Improving Library Book Retrieval By Using Topic Modeling

Skhumbuzo Dube
School of Computer Science and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
1608194@students.wits.ac.za

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

***Abstract*—Given the increasing number of books that libraries have, it becomes increasingly difficult for students to find books. We note that the current methodologies like the Dewey Decimal system, are becoming inefficient with book retrieval as the number of books increases. In this paper, we attempt to provide a content-based classifier to organize books to significantly improve retrieval over the current retrieval method. Support Vector Machine (SVM) was the best performing model achieving an accuracy of 79.8%, while latent Dirichlet allocation achieved an accuracy of 28.1%. We also note that the SVM model predicts each news headline in constant time. On average it takes 0.0029s to predict the category of a news headline.**

I. Introduction

As more documents, texts, articles, magazines, and information continue to be digitized and uploaded on the world wide web, it becomes increasingly difficult to search, classify and perform sentiment analysis by just using traditional techniques such as keyword search and manual text classification.

Libraries face a similar problem. Given the increasing number of books that the library has, it becomes increasingly difficult for students to find books. Attempts to solve the classification of books historically included separating them by author, genre, using manual methods that are not effective. These books are then sorted into a database and book retrieval is done linearly. The rise of machine learning provided content-based methods to organize their books for easy retrieval.

As the number of books increases, the more inefficient the current retrieval methodology becomes as book retrieval is done linearly ($O(n)$ runtime). This paper addresses this problem. This is achieved by organizing hundreds of texts by their respective topics using only content-based features. News headline texts are used to predict the category each headline text falls in.
We trained latent Dirichlet allocation (LDA) which is a generative model that explains the topics of observations by using a latent component, as well as a support vector machine(SVM) to predict the category of each news headline. Although LDA is an unsupervised model, we manually assigned categories to the topics generated by the keywords that appeared frequently, then measured the accuracy of the model (which in turn made it a semi-supervised model). Confusion matrices were used to gauge model performance. The best-reported accuracy was the support vector machine which achieved 79.8%.

II. Background and Related Work

A. Dewey Decimal system

The Dewey Decimal system was developed to organize and arrange collections of libraries. It was initially developed for Amherst College Library, in the 1800s, but over the years has been adopted by many libraries all over the world.

The Dewey Decimal system makes use of numbers to organize books and arrange books via the subject. Each book in the library is issued a shelf-number which is usually found on the spine of the book, and arranged in numerical order. The Dewey Decimal system is usually in this format: 945.805 TAB. The first set of numbers refers to the broad subject area and the next set of numbers after the decimal point refers to the sub-section of the subject area. After the numbers, there are usually 3 letters that refer to the author or title of the book. An example of the Dewey Decimal, taken from a library is shown on the next page. It should also be noted that

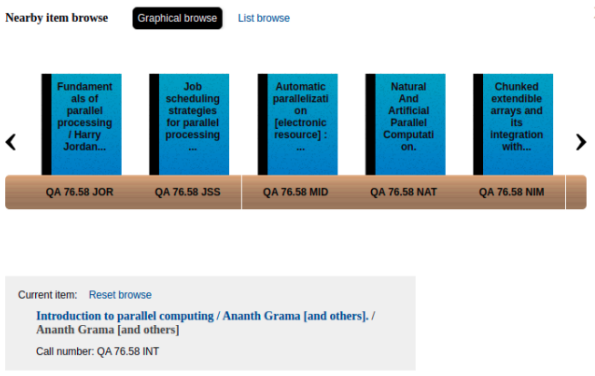

Fig. 1. Graphical representation of the Dewey Decimal system

the implementation of the system is usually done manually

by librarians. The books are then stored sequentially based on this method. Another widely used library system to manage books is called the Library of Congress Classification (LCC) [17].

III. Related Work

There have been multiple techniques that were used to solve questions or problems related to the research question proposed in this paper. One technique is to use a solution proposed by [4] to manage large sets of documents by using LDA as described in section 2.2. The main weakness of this paper is that how people would interact with the model output is not clearly defined and we are also not given any clear explanation on how to evaluate the performance of the algorithm.

1) Model Output Interaction: An important aspect of topic modeling is providing a way for users to be able to interact with the generated topics. One way to do this is by making use of network analysis, which enables us to find groupings of similar topics and provide navigation based on how the topics are related, as discussed by [8]. The key problem with this approach is that a link between topics is dependent on a user-defined value which will determine the effectiveness of the visualization and analysis. For example, having a document linked to just a few documents destroys much of the rich data that went into the topic model. [13] presents and provides an open-source implementation method for visualizing topic models by using browsing interfaces with two main types of pages: one for displaying generated topics and another for the documents. Selection bias on this method is a potential concern because a preliminary user study was only conducted on seven individuals. There are several similarities between the methods provided by [8] and [13], one being that both allow end-users to explore topics/documents via connections. Selecting features using an entropy weighting scheme to train an SVM model as done in [1], significantly improved the classification of text on the Reuter and TREC corpora.

2) Modern Implementations Of LDA: One disadvantage of using traditional topic modeling techniques is that they perform poorly on small datasets which could result in poor performance when organizing library books into various topics. [6] introduced a hierarchical topic modeling system with 2 stages named Dirichlet Multinomial Mixture mode (GSDMM) and latent features latent Dirichlet allocation (LFLDA) to gain competitive performance on small datasets. This technique performed better than LDA and topic models in clustering performance as well as topic coherence. However, this method was only applied on a twitter-tweet data set. The approach of using hierarchical topic modeling is similar to that used by [5]. The evidence presented thus far supports the idea that introducing some form of hierarchy in the implementation of LDA results in an improvement of the quality of topics generated. [11] is also an example of this.

Recently, researchers have shown an increased interest in merging deep neural networks and LDA. [7] provides a variation of LDA by using a deep neural network named two deep neural networks (2NN DeepLDA) and three deep neural network (3NN DeepLDA) to decrease computing processing in large corpora. This technique would potentially be useful in solving some of the computational challenges that might occur when developing a system to solve the proposed research problem.

IV. Methodology

In this paper, we attempt to classify categories in which news headline belong to by using content-based features, as a way of improving the retrieval time of a news article. We will train LDA as well as SVM to achieve this goal. More specifically, for SVM, we used a multiclass, error-correcting output codes (ECOC) model that reduces the problem of classification with three or more classes to a set of binary classification problems [2].

A. Data collection

The dataset used in this research is the "News of India" dataset, which consists of 3 features and contains over 2 million headlines. The initial dataset contained 3 features: "category", "publish date" as well as "headline text". Each data belongs to one of the 42 classes within "category". To train our models, we reduced the dataset to 4 classes, namely, "College", "Comedy", "Environment" as well as "Taste". The class distribution of each of these categories is even.

B. Preprocessing

To successfully implement our models the following steps were applied to the dataset to achieve a high degree of separation among classes:

1) Tokenization: Splitting text into word then converting all of those words to lowercase as well as removing punctuation.
2) Lemmatization - words in the third person will be converted to first person as well as converting past tense verbs to the future tense.
3) All words will be stemmed and stopwords will be removed

C. Feature selection

To train our models, we had to convert our text into a bag-of-words representation. This results in a 3576×1982 matrix (i.e, number of documents $\times$ number of words), which we then used to train our models.

V. Prediction and Evaluation

We use the following models to predict the category of a given headline: latent Dirichlet allocation (LDA) as well as support vector machines (SVMs).

A. LDA

[3] proposed a generative probabilistic model named LDA by introducing Dirichlet prior to Probabilistic Latent Semantic Analysis (PLSA), which is a statistical technique for the analysis of co-occurrence data. The concept behind latent Dirichlet allocation is that:

1) Each document can be represented by a distribution of topics.
2) Each topic can be described by a distribution of words.

The first step is to select the number of topics to be discovered and then once the number of topics has been selected, LDA will go through every word and randomly assign the word to one of the selected numbers of topics. After this step, we will have a non-optimal distribution of words in each topic as well as the documents represented in terms of topics, as stated in the 2 points mentioned above. Since this is not yet optimal, to better this representation LDA will analyze per document the percentage of words within the document that were assigned to a particular topic. For each word in the document, LDA will analyze over all the documents, the percentage of times that particular word has been assigned to a particular topic.
The algorithm will then go through each word in the document and calculate 2 probabilities:

1) $p(topict \mid documentd)$ = probability of words and that are assigned to t.
2) $p(wordw \mid topict)$ = probability of new assignments to t over all documents that come from the given word w.

After the above steps are repeated over multiple iterations, a "steady" state will be reached, where topic assignments are good.

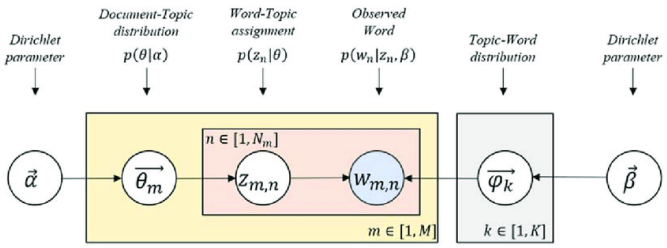

Fig. 2. Graphical representation of latent Dirichlet allocation proposed by [4]

We use LDA to explain the observed topics by using a latent component.

B. SVM

SVM is a supervised machine learning algorithm for regression and classification problems. SVMs objective is to find a hyperplane that best divides a data points into their respective classes. The equation of the hyperplane is given by:

$$f(x) = x^{'}\beta + b = 0$$

where $\beta \in R^d$ and $b \in \mathbb{R}$. The minimization problem is formulated by finding β and b that minimize $||\beta||$ such that for all data points (x_j, y_j):

$$y_j f(x_j) \geq 1$$

The support vectors are the data points that are the closest to the hyperplane as shown in the image below:

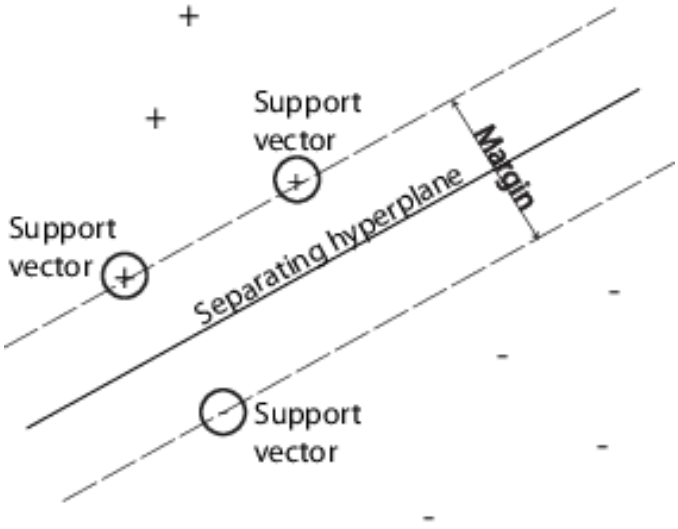

Fig. 3. Graphical representation of support vector machine

VI. Results and Discussion

In this section, we present the results of predicting the category of a new headline by using content-based features. In figure 4, we present our results for our SVM model and in figure 5 we present our results for our LDA model. For our LDA model, target classes are translated as follows: 1 - College, 2 - Comedy, 3 - Taste, 4 - Environment.

A. Classification

In this section, we will present the results of the classification algorithms.

We note that the SVM model outperformed the LDA model by a large margin. LDA could have performed poorly due to the short length of the headlines. We also note that the SVM model predicts each news headline in constant time. On average it takes 0.0029s to predict the category of a news headline.

Fig. 4. Confusion matrix illustration the SVM model on a set of test data. The SVM model achieves 79.8% accuracy with 319 correctly classified instances

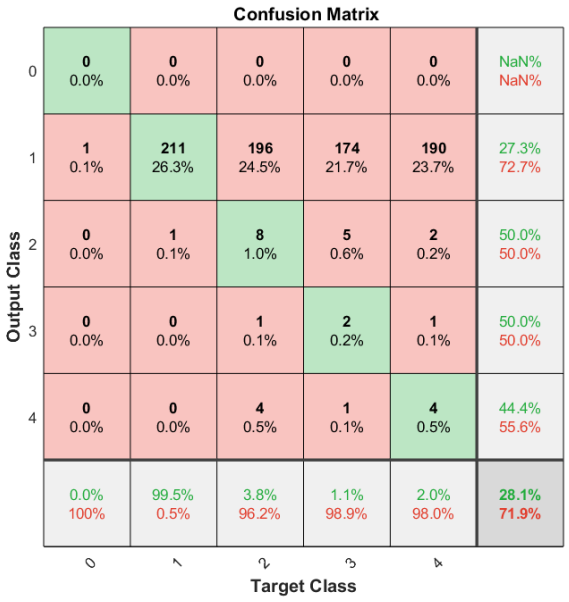

Fig. 5. Confusion matrix illustration the LDA model on a set of test data. The LDA model achieves 28.1% accuracy with 225 correctly classified instances

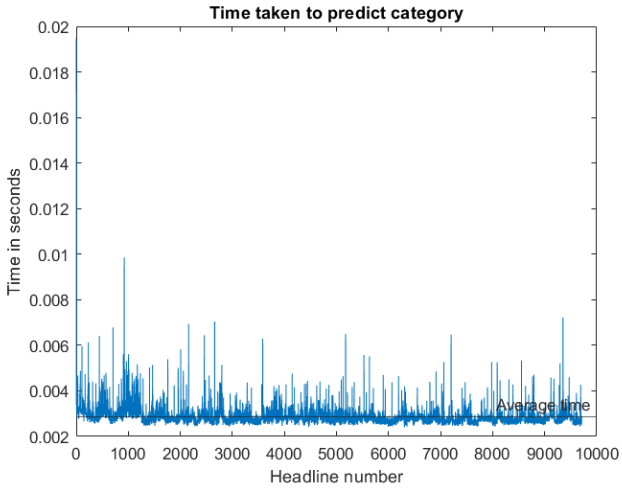

Fig. 6. Time it takes to predict respective categories across 9000 news headlines.

Although LDA achieved a low accuracy score, it can organize headlines into different categories based on just the content of the headline as well as make predictions. It is important to note that although LDA is an unsupervised model, we manually assigned categories to the topics generated by the keywords that appeared frequently, then measured the accuracy of the model (which in turn made it a semi-supervised model). The LDA model allows multiple topics for each document, by showing the probability of each topic. An example of this is shown in figure 7.

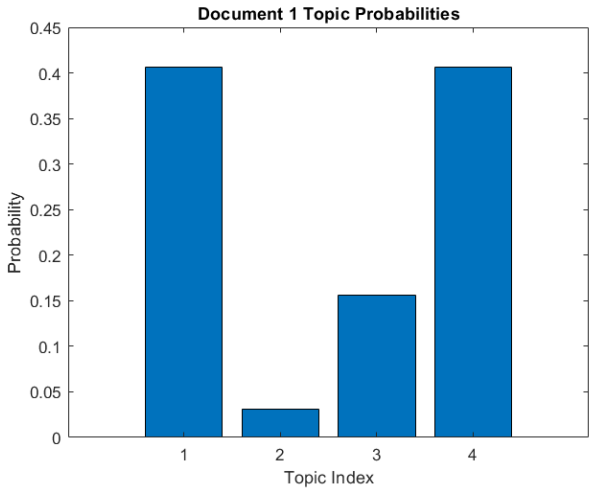

Fig. 7. Distribution of topics for document 1. Topic 1 and topic 4 have an equal probability while topic 2 has the lowest probability.

LDA is suitable when the categories of the headlines are not known, i.e, only the headlines are given. However, this method would require manual observation to label each of the clusters generated by the model.

VII. Conclusion

We note that the current library classification methodologies, like the Dewey Decimal system, are becoming inefficient as the number of books increases. In this paper, we provided a content based classifier to organize books to significantly improve retrieval time. We implemented LDA as well as a multi-class support vector machine ECOC model. The support vector machine is the best performing model achieving an accuracy of 79.8%, while LDA achieved an accuracy of 28.1%. LDA's relatively poor accuracy performance could be attributed to the length of the headline being short, resulting in a low degree of separation among the 4 classes. For future work, we suggest using the proposed method on digital library by running the model on digital content, such as PDF's, to organize and arrange the content digital library for faster retrieval. We also suggest introducing some form of hierarchy in the implementation of LDA results in an improvement of the quality of topics generated. [11] is also an example of this.

In section 1, a high-level overview of the problem domain, as well as the aim of this paper is stated. In section 2, a brief, high level, overview of the Dewey decimal system is given. In section 3, related work, as well as the results found in research papers related to the proposed research question was discussed in-depth and how those solutions could potentially solve the proposed question in this paper. In section 4, the methodology used is given as well as the steps taken to preprocess our dataset. In section 5, a brief description of the models used is given. Finally, in section 6 we present our results.

VIII. Acknowledgement

This work is based on the research supported in part by the National Research foundation of South Africa (Grant number: 121835).

References

[1] Z. Wang, X. Sun, D.Zhang and X. Li, "An Optimal SVM-Based Text Classification Algorithm", 2006 International Conference on Machine Learning and Cybernetics, 2006

[2] M. a. Bagheri, G. A. Montazer and S. Escalera,"Error correcting output codes for multiclass classification: Application to two image vision problems", The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012), pp. 508–513, 2012

[3] Blei. David M, Y. Ng. Andrew and Jordan, Michasel I., "Latent Dirichlet allocation", Journal of machine Learning research, pp. 993–1022, 2003.

[4] Blei, David M,"Probabilistic topic models", Communications of the ACM 55.4, pp. 77-84, 2012

[5] L. Rui, X. Wang, W. Deqing, Z. Yuan, Z He, and Zheng Xianzhu, "Topic Splitting: A Hierarchical Topic Model Based on Non-Negative Matrix Factorization", Journal of Systems Science and Systems Engineering, 2018

[6] W. Bo, L Maria, Z. Arkaitz, P. Rob, "A Hierarchical Topic Modelling Approach for Tweet Clustering", Journal of Systems Science and Systems Engineering, pp. 1004–3756, 2012

[7] M. Bhat, M. Kundroo, T. Tarray and B. Agarwal,"Deep LDA : A new way to topic model", Journal of Information and Optimization Sciences, pp. 1–12, 2019

[8] W. Scott,"Topic Modeling and Network Analysis", http://www.scottbot.net/HIAL/index.html@p=221.html, 2019

[9] L. Junseok, K. Ji-Ho, J. Sunghae, L. Hyunwoong, J. Dongsik, P. Sangsung, "Ensemble Modeling for Sustainable Technology Transfer", Sustainability, pp 2278, 2018

[10] X, Wei, L. Xin and G. Yihong, "Document Clustering Based on Non-Negative Matrix Factorization", Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 267—273, 2003

[11] A. Sharaff and N.K Nagwani,"Email thread identification using latent Dirichlet allocation and non-negative matrix factorization based clustering techniques", Journal of Information Science, pp. 200–212, 2016

[12] B. Pijush, "Modified Dewey Decimal Classification Theory for Library Materials Management", pp. 292–294, 2010

[13] C. Allison and B. David M, "Visualizing Topic Models", Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, 2012

[14] M. Rishabh, "News Category Dataset", 2018

[15] R.Mojarad, M. H. Scott, Ye, Zhan and Mayer,"Application of Clinical Text Data for Phenome-Wide Association Studies (PheWASs)", 2015

[16] A. Schofield, "Understanding Text Pre-Processing for Latent Dirichlet Allocation", pp.292–294, 2010

[17] B. Pijush,"Modified Dewey Decimal Classification Theory for Library Materials Management", pp. 292-294, 2010