Supervised Machine Learning and Deep Learning Classification Techniques to Identify Scholarly and Research Content

Huilin Chang
School of Data Science
University of Virginia
Charlottesville, VA
hc5hq@virginia.edu

Yihnew Eshetu
School of Data Science
University of Virginia
Charlottesville, VA
yte9pc@virginia.edu

Celeste Lemrow
School of Data Science
University of Virginia
Charlottesville, VA
ctl7t@virginia.edu

2021 Systems and Information Engineering Design Symposium (SIEDS) | 978-1-6654-1250-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/SIEDS52267.2021.9483792

***Abstract*—The Internet Archive (IA), one of the largest open-access digital libraries, offers 28 million books and texts as part of its effort to provide an open, comprehensive digital library. As it organizes its archive to support increased accessibility of scholarly content to support research, it confronts both a need to efficiently identify and organize academic documents and to ensure an inclusive corpus of scholarly work that reflects a "long tail distribution," ranging from high-visibility, frequently-accessed documents to documents with low visibility and usage. At the same time, it is important to ensure that artifacts labeled as research meet widely-accepted criteria and standards of rigor for research or academic work to maintain the credibility of that collection as a legitimate repository for scholarship. Our project identifies effective supervised machine learning and deep learning classification techniques to quickly and correctly identify research products, while also ensuring inclusivity along the entire long-tail spectrum. Using data extraction and feature engineering techniques, we identify lexical and structural features such as number of pages, size, and keywords that indicate structure and content that conforms to research product criteria. We compare performance among machine learning classification algorithms and identify an efficient set of visual and linguistic features for accurate identification, and then use image classification for more challenging cases, particularly for papers written in non-Romance languages. We use a large dataset of PDF files from the Internet Archive, but our research offers broader implications for library science and information retrieval. We hypothesize that key lexical markers and visual document dimensions, extracted through PDF parsing and feature engineering as part of data processing, can be efficiently extracted from a corpus of documents and combined effectively for a high level of accurate classification.**

***Index Terms*—machine learning, data modeling**

I. Introduction

Despite the constant exponential increase of written content on the internet and the expansion of online publication opportunities and platforms, there is wide variance in content visibility, accessibility, and longevity. Furthermore, there can be substantial discrepancies in access among potential consumers of content, particularly when it comes to access to legitimate library collections that provide quality material in support of education and research. The Internet Archive (IA), a non-profit founded in 1996 and one of the largest open-access digital libraries, aims to bridge those gaps, seeking to preserve digital content and enhance its visibility through curation and collection design, and provide wide access, especially to users who may not otherwise have library resources available to them. The IA's comprehensive collection offers 28 million books and texts, with ongoing efforts to curate and organize its vast trove of content. One element of that effort involves identifying research content from among documents culled from its web crawl activity, so that it can be further organized and made available for academic purposes. This identification effort helps to democratize the accessibility of scholarly content for educational and research activity and also provides a long-term archival home for content at risk of slipping through the digital interstices and vanishing from the web due to lack of funding, unclear provenance and line of responsibility for preservation, disruption to digital storage infrastructure, or other reasons [1] [5].

Given the massive volume of the IA's text data, and how much of it is unlabeled at ingest, an accurate and computationally efficient machine learning classification model is needed for initial identification of research and scholarly material. Building on a hypothesis that key lexical markers and visual, physical document elements can combine into a set of features that provide a high level of accurate classification, we developed and compared the performance of several machine learning classification algorithms to determine the best approach for identification of research documents. Using a dataset of 60,000 text documents, we deployed data extraction and feature engineering approaches to identify lexical and structural features for consideration within the models. We employed a dual-pronged strategy, developing models with both machine learning and deep learning methods such as Logistic Regression, XGBoost, and a custom 2-layer neural network, using text-based features. Additionally, we generated image data and built three Convolutional Neural Network (CNN) models for image classification for more unique cases, such as papers written in non-Romance languages. Text-based models included text- and document-based features, while image-based models extracted the first image from each

978-1-6654-1250-6/21/$31.00 ©2021 IEEE

document for analysis.

Developing models to identify and classify research requires several considerations, especially in the context of the IA's mission to address preservation and representation, in addition to open access. It was important to ensure that the model would capture a broadly inclusive body of scholarly material, reflecting the "long tail" distribution principle that includes a range from mainstream material to documents with lower visibility and usage [2]. In addition, for both the training data and feature engineering, it was important to consider ways to mitigate implicit bias that could prevent legitimate research products from being accurately identified [6]. Finally, when developing the lexical and content-based features, we considered key domain markers that would be representative of widely-accepted criteria and standards of rigor for research or academic work, to ensure credibility on the resulting repository for scholarship. While our work focused on the IA corpus, the underlying principles of the issue of research classification, the key elements of the model, and the results have implications and relevance for library science more broadly, in terms of how to efficiently identify scholarly work, ensure an appropriately diverse range of documents, and better match supply and demand across the full length of the "long tail" of available material [4]. Our approach can be leveraged and modified for other contexts by organizations engaged in building digital research collections.

II. Data Processing

A. Dataset

Our total dataset consisted of approximately 60,000 PDF text files provided by the IA, from four different text datasets ("fatcat_longtail_lang", "fatcat", "gwb_random", and "long-tail_crawl"), which contained research and non-research documents and were organized in groups that essentially provided labeling. Based on that initial curation, we created a combined, labeled dataset to use for training, validation, and testing of the models. In total, 48 languages are represented, with the most prevalent five languages being English (56% of all papers), followed by Portuguese (4%), Spanish (4%), German (3%), and French (3%). The dataset was balanced, with approximately 30,000 research papers and 30,000 non-research papers.

B. Data Pipeline

Our data pipeline consists of two main processes: extracting textual information from PDFs for the text-based model features and converting the first page of a PDF to an image for the image classification models. Using a Python package called *PyMuPDF*, we ingested PDF files and extracted metadata such as author, title, page dimension, and the number of pages. Furthermore, the package also allowed us to retrieve the text from each page. After obtaining text, the pipeline then detects the primary language and searches for a list of keywords that we specified based on domain knowledge and library science principles as likely markers of research products; these are discussed further in the next section. The language of the PDF was detected using the *langdetect* Python package. We initially used Google's Translate API to identify and translate keywords from non-English languages to English in real time, but soon realized that the API limits the number of requests in a given time interval, making that infeasible. Instead, we translated all of the keywords to the 40 non-English languages and stored the translations in an SQLite database. Initially, this process was completed on a single processor, but we noticed the run time for 60,000 PDFs was several hours. While that was manageable for purposes of our project, we wanted to identify a more efficient approach for the broader application to an entire large archive. Therefore, we concluded using multiprocessing and spreading the work over 40 processors significantly reduced the processing time from 8-10 hours to roughly 24 minutes for the same number of documents. Extrapolating to a much larger dataset, this sets the basis for a more efficient data engineering workflow. The second major data wrangling process consisted of using PyMuPDF to retrieve the PDF's first page and convert it into a PNG image. Initially, this process was also using one processor, but we used multiprocessing and reduced the processing time to 5 minutes.

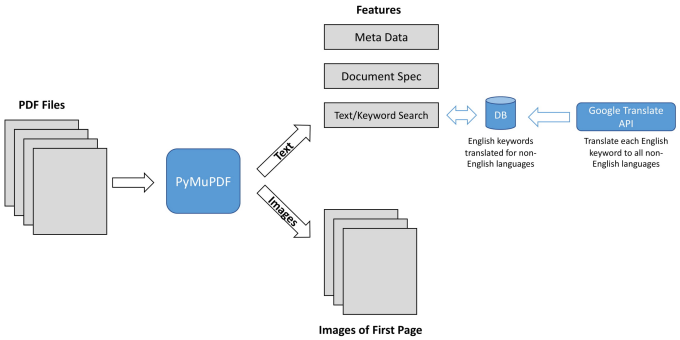

Fig. 1. Data Pipeline

Our dataset contained numerous missing data, though missing values were largely concentrated under the title, author, subject, and producer features, which were meta-data elements extracted with PyMuPDF and dependent on those being available in the original document. To solve this problem, we decided to convert their multi-valued features to a binary category feature, to provide transparency about whether a sufficient number of documents contained these data elements, to facilitate a decision about whether they could be used defensibly as a model feature. Using an encoder, we represented missing data for non-numerical columns as 0 and 1 for non-missing data. For numerical columns, we presented missing data as a "-1."

C. Feature Engineering

We extracted 11 features through PyMuPDF that focused on the length, format, and physical dimensions of the document, including number of pages, PDF format, title, author, subject,

producer, height, width, PDF size, text, and word count. We also created a language variable to denote the language of each document. Most of these features were easily retrieved through PyMuPDF and had a sufficiently low number of missing values to include credibly in a model. We also generated a series of binary and composite features that focused on key content dimensions of a research paper or product. We named the three composite features ”structure”, ”content”, and ”association.” The ”structure” variable included words that represented the typical structural framework of a research paper: abstract, introduction, conclusion, reference, and table of contents. The ”content” variable focused on typical language that would appear in a research paper, regardless of its specific topic: research, analyze, result, table, investigation, explain, theory, study, paper, data, and performance. Finally, the ”association” variable covered organizational and affiliation characteristics that can signal a research product: journal, association, organization, doi (digital object identifier), university, school, board, and publish. Our exploratory data analysis indicated that these various markers were fairly common across languages and geographic areas.

III. Exploratory Data Analysis

Our exploratory data analysis demonstrated that the dataset is balanced, with approximately 30,000 research papers and 30,000 non-research papers. The label for research papers is 1 and the label for non-research papers is 0, Fig. 2.

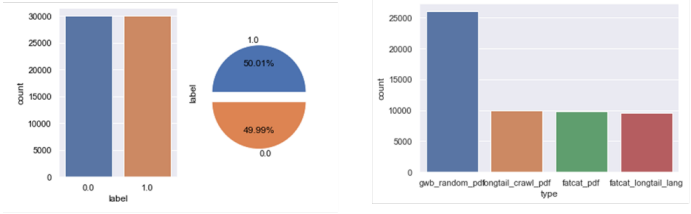

Fig. 2. Dataset Count and Document Type Distribution

Fig. 3 shows the width versus height of the four types of papers, illustrating clear binomial distributions for gwb_random and longtail_crawl, while fatcat_longtail_lang and fatcat show relatively more uniform distributions.

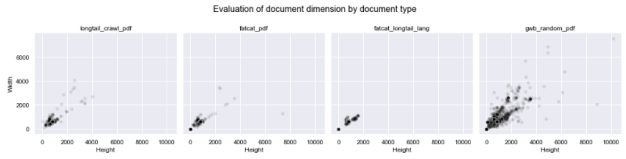

Fig. 3. Height vs Width

Fig. 4 presents distributions of the lengths of the four types of papers in pages. The gwb_random papers show a high probability of being a single page or zero pages, while the research paper datasets show more of a range, demonstrating a substantial difference between the broad document types. This also shows that, for our initial investigation, it was important to extract the entire text of each document, to ensure that all relevant information for feature engineering was available. We also assessed correlations between features.

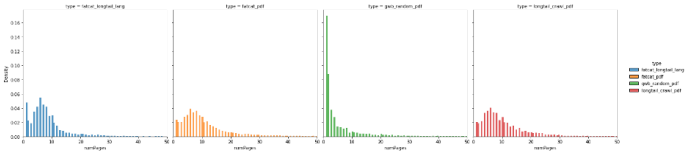

Fig. 4. Page Distribution by Document type

In the heatmap shown in Fig. 5, multicollinearity can be found in the text-based features among structure, content, and association. The correlation between structure and content is 41%, between structure and association it is 36%, and between content and association it is 58%. This was an instructive exploratory data analysis finding that demonstrated a potential linear relationship among one or more variables, which could affect eventual model performance depending on feature combinations. Given how each of the words could be associated not only with the other words within a given composite variable, but also associated across the groups for each variable, the multicollinearity was not necessarily surprising. This signaled that we would need to further assess and parse the relative significance of the different language-related features when constructing and experimenting with different models. The correlation between the label and the text-based features of structure, content, and association are 15%, 21%, and 14%, respectively. The correlation between label and the features of language and text length are 9% and 16%, respectively. For label and other features, the correlations were all less than 8%.

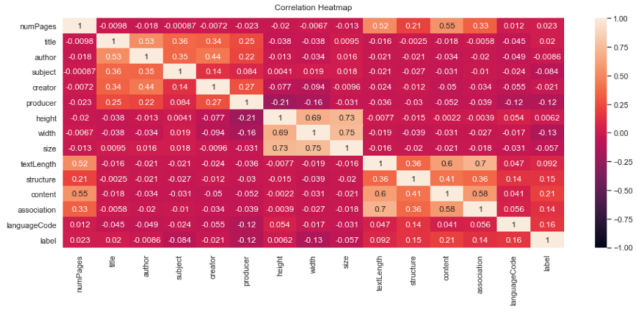

Fig. 5. Correlation Heatmap of Features

IV. Models and Results

The full set of models and results are summarized below in Table I, denoting the accuracy, F-score, precision, and recall for each of the models to show comparative performance.

TABLE I
MODEL RESULTS

Model	Accuracy	F Score	Precision	Recall
Text Based				
Logistic Regression	76.90%	79.00%	79.00%	79.00%
XGBoost	90.20%	92.50%	90.10%	90.60%
2-layer NN	89.10%	—	—	—
Image Based				
Xception	90.30%	90.11%	93.50%	86.66%
VGG16-1	88.92%	89.04%	89.70%	88.50%
VGG16-2	89.27%	89.68%	88.05%	91.80%

A. Text-Based Models

The text-based models, defined by their use of linguistic and physical document features, used Logistic Regression, XGBoost, and a custom two-layer neural network algorithm.

1) Logistic Regression and Bayesian Model Averaging (BMA): Logistic Regression combined with Bayesian Model Averaging (BMA) was used to address problems of model selection and eliminating unnecessary features. Table II shows the feature importance derived from BMA. The BMA results suggested keeping all features except author, creator, and text length, which had a probability of less than 100%.

TABLE II
FEATURE IMPORTANCE FROM BAYESIAN MODEL AVERAGING

Variable Name	Probability	Avg Coefficient
const	1	-0.701
numPages	1	-1.797
title	1	0.178
author	0.953	0.043
subject	1	-0.169
creator	0	0
producer	1	-0.314
height	1	0.635
width	1	-0.922
size	1	-0.551
textLength	0	0
language	1	-0.011
structure	1	0.065
content	1	0.013
association	1	0.003

2) XGBoost: XGBoost works with numerical data only; therefore, we conducted one-hot encoding of the data set into sparse matrices for categorical data. The XGBoost top 10 feature selection function was used to further optimize the model; the top 10 features were adopted for hyperparameter optimization, as shown in Fig. 6 (which also shows the ROC curve). XGBoost hyperparameters are divided into 4 categories: general, booster, learning task, and command line parameters. Three-fold cross validation was adopted, and hyperparameters were optimized by maximizing the average validation AUC across the three validation sets. The grid search uses the estimator, learning rate, maximum depth, and colsample_bytree to achieve the lowest mean squared error.

3) Two-Layer Neural Network: This neural network adopts two hidden layers and 50 epochs. We split the data into training, validation, and testing set before any modeling, with

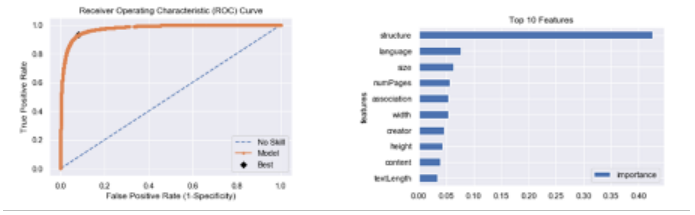

Fig. 6. **Left** Receiver Operating Characteristic (ROC) **Right** : Top 10 Features

80% for training, 10% for validation, and 10% for the testing set.The first hidden layer has 2028 neurons, and the second layer has 1024 neurons. The efficient ADAM optimization algorithm was used for model accuracy improvement. We used the sigmoid activation function, which is well-suited for binary classification in this study.

B. Image-Based Models

Image classification with deep learning for computer vision has made drastic improvements in the past decade and has been proven to achieve remarkable performance on complex visual tasks. Thus, we decided to use Convolutional Neural Networks (CNN), a deep learning neural network architecture, for classifying the images of the first page of PDFs. Instead of creating our own CNN architecture, we decided to leverage pre existing state-of-the-art CNN models, Xception and VGG16, using transfer learning. Transfer learning is a deep learning technique where a trained architecture is reused to solve a new computer vision problem. By using transfer learning, we can reduce the time needed to train and develop a model, obtain better model performance, and reduce the need for a lot of data.

We split our images into a training, validation, and testing set before any modeling, with 72% (42687) for training, 18% (10671) for validation, and 10% (5929) for the testing set. Before we downloaded each model architecture, we created an input layer with set image width and height. Then we passed the input layer into a Keras data augmentation sequential layer, which rescaled every image input from a [0, 255] range to a [0, 1] range.

1) Xception: Xception merges GoogLeNet and ResNet's architectural design, claiming the usage of fewer parameters, less memory, and fewer computations than a regular convolution layer, yet with better performance. Our Xception architecture included additional layers of GlobalAveragePooling2D, BatchNormalization, Dropout, and a Dense softmax layer with two classes. Once the model was designed, we compiled our model using a sparse categorical cross-entropy loss function and a Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.01. We also incorporated an EarlyStopping callback to stop training when the accuracy stopped improving and a ReduceLROnPlateau callback to reduce the learning rate when validation accuracy stopped improving.

2) VGG16: VGG16 thoroughly evaluated that increasing the depth to 16-19 weight layers with small convolution filters leads to improvement in comparison to prior configurations

[3]. We developed two different VGG16 models that differed in the number of layers and type of layers used. Our first VGG16 model (VGG16-1) had additional layers of GlobalAveragePooling2D, BatchNormalization, Dropout, and a Dense softmax layer with two classes. Our second VGG16 (VGG16-2) model had additional layers of GlobalAveragePooling2D, BatchNormalization, Dropout, Flatten, and a Dense softmax layer with two classes. Once both models were designed, we compiled our models using a sparse categorical cross-entropy loss function and an Adam optimizer with a learning rate of 0.0001. Like the Xception model, we included an EarlyStopping and ReduceLROnPlateau callback.

V. Discussion

Overall, we observed strong performance across the text-based and image-based models. XGBoost shows the best performance based on its gradient boosting machines and has the dominant performance among the text-based models. XGBoost is written to scale seamlessly with large datasets, the computation time is much lower compared to the 2-layer NN in this study, and the performance is substantially higher than the Logistic Regression model.

Within the XGBoost model, the finding that the structure variable was the strongest predictor seems to support the hypothesis that research paper formats are generally consistent enough such that widely-accepted structural elements (eg, abstract, introduction, conclusion, etc) are sufficient markers of a research product. The ranking of the language variable as the second-strongest is not surprising, given that English is the dominant language in the dataset. Although the content variable was one of the weakest predictors in the XG Boost model, the finding of multicollinearity between these two features suggests that this correlation may be captured within the structure variable. There may be less correlation between the structure and association variables, given the association variable's higher-ranked feature importance.

The image-based models had similar performance to the text-based XGBoost model. The Xception model had the highest accuracy of all of the models, but in this particular use case, we want a model with the highest recall, as it indicates the proportion of real research papers we are accurately labeling as such. Thus, we would conclude that the VGG16-2 model is the preferred model.

The text-based approach allows us to capture numerous data features from PDFs in a structured tabular format. In contrast, the image-based process creates one PNG image of the first page of a PDF. Although the text-based method allows us to obtain several data fields, it does have a significant drawback, the amount of time to pre-process the data. The text-based pre-processing takes 24 minutes, while the image pre-processing takes 5 minutes, nearly five times as fast. Even though the image pre-processing is faster, this does not translate to poor model performance as the image-based models had similar or even better accuracy than the text-based models. Ideally, the image-based method seems the most promising approach due to its pre-processing time and possible improvements that we can make to the CNN models.

VI. Conclusion

This study demonstrates that several types of classifiers can be used successfully to correctly distinguish research papers from non-research documents using text-based and image-based models. The results also demonstrate the potential for future work to further identify an even more pared-down set of features that provide greater efficiency while maintaining the same performance on accuracy and recall. For example, the 60,000 documents used in this study come in various lengths, suggesting the potential to investigate whether the text-based models can be constructed more efficiently by extracting fewer pages to save computation time. In addition, the use of a contextualized word-embedding such as ElMo could allow us to create word vectors on the fly by passing text through a deep learning model rather than having a set dictionary of keywords. The strong performance of image classification from only the first page of a document also demonstrates these models' potential. Future work could include further improvements by leveraging different techniques like Intra-Domain transfer learning. Using the concept Intra-Domain transfer learning, we can train region-specific models by cropping a whole image into sections. These regional-based CNN models for document classification have proven to improve accuracy on well-known datasets such as the popular RVL-CDIP document image dataset. This method is a simple and effective way of enhancing our CNN models. Furthermore, we can also increase the resolution of the images to improve the network performance potentially. The value of the dual-pronged approach demonstrates that both text and image-based model algorithms yield effective, accurate classification options for different contexts, depending on preference and feasibility to work with text or image data. Future work could also focus on expanding the size and diversity of the dataset to see how each model's individual performance, and comparative performance, changes, depending on distribution and representation of different topics, fields, and disciplines, as well as broader geographic range, languages, and diversity of authors. This is relevant not only for ensuring robust accuracy under broader parameters within the data, but also to ensure inclusive identification that mitigates implicit bias about what may constitute a research product [6]. Overall, the approaches and features leveraged in the study provide a foundation for continued work on binary classification of documents, to help institutions conduct initial assessments of archives or other document respositories.

Acknowledgment

We would like to thank Bryan Newbold, our sponsor and technical point of contact at the Internet Archive, for his thought partnership and support in providing data for us to work with. We would also like to thank the Internet Archive for sponsoring an SDS Capstone project. Finally, we would

like to thank Professor Rafael Alvarado for his guidance and support as we conceptualized and developed our project.

REFERENCES

[1] D. Kwon, "More than 100 scientific journals have disappeared from the Internet," Nature News, 10-Sep-2020. [Online]. Available: https://www.nature.com/articles/d41586-020-02610-z. [Accessed: 10-Apr-2021].

[2] C. Anderson, "The Long Tail," Wired, 01-Oct-2004. [Online]. Available: https://www.wired.com/2004/10/tail/. [Accessed: 10-Apr-2021].

[3] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv.org, 10-Apr-2015. [Online]. Available: https://arxiv.org/abs/1409.1556. [Accessed: 10-Apr-2021].

[4] L. Dempsey, Libraries and the Long Tail: Some Thoughts about Libraries in a Network Age, Apr-2006. [Online]. Available: http://www.dlib.org/dlib/april06/dempsey/04dempsey.html. [Accessed: 10-Apr-2021].

[5] M. Laakso, L. Matthias, and N. Jahn, "Open is not forever: a study of vanished open access journals," arXiv.org, 22-Feb-2021. [Online]. Available: https://arxiv.org/abs/2008.11933. [Accessed: 10-Apr-2021].

[6] T. Padilla, "Responsible Operations: Data Science, Machine Learning, and AI in Libraries," OCLC, 08-Dec-2019. [Online]. Available: https://www.oclc.org/research/publications/2019/oclcresearch-responsible-operations-data-science-machine-learning-ai.html. [Accessed: 10-Apr-2021].