover, working on automated test categorization system for Arabic articles & documents is a business that comes with many challenges, because of the complicated morphological principles of the Arabic language [15], which is considered as the most challenge feature through working in this language comparing it to English, thus why Arabic language has unique nature which distinguish it from other languages.

This model works on tracking changes in User profiles, it acts as a measurement tool of public opinion, and track their changes would be an effective indicator to measure their opinions and trends, for different types of articles as politics, Economic, etc.

The paper handles recommendation of Arabic News article on time by performing Arabic text preprocessing first, then gets its main keywords (that attract user to read it)[9], after that it captures user preference by observing the rate of each read article, and sorts his keyword preference by calculating its score, by doing this, the recommendation engine starts working, it applies a hybrid techniques using Content-based, Collaborative, and demographic filtering techniques to recommend a dynamic list of news article to user on time.

The rest of paper is organized to discuss the following sections, the second section discuss background of the followed techniques by Egyptian News Publishers, and samples from Google Analytics reports they used to visualize their customer data, the third section discuss personalization and recommendation system, and how to use hybrid different filtering techniques are merged together to produce highly recommendation engine to serve personalization, the forth section discusses Arabic Text Mining, and steps of text preprocessing for Arabic and the algorisms working with Arabic text, the fifth section discusses the proposed model in details, the sixth section discuss the data set, and seventh discusses method of validating model, the eighth discusses conclusion, the ninth discuss future work.

II. Background

During the last decade in Arabian countries, different Journals build their online journal website, for different business purposes like minimizing cost, expand their customer range, cover different events on time … etc., therefore some issues appear and face news' users like information overloading, that by time the amount of news articles are increased, and both of publisher and news user had new needs as following:

1) News Publishers → Need to enhance capturing the news' user profile, track its change and new trends.

2) News' Users → need to enhance the search process to get him the most relevant articles to his interests.

For example by reviewing "Google Analytics Reports" for the official news website of Egyptian Radio & TV union [www.maspero.eg], it was finding that, it provides some basic reports analyzing traffic on the website with simple category like gender or age, without ability to export a report reflect combined segmentation like age and gender together.

Fig. 1 reflects website usage of the "Website" and categorize users to segments based on Age in separate report and Gender in another one.

Fig. 2 displays three separate reports for three different indicators, which are percentage of used devices to access the website, time that users access the website, and percentage different countries, without ability to export a combined report includes two or more of the mentioned indicators together.

The way it creates a need to generate a model working on automatic abstracting 'User preference', and update 'User profile' based on the interested articles they review in real time.

Arabic online journals and news providers usually display three types of article recommendation queues to the news User based on traditional ways, First based on the most recent published news regardless whether s/he is interested in them or not, and the Second based on the most read article, Third is related news, which get some related news to the current read one from user (as per the common keywords between them) regardless this news is what s/he interested in it or not, the way it makes life-time of any article be short, unless the user go searching for news article with one or more of its keywords.

And to cover all those requirements, there is a need to have a 'Model' able to capture User profile and get him the relevant article to his preferences in real time, based on his updated profile using the highly rated keywords from his side to get relevant article includes similar subjects which he likes.

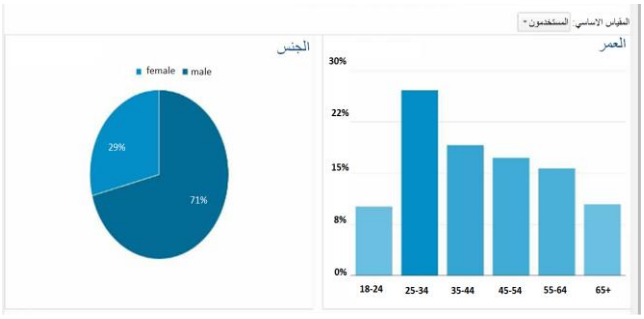

Fig. 1. Google Analytics Graph Reports of "Maspero.eg" it Displays Gender Segments in Separate Report, and Age Segmentation in another Report.

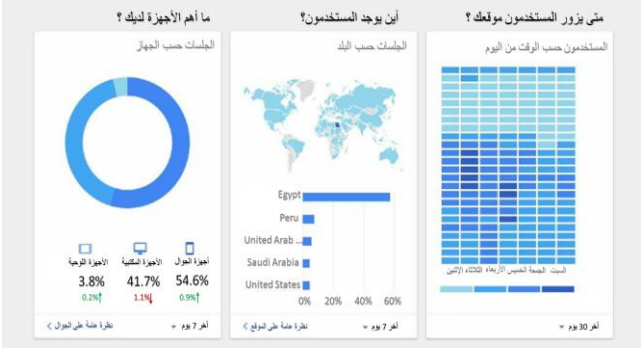

Fig. 2. Google Analytics Graph Reports of "Maspero.eg", it Displays three Separate Reports for each of Percentages of Devices, Countries, and Accessibility Time.

The paper discuss generating a model using machine learning to serve the online journalism for Arabic news provider, to personalize news and save interest of their users, and has the ability to achieve the following targets:

- Enhance capturing profiles of News' users, track their preference of viewed news, and trends of changes in their opinion.
- Enhance information Retrieval process by getting the relevant news article related to certain keywords using machine learning techniques.

III. Personalization and Recommender System

Recommendation system is an 'Information Filtering Technology' which assists to find the research paper and items what the user wants quickly and accurately [3].

Traditional recommendation systems were built based on the following items [5]:

Users: People who use the system to use the item or purchase it, (News Reader)

Items: They are main concern of system users, and based on user usage of those items, the model recommends him similar items (News Articles)

Preference: It records which items users like and which they dislike. In this paper, they are found in (highly rated keywords) of news, and system records preferred keywords and calculates its rate based on user rating of the viewed news.

In general, those components are used to build many algorithms in different scales, which could be categorized as following:

None-personalizes systems: it includes brief statistics were gathered based on common uses, like the best seller, most popular products, or popular service providers, such as this data is published based on common rating between all types of users, and that's what's followed by most of publisher under "Most Read" list.

Personalized systems: which can use one or more of the following filtering techniques:

- Content Based Filtering techniques (CBF): it recommends items to user based on early comparison between items, and finding similar features between them, then recommend those include similarities features to the user who highly rated the first one.

Fig. 3 presented how content based filtering technique is working, that when user read for example articles (A,B,& C), the model saves the main features of the selected items, which are the main keywords in the proposed model, and rating is automatically calculated based on the read parts of article by the user, and the time he spent on each of them, in the background the model saved user preferences, and when a new article is published, the model investigate whether it matches user preferences or not, and if yes it recommend it to user.

Collaborative Filtering techniques (CF): It predicts items based on the items previously rated by other similar users, act as recommending the items that are preferred by other people who are similar to the current one.

Fig. 4 displays how collaborative filtering techniques is work in recommendation system, it calculates users rating for certain item (articles & keywords), then predict rating for the current item (article & keyword) based on the usage of other users to this item, and recommend it to the current user based on the calculated prediction of rating .

The main difference between Content-Based Filtering and Collaborative Filtering is that Collaborative Filtering works on preferences of other users (users with similar preferences for some items) to recommend new items whereas Content Based Filtering is not at all concerned with preferences of the other users.

Fig. 5 displays how item is recommended each type of both recommendation techniques, in content based the item is recommended to the user if it contains similar features of a previous item was highly rated by the same user, regardless if this item was early selected or rated by other users or not, while the collaborative depends in recommending the item on rating by other users, that if 2 users select the same item with similar rate, then when first user selects an item with highly rate, so the item is recommended to the other user automatically, regardless if it contains similar feature of the previous selected items from the second user.

- Demographic Filtering techniques (DF)**:** it predicts items based on user demographic characteristics, such as age, gender, level of education, employment, income, and behavioral data, which refers to the customer dynamic data such as, location and activity status. This technique is widely used in e-commerce websites to recommends items to users who have the same demographic data, taking into consideration rules of privacy.
- Hybrid methods: these methods are a combination of two or more recommendation algorithms are used to take or maximize advantage of some techniques and avoid or minimize the drawbacks of another [10],[26].

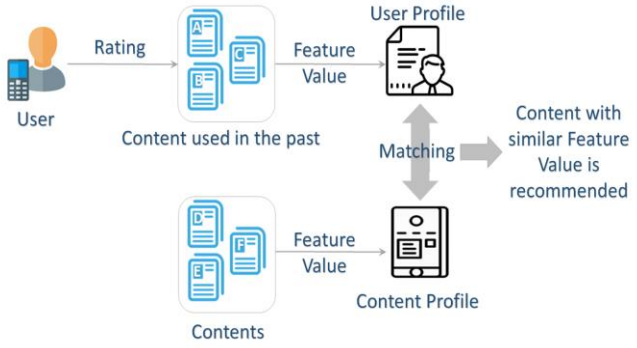

Fig. 3. Content based Filtering.

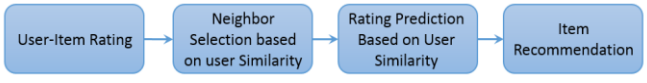

Fig. 4. Collaborative Filtering Recommendation Algorism Framework [7].

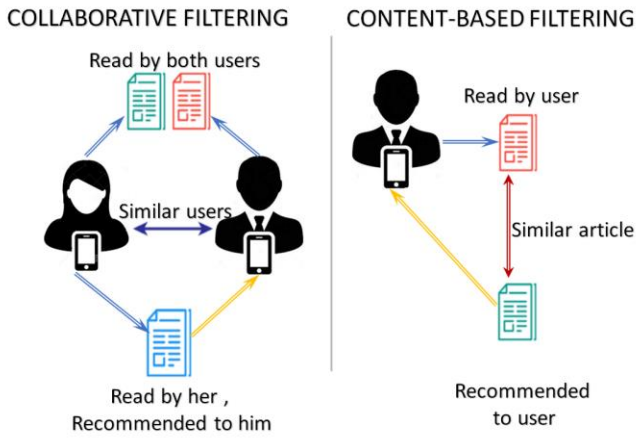

Fig. 5. Collaborative Vs. Content based Filtering [4].

Various ways of combining different algorithms are shown in Table I.

Thus, the current model uses the mixed technique in getting the recommended article to the current user using the content-based, Collaborative, and Demographic filtering techniques together.

Privacy Issue: Privacy protecting is one of the main factors that helps the recommender systems to success in personalization, therefore deploying such as personalized recommendation services typically requires the collection of users' personal data for processing and analytics. More data it collects about user means high accurate results.

It predict items to him/her, d and it's considered as despite the great benefits for both users and service/ product provider, but this matter makes users susceptible to serious privacy violation issues.

TABLE I. METHODS OF HYBRIDIZATION [6]

Hybridization method	Description
Weighted	The scores (or votes) of several recommendation techniques are combined together to produce a single recommendation.
Switching	The system switches between recommendation techniques depending on the current situation.
Mixed	Recommendations from several different recommenders are presented at the same time
Feature combination	Features from different recommendation data sources are thrown together into a single recommendation algorithm.
Cascade	One recommender refines the recommendations given by another.
Feature augmentation	Output from one technique is used as an input feature to another.
Meta-level	The model learned by one recommender is used as input to another.

IV. RELATED WORKS

Personalization for News became the general trend from different international news agencies during the last few years, that recommendation systems in that field was a hot topic for investigation during the last decade [33].

Hence the journalism industry in Arab country is changed to a new business model which provide both electronic version beside the hard copy version, then they need to attract new segment of customers who prefers to use the electronic version, therefore they need to know their customer preferences and provide him what he needs to read,.

Personalization for customer who prefers to read electronic version of news is the current trends from international news agencies, and the following are a brief review of related work on three different news & Journalism industries worldwide with examples.

A. English News Research

BBC is the most famous agency at United Kingdom, it worked on knowing its customer since long time , and now it defines its own personalizing settings for its audience , and clarifies its way in doing personalization by collecting user data and reason of doing this on its website ,it also divided its audience to different segments to target their needs through different channels of media and increase its audience/customer numbers, using in this audience demographic data like age, gender, income, socio-economic group or parental status , and its R&D (research and development) team build a customized system that apply three different model of filtering techniques (first is :Weighted average of item embedding, second is : Cosine-based collaborative filtering., third is :Rank-optimized neural network), to suit the huge amount of data are daily published in different channel and provide personalization service to its huge number of audience accurately based on audiences' history of reading (behavior)[22].

B. Chinese News Research

A sample of Chinese research is handled in this section to present the method of dynamically personalizing news for end user, that they create their own methodology which is called BAP (Behavior and Popularity) to update on time user profile including his preferences [23], and they divide their framework into three phases, first collect News and perform preprocessing on it using vector space model using TF-IDF, second generate user profile depending not only on the interested keywords he likes, but extends to article topic, and his behavior on each article, third dynamically personalize and recommend news to end user using the long-term and short-term of user's preferences.

C. Indonesian News Research

Indonesian research provide a model of personalization using collaborative filtering techniques, user behavior pattern, and KNN K-nearest neighborhood algorithm is used to predict user preference based on his history of selection and his behavior on them , and based on similar user matched his preferences [24].

The previous examples presented distinguished approaches for produce personalization that they were produces in three

countries with different languages and disparate sizes of news agencies, and it was finding that the first solution required a big fund to be implemented, and in addition of that, working in Arabic Language need more time for doing experiment and insure about result before applying similar solution as it does,

the second example present a solution depends on multilevel of profiling even for the user or for the news using TF-IDF, the way it will cause slowness in retrieving data when storage enlarge during time, the third approach neglect the importance of content-based filtering techniques and demographic data, which helps in fast recommend news for user from the first moment he creates his profile.

V. Arabic Text Mining

Arabic is spoken by nearly 400+ million people worldwide[31], and it's the official language for 23 country in middle east, and one of the six official UN language, despite its political, religious, and cultural importance, but researches in modern computational linguistics are limited compared with other languages.

Text mining is a method of getting valuable information from a text, which is written with human Natural Language, and requires preprocessing to get useful information from it, and that's what represented in keyword(s) s/he's interested in them.

Natural Language Processing (NLP) is a subfield of artificial intelligence, it's used to apply machine learning algorithms to text and speech that focus on interaction between computer and human natural language. There are different platforms that allow machine learning models to work with human NLP, and build a useful application in different business fields with different languages, i.e. in Python, there are NLTK, TextBlob, and PyNLPI, and all of them follow the same steps to perform the preprocessing as mentioned in the preprocessing module.

- Text Preprocessing

Text preprocessing plays a major role in text mining techniques, it is the first step in text mining to extract a useful information from it, but it has to go through the following steps in general to perform extraction of major keywords from article headline and body[11].

For example, we have the following sentence 'News Headline' and need to apply preprocessing operation on it

الكهرباء: بدء بناء محطة الضبعة النووية عام 2020

Tokenization → It refers to sentence segmentation, that breaks a sentence into words, after selecting news headline 'a' from training data 'A', program tokenizes to and extracts words on basis of delimiters (i.e. whitespace, comma, semi-colon....etc.)

Sanitization → it removes non-letter characters that includes special character, quotation, punctuation, numbers,…..etc., therefor, sentence become ready for the next step.

الكهرباء، بدء، بناء، محطة، الضبعة، النووية، عام.

Stop Words Removal → stop words are words which are filtered out before or after processing of natural language data (text).[26], they are some common words are used in the natural language, which add little meaning to the sentence, and they are saved in separate dataset for every program use text mining techniques, taking into consideration the dialects of the used Language before working on text preprocessing .

So when system works on this step, it checks the "stop words" Data set and compares it versus the new News headline record then removes stop words from it .

الكهرباء، بدء، بناء، محطة، الضبعة، النووية، عا

In the proposed module, the stop words dataset is stored as a separate dataset, and when system starts working on a new record of News Headline to extract keyword, system compares it with the stop words data set and removes any of them from the corpus such as the following character in the Arabic example

ة ، ال، عام

Stemming & Lemmatization → The main difference between stemming and Lemmatization lays into their way of work therefore the result they return for each of them, Stemming algorithms work by cutting the beginning or the end of the word[19], taking into consideration a list of common prefixes and suffixes that can be found in the word, but this way doesn't work well in some cases, even it can get fast results than lemmatization.

The lemmatization works on the morphological analysis of the words, and the used algorithms in it should have detailed dictionaries to get the word's form back to its Lemma, in addition of that, the lemma is the base form of all its inflectional forms, whereas stem could be the same as of the inflectional form of different lemma.

Apply N-Gram Model→ N-gram is a contiguous sequence of n items from a given sequence of text. Given a sentence, s, we can construct a list of n-grams from s by finding pairs of words that occur next to each other[4], applying N gram model is not directly related to text preprocessing steps, but it works on getting composite keywords based on its occurrence in the text together[20], for example " الواقع الافتراضى", "محمد صلاح"

After performing the preprocessing on every news article, the system would have a bag of word for each of them, and the second step would be extraction the main keyword from it[8].

VI. Proposed Model

The model works on capturing ‘User Preference’ through using mobile app to read a news article, it registers his actions on the app to read a part or full of the article, and calculates the spent time in reading each part, that news article in any mobile app is divided to two parts, the first is abstract and the second is the full article.

Recommender system plays vital role in this model, hence the proposed recommendation system includes the demographic data, which is classified as critical for some user, and it should cover the following role to gain customer trust.

Table II present the most important features that are required to be in recommendation system.

Fig. 6 describes the proposed News Personalization Model and describes how it works, the Model is divided into 4 modules:

The first and second two modules work on each published news article, and search for major keywords in article, and it's considered as main core of finding user preferences, that headline always contains the major keywords that attract users to read the article, thereby user preferences is measured by calculating the rate of articles and the contained keywords in it.

TABLE II. METHODS OF HYBRIDIZATION [25]

Role	Explanation
Effectiveness	good decisions can be taken with the help of effectiveness by user
Satisfaction	Ease of use or enjoyment can be increased with the help of satisfaction
Securitability	it it's wrong then the user have the option to tell the system
persuasiveness	try and buy are the two convincing power of the feature
trust	it increases the confidence level of the user in the system
Transparency	working of the system is explained by implementing transparency
Efficiency	user can take decisions faster with the help of efficiency

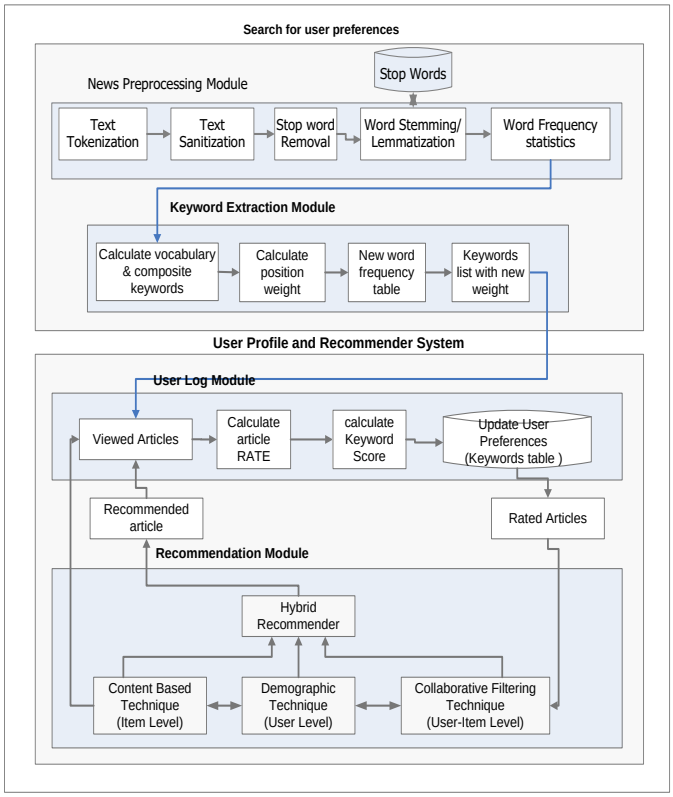

Fig. 6. Proposed News Personalization Model.

The first module "News Processing Module" performs preprocessing on article headline and the abstract, then it calculates the frequency of repeated words in both of them,

The second module "Keyword extraction Module" works on finding composite keywords, that contain two or three Arabic words, and calculates its weight based on its position in each of headline, abstract, and article body.

The third and fourth modules works on capturing user preferences by calculating the rate of read article, and recommends him similar articles using hybrid method of recommendation techniques.

The third module "User Log Module" calculates score of read keyword with dynamic method that measures user rate through the done actions using the mobile app [14], and saves the highly rated keywords after sorting them descending in user preferences list.

- The fourth module used a hybrid recommendation techniques includes content-based, collaborative, and demographic filtering techniques that recommends user news articles that matches his preferences of keywords, or matches similar users like him in demographic data, or matches similar users reads similar articles like he early reads.

The following sections explain the mechanism of the proposed model, and how it works between the published news articles and updating users' profiles module in details:

A. Search for user Preferences

Searching for user preference is not limited to the topics she's interested in, such as politics, business, arts, … etc., or using the filtering techniques as early mentioned only, the model depends on analyzing the news article and gets the bag of words from it after text preprocessing, then works on extracting major keywords, and define its weight using its TF (Term Frequency) in the text, and after user select the article, the model gets its pre-extracted keywords with its TF, and calculate rate for its word as per user's actions on this article that defines the level of his interest, therefore, the single keyword would have a combined calculated rate, first based on its weight in text using TF, and the second based on user actions on this news article using the system (word TF * Action rate).

This part includes three steps; Article/Text preprocessing, Keyword Extraction, and Rating.

1) News headline and text preprocessing module: The model uses text mining techniques with each published article as explained in text mining steps, from NLP perspective, Arabic as a language is characterized with a number of challenges, like direction of writing, diactritized and none-diactritized, using the same letter as it is for nouns and pronouns, not like other languages that use capital letter for noun and small for pronouns… etc., and because of reasons like this, and the way we need to capture user preference, the model uses FARASA Package for Segmenting and Lemmatization to get accurate bag of words from the corpus before working on keyword extraction, it gives the best result

comparing with other tools for doing the same purpose like NTLK & MADAMIRA, and uses NTLK to exclude stop words from the text.

For example, we have the following sentence and need to apply preprocessing operation on it

6أرقام قياسية حققها محمد صلاح بعد صاروخية تشيلسي

Fig. 7 describes text preprocessing steps in details starts from Tokenization, sanitation, then remove stop words, and update the stop words dataset during preprocessing, and apply lemmatization or stemming on the text, at final it calculates word "term frequency after performing text preprocessing steps on the text and using FARASA, for segmentation & lemmatization and remove 'stop words', it gets the following bag of words.

أرقام - قياسي- حقق – محمد- صلاح - صاروخي – تشيلسي

2) Key word extraction module: After performing the preprocessing on news text, the model would have a lot of segmented and candidate words (bag-of-words) that demonstrate the article document as numerical vectors, which are not enough to go further than enumeration, and not to select from those words the accurate and relevant keywords to the news text.

Extracting keyword in this section would mainly use TF formula to find the effective keyword(s) related to news text, and because TF is a scoring measure widely used in information retrieval (IR) or summarization, TF formula is intended to reflect how relevant a term is in a given document.

TF (Term Frequency) measures the frequency of a word in a document. TF = (Number of time the word occurs in the text) / (Total number of words in text), where *t*, refers to term/word, and *d*, refers to document.

$$TF(t,d) = \#\ t \text{ in } d \quad (1)$$

But calculating weight/score using traditional TF of keywords is not accurate enough to extract the keywords that attract user to read the article, so the position of keyword should be considered also in calculating keyword weight/score [14].

In Arabic News, the major keywords that attracts users to read the article are always found in news headline [12], and also in the first paragraph (abstract) more than other paragraphs in article body, in news mobile applications, the abstract (first page) always includes the first paragraph of news article, which describes the head line in more details and contains the major keyword, which is mainly considered in position weight in this paper.

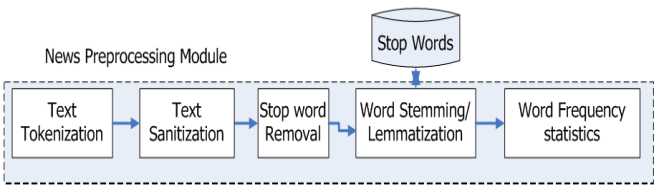

Fig. 7. Proposed News Headline and Text Pre-Processing Module.

Based on the position of the shared keyword in the headline and the first paragraph, the model works as following to extract major keywords from article to get its TF on article level as following:

Step 1- Perform preprocessing on the headline text and get a bag of word of headline text.

Step 2- Perform preprocessing on the first paragraph, and get the bag of words of it also.

Step 3- Compare between the gotten two bags of words from headline and the first paragraph, then get the *shared* keywords between the two lists.

Step 4- Apply N-gram Model on both the headline's and the first paragraph's bag of words to get the composite keyword from article.

Step 5- Calculate TF for the extracted keyword from the article.

Step 6- Sort the extracted Keywords descending from higher to lower then saved them to its related article with their term frequency.

By applying N-gram Model, the model is not working to get single keyword only, it works on composite keywords, and it could get composite keywords from two words like" محمد صلاح", or three words like " كأس الأمم الأفريقية"also.

Fig. 8 shows the mechanism of extracting keyword, once a new article is published, the model starts extracting its keyword in the back-ground from article title, and calculates its initial weight based on its position (as described in steps 3,4) in article title and abstract, after that it recalculates its frequency based on position, then calculates its new frequency and weight before any user selects it using the application and does any action on it.

By applying Preprocessing and Keyword extraction Modules on each new article, every news article would be saved with its extracted keywords and their term frequency, to be ready for use from users who are interested in similar topics.

But calculating user interest in a keyword should consider rating from user on the selected article also, so when user starts selecting an article to read, the application would measure his interest based on the actions s/he does on the application with a predefined rate.

B. Rating Mechanism

Every news mobile application has a similar flow to allow user navigating and selecting article on the application, then open it to be read, user can definitely click on article headline from the news home page, then it forwards him to the article abstract (first page), and if user need to read the full article, there is a link in the abstract page forwards him to publisher website who published this article.

Also user should have some other features like receive notification on mobile desktop with the news s/he is interested in them, search with a word, or keep an article to be read later on the application, and all the proposed flow to read an article from different sources are shown in Fig. 9 as following:

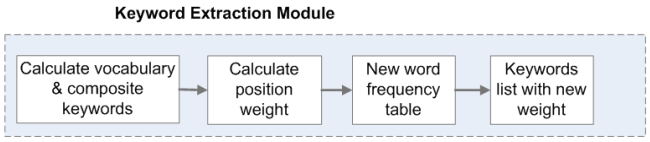

Fig. 8. Proposed Keyword Extraction Module.

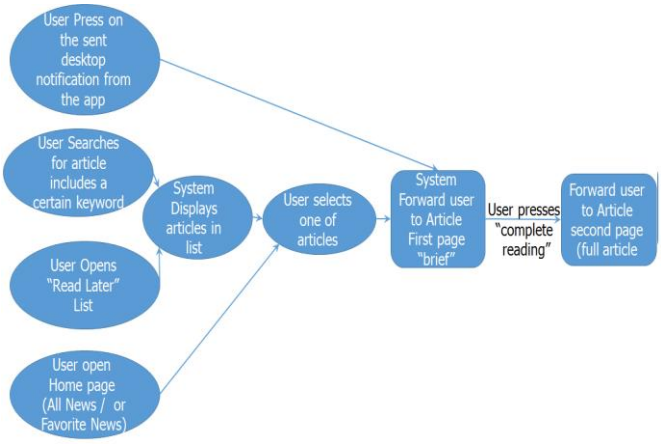

Fig. 9. All Proposed Flows to Read Full Article using Mobile App.

The model registers every single action is done by the end user on the app, and gives it a rate, then saves the extracted keywords to his preferences with rates given to the article.

The following matrix displays the calculated Rate with all proposed consecution of actions that falls between 1 to 31 and measures user interest in the selected article to read.

Table III is a matrix that displays the whole proposed actions could be done from user on the app to read or save an article, it measure the degree of interest by giving a number for each action, time of each action, and it defines the dependency between actions.

The matrix rows present the whole proposed actions with its rate, and dependency between them, and columns displays the all proposed series of actions with its final rate, and it was finding that they are all 16 proposed series of action could be done using the app to read or maintain a news Article on it, and the minimum rate for doing a single action is 1, and the maximum number for doing a series of actions are 31.

The matrix presents, that the minimum degree of interest using any of the consecution actions will meet rate of "5", that if user is interested in an article, he will do action with rate more than "5".

The model works on measuring user's interest in keyword using its term frequency in the article multiplied by the automatic calculated rate of the total actions happened on that article, as following.

$$KWs = TFKW * RA \quad (2)$$

Where KW_s refers to Keyword Score on article level, TF_{KW} refers to term frequency of extracted keyword from that article, and R_A refers to Rate of the selected Article.

As per the previous rating mechanism, user would have a list of keywords for each article with different rating values, and in most cases, the user log would include the same keyword repeated with different articles and different rates.

After calculating the keyword score for every read article from user, the model aggregates the total score of every keyword s/he read in user preference, then finds the top ten keywords related to his preference as following.

In case the user is interested in an article and do actions exceed its rate more than 5, the model will add the KW_s to the previous calculated Total Keyword Score $PTKW_s$.

$$TKW_S = PTKW_s + (TF_{KW} * R_A) \quad (3)$$

TABLE III. A MATRIX DISPLAYS ALL PROPOSED ACTIONS ON THE APP WITH RATE OF INTEREST

S	Action	Dependency	Weight /Rate	Assump 1	Assump 2	Assump 3	Assump 4	Assump 5	Assump 6	Assump 7	Assump 8	Assump 9	Assump 10	Assump 11	Assump 12	Assump 13	Assump 14	Assump 15	Assump 16
E	Select Reading Later	o	1	1	1	1	1	1	o	o	o	o	o	O	o	o	1	1	1
A	Click on article mobile desktop notification	o	2	o	2	2	2	2	2	o	2	o	2	o	2	o	o	o	o
B	Open article first page and spend more 5 second on	o	4	o	o	4	4	4	4	4	4	4	4	4	o	o	4	4	4
C	Press'Complete Reading'	B	8	o	o	o	8	8	8	8	8	8	o	o	o	o	o	8	8
D	Spend more than 30 second page	C	o	o	o	o	16	16	16	o	o	o	o	o	o	o	o	o	16
All Probability Counts			1	3	7	15	31	30	28	14	12	6	4	2	o	5	5	13	29

While the score of keyword is deducted from the PTKWs when the total actions' rate is equal to/or doesn't exceed 5 in the selected article as following:

$$TKW_S = PTKW_s - (TF_{KW} * R_A) \quad (4)$$

By implementing that model, the top ten keywords for each user would be updated automatically as per his/her daily read news and the rate s/he did by application actions.

Fig. 10 explains the mechanism of extracting keyword from Article first, then when a user selects it and does some actions on the app, model starts calculate rating based on the actions were done by the user on the app, and the rate of each action, as when user is interested in an article with a Rate more than 5, system calculates its keyword score and add it to user preference, but if he did actions less than 5 rate, system deduct it's score from the total user keyword score table.

However, user log is not limited to job in the model, but it works on building a list with the interested keywords from user. Every time the user read an article includes a certain keyword, it would be added to the last number counted to this keyword regardless the selected topic, the following is an example of a reader records with some keywords were shared between different topics in Arabic, and the model displays results of his interaction on the app in details.

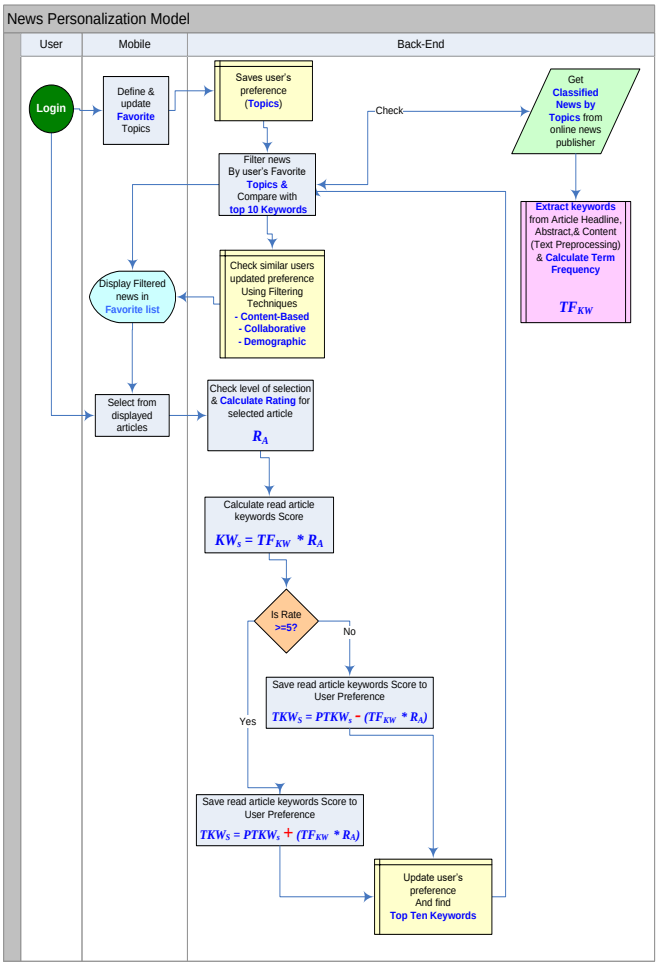

Fig. 10. Flow Diagram of the Proposed News Personalization Model.

TABLE IV. EXAMPLE ON USER PREFERENCE WITH RATED KEYWORDS BASED ON THE SUGGESTED MECHANISM

Keyword	Regardless topic name	Last read date
محمد صلاح	215	15/7/2019
كاس الامم الافريقية	187	16/7/2019
أسلحة الذكية	142	10/6/2019
طائرة بدون طيار	92	11/7/2019
الواقع الافتراضى	76	10/6/2019
محطة فضاء	64	21/6/2019
طباعة ثلاثية الابعاد	23	11/5/2019
نووى	23	10/7/2019
المنتخب	21	11/7/2019
الاهلى	19	14/7/2019
الزمالك	17	13/7/2019

Table IV displays example of user Keyword preference list, every keyword in the list was cumulative calculated using the mentioned equations and mechanism of rating and after finding the top ten keywords' score for the current user, and the list was ordered descending from the keyword which has biggest score to the one that has smaller one.

By reading numbers of the captured preference, the publisher can track changes in his customer (news user) opinion, and translates it into 'number' to analyze it anytime, From the displayed number it was finding that user is most interested in reading articles contains " محمد صلاح", and the last time he read an article was on 15/Jul/2019, and the lowest keyword was "الزمالك" till 13/Jul/2019.

On the other hand, the model suggests the similar articles to other users using different filtering techniques and machine learning algorithm and measures their trends in reading also, the way change management of Arabic online journalism at all to be managed based on the new analysis of their customer profiles.

C. User Profiling and Recommendation System

Personalizing and recommending article in the proposed model works based on two main modules, first is User Log Module, and the second is Recommendation Module that contains recommendation process & filtering techniques.

1) User log module: User log module is keeping user log, and tracking his/her daily transactions of reading articles with automatic calculation for each article's rate, then finds its keywords' score, and saves those keywords in a private dataset for each user, after that it updates the total keyword score based on the early mentioned rating mechanism, and finally it finds the top ten scored keywords for each user.

2) Recommendation module: Based on the user log that was kept by User Log module, the recommendation module starts working, first it uses the Content-based technique to recommend user similar articles to his/her preference (preferred keywords).

While the model doesn't limit to keep user log only, but it records user rating for each article, the way it allows the system using Collaborative-based technique to recommend user similar article to his similar neighborhoods users regardless their personal data.

News recommender model balance between long term user preferences – driven by professional activity, education, etc., which are best captured by content based recommendation systems. and between short term trends – driven by some discontinuity in the public or personal context, which are best captured by collaborative systems, while the new approach in this model considered also user personal/ demographic data, like Gender, age, education level, class of residency, having children or not, etc.

Using demographic data is mostly used in recommendation system for retails, the new approach here works on classifying users based on the mentioned factors, and recommend them articles based on their demographic data.

The recommendation module in this model used hybrid recommendation techniques, it combines Content-Based filtering Technique, Collaborative Filtering Technique, and Demographic Technique to enhance the recommendation process, and update user profile on real time, considering his/her reading trends, then 'recommend' user the relevant news to his/her preferences. Fig. 11 describes the interaction between user log module, rating mechanism, and recommendation module.

a) Collaborative Filtering Techniques: It focuses on User-Item preferences based on the previous users rating for the current item (article), it predicts rating for the current article based on the usage of other user to this article and the rating degree of it. For this technique, the model works as following.

1) Build a matrix of articles that user viewed and rate them.

2) Compute similarity rate between users.

3) Find users similar (reads some articles like each other) to the current user.

4) Recommend articles they viewed and rated to the current user.

Collaborative techniques here is not limited to recommend the most read news article to user, but it extend to measure the public opinion by observing the most attractive articles that are rated from different segments of users with closed /similar rate.

b) Content-Based and collaborative technique Filtering Techniques: Content-based filtering techniques in the model works on comparing any new article versus the top ten keywords for every user feature, and compare them with users' profiles, in order to find similar articles that matched user interest.

After that the system track 'User behavior' with the recommended articles, whether he reviewed it or not, and the actions he did on the article page, system would update his profile in real time with the degree of his interest on the recommended article.

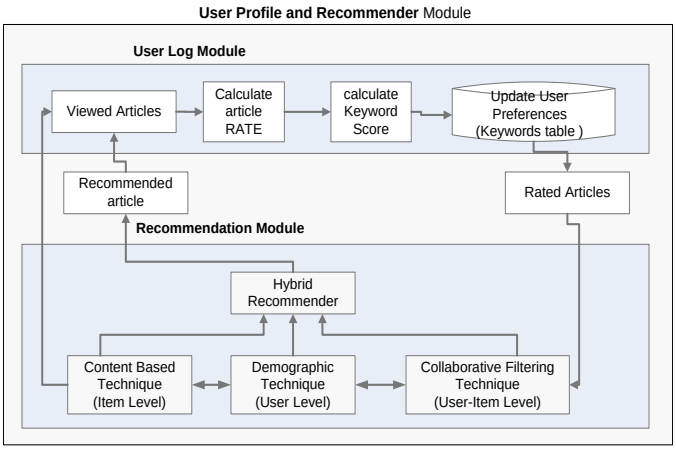

Fig. 11. Proposed user Profile and Recommendation Modules.

The article keywords are considered as item feature in the model, that when user selects an article to read, system measure his degree of interest by calculating the keyword score, then add/deduct it from his total interest in the keyword at all by calculating its total score.

c) Demographic Filtering Techniques: The model uses users' demographic data like, age, gender, economic status using in this residency class (the residency includes 20 areas classified to 3 levels [A,B,C] from 10 cities of 6 governorates in Egypt), and percentage of deducted Taxes from user as another indicator of his economic level, marital status and number of dependency, education level, working status and all of the mentioned demographic attributes are used to recommend articles to similar users, this approach is mostly used in online book store/libraries, movies, or retails websites, those online stores keeps users data and classify them into segments to recommend them the most relevant item to their preferences, even they face a privacy problem, but the most gathered data about customer, the most powerful recommendation engine they would have .

By having User profile and defined preference in Arabic online news, system can classify users in groups & segments, and track changes in their opinion and inclination.

The model uses KNN machine learning algorithm in the recommendation process, works on finding the likely users to the current one, who are similar into their demographic data to him, then find the most interested keywords to them and recommend him articles includes the same keywords in their preferences.

- Using k-nearest neighbor algorithm to get most similar users in their demographic attributes to the acquired new 'User E', Table V is an experiment result as was gotten from applying KNN algorithm in the model.

'User E' is new user, so most similar user like him is 'User A', the model recommends 'User A's' keywords to user E, that most properly 'User E' will be interested in them like 'User A', as both of them are similar to each other in their demographic features like [govern_id, age, Education level, area, working, children number].

TABLE V. EXPERIMENT RESULTS SHOWS HOW K-NEAREST RECOMMENDS NEW REGISTERED USER E ARTICLES AS PER HIS DEMOGRAPHIC DATA

id	govern_id	age_id	gender_id	Edu_Lvl_id	area_id	working	children number	Keywords
User A	1	2	1	2	2	1	4	نفى غلق محطة,دولي دور,إبرة داخل,مطروح نهاية,اعتقل شخص معارضة,معرض ايجي بلاس,قاد نيوكاسل فوز,عباس أدان,بترولي أسطوانة,ل فبراير,طبق سري خط,نهم شمال شرق,حديدي تجاوز مليار,جرام ذهب دقهلي,
User B	5	3	2	3	1	1	2	قصف آخر معقل,عاد قائمة اهلي,رئيس فخري ل,في انجاز,خطف جريزمان قبل,قوة إيهاب,قضاء إرهابي,و ميجاوات,مراهق صداع,ولاية جار أفغاني,قفز تاريخي احتياطي,إداري نجح فض,
User C	2	1	2	2	9	0	0	قطعة سلاح,رئيس فخري ل,عاطل سرق,انفجار فيتنام,أمام حوار,أول نائب,اتشيمبونج قاد هجوم,مركب نيلي,أعلن عودة,مطروح نهاية,زيت نبات,مقر مؤسسة هرم,
User D	1	2	1	3	4	1	2	مليار دولار عام,مليار جنيه,سوري أحمد,شخص سبب,غردق زور,مبنى علم طبي,قرب أعلى,يوم رائعة,أكبر لخامنئى,طن دقيق,دولى حماية,
User E	1	2	1	2	2	1	3	؟؟؟؟

- After the model got recommended interested keywords for new 'User E', the model search for articles that those words mentioned in their contents using articles table.
- So, based on that table after matching Article keywords with interested user keywords the model got article 1 and article 3 from matching.
- Table VI produce an example for each article and the main keywords for each of them, and based on the explained example therefore the model recommend article 1 and article 3 to new 'User E'.

TABLE VI. MATCHED K-NEAREST NEIGHBORHOOD IN INTERESTED KEYWORDS

Article id	Article keywords
1	دور,إبرة داخل,مطروح نهاية,اعتقل شخص أسطوانة, فبراير,طبق سري
2	مليار دولار عام,مليار جنيه,سوري أحمد,شخص سبب,غردق زور,مبنى علم طبي,قرب أعلى,يوم رائعة,أكبر لخامنئى,طن دقيق,دولى حماية,
3	بلاس,قاد نيوكاسل فوز,عباس أدان,بترولي أسطوانة,ل فبراير,طبق سري خط,نهم شمال شرق,حديدي تجاوز مليار,جرام ذهب دقهلي,

VII. DATASET

- The model uses a data set of 6000 Arabic article from Egyptian Radio & TV official website(www.maspero.eg), its articles were already classified into 10 topics (Egy-New, Arab & world, sports, entertainment, cultures, Accidents, interview & discussions, Healthcare & beauty, Science and Technology, economics & bourse).
- Hence the size of the dataset was 6000 news articles, which mean the average of displayed articles on the application were 200 article per day.
- The model work on the daily articles and collect about 100 readers data for 30 days, it was found that the average of readers reading *full article* was 4 per day and average of reading article abstracts were 7 per day, which reflected on calculating keyword score as per the suggested rating model, considering variance in weekdays and weekends.
- Hence the model works on finding the most relevant articles to the audience, it was designed to find the audience the most 6 article that matches his profile in the moment he opens the app, and put them at the top of recommended list, once the user select from them to read, system updates user keywords preferences with the new updated score based on the automatic calculated rate of his reading.

VIII. VALIDATING MODEL

Researches in natural language processing were increased during the last decade, and lately researches in Arabic language processing begins to take its share in this field , the way make it a big challenge to extract the main keywords from news headline and article body, cause of special properties of Arabic language, and its morphology that required certain rules in handling the internal structure of its words, then insure that it was the right one(s) that attract the user to read the selected article, and start calculating its dynamic rank based on continuous changes in user behavior, and build a recommender system on this calculation.

The challenge is not limited to this , but extends to sorting user's keyword preferences in real time ,and predicts &retrieves the related news article to his preference in few seconds, and sorts them to put the articles related to his preferences descending in the top of news list when he start his online session.

A. Comparing the Provided Solution with the Related Work

The discussed three examples in the related work were chosen because they were presenting the following points from business and technical perspective:

1) First example presents: How big organization news agency handle recommendation system, taking into consideration the thousand numbers of producing news articles per week, and million numbers of daily news users who need personalization; And it was finding that even they produced valuable solution, but their solution was suitable with BBC serving infrastructure which means that in case there is another medium organization need to apply one of

their proposed solution, it should have same as BBC infrastructure to produce this personalization method.

2) Second example presents: How news agencies that works with the most used languages treats with extracting its keyword (Chinese language is the mother tongue of 1.3 billion people worldwide), and what the used methodology to extract keyword - regardless its grammar or morphology rules - and match it with dynamically with changeable user preferences on real time; and it was finding that even they produced valuable solution, that provides almost near results in recall, & precision, but using TF-IDF across time with huge number of documents, it will effect on response time of retrieving the needed information (News articles) which will not serve our purpose with Arabic Language treatment for now.

3) Third example presents: a proposed model for using KNN for recommendation system to get successful personalization model, and it was finding that, even the presented solution is simple and applicable to be implemented from different agencies with low cost and avoid any concern about privacy and keeping their user data, but it mainly depends on user behavior on the website, using in this just the device IP address, and track user behavior on that device to define his profile, but the experiment was only applied on 39 users, in addition of that, obtaining the user behavior through keeping his IP address of the used device and track his behavior doesn't mean that there is a single user on that device, especially for those PCs / devices in work area or a usable device from most of family members, which means that personalization requires predefined profile from each user, especially with medium and big agencies who provides news article.

B. Advantage of using a Hybrid Filtering Techniches for the Proposed Recommendation System

As the proposed model targets to recommend user suitable articles to his preferences, the first time he login after registration, the model user his demographic filtering techniques to present him articles were interested to read from other user they have the same demographic data , then after he start reading the first article, the model capture his preference and starts building his dynamic profile on time using the content based filtering techniques, and start predict other articles were read by other users who are interested in the one he read, but the question here is why using a mix between those techniques? And why one or two of them are not enough to produce this recommendation system?

The following is the Cons& Pros of using each filter technique alone, and why the proposed model used them together.

1) Demographic filtering approach:

a) Pros

- It's easy and quick for getting results based on small action of observation.

- It's doesn't ask for user rating like happened in collaborative and content based.

b) Cons

- Privacy problem for using user data safety

- Ability to recommend the same item to the users who have the same demographic data

2) Content Based Approach

a) Pros

- Its method depends mainly depends on exploring the user profile and item for doing recommendation.

- Items can be recommended to user regardless it was rate by other user or not.

b) Cons

- If content was not defined well with enough amount or required information to build user profile

- The recommendation could be inconsistent to user preferences
- The recommendation will not provide reliable recommendation

- In case other items were defined by accident with highly rate, and it matched user profile, then the system can recommend it to the user, which is considered as drawback

3) Collaborative approach

a) Pros

- It doesn't depend on item features or content, but it depends on popular ranking of this item

- Scalability of item database is large, as it doesn't require human involvement

- It save times, as recommending it's item doesn't require previous knowledge of item field.

b) Cons

- Item should be recommended first by a user or more to be recommended automatically to other user(s)

- Active users usually rate limited amount of items, the way doesn't present the unrated item its importance.

- The approach is expensive according to the required time to rate items to be recommended to another users

And to avoid all those cons in case of using single filtering techniques, and gain the presented pros for each approach, a hybrids techniques was developed -even that it was complex- to provide such as the news list related to each user preference, but it finally can be produced to different sizes of news agencies, taking into consideration the importance of rapidity of retrieving information to end user.

C. Validating the Porposed Model

To validate the proposed model on news audience, testing contains occurred on a mobile application for News (Your News Today), and it tracks the preference of about 100 registered users during 30 days, registration contains the needed demographic data and defining their favorite topics is used for defining their interests in topics in general.

- Pick User preference based from the first time he access an article and start reading it
- Matching the coming news to his preference (content-based), other users preferences similar to him (Demographic), or Preferences of users are interested in similar current user's preferences (collaborative), and evaluate whether he is interested in them or not
- Track changes in their profiles and interest and put them in advanced segments based on some other factors based on their demographic data.
- Put the relevant article to their interest, based on a hybrid filtering techniques combined (content-based, collaborative m and Demographic filtering techniques) together and retrieve the most relevant article to the current user based on them.
- Recommendation system in this model is evaluated using 3 metrics, Accuracy, Precision, and Recall[16-18].
  - ✓ Where Accuracy is measured by dividing the relevant recommendations to the Total Possible Recommendations using this equation.

$$Acc = ARR / TPR \quad (5)$$

Where Acc refers to Accuracy, and ARR refers to the Actual Relevant Recommendations to the News user, and TPR refers to Total Possible Recommendations which are candidate to user to be read.

- ✓ And precision is measured by dividing the relevant recommendations to the Total Recommendations using this equation:

$$P = ARR / TPR \quad (6)$$

Where P refers to Precision, and ARR refers to the Actual Relevant Recommendations, and TPR refers to Total Possible Recommendations which are candidate to user to be read.

- ✓ And at final Recall is measured by dividing the relevant recommendations to the Total Recommendations using this equation:

$$R = ARR / TR \quad (7)$$

Where R refers to Recall, ARR refers to Actual Relevant Recommendations, and TR refers to the Total number of Relevant Recommendations articles to user interests.

On the basis of metrics discussed above, comparative analysis of Content-based, collaborative filtering, and Demographic filtering (as hybrid Techniques) have been measured automatically in the system back end as found in the following chart.

The chart in Fig. 12 displays the weekly result of the used model, and how accuracy is measured for using each filtering techniques, where Col, refers to Collaborative, and D. Refers to Demographic, and C.Based refers to Content Based, that Recommended = Sum of all (recommendation number happened on a single article per Day1 +Day2 +Day 3+ Day4 + Day5 +Day6 + Day7) for each filtering techniques, and Actual Read = Sum of All selected articles from all users (for Day1 +Day2 +Day 3+ Day4 + Day5 +Day6 + Day7) based on certain filtering technique, and actual read in this case refers to those users who select an article and read the article abstract.

And it was finding that Accuracy of proposed Model = Relevant recommendations / Total Possible Recommendations = 0.7847.

The chart in Fig. 13, it was displays how Precision was calculated, it consider the full reading as a degree of more relevancy degree to user preferences as following

However, Recall for each type of filtering techniques is calculated using the following table.

Table VII present the actual results of the experiment that happened during the experiment time.

Hence Recall for each filtering technique is calculated using the equation no 7

Then Recall For Content Based = 134/626 = 0.2140

Recall for Collaborative = 403/626 = 0.6437

Recall for Demographic = 151/626 = 02412

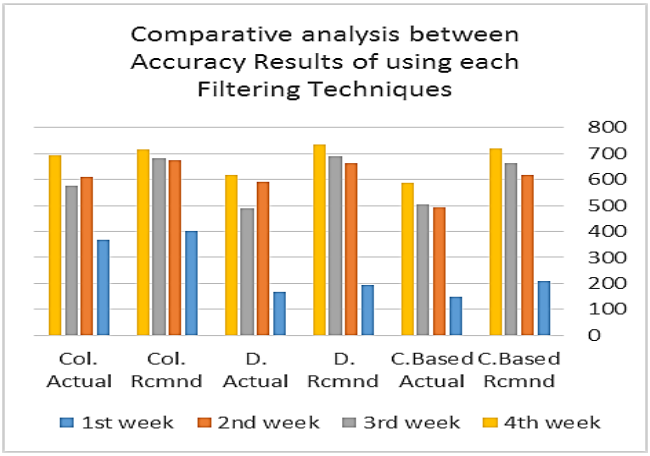

Fig. 12. Graph Shows Comparative Analysis between Accuracy Results of using each Filtering Techniques.

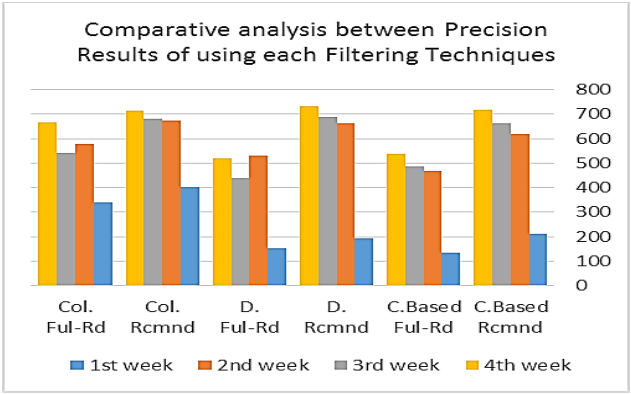

Fig. 13. Graph Shows Comparative Analysis between Precision Results of using each Filtering Techniques.

TABLE VII. A TABLE SHOWS HOW RECALL FOR EACH TYPE OF FILTERING TECHNIQUES IS CALCULATED

Week no.	Content Based		Demo-graphic		Collabor-ative		total recommended	total Fully read	Precision
	Recommended	Fully read	Recommended	Fully read	Recommended	Fully read			
1st week	210	134	194	151	403	341	1286	626	0.4868
2nd week	618	467	661	530	672	578	1951	1575	0.8073
3rd week	663	486	689	438	682	541	2034	1465	0.7203
4th week	718	537	734	519	714	667	2166	1723	0.7955
Totals							7437	5389	0.7246

IX. CONCLUSION

The paper presents a well-round investigation on presenting a recommender system for online Arabic news on time.

The model establish first user profile based on the favorite topics, and extracted article's composite keywords, then update user profile and capture his preferences on time, by calculating a dynamic score for his preferred keywords, then recommend him news article using a hybrid filtering techniques: Content-Based, Collaborative, & Demographic based techniques that recommend the existing user a list of news article that most matches his preference.

The challenge didn't lay in using hybrid techniques for recommendation, or in providing a high quality dynamic recommendation result with short term & long term preference of users, as match as finding and extracting a composite Arabic Keywords from the Arabic news text, that preprocessing Arabic text and find the best way to get Arabic word to its root without differentiating its meaning [17].

Using this way in defining News user, and segmenting them could lead Journalism industry in Arabian countries to produce a new model of business and follow the international trends as happened in other Journalism organization, that track their customer profiles and produce the articles they are really interested to follow it, then increase its profitability and became a data driven organization.

X. FUTURE WORK

The research can be extended later to include Search for keyword as one of the main methods to consider it in user preferences with a high score.

REFERENCES

[1] Riyad Alshammari, King Saud Bin Abdulaziz University for Health Sciences,College of Public Health and Health Informatics,"Arabic Text Categorization using Machine Learning Approaches",(IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 9, No. 3, 2018.

[2] Cedric De Booma, Steven Van Canneyta, Thomas Demeestera, Bart Dhoedta, aGhent University – iMinds, Department of Information Technology,"Representation learning for very short texts using weighted word embedding aggregation", arXiv:1607.00570v1 [cs.IR] 2 Jul 2016.

[3] Kwanghee Hong,2 Changho Jeon, 3 Hocheol Jeon, 1,2 Hanyang University, Korea,3 Agency for Defense Development Seoul, Korea "UserProfile-Based Personalized Research Paper Recommendation System", 2012, 8th International Conference on Computing and Networking Technology (INC, ICCIS and ICMIC).

[4] Marwa Hussien Mohamed , Mohamed Helmy Khafagy , Mohamed Hasan Ibrahim , Fayoum University , "Recommender Systems Challenges and Solutions Survey " , 2019 International Conference on Innovative Trends in Computer Engineering (ITCE'2019), Aswan, Egypt, 2-4 February 2019.

[5] R. Burke. Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction 12(4), 331–370 (2002).

[6] Harbhajan Kaur, Dr. Mohita Garag, Amanjot Kaur, North West Institute of Engineering &Technology, Moga, India, " Review of Techniques for Recommender Systems", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 7, Issue 4, April 2017.

[7] Yi Liu, Jun Feng, Jiamin Lu, Hohai University ,Nanjing, China , "Collaborative Filtering Algorithm Based on Rating , Distance , Published in IMCOM '17 2017 , DOI:10.1145/3022227.3022292.

[8] Wei Wang and Yongxin Tang, Hebei University of Engineering - School of Information and Electric Engineering,China, "Improvement and Application of TF-IDF Algorithm in Text Orientation Analysis",International Conference on Advanced Material Science and Environmental Engineering (AMSEE 2016).

[9] Lluís Codina, Mar Iglesias-García, Rafael Pedraza&Lucía García-Carretero─A DigiDoc–UPF Research Group Publication,"Search Engine Optimization and Online Journalism: The SEO - WCP Framework",April 2016.

[10] Manisha Chandak, Sheetal Girase, Debajyoti Mukhopadhyay, Maharashtra Institute of Technology, India,"Introducing Hybrid Technique for Optimization of Book Recommender System",International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).

[11] Armughan Ali 4, Mehr Yahya Durrani 5,and Farhan Aadil 6, Department of Computer Sciences, Comsats Institute of Information Technology, Attock, Pakistan,"News Headlines Classification Using Probabilistic Approach", ISSN: 2090-4274 Journal of Applied Environmental and Biological Sciences, March 2016.

[12] Terrence Szymanski, Claudia Orellana-Rodriguez, Mark T. Keane,University College Du8blin,Insight Centre for Data Analytics & School of Computer Science, "Helping News Editors Write Better Headlines:A Recommender to Improve the Keyword Contents & Shareability of News Headlines", arXiv:1705.09656v1 [cs.CL] 26 May 2017.

[13] Fawaz S. Al-Anzi1,and Dia AbuZeina2, "Big Data Categorization for Arabic Text Using Latent Semantic Indexing and Clustering", International Conference on Engineering Technologies and Big Data Analytics, (ETBDA'2016) Jan. 21-22, 2016 Bangkok (Thailand).

[14] Tom Nicholls,Reuters Institute for the Study of Journalism, University of Oxford, and Jonathan Bright Oxford Internet Institute, University of Oxford, "Understanding news story chains using information retrieval and network clustering techniques", 25th January 2018.

[15] Fatma Mallek, Billal Belainine, Fatiha Sadat,Université du Québec À Montréal, "Arabic Social Media Analysis and Translation", 3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5-6 November 2017, Dubai, United Arab Emirates.

[16] Ismail Hmeidi,Mahmoud Al-Ayyoub,Nawaf A. Abdulla,Abdalrahman A. Almodawar, Raddad Abooraig, Nizar A. Mahyoub,Computer Science Dept., Jordan University of Science and Technology,"Automatic Arabic Text Categorisation: A Comprehensive Comparative Study", Journal of Information Science 2015, Vol. 41(1)1 14 – 11© The Author(s) 2014 Reprints and permissions : sagepub.co.uk/journalsPermissions.nav, DOI: 10.1177/0165551514558172 jis.sagepub.com.

[17] S. Al-Fedaghi and F. Al-Anzi, “A new algorithm to generate Arabic root-pattern forms,” in proceedings of the 11th national Computer arabic information retrieval,” Arabic computational mor- phology, pp. 221–243, 2007.

[18] S. Vinodhini¹, B.Govindarajalu², V. Rajalakshmi³,College of Engineering, Chennai, India, "Building Personalized Recommendation System Witrh Big data Hadoop Mapreduce ", International Journal of Engineering Research & Technology (IJERT),IJERTIJERT,ISSN: 2278-0181, Vol. 3 Issue 4, April – 2014.

[19] S. Khoja and R. Garside, “Stemming arabic text,” Lancaster, UK, Computing Department, Lancaster University, 1999.

[20] Jiwei Guan,Macquarie University,"A Study of the Use of Keyword and Keyphrase Extraction Techniques for Answering Biomedical Questions",Jan 2016.

[21] The Campus Herald,the Student-Run Newspaper of Johnson & Wales university, Vol XXVII, No 11, Wednesday Feb 14-2007,

[22] Maria Panteli, Alessandro Piscopo, Adam Harland, Jonathan Tutcher, Felix Mercer Moss , British Broadcasting Corporation , INRA’19September, 2019Copenhagen, Denmark ,

[23] Zhiliang Zhu, Deyang Li, Jie Liang, Guoqi Liu, Hai Yu "A Dynamic Personalized News Recommendation System Based on BAP User Profiling Method". IEEE Access 6: 41068- 41078 (2018).

[24] Arie Satia Dharma ,Faculty of Informatics and Electrical Engineering,Institut Teknologi Del Toba Samosir 22381, Indonesia "The User Personalization with KNN for Recommender System" , Journal Publications & Informatics Engineering Research ,Volume 3, Number 2, April 2019.

[25] Mohammad Aamir , Mamta Bhusry ,AKG Engineering College Adhyatmik Nagar, GZB (UP) India "Recommendation System: State of the Art Approach" , International Journal of Computer Applications (0975 – 8887) Volume 120 – No.12, June 2015.