Library Book Classification Using an LDA Model

Nasiem Ayob
School of Computer Science and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
1825850@students.wits.ac.za

Ritesh Ajoodha
School of Computer Science and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
Ritesh.ajoodha@wits.ac.za

Ashwini Jadhav
Science Teaching and Learning Unit
Faculty of Science
University of the Witwatersrand
Johannesburg, South Africa
Ashwini.Jadhav@wits.ac.za

***Abstract*—As the number of books increases, there is a need for books to be digitized and made available on the internet for scholars. Due to the advances in computing power and technology, this can be done efficiently. The problem arises when libraries attempt to organize and retrieve these books using traditional methods such as the Library of Congress Classification (LC), a system that manually categorizes texts. This method has been proven to be resource expensive and inefficient. Few kinds of research have been conducted on improving book classification using topic modeling. In this paper, we explore topic modeling (LDA) as an approach to classify books into their respective genres. Using the LDA, each word of a document will be connected to one or more topics through a probability. This will be used to determine whether a document is strongly associated with a topic. The hyperparameters of the LDA model have been tuned using the topic coherence score as a metric. The LDA model achieved an accuracy of 28.1% and 28.0% after tuning the parameters. Both the SVM and NB Classifier produced better results with an accuracy of 87.2% and 89.2% respectively.**

***Index Terms*—Libraries, Classification, Topic Modeling, LDA.**

I. Introduction

Information retrieval plays a fundamental role in our daily activities. Whether it is spending hours searching the web for research related material or simply using online databases to organize and retrieve information on hobbies and interests, you are partaking in some form of research retrieval. In the age of information, this has become a common practice as more of our information such as text, audio, video and images become digitized. The problem arises when classifying and retrieving this information.

Since the amount of books in libraries is rapidly increasing, libraries face a similar problem. Conventional retrieval methods such as the Library of Congress Classification (LC) and latent semantic indexing has been proven to be inefficient when attempting to organize, classify and retrieve online information. The Library of Congress Classification (LC) is outdated and often done manually which requires both financial and human resources. The latent semantic indexing is limited to the size of the dataset and requires indexing the entire corpus once new content has been added, thus making it unsuitable for web searching.

[5] Addresses a comparison study on unstructured text by applying both the LDA and LSA models to classify e-books based on full text.The authors used the topic coherence value as a metric with the LDA obtaining a coherence value of 0.592179 and the LSA had a coherence value of 0.577302. [4] Attempted to use the latent Dirichlet allocation and support vector machine (SVM) in their research to classify hundreds of news headlines into their respective topics using content-based features. The authors trained the LDA model and achieved an accuracy of 28.1% and the SVM performing better with an accuracy of 79.8%.

The purpose of this research is to explore a form of topic modeling (LDA) to assist libraries when attempting to organize, classify and retrieve books using unstructured text. Topic modeling is defined as the process of analyzing the relationship between words and phrases to determine which content topics they may belong to. Topic models became popular in the late 1990s and were developed to help organize and understand large collections of unstructured data.

The LDA model is trained on a dataset of book titles and their respective genre's. The words of each title text together with their frequencies are converted into a term-frequency matrix known as the corpus. For each word of the corpus, the model connects it to one or more topics through its probability. The final model is obtained by finding the most suitable hyperparameters based on the the topic coherence score of different models. To evaluate the performance of the LDA model, it will be compared to the SVM and NB classifier on the same dataset.

The base LDA model achieved an accuracy of 28.1%. We then found the best parameters from various models of different topic coherence scores. The final LDA model dropped slightly with an accuracy of 28.0%. Both SVM and NB classifier outperformed the LDA models with accuracies of 87.2% and 89.2% respectively.

II. Background

A. Information Retrieval

Information retrieval (IR) is the activity of obtaining information from large collections of information resources in response to a user query. The earliest traditional methods of managing large collections of information originates from librarianship. A commonly used method in research and academic libraries is the Library of Congress Classification (LC) System. This system uses what is known as a "call number" to

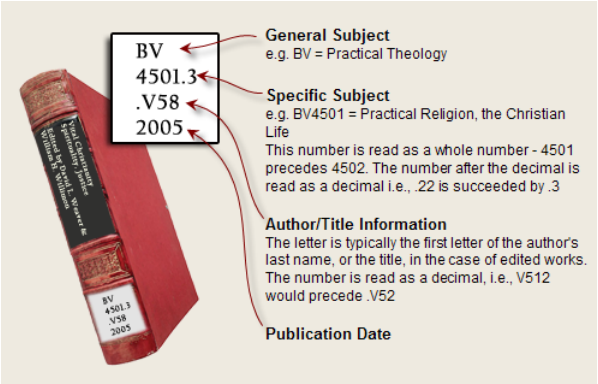

Fig. 1. Example of call number in LC system. Adapted from [16].

locate books in a library, as shown in Fig. 1. As the information increased over the years and traditional cataloging methods could not cope with this change, IR systems were developed.

In the early 1990s, the world wide web was created. Due to its popularity, the number of websites doubled in only months apart from being created. This led to the rise of web search engines. Information retrieval was now developed into what is known as web mining or web information retrieval (wIR). This is when a search engine responds to a given query by a user with a ranked list of documents that could be relevant to the query. To retrieve the information, topic models were used.

B. Topic Modeling

Topic modeling is an unsupervised machine learning technique that discovers topics from a collection of documents. Topic modeling became popular in the late 1990s as a way to help organize and make sense of large collections of unstructured data. Since then, these models have been developed and used in various fields and industries to optimise information retrieval.

Topic models are also known as probabilistic models. One of the most basic probabilistic topic models is the probabilistic latent semantic indexing (pLSI). PLSI is a statistical technique that finds a probabilistic model used to generate data from observations (words). The pLSI model has been reformulated to develop a model known as the latent Dirichlet allocation (LDA) by [1].

C. Latent Dirichlet Allocation

The LDA model is a topic model that automatically discovers hidden topics from words and documents. It is different from the pLSI model as it uses Bayes estimation instead of maximum likelihood.

There are two parts in LDA:

- Words that belong to a document. This is known.
- The probability of each word belonging to a topic. This needs to be calculated.

Algorithm to find the 2nd part can be followed using Fig. 2:

- First decide the number of topics, this will be k.
- For each document and word in it, randomly assign a topic to the word.

For each document d, go through each word w and compute:

- $p(topic\ t \mid document\ d)$: This is the proportion of words assigned to topic t in document d. This is to capture the number of words belonging to a particular topic in document d. Hence if most words in d belong to a topic t, then the current word w is likely to belong to t as well.
- $p(word\ w \mid topic\ t)$: This is the proportions of assignments to topic t over all the documents that comes from the word w. This is to capture how many documents are in topic t because of word w.

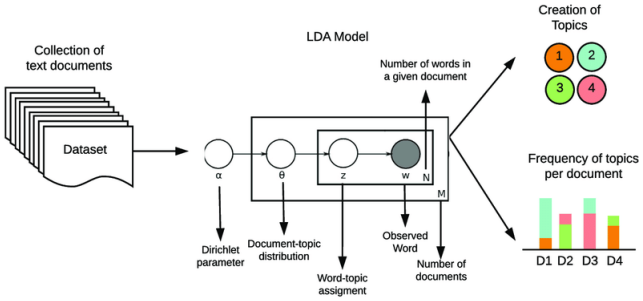

Fig. 2. Schematic of LDA algorithm. Adapted from [15].

The next step is to update the probability of the word w belonging to topic t. This continues until the model gives the most optimized representation of the document-term matrix.

III. Related Work

This section summarizes previous research conducted in using the LDA model in various fields including library book classification and retrieval.

In the field of medical science, the authors [12] introduced three LDA models to predict ADR from a large number of ADR candidates. For their study, they created drug features (structural descriptors of drugs) and used it together with a database of ADR terms (ABReCS). Their model even proved to obtain accuracies higher than the conventional LDA models.

Researches have also proposed LDA approaches for analyzing social networks such as tweets on Twitter. In [17], the authors investigated the cold start, when a recommendation finds it difficult to recommend items to a new user. They trained a LDA model to discover latent groups from pseudo-documents. These latent groups were a representation of "Twitter personalities". For their study, they used a Twitter dataset together with Apple's iTunes App Store. Their experimental results shows that their approach is significantly better than the previous recommendation systems.

Researches have also applied topic modeling methods as a way to predict and analyze crime. In [14], the authors introduced an early warning system which would identify safe transit paths by finding crime activity intention. Their approach involved the LDA model and collaborative representation classifier. The LDA model was used to learn and extract features for given articles. These features were used

by the collaborative representation classifier to classify a new document.

In [13], the authors investigated tweets related to criminal incidents. The authors attempted to use the LDA model together with a generalized linear regression model to analyze and also understand the tweets. By only using the information that existed in the tweets, their evaluation showed that their approach was capable of predicting hit-and-run crimes.

Few kinds of research have been conducted in developing techniques to solve questions or problems related to our research question. In [18], the authors focused on extracting topics from research papers. In their research, they used the Cora dataset. This dataset contains a collection of scientific papers. They used the Gibbs sampler to fit three models, namely the LDA, an extension known as the supervised LDA (sLDA) as well as the mixed-Membership Stochastic block model. Their results proved that the LDA performed the task with high efficiency.

[5] Addresses a comparison a study on unstructured text by applying both the LDA and LSA models to classify e-books. For their study, the authors used a dataset of 300 books which they downloaded from various sites such as the categorized E-library. They selected 10 scientific fields to evaluate the performance of the topic modeling techniques by choosing the coherence score as the metric. The LDA model had a coherence value of 0.592179 and the LSA had a coherence value of 0.577302.

The most recent research related to our research question has been carried out by [4]. For their research, they used the "News of India" dataset that consists of 3 features and over 2 million headlines. The authors attempted to use both the LDA and SVM to classify news headlines into their respective classes, namely "College", "Enviromnet" and "Taste". This resulted in the LDA model achieving an accuracy of 28.1% with the SVM achieving an accuracy of 79.8%.

Thus the work represented in this paper should be viewed as complimentary to the work reported by [4].

IV. Research Methodology

The aim of this research is to train models using a book related dataset with the goal that the models will be able to classify these books into their respective topics. The models that will be trained are the latent Dirichlet allocation (LDA) model, the Naive Bayes (NB) classifier and the Support Vector Model (SVM). The LDA model's parameters will be tuned based on the topic coherence score.

A. Dataset and feature selection

The dataset used in this research is the "Judging a book by its cover" dataset that was collected by [8]. This dataset has 4 features and over 2000 000 books belonging to 32 different categories. The features are "book cover images", "title", "author" and "category". The selected features for this experiment are "Title" and "Genre". To train our model, the dataset has been reduced into 4 genres/classes of even size (7979). These genres are "Science & Math", "Travel", "Medical Books" and "Computers & Technology".

B. Data Preprocessing

1) **Punctuation's and lowercase**: Initially the punctuation's "[,!.?]" are removed from the title text and then converted into lowercase.
2) **Tokenization**: Each title text is divided into smaller pieces while removing punctuation's and unnecessary characters altogether.
3) **Stopwords and bigrams**: Stopwords are first removed then bigrams are created using 2 consecutive words frequently appearing in a document. (For LDA)
4) **Lemmatization**: The data is lemmatized while keeping verbs, adverbs , nouns and adjectives.
5) **Corpus**: The words of the title texts together with their frequencies are converted into a bag of words (BOW) representation or corpus that will be used as an input to the models during training.

C. Experimental setup

- Gensim's "LdaMulticore" will be used to train the LDA model on a corpus for a number of 4 topics.
- Sklearn's "MultinomialNB" and "SGDClassifier" will be used to train the Naive Bayes classifier and support vector model respectively as well as calculate the accuracy of both models.

D. Hyperparameter Tuning

Model parameters are known as the settings of a machine learning model that are tuned by a data scientist before the model is trained.

In this experiment, 2 model parameters are determined:

- Dirichlet hyperparameter α: Document-Topic Density.
- Dirichlet hyperparemeter β: Word-Topic Density.

These tests will run in sequence, one parameter at a time while the choice of metric for performance being the coherence score of the model.

E. Evaluation

The **topic coherence metric** is a metric that measures to what extent the top topic words, or the words that have high probability in each topic are semantically coherent [3].

$$coherence(V) = \sum_{(w_i, w_j) \in V} score(w_i, w_j, \epsilon) \qquad (1)$$

V is a collection of words used to describe the various topics while ϵ is the smoothing factor.

The **accuracy metric** together with the **confusion matrix** will be used to summarize the prediction results of the models.

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP} \qquad (2)$$

The evaluation metrics used here has also been used in [4] and [5].

V. Results and Discussion

In this section, the results of each model's predictions will be presented and discussed. Table I shows the parameters and coherence score of the LDA model before and after hyperparameter tuning. The base model with default parameters obtained a coherence score of 0.303712. After hyperparameter tuning, we found that only the alpha value changed from "symmetric" to "asymmetric". This resulted in the final model obtaining a coherence score of 0.46913, a 54% increase from the base model.

TABLE I
Parameters and Topic Coherence Score for LDA

Model	*Alpha* (α)	*Beta* (β)	*Coherence*
Base	Symmetric	Symmetric	0.303712
Final	Asymmetric	Symmetric	0.469313

Fig. 3 represents the confusion matrix for the base LDA model. This model achieved an accuracy of 28.1% which is the same accuracy as [4]. Fig. 4 is the confusion matrix for the final LDA model after hyperparameter tuning. This model achieved an accuracy of 28.0% which is sightly less than that of the base model. This is because during hyperparameter tuning, we chose the topic coherence score as a metric instead of accuracy.

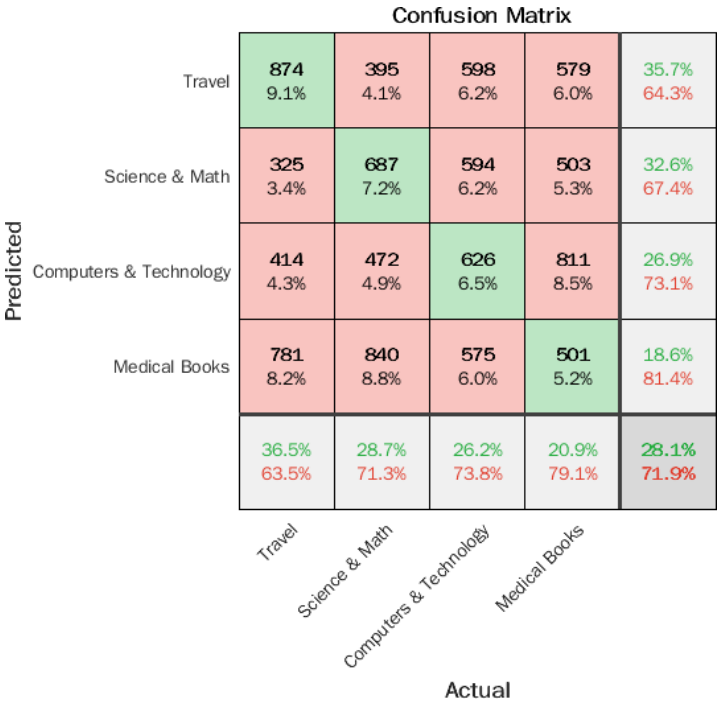

Fig. 3. Confusion matrix for LDA (base) model on a set of test data.

Fig. 5 is the confusion matrix for the SVM which achieved an accuracy of 87.2% which is higher than that of [4]. Fig. 6 is the confusion matrix for the NB classifier which achieved the highest accuracy of 89.2%.

Even though both LDA models were able to categorize the books into different topics by using only the title texts of the

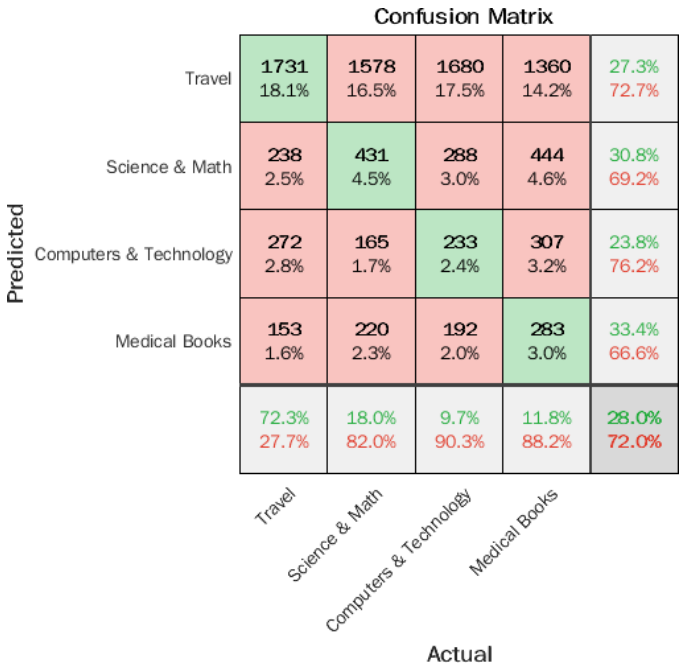

Fig. 4. Confusion matrix for LDA (final) model on a set of test data.

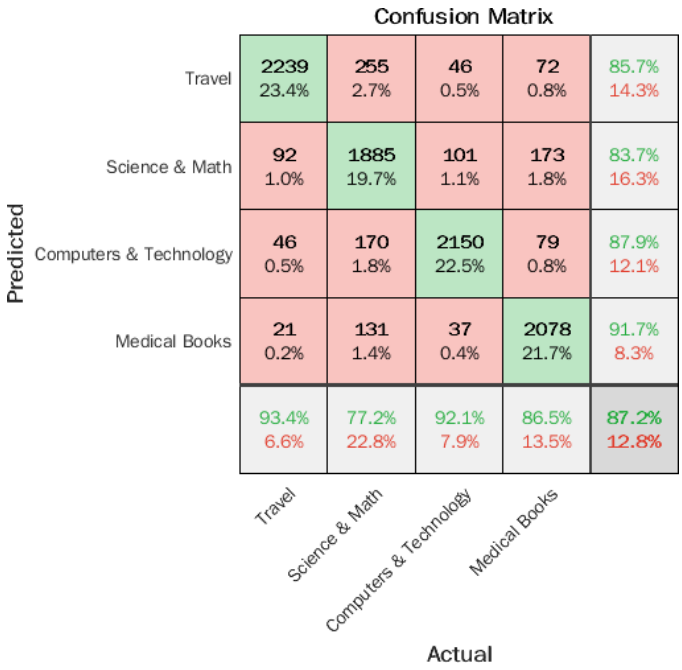

Fig. 5. Confusion matrix for SVM model on a set of test data.

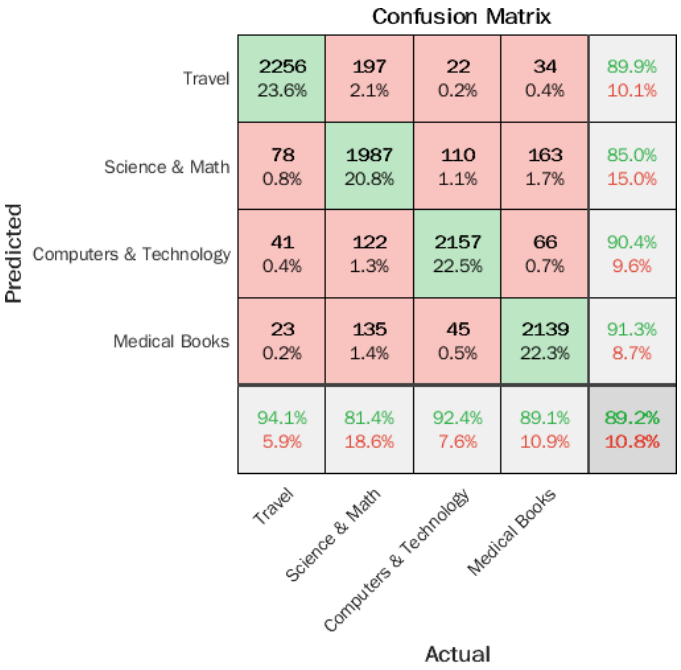

Fig. 6. Confusion matrix for NB classifier model on a set of test data.

books, the models performed poorly compared to the SVM and NB classifier.It is also worth noting that although the LDA is unsupervised, we still had to decide which genres corresponds to the different topics produced by the model. Not only were we able to improve the coherence score of the LDA model, we were also able to represent the accuracy of the model.

VI. Conclusion

As the number of books rapidly increases, libraries are finding it difficult to organize, classify and retrieve these books. In this paper, we attempted to use topic modeling as an approach to classify and retrieve these books more efficiently. A LDA model was trained and tuned using a dataset of book titles and genres.

Even though the LDA model was able to categorize the books into their respective genres, the models performed poorly compared to the SVM and NB classifier. The SVM model achieved an accuracy of 87.2% while the NB classifier achieved the highest with an accuracy of 89.2%. The LDA model achieved an accuracy of 28.1% and dropped slightly to 28.0% after selecting the best parameters. The reason for the slight drop in accuracy is due to the study's limitations. We had to select the best parameters based on the model with the highest topic coherence score instead of accuracy. For future work, during hyperparemeter tuning, the accuracy may be chosen as a metric to measure performance instead of the topic coherence score.

References

[1] David M. Blei and Andrew Y. Ng and Michael I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research, pp. 993–1022, 2003.

[2] Muzafar Rasool Bhat and Majid A Kundroo and Tanveer A Tarray and Basant Agarwal, Deep LDA : A new way to topic model, Journal of Information and Optimization Sciences, 823–834, 2020.

[3] Wang, Bo and Liakata, Maria and Zubiaga, Arkaitz and Procter, Rob, A Hierarchical Topic Modelling Approach for Tweet Clustering, 1004–3756, 2017.

[4] Dube, Skhumbuzo and Ajoodha, Ritesh, Improving Library Book Retrieval By Using Topic Modeling, IRTM 2021 (International conference on Interdisciplinary Research in Technology and Management in association with Taylor & Francis, Kolkata, India, 2021.

[5] Mohammed, Shaymaa and Al-augby, Salam, LSA & LDA Topic Modeling Classification: Comparison study on E-books, 2502–4752, 2020.

[6] Lei, Hao and Chen, Ying, "Exclusive Topic Modeling", 2021.

[7] Hsiang-Fu Yu AND Cho-Jui Hsieh AND Hyokun Yun AND S.V.N. Vishwanathan AND Inderjit S. Dhillon, A Scalable Asynchronous Distributed Algorithm for Topic Modeling, International World Wide Web Conference (WWW), 1340–1350, 2015.

[8] Iwana, Brian Kenji and Raza Rizvi, Syed Tahseen and Ahmed, Sheraz and Dengel, Andreas and Uchida, Seiichi, Judging a Book by its Cover, 2016.

[9] Sanderson, Mark and Croft, W, The History of Information Retrieval Research, Proceedings of The IEEE - PIEEE, 1444–1451, 2012.

[10] Rana, Mazhar Iqbal and Khalid, DR and Abid, Fizza and Ali, Armugh and Durrani, Mehr and Aadil, Farhan, News Headlines Classification Using Probabilistic Approach, 2015.

[11] Bogery, Raghad and Al, Nora and Aslam, Nida and Alkabour, Nada and Al, Yara and Ullah, Irfan, Automatic Semantic Categorization of News Headlines using Ensemble Machine Learning: A Comparative Study, International Journal of Advanced Computer Science and Applications, 2019.

[12] C. Xiao, P. Zhang, W. Chaovalitwongse, J. Hu, and F. Wang, "Adverse Drug Reaction Prediction with Symbolic Latent Dirichlet Allocation", AAAI, vol. 31, no. 1, Feb. 2017.

[13] Wang, Xiaofeng and Gerber, Matthew and Brown, Donald, Automatic Crime Prediction Using Events Extracted from Twitter Posts, Social Computing, Behavioral-Cultural Modeling and Prediction, 231–238, 2012.

[14] Sharma, Vasu and Kulshreshtha, Rajat and Singh, Puneet and Agrawal, Nishant and Kumar, Akshay, Analyzing Newspaper Crime Reports for Identification of Safe Transit Paths, 17–24, 2015.

[15] Buenaño-Fernández, Diego and Gonzalez, Mario and Gil, David and Luján-Mora, Sergio, Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach, IEEE Access, 2020.

[16] Runge, Steve, Boston College Librariee [Online image], https://answers.bc.edu/faq/136371, 2021.

[17] Lin, Jovian and Sugiyama, Kazunari and Kan, Min-Yen and Chua, Tat-Seng, Addressing cold-start in app recommendation: latent user models constructed from twitter followers, 283–292, 2013.

[18] Kolla, Bhanu, Categorizing Research Papers By Topics Using Latent Dirichlet Allocation Model, International Journal of Scientific & Technology Research, 2019.