Machine learning for document classification in an archive of the National Afrikaans Literary Museum and Research Centre

Susan Brokensha
University of the Free State
broksha@ufs.ac.za
https://orcid.org/0000-0001-6166-3981

Eduan Kotzé
University of the Free State
kotzeje@ufs.ac.za
https://orcid.org/0000-0002-5572-4319

Burgert Senekal
University of the Free State
burgertsenekal@yahoo.co.uk
https://orcid.org/0000-0002-1385-9258

Received: 02 November 2021
Revised: 14 August 2022
Accepted: 28 September 2023

Abstract

Most archives were established before the digital age, where hardcopies of much smaller volumes were archived. In the information age, archives struggle to accommodate the large volumes of material produced. In addition, many archives, including in South Africa, had to contend with budget cuts that reduced the number of staff available. If digital material is not archived now, it creates the risk of gaps in the historical record in the future. In addition, with digital humanities gaining wider acceptance, large corpuses of digital material are needed, which archives could provide. This study's aim was to investigate whether document classification using machine learning classifiers is feasible in a South African archive context, with a focus on the National Afrikaans Literary Museum and Research Centre (NALN). The researchers created and trained a document classification model and tested it for accuracy against human classifiers. It followed a basic linguistic approach to prepare specific text documents for text classification, in terms of Galloway and Roux's (2019) six categories, namely articles, media reports, books, interviews, reviews, and dissertations and theses. The classification was done using two annotators, after which the annotated corpus was employed as training data for machine learning models. Following Rolan et al. (2018), Suominen (2019), and Connelly et al. (2020), Python libraries were used for document classifications. The researchers show that machine learning classifiers can accurately categorise documents into different types. If implemented, this means that archives can improve their collection efforts without spending more on salaries. One way of coping with the information explosion is to develop metadata generation tools, like machine learning and artificial intelligence. If metadata could be automatically generated, it would reduce the pressure on archival personnel by providing a way to handle larger volumes.

Keywords: Digital humanities, archives, heritage preservation, information age, machine learning, artificial intelligence, language models

https://dx.doi.org/10.4314/jsasa.v56i1.9
ISSN: 1012-2796

1. Introduction

The world has increasingly moved online over the last three decades. In 2016, Mitchell, Shearer, Gottfried and Barthel (2016) found that 38% of Americans received their news online, with only 20% using print media, and only 5% of 18- to 29-year-olds receiving their news from print sources. The latter finding suggests that people will increasingly rely on digital sources in future, as this age group grows older. Newman et al. (2020) also indicate the escalating use of digital media for news consumption and the increasing use of mobile devices. In addition, social media has played an important role in spreading news over the past decade and a half. People no longer rely solely on mainstream media to access information, and platforms such as Facebook, Twitter, and WhatsApp have become more important sources of news.

The information age brought about challenges to archives, particularly through the volume, variety, and velocity of material that is now being produced (Rolan et al. 2018: 180) (so-called big data, see Senekal & Brokensha (2014)). This results in a tsunami of digital material that archives, which were mostly designed to handle hardcopies with much smaller volumes, struggle to cope with. In addition, archives' staff budgets have not kept pace with the information explosion (Park & Brenza 2015: 22). This issue is particularly acute in South Africa, where an expanding social welfare system, coupled with a lack of economic growth, have meant that less funding is available for archives in the public sector.

Archiving born-digital content is crucial, because digital records can easily disappear (Debuysere et al. 2010; Consultative Committee for Space Data Systems (CCSDS) 2012; De Souza et al. 2016). With most information now available in a digital format, the risk that there will be gaps in the historical record in the future will be created. De Souza et al. (2016: 49) write, "'born digital' materials must be proactively collected 'now' so there won't be a 'black hole' in heritage collections in the future".

Artificial intelligence (AI) is a new tool in the archival context and Rolan et al. (2018: 195) claim that it "appears to be an emerging capability and certainly not a production-ready 'silver bullet'". The aim of the current study was to investigate the feasibility of this "emerging capability" in a South African context. The researchers develop and train a machine learning classifier, test it against human classifiers, and show that machine learning can aid the South African archive by handling large numbers of documents without additional investment in staff salaries. While the current experiment is primarily centred on the clippings collection at the National Afrikaans Literary Museum and Research Centre, the methods employed should be equally suitable for other archives.

1.1 Contextual setting to the National Afrikaans Literary Museum and Research Centre

The National Afrikaans Literary Museum and Research Centre was officially opened in 1973 and the brainchild of PJ Nienaber (Nienaber 1976), the National Afrikaans Literary Museum and Research Centre (Nasionale Afrikaanse Letterkundige Museum en Navorsingsentrum (NALN)) in Bloemfontein archives material relating to Afrikaans literature. According to Nienaber (1976: 18), such documentation centre should include everything relating to any literary or linguistic writer in Afrikaans, everything by and about him, his writings, photographs, manuscripts, documents, studies, and textbooks about him, clippings from newspapers, studies in journals, and the like.

NALN currently archives books, manuscripts, magazines, reviews, interviews, academic studies, and theses and dissertations (Lategan 2004: 119). NALN's most used collection is the clippings collection, which consists of material published about Afrikaans authors and their works. This is an extensive collection with over 5 000 clippings collected each year (own calculations from NALN's database).

However, budget cuts meant that the staff responsible for collecting, archiving, and indexing this collection were only one. The current manual way requires this staff member to read through every publication and assign keywords, type of publication and other metadata. This metadata is entered by hand into a database in InMagic/DBText. The volume and use frequency of the clippings collection suggested that this is where the introduction of AI could be most useful to lighten staffs' workload. In addition, most of this material is currently published in digital format, whereas other collections such as manuscripts will have to be digitised first. During a previous project (Senekal 2011), large parts of this existing collection were also digitised, while currently, some material is collected in digital format. NALN shares the premises with the Sesotho Literary Museum (SLM), which provides an opportunity for cooperation between the literary archives of two of South Africa's official languages. During the previous digitisation project (Senekal, 2011), the clippings collection of the SLM was digitised, and the researchers hope that the current study could benefit the SLM in a similar way.

1.2 The need for automated metadata generation

While storage is a core function of archives and the archiving of digital material brings its own set of challenges, adding metadata is crucial in facilitating information retrieval. Thurlow (2020: 80) contends, "Metadata underpins the management and use of collections, as it is needed for discovery and interpretation, providing provenance, context, and structure" (see also Kleppe et al. 2019: 14)). However, adding suitable metadata becomes problematic when large volumes of material have to be indexed, since metadata generation is one of the most time-consuming endeavours of archives it puts a significant strain on human capacities if archives' staff have to read through all material and assign metadata, as is currently done at NALN.

One way of handling large numbers of records is through automatic and semi-automatic metadata generation tools, as argued by Park and Brenza (2015:23):

> Through the use of semi-automatic metadata generation tools, the library community has the potential to address many issues related to the increase of information resources, the strain on library budget, the need to create high-quality, interoperable metadata records, and, ultimately, the effective provision of information resources to users.

For instance, at the Freedom of Information Archive (FOIArchive), archivists were confronted with a corpus of almost 3,8 million documents, the full text of which is available for 2,9 million records (Connelly et al. 2020). Since this archive was designed as a digital archive from the outset, automatic metadata generation was developed to index material. This includes tagging material with entities (i.e. people and countries mentioned in the document) and automatically assigning a topic to the document (Connelly et al. 2020: 4). In doing so, the archivists made use of Latent Dirichlet Allocation (Blei, Ng & Jordan 2003), which is a form of machine learning and artificial intelligence in the Natural language Processing (NLP) field.

Artificial Intelligence (AI) in particular holds great promise in generating metadata automatically. Rolan et al. (2018: 180) argue, “if we are to meet the challenges of managing records in the digital age, such technological aids [AI] – together with the skills and knowledge required to wield them – will be necessary” (see also (Lee 2019:514). Similarly, Goudarouli et al. (2018: 176) contend, “We see a future where AI and emergent technologies become part of our everyday recordkeeping practices”. With the amount of data produced set to increase in the foreseeable future, integrating AI into archival practices will become increasingly necessary if suitable metadata is to be assigned to material. Before proceeding to the methods used in metadata generation, the following section provides a short overview on artificial intelligence.

1.3. Background to artificial intelligence

Although historical accounts of AI are somewhat obscure, most scholars and technology experts agree that as a branch of computer science, it can be traced back to Alan Turing who published a paper in *Mind* in 1950 entitled ‘Computing machinery and intelligence’ in which he conceived of a machine that could communicate (Nadin 2020), considering the question, “Can machines think?” (Turing 1950). AI enjoyed a seminal moment five years later when it was conceptualised by John McCarthy (Dartmouth College) in ‘A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence’. In conjunction with computer scientists Marvin Minsk (Harvard University), Nathaniel Rochester (IBM Corporation), and Claude Shannon (Bell Telephone Laboratories), McCarthy et al. (1955) conjectured “that every aspect of learning or any feature of intelligence can in principle be so precisely described that a machine can be made to simulate it”. Their ambitious aim was to “find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves” (McCarthy et al. 1955).

The modern definition of this nascent technology entails describing AI as a computer or machine that has been designed in such a way that it simulates human behaviour – that is, that models cognitive capabilities with minimal human intervention (Russell & Norvig 2021). In terms of this traditional classification, ‘narrow AI’ refers to a machine that performs simple and predefined tasks, while ‘strong AI’ is hypothesised to be constrained neither by task nor by context. The latter type of AI is also referred to as Artificial General Intelligence (AGI), and has been described as constituting “[the] holy grail of AI research” (Ng & Leung 2020: 63). Well-known examples of narrow AI include Google Assistant, Apple’s Siri, Amazon’s Alexa, and IBM’s Watson. Examples of strong AI are non-existent at this stage, as “a phenomenally aware cognition that possesses and understands its own mental states and subjective experiences” (Flowers 2019:2) has not been developed yet. Machine learning is an example of narrow AI technology (Raschka & Kaufman, 2020) because it exhibits specific rather than general intelligence. The same can be said of deep learning: although it has developed by leaps and bounds, its capabilities remain limited, with both its effectiveness and accuracy depending on a huge amount of training data.

Not to be conflated with AI, machine learning and its more advanced counterpart, deep learning, are classified as subfields of AI, with other subfields including computer vision, robotics, fuzzy logic, neural networks, and natural language processing, to name just a few. Machine learning and deep learning are sometimes used interchangeably, yet are fundamentally different: the term ‘machine learning’, which was coined in 1952 by AI research pioneer Arthur Samuel, may be defined as a discipline that “[focuses] on how computers learn from data” (Balyen & Peto 2019:264). By contrast, deep learning is defined as “a type of machine learning that trains a computer to perform human-like tasks, such as recognizing speech, identifying images or making predictions" (Wong 2021: 1). Significantly,

> The main difference between deep learning and traditional machine learning is its performance as the amount of data increases. Deep learning algorithms do not perform as well when the data volumes are small, because deep learning algorithms require a large amount of data to understand the data perfectly. Conversely, in this case, when the traditional machine-learning algorithm uses the established rules the performance will be better (Xin et al. 2018:35367).

Machine learning has a myriad of applications in fields that require classification, image processing, and diagnosis of disease, for example, while deep learning is widely applied in sectors that require sentiment analysis, facial recognition, and supercomputing, among other things. Both machine learning and deep learning algorithms reflect supervised, unsupervised, and semi-supervised learning. Supervised learning means that the algorithm learns on labelled training samples (as done in the current study), while unsupervised learning means that it is trained on unlabelled data. A semi-supervised approach integrates the two types of learning.

3. Problem statement

Given that archives are overwhelmed in the face of the information explosion and budget cuts (cf. De Vaal-Senekal, De Kock & Putter 2018), technology will have to be implemented for archives to continue functioning as repositories of knowledge. At the moment, staff are overwhelmed when collecting and indexing material, which puts an enormous strain on human capacity and negatively affects the archive's ability to handle queries submitted by the public. In addition, there is often a backlog of unindexed collections, meaning that staff cannot find suitable sources relating to queries, which negatively impacts the quality of research that relies on archival material. Artificial intelligence has emerged over recent years as a possible solution to this dual problem of more documents being published and having fewer staff to index them (Jaillant 2022).

4. Aims and objectives

In view of the above challenges, the current study aimed to explore the use of machine learning for document classification with a view to demonstrating how it can assist archivists who have to manage voluminous amounts of data. To this end, the objectives were to develop and train a machine learning classifier to manage the given South African archive and test it against human classifiers. The literature themes in the following sections cover previous work carried out in the area of developing and training AI for document classification.

5. Literature review

Apart from assigning topics as done at the Freedom of Information Archive (Connelly et al. 2020), another area where AI can be employed for metadata generation is document classification. Rolan et al. (2018:83) call classification, "the task, perhaps, of most interest to recordkeeping knowledge work today". While the topic and the entities mentioned in a document carry important information to allow researchers to find the right material in an archive, the type is equally significant. Johl (2020:551) refers to Galloway and Roux's (2019) bibliography around the Afrikaans author, Breyten Breytenbach, as the ideal model for a bibliography, and Galloway and Roux (2019) classify publications around Breytenbach by type. Types they used included books, theses and dissertations, academic articles, reviews, interviews, and general media reports.

Classifying text documents was successfully performed using machine learning (Brygfjeld, Wetjen & Walsøe 2017; Kleppe et al. 2019; Lee 2019; Suominen 2019; Kotzé, Senekal & Daelemans 2020) and deep learning (Lai et al. 2015; Zhang, Zhao & LeCun 2015; Yin et al. 2017; Minaee et al. 2021), while deep learning was also shown to aid information retrieval for visual material in an archival context (Yasser, Clawson & Bowerman 2017). Static word embeddings (Mikolov, Chen, Corrado & Dean 2013; Le & Mikolov 2014), pre-trained word embeddings (Pennington, Socher & Manning 2014; Bojanowski et al. 2017), and contextualised word embedding models such as BERT (Devlin, Chang, Lee & Toutanova 2018) and GPT-2 (Radford et al. 2019) were used to automatically classify text documents.

6. Methodology

This section presents research methodology for the study.

6.1. Data gathering

The researchers gathered a diverse corpus of digital documents around Afrikaans literature to act as training data, which included book reviews, theses, dissertations, scholarly articles, interviews, and scholarly books. Documents were collected from online news sources (like *Maroela Media* and *Netwerk24*), literary discussion platforms (such as *LitNet* and *Versindaba*), scholarly journals (like *Stilet* and *Tydskrif vir Letterkunde*), and universities' websites. Documents were collected in Portable Document Format (PDF), and where reviews or media reports were not available in PDF, such as reviews from *Maroela Media* or *Netwerk24*, these were stored as PDF during the collection phase. Note that while most of the corpus is in Afrikaans, English publications were included to ensure the cross-language applicability of the techniques employed.

These documents were classified into six categories that were informed by the categories used by Galloway and Roux (2019) and that are used by NALN: articles, media reports, books, interviews, reviews, dissertations, and theses. The classification was done using two annotators, after which the annotated corpus as training data for machine learning models was used. Following Rolan et al. (2018), Suominen (2019), and Connelly et al. (2020), Python libraries for document classifications was used.

6.2 Text extracting

For the documents to be usable for this study, text should be easily extractable for analysis. The PDF documents were converted to text format using *pdfplumber*, which is an open source Python library to convert machine-generated PDFs. The text documents were then stored in a format to ensure all special characters that often occur in Afrikaans (i.e. *ê* and *ë*), are retained for creating text features.

6.3 Data preparation

A basic linguistic approach was followed in terms of text processing to prepare the text documents for text classification. The steps included data cleaning, tokenisation, and conversion of all letters to lowercase. For data cleaning, all tags (i.e. <*i*> and <*html*>), punctuation, multiple whitespaces, and numeric characters were removed using the Gensim open source Python library (Řehůřek & Sojka 2010). Stop words (i.e. the most frequent words such as the definite and indefinite article) were not removed, given that they may have meaningful features in this corpus. The text documents were transformed into a single Pandas data frame (McKinney 2011) which is a user-friendly data structure for text analysis. The data frame was stored in a single Comma Separated Value (CSV) file. The CSV file was divided into four columns, namely the file name (unique identified), classification, clean text, and original text.

6.4 Text vectorisation and feature extraction

Text vectorisation and feature extraction are necessary tasks before any text classification experiment can be conducted (Aas & Eikvil 1999). Text vectorisation is the process of turning text into numerical feature vectors that are usable for machine learning. For this study, the Bag-of-Words (BoW) and the Term-Frequency Inverse Document Frequency (TF-IDF) text vectorisation approaches (Manning, Raghavan & Schütze 2008) were employed. BoW is a simplistic approach that disregards the location of words in a document and instead focuses on word occurrences and creates a frequency count. The drawback of this approach is that certain words will have a very high occurrence (usually articles and pronouns) while carrying very little meaningful information about the actual content of the document (Manning et al. 2008). In TF-IDF, weights are created instead of frequency counts. The rationale behind this is that terms that occur frequently in a document relative to the number of times they occur in the entire corpus, are more important than terms that occur commonly (Manning et al. 2008).

Single words (unigrams) were used in the current study for both BoW as well as TF-IDF. A maximum of 10 000 features were applied for each vectorisation approach when building the vocabulary from the corpus. A cut-off value of 2 (min_df=2) was used to remove terms that appeared too infrequently when building the vocabulary. In other words, terms that appear in fewer than two documents when building the vocabulary were ignored since they were not deemed useful for the given experiments.

6.5 Classification

The current study followed the suggestions by Rolan et al. (2018), Goudarouli et al. (2018), and Suominen (2019) and focuses on developing metadata generation tools for document type classification by using machine learning. For this study, the researchers employed two traditional machine learning models (SVM and MLP). No deep learning models were employed due to the relatively small dataset size (n=621), as the researchers were concerned that the deep learning models would accidentally learn features by over-fitting. The handcrafted linguistic features included surface and syntactic features (Daelemans et al. 2019; Rangel & Rosso 2019) for the traditional machine learning models.

From text classification literature, Support-Vector Machines (SVMs), naïve Bayes, decision trees, logistic regression, and multi-layer perceptron (MLP) are popular machine learning algorithms used to classify text (Mirończuk & Protasiewicz 2018). For the project, SVM and MLP were selected.

- SVMs, also known as Support-Vector Networks, use a statistical learning framework to perform classification tasks. In a multidimensional space, the SVM creates a hyperplane to best divide the distinct classes (in this case, there were six classes). The data points that are closest to this hyperplane are called support vectors.
- Multi-Layer Perceptron (MLP) is a form of machine learning network that performs the text classification task. Unlike SVM, MLP uses a feedforward artificial neural network (ANN) that maps input data sets to a set of appropriate outputs. The MLP consists of several hidden layers and each layer is fully connected to the following one. For the experiment, the default, which is two hidden layers, was used.

Two versions of the corpus with each classifier were used: an original version and a cleaned version. The dataset was split 80/20 for training (*n*=496) and testing (*n*=126) data. The number and classification of each document type are shown in Table 1.

Table 1: Breakdown of the corpus

	Training	**Testing**	**Total**
Articles	80	21	**101**
Reports	94	24	**118**
Books	83	21	**104**
Interviews	80	19	**99**
Reviews	90	18	**108**
Theses and dissertations	69	22	**91**
Total	**496**	**126**	**621**

Scikit-learn (Pedregosa, Varoquaux & Gramfort, 2011) and Python 3.7 were used to build and train the classification models. Both SVM and MLP classifiers were trained with the training set (80% of the corpus) and the training was repeated for both text vectorisation approaches. Once both classifiers were trained, the same algorithm and model (TF-IDF & BoW) were used on the 20% test set to classify unseen documents. The predicted labels were then compared to the original class label to calculate the necessary performance metrics.

7. Findings

The results from the four cases (clean/original corpus, SVM/MLP algorithm) are shown in Table 2, which includes the feature type, feature vector size, accuracy, precision, and recall measures with the F1-Score for each case using testing data, and, finally, the training time.

Table 2: Test results from the four cases (clean/original corpus, SVM/MLP algorithm)

Support Vector Machines				**Multi-Layer Perceptron**			
No data cleaning		**Data cleaning**		**No data cleaning**		**Data cleaning**	
Features:	**TF-IDF**	**Features**	**TF-IDF**	**Features**	**TF-IDF**	**Features**	**TF-IDF**
Size:	10002	Size	10002	Size:	10002	Size	10002
Accuracy:	0.928	Accuracy	0.896	Accuracy:	0.904	Accuracy	0.896
Precision:	0.928	Precision	0.897	Precision:	0.908	Precision	0.896
Recall:	0.931	Recall	0.898	Recall:	0.908	Recall	0.899
F1 score:	0.929*	F1 score	0.896	F1 score:	0.907	F1 score	0.896
Train time:	62.4s	Train time	53.83s	Train time:	235.29s	Train time	237.7s
Features:	**BoW**	**Features**	**BoW**	**Features**	**BoW**	**Features**	**BoW**
Size:	10002	Size	10002	Size:	10002	Size	10002
Accuracy:	0.888	Accuracy	0.904	Accuracy:	0.832	Accuracy	0.824
Precision:	0.890	Precision	0.906	Precision:	0.845	Precision	0.829
Recall:	0.891	Recall	0.904	Recall:	0.838	Recall	0.827
F1 score:	0.889	F1 score	0.904	F1 score:	0.831	F1 score	0.827
Train time:	58.58s	Train time	52.8s	Train time:	141.33s	Train time	151.15s

(*) denotes best test result
TD-IDF = Term Frequency Inverse Document Frequency
BoW = Bag-of-Words

The SVM classifier with no data cleaning scored the highest F1-score (92,9%) in the test portion of the corpus. This means that the SVM classifier was able to classify the type of document with a 92,9% accuracy, making it suitable for classifying documents in this archive. An interesting result was that the SVM classifier with TF-IDF vectorisation performed better on no data cleaning (92,9%) than on cleaned data (89,6%). On the other hand, the SVM classifier with BoW vectorisation performed slightly better on cleaned data (90,4%) than with no data cleaning (88,9%). Similarly, the MLP classifier with TF-IDF vectorisation performed slightly better on no data cleaning (90,7%) than on cleaned data (89,6%). The score was very similar for BoW vectorisation using no data cleaning (83,1%) and cleaned data (82,7%). This would suggest that the MLP classifier was less sensitive to noise in the data than the SVM classifier. These results suggest that the algorithms are capable of assisting with the classification of documents, even if the data are not cleaned. However, the SVM classifier with TF-IDF vectorisation with no data cleaning performed best and was used in future deployments.

8. Conclusion and recommendations

The information explosion and the digital revolution have generated major challenges for archives as they pertain to the collection, storage, and indexing of material. In addition, budget cuts have significantly strained human resources, which effectively means that archival personnel have to accomplish more with less human capital. If the archive is to maintain its function of archiving and indexing material within its collection scope, the only viable way of maintaining this function is by employing automated metadata generation tools. In this field, artificial intelligence holds specific promise, as the current study showed.

The study investigated the use of machine learning for document classification. Techniques explored included SVMs and MLPs, and it was shown that SVMs could be employed for document classification with almost 93% accuracy, even without data cleaning. This means that using machine learning for document classification is a viable alternative to human indexing, at least in the context of the NALN’s clippings collection.

While document classification adds some useful metadata for future document retrieval, future research could experiment with Named Entity Recognition (NER), as successfully deployed by Connelly et al. (2020). In addition, future work will include topic modelling, similar to what Connelly et al. (2020) employ. Investigating deep learning for document classification is also critical. Lastly, employing the techniques discussed in the current study and integrating it into a workflow will also require further research, as this will entail overcoming technical challenges.

Declarations

The authors declare that:

- This manuscript has not been previously published and is not under review with any other publication or copyrighted publishing platform.
- Consistent with the journal’s ethical requirements and the legal requirements of South Africa, we obtained ethical clearance from the University of the Free State’s General/Human Research Ethics Committee (UFS-HSD2021/0101/162).
- Human participants were not used; the study did not involve the analysis of secondary data; and the texts used are publicly available.
- We have avoided unlawful statements that breach existing copyrights.

- The tables are our own, so it was not necessary to obtain copyright permission from third parties.
- We declare no potential conflict of interest for the research.
- We are familiar with the manuscript's content and have all contributed to drafting it.
- We give consent to the Journal of the South African Society of Archivists to publish the manuscript.

Acknowledgements

The authors would like to thank the University of the Free State for an interdisciplinary research grant awarded by the Faculty of the Humanities that made this research study possible.

References

Aas, K. & Eikvil, L. 1999. *Text categorisation: a survey*. Oslo: Norwegian Computing Center.

Balyen, L. & Peto, T. 2019. Promising artificial intelligence-machine learning-deep learning algorithms in ophthalmology. *Asia-Pacific Journal of Ophthalmology* 8(3): 264-272.

Blei, D.M., Ng, A.Y. & Jordan, M.I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3: 993-1022.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5: 135-146.

Brygfjeld, S.A., Wetjen, F. & Walsøe, A.U. 2017. Machine learning for production of Dewey Decimal. Paper read at the World Library and Information Congress, 84th IFLA General Conference and Assembly, 24-30 August 2018, Kuala, Lumpur. Available at: http://library.ifla.org/2216/1/115-brygfjeld-en.pdf (accessed: 1 May 2022).

Connelly, M., Hicks, R., Jervis, R., Spirling, A. & Suong C.H. 2020. Diplomatic documents data for international relations: the Freedom of Information Archive Database. *Conflict Management and Peace Science* 38(6): 762-781.

Consultative Committee for Space Data Systems (CCSDS). 2012. *Reference model for an open archival information system (OAIS)*. Washington: Space Operations Mission Directorate.

Daelemans, W., Kestemont, M., Manjavacas, E., Potthast, M., Rangel, F., Rosso, P., Specht, G., Stamatatos, E., Stein, B., Tschuggnall, M., Wiegmann, M. & Zangerle, E. 2019. Overview of PAN 2019: bots and gender profiling, celebrity profiling, cross-domain authorship attribution and style change detection. In Crestani, F., Braschler, M., Savoy, J., Rauber, A., Müller, H., Losada, D.E., Bürki, G.H., Cappellato, L. & Ferro, N. (Eds.). *Experimental IR Meets Multilinguality, Multimodality, and Interaction:* 10th International Conference of the CLEF Association. Cham: Springer International Publishing (Lecture notes in computer science), 402–416. doi:10.1007/978-3-030-28577-7_30

Debuysere, S., Moreels, D., Van de Walle, R., Van Nieuwerburgh, I., & Walterus, J. 2010. *Bewaring en ontsluiting van multimediale data in Vlaanderen: perspectieven op audiovisueel erfgoed in het digitale tijdperk*. Leuven: Lannoo Campus.

De Vaal-Senekal, P., De Kock, C. & Putter, M., 2018. Challenges in the archives of the Afrikaans Language Museum, Paarl, Western Cape, South Africa: a case study. *Atlanti* 28(1): 195-205.

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. 2018. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

De Souza, P., Edmonds, F., McQuire, S., Evans, M. & Chenhall, R. 2016. *Aboriginal knowledge,* *digital technologies and cultural collections. policy, protocols, practice*. Melbourne Networked Society Institute (Research Paper 4). Melbourne: Melbourne Networked Society Institute.

Flowers, J.C. 2019. Strong and weak AI: Deweyan considerations. In *AAAI Spring Symposium: Towards Conscious AI Systems*. Stanford, CA. Available at: http://ceurws.org/Vol-2287/paper34.pdf (accessed: 24 April 2022).

Galloway, F. & Roux, A. 2019. Uitgesoekte bibliografie. In Galloway, F. (ed.) *Breyten Breytenbach: woordenar, woordnar – 'n huldiging*. Pretoria: Protea Boekhuis, 344-408.

Goudarouli, E., Sexton, A. & Sheridan, J. 2018. The challenge of the digital and the future archive: through the lens of the national archives UK. *Philosophy & Technology* 32(1): 1-11.

Jaillant, L. 2022. Introduction. In Jaillant, L. (Ed.). *Archives, access and artificial intelligence: working with born-digital and digitized archival collections*. Verlag, Bielefeld: Bielefeld University Press, 7-28.

Johl, R. 2020. Skrywersbibliografieë, grootdatanetwerke en die posisie van skrywers soos NP van Wyk Louw in die literêre kanon. *Tydskrif vir Geesteswetenskappe* 60(2): 534-555.

Kleppe, M., Veldhoen, S., Waal-Gentenaar, M.V.D., Oudsten, B.D. & Haagsma, D. 2019. Exploration possibilities automated generation of metadata. *Zenodo*. doi:10.5281/zenodo.3375192

Kotzé, E., Senekal, B.A. & Daelemans, W. 2020. Automatic classification of social media reports on violent incidents in South Africa using machine learning. *South African Journal of Science* 116(3/4): 1-8.

Lai, S., Xu, L., Liu, K. & Zhao, J. 2015. Recurrent convolutional neural networks for text classification. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Twenty-Ninth AAAI Conference on Artificial Intelligence. Austin, Texas: AAAI Press, 2267-2273.

Lategan, L. 2004. Die Letterkundemuseum as Navorsingsbron in die Edisiewetenskap. *Stilet* 26(2): 119-124.

Lee, B.C.G. 2019. Machine learning, template matching, and the International Tracing Service digital archive: automating the retrieval of death certificate reference cards from 40 million document scans. *Digital Scholarship in the Humanities* 34(3): 513-535.

Le, Q. & Mikolov, T. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, 1188-1196.

Manning, C. D., Raghavan, P. & Schütze, H. 2008. *An introduction to information retrieval*. Cambridge: Cambridge University Press, 117–119; 268–269.

McCarthy, J., Minsky, M.L., Rochester, N. & Shannon, C.E. 1955. *A proposal for the Dartmouth Summer 542 Research Project on artificial intelligence*. Available at: http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html (accessed: 12 October 2021).

McKinney, W. 2011. Pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing* 14(9): 1-9.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. 2013. Efficient estimation of word representations in vector space. arXiv preprint:1301.3781.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M. & Gao, J. 2021. Deep learning-based text classification. *ACM Computing Surveys* 54(3): 1-40.

Mirończuk, M.M. & Protasiewicz, J. 2018. A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications* 106: 36-54.

Mitchell, A., Shearer, E., Gottfried, J. & Barthel, M. 2016. *The modern news consumer*. Pew Research Centre.

Nadin, J. 2020. *A life story: Alan Turing*. Scholastic UK.

Newman, N., Fletcher, R., Schulz, A., Andı, S. & Nielsen, R.K. 2020. *Reuters Institute Digital News Report 2020*. Reuters Institute.

Ng, G.W. & Leung, W.C. 2020. Strong artificial intelligence and consciousness. *Journal of Artificial Intelligence and Consciousness* 7(01): 63-72.

Nienaber, P.J. 1976. *Die Nasionale Afrikaanse Letterkundige Museum en Navorsingsentrum, Musiek en Toneel. Sy ontstaan, stigting en vestiging*. Bloemfontein: NALN.

Park, J. & Brenza, A. 2015. Evaluation of semi-automatic metadata generation tools: a survey of the current state of the art. *Information Technology and Libraries* 34(3): 22-42.

Pedregosa, F., Varoquaux, G. & Gramfort, A. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12: 2825-2830.

Pennington, J., Socher, R. & Manning, C. 2014. Glove: global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 1532-1543.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1(8).

Rangel, F. & Rosso, P. 2019. Overview of the 7th author profiling task at PAN 2019: bots and gender profiling in Twitter. In Proceedings of the CEUR Workshop. Lugano, Switzerland, 1-36.

Raschka, S. & Kaufman, B. 2020. Machine learning and AI-based approaches for bioactive ligand discovery and GPCR-ligand recognition. *Methods* 180: 89-110.

Řehůřek, R. & Sojka, P. 2010. Software framework for topic modelling with large corpora. In *LREC workshop on new challenges for NLP frameworks*. Valletta, Malta: ELRA, 45-50.

Rolan, G., Humphries, G., Jeffrey, L., Samaras, E., Antsoupova, T. & Stuart, K. 2018. More human than human? Artificial intelligence in the archive. *Archives and Manuscripts* 47(2): 1-25.

Russell, S. & Norvig, P. 2021. *Artificial intelligence: a modern approach*. Global edition. *Foundations* 19: 23.

Senekal, B.A. 2011. Die digitalisering van NALN se knipselversameling: die bemiddeling van 21ste-eeuse navorsing in die Afrikaanse letterkunde. *LitNet Akademies Geesteswetenskappe* 8(2): 46-65.

Senekal, B.A. & Brokensha, S. 2014. *Surfers van die tsunami: navorsing en inligtingstegnologie binne die geesteswetenskappe*. Bloemfontein: Sun Press.

Suominen, O. 2019. Annif: DIY automated subject indexing using multiple algorithms. *LIBER Quarterly* 29(1): 1.

Thurlow, E. 2020. Preserving an emerging digital arts landscape: digital preservation at University of the Arts London. *Art Libraries Journal* 45(2): 78-82.

Turing, A.M. 1950. Computing machinery and intelligence. *Mind* 59(236): 433-460.

Wong, Y.K. 2021. Advanced deep learning approach and applications. *International Journal of Information Technology (IJIT)* 7(5).

Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., Gao, M., Hou, H. & Wang, C. 2018. Machine learning and deep learning methods for cybersecurity. *IEEE Access: Practical Innovations, Open Solutions* 6: 35365-35381.

Yasser, A.M., Clawson, K. & Bowerman, C. 2017. Saving cultural heritage with digital make-believe: machine learning and digital techniques to the rescue. In *Electronic visualisation and the arts (EVA 2017), BCS learning & development (Electronic workshops in* *computing*). doi:10.14236/ewic/HCI2017.97

Yin, W., Kann, K., Yu, M., & Schütze, H. 2017. Comparative study of CNN and RNN for natural language processing. *arXiv preprint*:1702.01923.

Zhang, X., Zhao, J. & LeCun, Y. 2015. Character-level convolutional networks for text classification. *Advances in neural Information Processing Systems* 28: 649-657.