2021 6th International Conference for Convergence in Technology (I2CT)
Pune, India. Apr 02-04, 2021

Comparative Review on Sentiment analysis-based Recommendation system

Aditya Nair*, Christopher Paralkar*, Janya Pandya*, Yash Chopra*, Deepa Krishnan†
*Department of Computer Engineering, Mukesh Patel School of Technology Management & Engineering,
NMIMS University, Mumbai India.
{1,2,3,4}nairaditya1812@icloud.com, chrisparalkar@gmail.com
janya.pandya@gmail.com, ynchopra@gmail.com

†Assistant Professor, Department of Computer Engineering, Mukesh Patel School of Technology Management & Engineering,
NMIMS University, Mumbai India.
{5}Deepa.Krishnan@nmims.edu

2021 6th International Conference for Convergence in Technology (I2CT) | 978-1-7281-8876-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/I2CT51068.2021.9418222

***Abstract*—Recommendation systems are ubiquitous these days and are used in nearly every domain; from learning which videos could be recommended to users on streaming websites, to products that can be sold on e-commerce platforms. These systems are driven by the copious amount of data that is scraped and collected from sources such as review platforms and social media websites. On this collected data, sentiment analysis can be performed to recommend products to users based on an overall analysis of sentiments conveyed using reviews, comments, or opinions. The information thus obtained is provided to already existing machine learning-based filtering techniques which include content-based, collaborative, and hybrid filtering. The aim of this paper is to provide a detailed review of various techniques used for sentiment-based recommendation systems and the inherent challenges in these techniques.**

***Index Terms*—Recommendation system, web scraping, sentiment analysis, machine learning, content-based filtering, collaborative filtering, hybrid filtering**

I. Introduction

This era of growing technology is rightly coined as the "Digital Age", which is often characterized by the abundance and availability of information. Online shopping is becoming popular day by day because of the low cost, effective logistic systems, and variety. However, this abundant diversity leads to uncertainty in quality and indecisiveness which is the source of confusion for many customers. To find an answer to this question customers generally lookup reviews and opinions on websites and analyze them manually. This is not proportionate to the energy and time which it requires. In recent years lots of research has been done for providing useful recommendations to the customers. Many of these recommendation systems are based on the sentiment analysis of reviews and social media content. During this period, the researchers have come across problems such as data sparsity and cold start. The realization of the problems being factor-dependent has led to the development of techniques that are application-specific or problem-specific.

Data Gathering or data extraction is an essential part of a recommendation system; surveys, social media platforms, and review websites have become important sources of such data. There are two different approaches to achieve this; one is by using a programming language script to extract data, the other is using pre-built software to provide customization to the same.

The field of sentiment analysis bridges the gap between the extracted data, NLP (natural language processing), and AI (artificial intelligence). It is often referred to as "opinion mining". As the name suggests its main purpose is to determine the opinion or the sentiments of a person/group on a particular subject. The paper compares ML (machine learning) based sentiment analysis techniques such as Naive Bayes [1], Support Vector Machine [2], Linear Discriminant Analysis (LDA) [3], [4], Logistic regression [3] and Decision Tree [2]. These techniques can be applied to the extracted textual data based on the type of data, the volume of data, and time complexity.

Recommendation systems predicts the rating or the preference, that a user is likely to give to an item. With the help of these predictions, recommendations can be provided to the users about the items they might like. It is broadly divided into three approaches, namely: Content-based approach [5], Collaborative approach [5] [6]and hybrid approach [7]. ML approaches such as KNN (K nearest neighbor is used with the sentiment analysis output procured previously.

This paper is organized as follows: First, we have provided a summary of data collection techniques. The next section discusses approaches for sentiment analysis, filtering techniques, and various evaluation metrics. Following that a qualitative and quantitative evaluation of those techniques is presented based on those metrics. Finally, we have concluded the paper by stating inferences from the comparisons done in the paper.

978-1-7281-8876-8/21/$31.00 ©2021 IEEE

thorized licensed use limited to: SVKM¿s NMIMS Mukesh Patel School of Technology Management & Engineering. Downloaded on May 11,2021 at 07:03:30 UTC from IEEE Xplore. Restrictions app

A. Background

Today recommendation systems are being used in almost all sectors of society to provide useful user-centric predictions. The past few years of research in this domain has brought light upon many different approaches which include content-based, collaboration-based, and hybrid systems. These approaches tend to use user-user, product-user, and product-product similarities that are derived from the rating data that users provide as well as the product metadata [8] . These approaches being information-dependent seem inept when the information is insufficient, or the data is sparse. To tackle such problems, information can be ameliorated by using the textual review data along with the rating data. Such information is generally achieved by performing sentiment analysis on the comments and review data which is abundantly available on the web. Although, this abundance in data doesn't always mean that the data is directly usable, hence useful data must be extracted and pre-processed.

The general theoretical framework for a sentiment based recommendation is demonstrated in the figure:

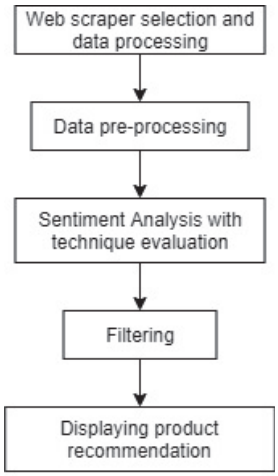

Fig. 1: Framework

II. Data Extraction

The preliminary step to initiate any process always involves the gathering of relevant data. There are several techniques out of which we will be summarizing ones which have been used for collection of data from various websites.

There are browser extensions such as Spider, data scrapers like Weboob, and import.io which do not require much knowledge in programming and have a flexible interface for converting them into various formats such as JSON (JavaScript Object notation) or CSV (Comma Separated Values). Alternatively, there are programming language libraries such as those of NodeJS and Java, a minor drawback of them is that they are not suitable for a layman and only a seasoned specialist may be able to use them to their full potential. [9]

Data analysis using python [10] is also an efficient way not just to scrape data, using software like Scrapy but also using a snippet of code to analyze the stored data. A minor problem is the lack of uniformity and the dynamic nature of web pages from which data needs to be extracted, which makes extraction a hard process.

There are several different ways to further refine data extraction, by understanding the parsing structure of HTML (Hypertext Markup Language) pages. This involves the DOM (Document Object Model) based tree structure as well as Beautiful Soup, a python library that can extract certain parts of the content from the web that can eliminate the HTML tags from them. Another method is wrapper classes, which allows the user to specify a particular algorithm, the wrapper finds the relevant web pages and can convert unstructured to structured data.

All these techniques mentioned above [11] can be succeeded by ML algorithms like Rapier (Robust Automated Production of Information Extraction Rules) and SRV(Sequence Rules with validation) which have their own rules for learning how to efficiently extract data from web pages.

TABLE I: Data Extraction Techniques.

Extraction tools	**Techniques used**
Software platforms like import.io, easy web extract, Weboob etc [11]	• Mimicry • Weight measurement • Differential approach
Machine learning based tools Rapier and SRV [9]	• HTML Parser • Semantic annotation • tree-based technique
Python based Web scrapers/crawlers [10]	• Scrapy • Beautiful Soup

III. Sentiment Analysis

Sentiment analysis also known as emotion detection is a process which uses NLP (Natural Language Processing) and classification models to determine the sentiments behind verbal or textual data. It has various applications such as mining social data, determining product reputation, and understanding customer requirements. It does so by detecting the polarity of the data that can be in the form of document, paragraph or a sentence. .

A. Types of Sentiment analysis

1) Fine grained: : It is performed at a sentence or subsentence level,it concerns with the polarity of the data which can categorized into strongly negative/positive, weakly positive/negative, neutral, positive, negative. For example, product

horized licensed use limited to: SVKM¿s NMIMS Mukesh Patel School of Technology Management & Engineering. Downloaded on May 11,2021 at 07:03:30 UTC from IEEE Xplore. Restrictions app

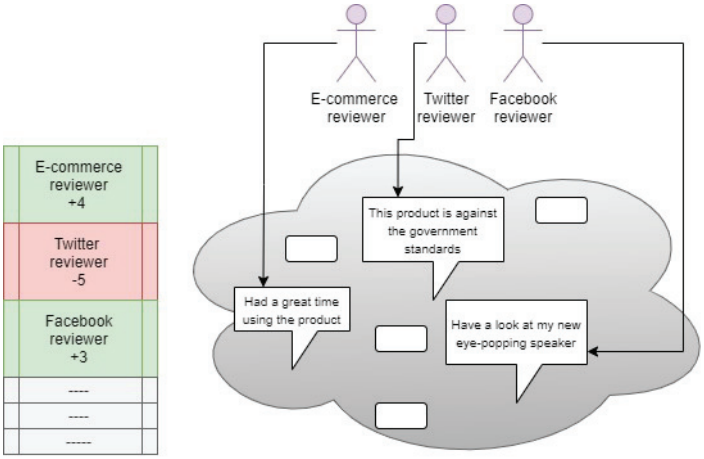

Fig. 2: Sentiment analysis

review ratings which range from 1-5(negative-positive). [12] [13]

2) Emotion detection: This is used to classify basic and complex natural human emotions involving happiness, sadness, and anger. It is performed from textual or verbal data using lexicons or classification algorithms.

3) Aspect-based Sentiment Analysis : Also known as context-based sentiment analysis, it is mainly used in analyzing the sentiment of texts, for example, this model aims to determine which aspect or feature of a product in a product review is being mentioned in a positive, neutral, or negative way. [2] [18]

B. NLP methods and algorithms

1) Rule Based: Rule-Based: Use a set of human-crafted rules such as stemming, lexicons, counts, and so on, that help in identifying the contextuality, polarity, or emotion. [6] [14]

2) Automatic Approaches: These are machine learning-based approaches that use classification algorithms to generate rules which help in identifying the sentiment behind a review or a speech. Some of the classification algorithms used are.

- Naïve Bayes: These classifiers are a simple family of probabilistic classifiers, that predict the sentiment of the text using Bayes's theorem. [1] [15]
- Logistic Regression: It is a statistical supervised machine learning model that uses a logistic function to classify a binary dependent variable [16]
- SVM (): A similarity driven model that, it takes the input data as points and creates an N-dimensional space (N-Number of features), which are used to find a hyperplane that precisely classifies all the data points, for example, different sentiments are mapped to different regions and new texts, are assigned classes based on its similarity to a particular region. [4] [15] [16]
- Decision Tree: Use a top-down tree-like structure that classifies the data points based on "if-then" rules, these rules are generated sequentially from input data. The features at the top greatest impact on the decision. [21]

TABLE II: Sentiment analysis techniques.

Algorithms	**Techniques used**
Rule based	• Hu and Liu Opinion Lexicon (Dictionary of good and bad words) based frequency model [6]
SVM	• Sentence level categorization and feature vector formation [21] • Feature extraction and emotionalism [16] • Aspect/context-based sentiment analysis [2]
Naïve Bayes	• Sentence level categorization and feature vector formation [21] • Segregation based on how many "bad" and "good" words show up [15]
Logistic Regression	• Used for probabilistic classification [8]
Decision tree, Random Forest	• Division by appearance or absence of a word to classify a document [8] • Random forest is an ensemble method that generates a multitude of decision trees classifies based on the aggregated decision of those trees. [21]

IV. Recommendation Systems

Recommendation systems are algorithms aimed at suggesting relevant items to users. These systems use algorithms or techniques like collaborative filtering or content-based filtering to suggest personalized items to the users or items which user might like or prefer. [17]

A. Collaborative filtering

Collaborative filtering as the name suggests uses the basis of collaboration to suggest or recommend products. This helps to provide users with a personalized choice of products or items which they may like. [5] [20]

It uses the nearest neighborhood algorithm to find the most similar items for any user. Users with similar tastes in products

horized licensed use limited to: SVKM¿s NMIMS Mukesh Patel School of Technology Management & Engineering. Downloaded on May 11,2021 at 07:03:30 UTC from IEEE Xplore. Restrictions app

and who have rated similar items or products are put within the same neighborhood. This is useful in recommending items which have not been rated by one user but has been rated by another user within the same neighborhood. [7]

They are different approaches to collaborative filterings, such as the user-based approach, item-based approach, which have been explained below

1) User based approach: This approach works with users acting in the main role. Recommendations are made to the user based on products that have been viewed and rated highly by other users having similar preferences. If the item has been rated highly and given a positive review by the other users within the neighborhood, then that product would be recommended to the user.

For this algorithm, two tasks are required

- Finding K nearest neighbors: A similarity function identifies the neighborhood to which a particular user belongs.
- Rating prediction: With the help of user-item matrix predict the rating the said user will give to the items which are rated by his neighbors but not by him. [5]

2) Item based approach: In this approach, rather than the user being in focus, it's the items that have been rated by the users which come into focus. Items, e.g. which have been brought in e-commerce websites or streamed in streaming services, are compared with other items within the same neighborhood and the similar items which have been rated highly are recommended to the user. [6] [19]

B. Content based filtering

Content-based filtering is a filtering technique based on profile attributes. The recommendations are made while looking at the profiles as the basis of these recommendations. These profiles contain the user's details and preferences. The preferences rely upon what the user has rated and what he or she has viewed or brought. The system using this technique will check the profile of the user for items that have been highly rated and then compare this with items that have not been rated by the user. Based on this comparison, similar positively rated items will be recommended to the user.

C. Hybrid Approach

A hybrid approach is part of a collaborative filtering technique when two or more techniques are combined to get better results. There are several different criteria under which a hybrid approach can be used. [7] Some are

- Finding predictions using different content-based methods and then combining the predictions made.
- Incorporate the characteristics of a content-based approach along with the collaborative approach.
- Incorporate the characteristics of a collaborative approach along with the content-based approach.
- A general model that will merge the characteristics of a content-based method along with the characteristics of a collaborative filtering method. This would help reduce the problems arising due to filtering as the disadvantages of one technique can be overcome by the advantage of another.

V. Performance evaluation metrics

A. Mean Absolute Error (MAE)

The MAE calculates the average magnitude of the errors in a dataset that can effectively measure accuracy for continuous variables. It is calculated by taking the average of the absolute values of the differences between forecast (xi) and the corresponding true value (x) over the sample with n data values. Furthermore, all the individual differences are weighted equally in the average making MAE a linear score.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x - x_i| \tag{1}$$

B. Precision

Precision is the fraction of relevant (True Positives) instances divided by the total number of relevant instances (true positives + false positives). It talks about how precise your model is, that is out of those predicted positive, how many of them are actually positive. It is computed as

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{2}$$

C. Recall

Recall can also be called sensitivity. It is the fraction of the total amount of relevant instances that were retrieved. It indicates the portion of actual positives that were identified correctly.

It can be computed by

$$Precision = \frac{TruePositive}{TruePositive + FalseNegative} \tag{3}$$

D. Root Mean Square Error (RMSE)

The RMSE is a quadratic scoring rule that finds the average magnitude of the error. The difference between the forecast values (Yi) and corresponding values that are observed (xi) are firstly squared and then averaged over the sample and finally, the square root of the average is taken. Since the errors are squared before they are averaged, the RMSE gives a relatively higher weight to large errors. We can thus say that the RMSE is useful when large errors are undesirable.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(\frac{d_i - f_i}{\sigma_i}\Big)^2} \tag{4}$$

horized licensed use limited to: SVKM¿s NMIMS Mukesh Patel School of Technology Management & Engineering. Downloaded on May 11,2021 at 07:03:30 UTC from IEEE Xplore. Restrictions app

VI. Comparison and Analysis

Although there are a lot of models present for sentiment analysis, it cannot be affirmed that a single model is the best because the recommendation system, in general, is domain-specific and depend upon qualitative and quantitative factors.

Table III presents the comparison of classification algorithms based on qualitative factors such as type, quality, the quantity of data, and time complexity.

TABLE III: Comparative Analysis

Algorithm]	Qualitative Analysis
Naïve Bayes	Supports effective and highly scalable model building along with scoring scales linearly with the number of predictors and rows.
Support Vector Machine	SVM is found effective in high dimensional spaces, it is not suitable for large data sets. Conversely, it doesn't work well when the dataset has a lot of noise.
Decision tree	Scaling of data is not required. Decision trees require less effort for data preparation during pre-processing compared to other algorithms
Logistic regression	It makes no assumptions about distributions of classes in feature space. It is very fast at classifying unknown records. Conversely, it is tough to obtain complex relationships using logistic regression.

Fig 3 presents the quantitative analysis involving precision, recall, and accuracy of these classification algorithms when applied on the IMDB data set consisting of 1568200 data values.

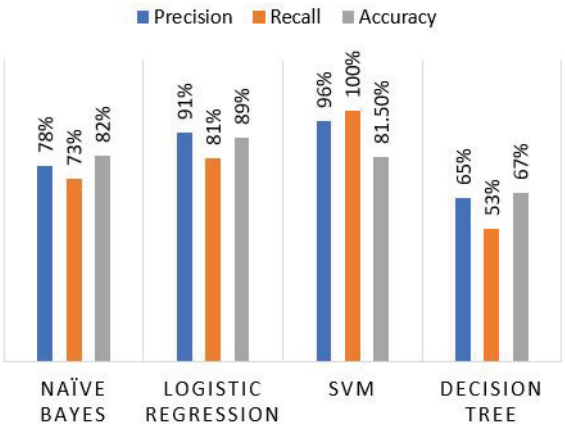

Fig. 3: Comparison of precision, recall and accuracy on IMDB data set

Similarly, information filtering techniques are compared based on their efficiency in dealing with problems such as data sparsity and cold start. Simultaneously we have compared them when they were applied to the same data set

TABLE IV: Result and analysis of filtering techniques.

Algorithm	Qualitative Analysis
User-based KNN	It is not context-dependent making it more reliable whereas sparsity is a major issue because most of the percentage of people who rate items is really low.
Item-based KNN	Domain knowledge is not required because the embeddings are automatically learned and Cannot handle fresh items and hard to include side features for query or item

Fig 4 below represents the quantitative analysis of RMSE and MAE of the algorithms when applied to Movie lens data set with 9000 instances.

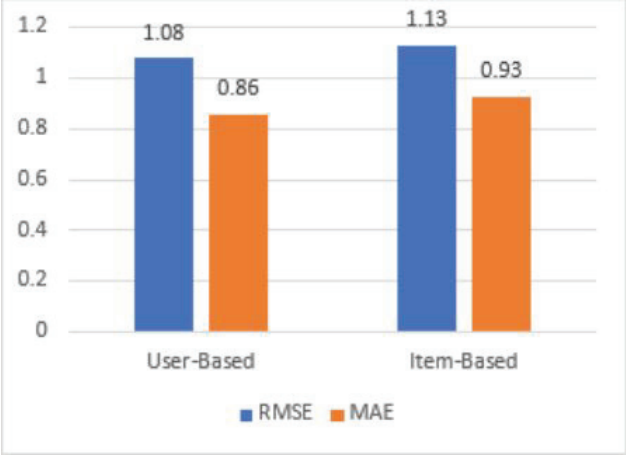

Fig. 4: Comparison of MAE and RMSE values for Movie Lens data set

VII. Conclusiom

Data collection is the initial phase of the process. Various web scrappers and data extractors help in effectively collecting data and finally choosing appropriate data for further processing. Various performance evaluation metrics such as RMSE, MAE, precision, and recall portray an effective way to compare algorithms and chose the best one. Sentiment analysis uses various ML algorithms and segregates data assigning them a sentiment score. SVM stands out as more effective in most cases but results are application dependent. Deep learning algorithms can be used for better accuracy. The major obstacle faced in a sentiment analysis-based recommendation system is data sparsity. Content-based filtering or collaborative filtering solely

horized licensed use limited to: SVKM¿s NMIMS Mukesh Patel School of Technology Management & Engineering. Downloaded on May 11,2021 at 07:03:30 UTC from IEEE Xplore. Restrictions app

cannot solve this problem. A hybrid approach is a solution to this problem. A hybrid approach using the advantages of both techniques solves the problem of data scarcity effectively.

REFERENCES

[1] Surya Prabha PM, Subbulakshmi B, "Sentimental Analysis using Naive Bayes Classifier,' 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN) .

[2] K. Schouten and F. Frasincar, "Survey on Aspect-Level Sentiment Analysis," in IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 3, pp. 813-830, 1 March 2016, doi: 10.1109/TKDE.2015.2485209.

[3] X. Lei, X. Qian and G. Zhao, "Rating Prediction Based on Social Sentiment From Textual Reviews," in IEEE Transactions on Multimedia, vol. 18, no. 9, pp. 1910-1921, Sept. 2016, doi: 10.1109/TMM.2016.2575738

[4] R. M. Gomathi, P. Ajitha, G. H. S. Krishna and I. H. Pranay, "Restaurant Recommendation System for User Preference and Services Based on Rating and Amenities," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, 2019, pp. 1-6, doi: 10.1109/ICCIDS.2019.8862048.

[5] P. Venil, G. Vinodhini and R. Suban, "Performance Evaluation of Ensemble based Collaborative Filtering Recommender System," 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 2019, pp. 1-5, doi: 10.1109/ICSCAN.2019.8878777.

[6] R. Pradhan, V. Khandelwal, A. Chaturvedi and D. K. Sharma, "Recommendation System using Lexicon Based Sentimental Analysis with collaborative filtering," 2020 International Conference on Power Electronics IoT Applications in Renewable Energy and its Control (PARC), Mathura, Uttar Pradesh, India, 2020, pp. 129-132, doi: 10.1109/PARC49193.2020.236571.

[7] Sanya Sharma, Aakriti Sharma, Yamini Sharma, Ms. Manjot Bhatia, "Recommender System using Hybrid Approach", International Conference on Computing, Communication and Automation (ICCCA2016)

[8] Alia Karim Abdul Hassan Ahmed Bahaa Aldeen Abdulwahhab "Reviews Sentiment analysis for collaborative recommender system" Kurdistan Journal of Applied Research (KJAR) — Print-ISSN: 2411-7684 –Electronic-ISSN: 2411-7706, kjar.spu.edu.iq Volume 2, Issue 3. August 2017— DOI: 10.24017/science.2017.3.22

[9] R. R. Diouf, E. N. Sarr, O. Sall, B. Birregah, M. Bousso and S. N. Mbaye, "Web Scraping: State-of-the-Art and Areas of Application," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 6040-6042, doi: 10.1109/BigData47090.2019.9005594.

[10] D. M. Thomas and S. Mathur, "Data Analysis by Web Scraping using Python," 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2019, pp. 450-454, doi: 10.1109/ICECA.2019.8822022.

[11] M. S. Parvez, K. S. A. Tasneem, S. S. Rajendra and K. R. Bodke, "Analysis Of Different Web Data Extraction Techniques," 2018 International Conference on Smart City and Emerging Technology (ICSCET), Mumbai, 2018, pp. 1-7, doi: 10.1109/ICSCET.2018.8537333.

[12] N. Aida Osman and S. Azman Mohd Noah, "Sentiment-Based Model for Recommender Systems," 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP), Kota Kinabalu, 2018, pp. 1-6, doi: 10.1109/INFRKM.2018.8464694.

[13] S. Hu, A. Kumar, F. Al-Turjman, S. Gupta, S. Seth and Shubham, "Reviewer Credibility and Sentiment Analysis Based User Profile Modelling for Online Product Recommendation," in IEEE Access, vol. 8, pp. 26172-26189, 2020, doi: 10.1109/ACCESS.2020.2971087.

[14] Z. Singla, S. Randhawa and S. Jain, "Statistical and sentiment analysis of consumer product reviews," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, 2017, pp. 1-6, doi: 10.1109/ICCCNT.2017.8203960.

[15] Shipra Goel, Muskan Banthia, Adwitiya Sinha "Modeling Recommendation System for Real Time Analysis of Social Media Dynamics" Proceedings of 2018 Eleventh International Conference on Contemporary Computing (IC3), 2-4 August, 2018

[16] Amel Ziani, Nabiha Azizi, Didier Schwab, Monther Aldwairi, Nassira Chekkai, et al.. Recommender System Through Sentiment Analysis. 2nd International Conference on Automatic Control, Telecommunications and Signals, Dec 2017, Annaba, Algeria. ffhal-01683511

[17] Q. Ren, Y. Zheng, G. Guo and Y. Hu, "Resource Recommendation Algorithm Based on Text Semantics and Sentiment Analysis," 2019 Third IEEE International Conference on Robotic Computing (IRC), Naples, Italy, 2019, pp. 363-368, doi: 10.1109/IRC.2019.00065.

[18] N.A. Osman, S.A.M. Noah,and M. Darwich "Contextual Sentiment Based Recommender System to Provide Recommendation in the Electronic Products Domain" International Journal of Machine Learning and Computing, Vol. 9, No. 4, August 2019

[19] N. S. Barley1 Asst. Prof. R. R. Keole2 "Review of Recommendation System in Domain Sensitive Manner" Imperial Journal of Interdisciplinary Research (IJIR) Vol-3, Issue-5, 2017

[20] R. Lydia Priyadharsini, M. Lovelin Ponn Felciah "Recommendation System in E-Commerce using Sentiment Analysis", International Journal of Engineering Trends and Technology (IJETT), V49(7),445-450 July 2017. ISSN:2231-5381. www.ijettjournal.org. published by seventh sense research group

[21] Fang, X., Zhan, J. Sentiment analysis using product review data. Journal of Big Data 2, 5 (2015). https://doi.org/10.1186/s40537-015-0015-2

horized licensed use limited to: SVKM¿s NMIMS Mukesh Patel School of Technology Management & Engineering. Downloaded on May 11,2021 at 07:03:30 UTC from IEEE Xplore. Restrictions app