Library Book Classification Using Topic Modelling

Umair Zunaid Bham
School of Computer Science and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
1871892@students.wits.ac.za

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

***Abstract*—With an increasing number of books being available in the public domain, there is a need for books to be classified in categories in inexpensive and efficient way. Whilst traditional classification systems such as the Dewey Decimal Classification, and the Library of Congress Classification are still made use of today, the process could be made simpler with the use of technology. This paper attempts to use Latent Dirichlet Allocation (LDA) - a form of topic modelling - to classify books into their respective categories, as well as explore how other machine models compare to the LDA model. The LDA model achieved an accuracy of 32.29%. The Naive Bayes, Support Vector Machine, and multinomial Logistic Regression models performed better achieving an accuracy of 85.45%, 86.12%, and 85.12% respectively.**

***Index Terms*—topic modelling, lda, book classification**

I. Introduction

Living in a digital age has made access to books, magazines, news articles among other texts increasingly available to the public in a digital format which in turn has made the process of storing, classifying, and retrieving information a problem.

Considering the traditional processes of classifying books in a library, such as the Dewey Decimal Classification and the Library of Congress Classification, the process is often time-consuming and requires numerous resources (since it is a manual process). Libraries face challenges similar to digitized texts with regards to information retrieval. Advances in the field of Natural Language Processing have now made it possible to apply machine learning techniques to automate the task of text classification using topic modelling to address this issue.

The purpose of this research is to apply Latent Dirichlet Allocation, a form of topic modelling, to assist with the classification of books in a library, as well as to see how well topic modelling compares to other machine learning techniques.

Headlines from articles between the years 2012-2022 were used to train an LDA model which achieved an accuracy of 32.29%. To provide a comparison of how well the LDA model performed, a Naive Bayes model, Support Vector Machine, and a multinomial Logistic Regression model were also trained on the same data. The Naive Bayes model achieved an accuracy of 85.45%, and the multinomial Logistic Regression achieved an accuracy of 85.12%. The Support Vector Machine performed the best overall, with an accuracy of 86.12%.

II. Background

A. Dewey Decimal Classification

The Dewey Decimal Classification is a library classification system that was adopted in libraries in the late 1800's to allow for new books to be placed in specific locations of a library based on its category/genre.

The system makes use of numbers to classify the books into categories and sub-categories with the first three digits before the period representing the category and the three digits after the period representing the sub-category. The three letters that follow the digits are used to represent the name of the author or the title of the book. After being manually classified by a librarian, the books are sorted linearly and stored in specific parts of a library [8].

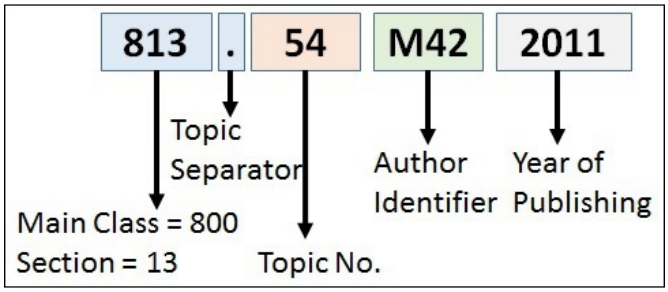

Fig. 1. Example of classifying a book using the Dewey Decimal Classification.

B. Topic Modelling

Topic modeling is an unsupervised classification method for unstructured data, akin to numeric data clustering, that finds natural groups of objects even when we do not know what we are looking for. One of the earliest topic models was the Probabilistic Latent Semantics Analysis (PLSA) model. As [4] noted, the PLSA model was subject to overfitting, and [5] called the model "inherently transductive" as there is no way of applying a learned PLSA model to new documents.

C. Latent Dirichlet Allocation

To address the issues of earlier topic models, [1] proposed the Latent Dirichlet Allocation (LDA) model in 2003. The LDA model is a generative model that attempts to find the topic a document in corpus belongs to based on the words in the document. More intuitively, each document in a corpus can be represented as a distribution of topics, and each topic can be represented by a distribution of words [2].

The LDA makes the following assumptions:

- bag of words assumption - in which the order of words does not matter but rather the frequency of words,
- and there is a predefined number of classes k.

After having a selected predefined number of topics k and randomly assigning them to each word in a document, the model goes through each document in the corpus and calculates the following probabilities:

1) $p(t|d)$: that determines the probability that a word w belongs to a topic t given a document d.
2) $p(w|t)$: that determines the proportion of documents that belong to a topic t given a word w.

Which are then repeated multiple times and the model updated.

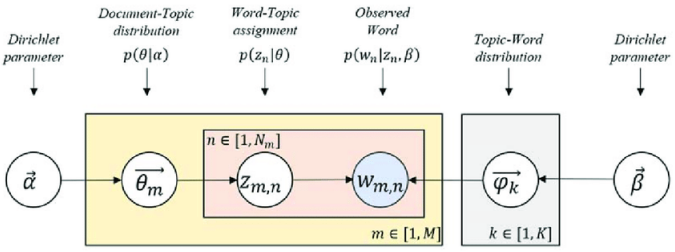

Fig. 2. Graphical representation of the LDA model.

III. Related Work

This section looks at previous work done with regards to using to the LDA model in various fields, as literature with regards to using topic modelling for book classification is limited.

Researchers such as [6] have shown that the LDA model has useful applications for analyzing social media posts. Reference [6] have used the LDA model to address the issue of a cold start - where recommendation systems struggle to new users. The authors proposed using 'pseudo-documents' where the followers of an app that a user may like are the 'pseudo-words' and one of their goals was to show how likely a target user will like an app. The authors used data from twitter between September to December 2012 and the results show that applying the LDA model yielded better results compared to other recommendation systems.

Reference [7] has also attempted to use LDA model for analyzing social media posts. Reference [7] extended the LDA model to come up with a model called the Multi-Aspect Sentiment Analysis for Chinese Online Social Reviews (MSA-COSRs) which was used to perform sentiment analysis on reviews. It differs from the traditional LDA model in that LDA is first applied to the data and thereafter a sliding window is used over the text to extract a local topic and associated sentiment by using sliding window approach. Although the model was specifically made for Chinese Reviews, the same methods can be applied to different data in order to gauge an idea of sentiment/polarity and associated local topic.

The LDA model, despite being useful for topic modelling, also has limiting factors as [5]points out that the standard LDA model produces topics that respond to global properties for an object (which means the results produced could lack nuance). Similar to the PLSA mode, the LDA model produces results which are global. In order to localise results, [5] proposed the Multi Grain LDA (MG-LDA) model. The MG-LDA model differs from the original LDA model by using a sliding window approach in which K^{gl} word distributions are drawn for global topics, and K^{loc} word distributions are subsequently drawn for local topics. Each window in the document is thereafter associated with a distribution to define the preference of local vs global topics.

Reference [5] trained an LDA and MG-LDA model using Gibbs sampling using three sets of data, two of which were a subset of reviews for MP3's and a subset of reviews for restaurants from Google Product and Local Search. For the MP3 dataset the MG-LDA model produced results which were more semantically coherent relative to the LDA model. However, both models did not produce great results for the restaurant dataset. The likely cause of this was due to the length of restaurant reviews, which were 4.2 sentences on average, which reveals the model may struggle on documents of shorter length.

[9] is another author that attempted to use the LDA model for performing library book classification. They also note that one of the limiting factors the LDA model has is that it struggles when content of the documents in the training corpus are too small and proposed using a hierarchical topic modeling system with 2 stages named Dirichlet Multinomial Mixture mode (GSDMM), and latent features latent Dirichlet allocation (LFLDA) on smaller datasets [10]. The authors in [10] have also used applied their model to social media posts similar to the authors mentioned above.

Furthermore, similar to this research, [9] used a semi-supervised approach by applying the LDA model, in addition to a Support Vector Machine, to headlines from the "News From India" dataset to categorize the headlines into 4 categories. The results of the research suggested that the LDA model performed poorly achieving an accuracy of only 28.1% whilst the Support Vector Machine achieved an accuracy of 79.8%.

IV. Research Methodology

This paper attempts to classify news articles to their appropriate categories using their headlines as features. To achieve this, a Latent Dirichlet Allocation (LDA) model, Naive Bayes model, Support Vector Machine (SVM) model, and multimodal Logistic Regression model are used.

A. Dataset and Feature Selection

The "News Category Dataset" is the dataset used in this research [11]. The dataset comprised of 6 features which included the 'category', 'headline', 'authors', 'link', 'short description', and 'date', with approximately 210000 headlines collected from the Huffington Post between the years 2012-2022. For the purpose of this experiment, only the 'headline' and 'category' were selected as features. The dataset was then reduced a balanced dataset to include only 5 categories with an equal number of training and testing data. The 5 categories used in this research include: 'BEAUTY', 'ENTERTAINMENT', 'POLITICS', 'TRAVEL', and 'WELLNESS'. The final training corpus comprised of 6750 samples for each category whilst the test corpus contained 2250 samples for each category.

B. Data Preprocessing

1) **Lowercase and Punctution Removal:** Each word in the headline is converted to lowercase and thereafter all punctuation marks are removed.
2) **Word Tokenization:** Each headline is split into tokens (i.e. each headline is broken down into individual words).
3) **Stopword Removal:** All tokens that are 'stopwords' are removed (e.g. words such as 'are', 'and', 'as', etc.).
4) **Lemmatization:** All the tokens are now reduced to their inflected forms.
5) **Bag of Words:** To make the corpus interpretable for the models, the training and test corpora were converted into a bag of words representation.

C. Theoretical Setup

The following python packages were used for the models used in this research:

- **Gensim:** for the LDAmodel.
- **SKLearn:** for the Naive Bayes (MultinomialNB), Support Vector Machine (SGDClassifier), and multinomial Logistic Regression (LogisticRegression with 'sag' solver) models.

D. Evaluation Metrics

The following evaluation metrics were used for the Latent Dirichlet Allocation model:

1) **Perplexity:** is a measurement that is used to determine how well a probability model can predict a sample [14]. The lower the perplexity of a model, the better it is at predicting a sample.
2) **Topic Coherence:** is an evaluation metric used to measure the similarity between the words most frequently associated with a topic [13]. Models that perform well tend to have a coherence score closer to 1.

Lastly, to evaluate how well each model performed, the **accuracy** was calculated with the formula below:

$$Accuracy = \frac{Predicted}{Total} \tag{1}$$

where $Predicted$ is the number of accurately predicted samples, and $Total$ is the total number of samples.

V. Results

The outcomes of the models used in this research will be presented and discussed in this section. The LDA model had a perplexity of -9.64866, and a coherence of 0.30841. The LDA model achieved an accuracy of 32.29%, whilst the Naive Bayes, Support Vector Machine, and Logistic Regression models achieved an accuracy of 85.45%, 86.12%, and 85.12% respectively. A summary of how well the models predicted the correct category is presented in the form of confusion matrix below where, Fig. 3 is the confusion matrix for the LDA mode, Fig. 4 is the confusion matrix for the Naive Bayes model, Fig. 5 is the confusion matrix for the Support Vector Machine model, and Fig. 6 is the confusion matrix for the Logistic Regression model.

TABLE I
Summary of Models and Results

Model	Accuracy
LDA	32.29%
Naive Bayes	85.45%
Support Vector Machine	86.12%
Logistic Regression	85.12%

The evaluation metrics suggest that the LDA model performed comparatively worse than the other models used in this research. It is important to note since the LDA model is an unsupervised model, labels had to be manually assigned to each category based on the word association for each topic.

VI. Conclusion

Acknowledging that libraries may face issues with information retrieval as there is an increase in books, using the Latent Dirichlet Allocation topic model to attempt to make the process more efficient has proved to be ineffective. The model performed the worse out of the machine learning models used with an accuracy of 32.29% compared to the best performing model, which was the Support Vector Machine, which achieved an accuracy of 86.12%.

The failure to produce more accurate results using the LDA model could be a consequence of the size of the headlines in the training corpus as they were relatively small. In order to improve the accuracy of the LDA model, extensions to model such as the MG-LDA model may be needed, or the model should be trained on a corpus with larger document size. This should be explored in future research.

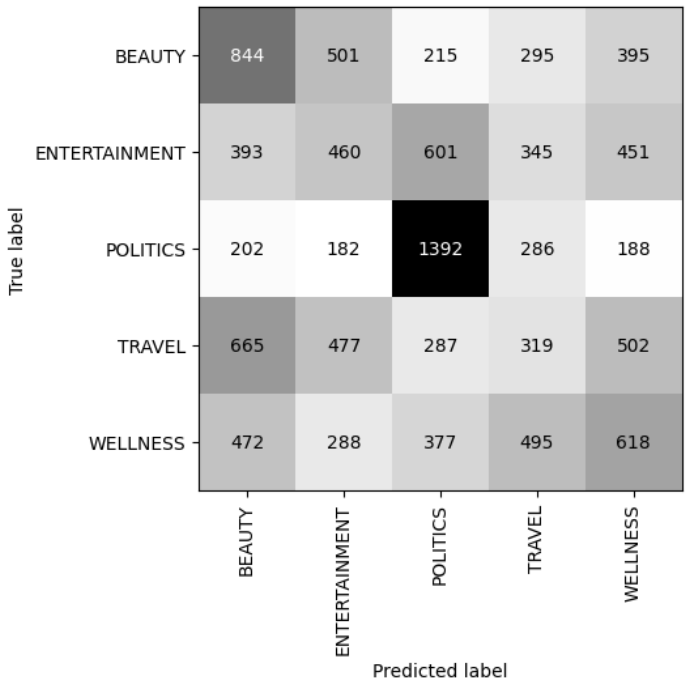

Fig. 3. Confusion matrix for LDA model on test corpus.

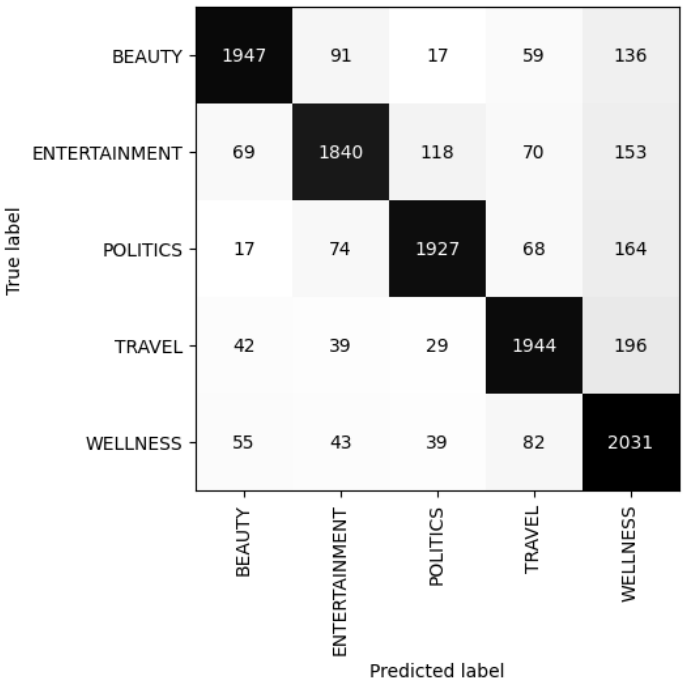

Fig. 5. Confusion matrix for Support Vector Machine model on test corpus.

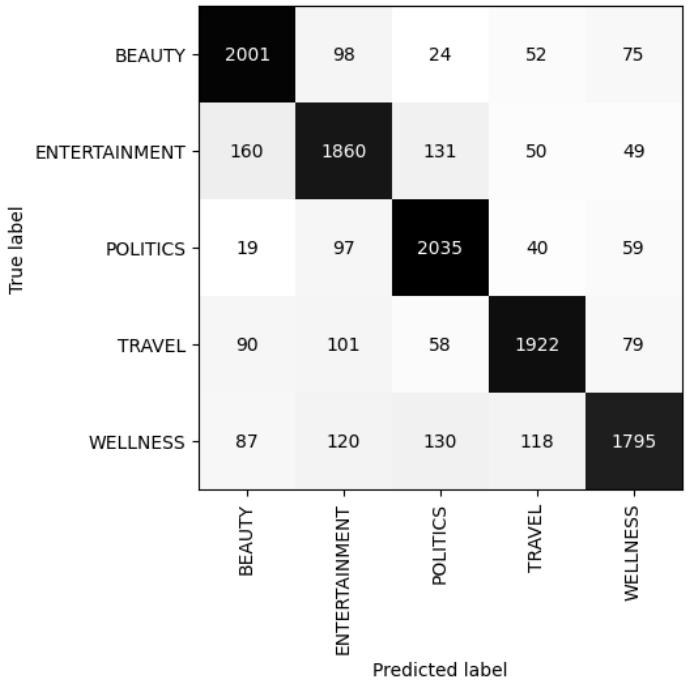

Fig. 4. Confusion matrix for Naive Bayes model on test corpus.

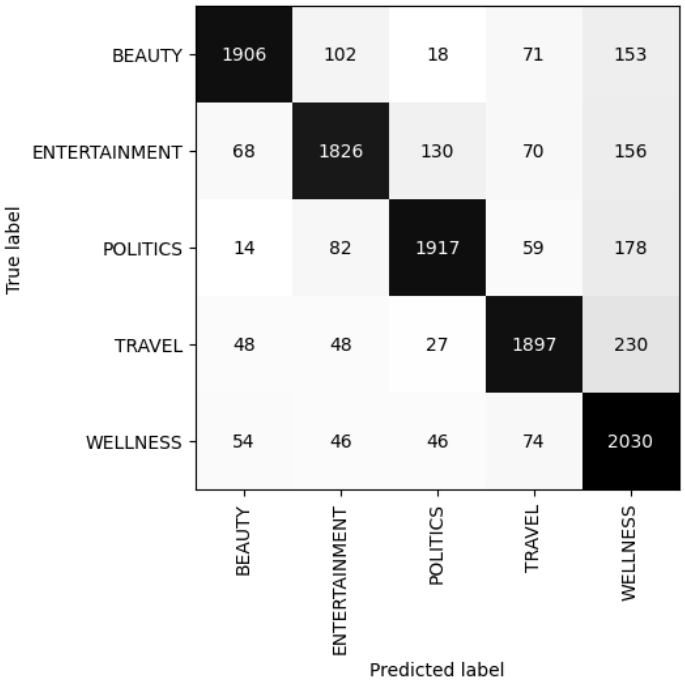

Fig. 6. Confusion matrix for Logistic Regression model on test corpus.

REFERENCES

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

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

[3] Blei, David M., and John D. Lafferty. "Dynamic topic models.", Proceedings of the 23rd international conference on Machine learning. ACM, 2006.

[4] Wu, Hao and BU, Jiajun and Chen, Chun and Zhu, Jianke and Zhang, Lijun and Liu, Haifeng and Wang, Can and Cai, Deng, "Locally discriminative topic modeling", Pattern Recognition, vol. 45, pp. 617-625, 2012.

[5] Titov, Ivan and McDonald, Ryan, "Modeling Online Reviews with Multi-Grain Topic Models", Proceedings of the 17th International Conference on World Wide Web, pp. 111–120.

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

[7] Fu Xianghua and Liu Guo and Guo Yanyan and Wang Zhiqiang, "Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon", Knowledge-Based Systems, vol. 37, pp. 186-195, 2013.

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

[9] 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 and Francis, Kolkata, India, 2021.

[10] 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.

[11] Misra, Rishabh. "News Category Dataset.", arXiv preprint arXiv:2209.11429, 2022.

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

[13] S. Syed and M. Spruit, "Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation", 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 165-174, 2017.

[14] Hoffman, Matthew, Francis Bach, and David Blei. "Online learning for latent dirichlet allocation", Advances in Neural Information Processing Systems 23, 2010.