INTERNATIONAL REVIEW OF LAW, COMPUTERS & TECHNOLOGY
https://doi.org/10.1080/13600869.2025.2557640

The fair dealing/fair use landscape for artificial intelligence innovation and computational research in Africa

Chijioke I. Okorie

Department of Private Law, Faculty of Law, University of Pretoria, Pretoria, South Africa

ABSTRACT

This paper sets out the issues of copyright ownership and risk of copyright infringement liability raised by data science research use of data held by public bodies (in particular, public service broadcasters) in South Africa and Nigeria. Considering both the fair dealing exception in South Africa's Copyright Act of 1978 and Nigeria's Copyright Act, 2022 as well the proposed fair use provision in South Africa's Copyright Amendment Bill B13F-2017, the paper discusses these issues elaborating on the reasons why data science researchers in public research institutions should not and do not require a copyright licence or be considered to be infringing copyright when they use copyright-protected materials held by public bodies for data science and artificial intelligence or machine learning research. The paper also suggests that where the outputs of data science research are copyright-protected, they should be made available in an open and accessible manner with reasonable safeguards.

ARTICLE HISTORY
Received 5 March 2025
Accepted 2 September 2025

KEYWORDS
African natural language processing; artificial intelligence; fair dealing

Introduction

This paper explores the copyright issues raised by data science researchers' access to, and use of copyright-protected data held by public bodies, especially public service broadcasters (PSBs). It considers, from a South African and Nigerian perspective, the question of whether data science researchers in a public research institution as defined in South Africa's Higher Education Act and Intellectual Property Rights from Publicly-financed Research and Development Act, using copyright-protected broadcast news content to train and/or develop natural language processing (NLP) models infringe on copyright in those materials. Given the possibility that technological tools and materials including annotated datasets may be created as a result of licensed or unlicensed use of copyright-protected materials for data science research, this paper also explores whether copyright (or other property rights) subsist in such materials as to enable the creators exercise proprietary rights thereon. The paper also takes a normative approach to examine whether and how such proprietary rights should be exercised.

CONTACT Chijioke I. Okorie chijioke.okorie@up.ac.za Department of Private Law, Faculty of Law, University of Pretoria, Cnr Lynwood & Roper Street, Hatfield, Pretoria 0028, South Africa

Data scientists and researchers seeking to train and/or develop NLP models to learn language tasks (e.g. part-of-speech tagging, named entity recognition, translations, etc.), require access to a significant amount of data in the relevant language(s) in order to do so. While accessing publicly available language data is significantly easier for high resource languages such as English, French, Chinese, and medium resource languages such as Greek, Dutch, Urdu (Joshi et al. 2020; Kruit 2023), it is quite difficult to find sufficient publicly accessible data and/or datasets, especially annotated datasets in languages spoken across Africa (generally classified as low resource languages) (Braun and Ong 2018; Marivate 2021). This is the case for various reasons including colonial legacies of African countries, inequality of language use in business and public settings, the geographical location and language identity of the developers of AI systems and NLP models, etc (Birhane 2020; Sanneh 2015). South Africa and Nigeria and indeed other African countries are no exceptions (Hu et al. 2025; Babirye et al. 2022; Agic and Vulic 2019). In such circumstances, it becomes imperative to find ways to improve the availability of, and access to datasets, and take steps to increase innovation around collection, curation, annotation and classification of datasets in those languages when and where found (Marivate 2021). In this regard, public bodies (including PSBs) and public funding (Marivate 2021; Ncube, Abrahams, and Akinsanmi 2014) can play a significant role. As part of the daily functions of public bodies, they create and collect a significant amount of data covering many types of information including language information (Lee 2016; Marcowitz-Bitton 2015). The nature and mandate of PSBs across Africa make them a significant resource for datasets in African languages (Marivate 2021). Laws establishing PSBs and the practices established by PSBs themselves require that they broadcast local content including content in the languages spoken in the relevant country. This is evident in section 2(r) of South Africa's Broadcasting Act 1999 and section 6(2) of the Nigerian Television Authority Act 2004.

The use of public sector data (particularly those from PSBs) in NLP research implicates various legal frameworks such as copyright, privacy and data protection, competition law, contract law, etc. Chief among these is copyright since language data are represented in text, speech and audio-visual format, which by their nature, may be subject of copyright protection. Copyright law grants a bundle of exclusive rights to authors of protectable subject matter such as literary, musical and artistic works, sound recordings, cinematograph films, computer programs, broadcasts, etc. These exclusive rights attaching to these works could differ depending on the work in question but generally include the rights of reproduction, adaptation, broadcasting or rebroadcasting, transmission in a diffusion service, publishing, etc. The exclusive nature of these rights means that anyone wishing to engage in activities covered by such rights with respect to a given work must obtain permission (i.e. a licence) from the relevant copyright owner of such work to avoid potential liability for infringement. From a copyright perspective, the processes involved in training and developing NLP models implicate the selection and use of copyright works (news articles, books, movies, television and radio broadcasts, etc.) in circumstances involving the exclusive rights of reproduction, adaptation, etc (de Castilho et al. 2018). It follows, therefore, that data science researchers seeking to train and develop NLP models including in African languages may require a licence from the relevant copyright owner(s) of the materials represented in the training data. Copyright owners (even PSBs) could, on the basis of such copyright, refuse access to the data resources or require

payment of licensing fees for NLP research activities unless copyright exceptions apply (P1 2023). Alongside these issues of the scope of, and tensions between copyright protection and copyright exceptions are concerns about reusability, as well as endorsement when it comes to use and application of such data as secondary data. These issues are explored in this paper.

To provide a factual context for the discussion of these issues, the analysis in this paper is framed around the experiences shared by a number of data science researchers from South Africa, Kenya and Senegal during an academic conference held at the University of Pretoria, South Africa in 2023 (P1 2023). There are other NLP projects across Africa that have grappled with the issue of using copyright-protected materials as training data. These include the JW 300 datasets created by Agic and Vulic from copyright-materials owned by the Jehovah's Witnesses (Agic and Vulic 2019). The JW 300 corpus covers 343 languages and the Masakhane Foundation (a grassroots NLP community for Africa, by Africans) used this corpus to develop 45 translation models translated from English into 32 unique African languages spoken in Kenya, Uganda, Ghana, etc. (Nekoto et al. 2020). Due to the refusal of the Jehovah's Witnesses organisation to grant a licence to ensure lawful use of the datasets, the Masakhane Research Foundation and other African NLP researchers have had to stop using the JW300 datasets. (Moody 2023). For ease of reference, the analysis in this paper zeroes in on a research proposal entitled 'Improving News Categorization, Translation, Named Entity Recognition, and Part-of-Speech Tagging with NLP Techniques' (Annex 1) prepared by the Data Science for Social Impact (DSFSI) research group of data scientists at the University of Pretoria, South Africa. This involved an attempted access by data science researchers in South Africa to use language data from PSBs for research purposes. This attempt was met with a licence as a condition for access and use, raising questions around the scope of copyright ownership and the application of copyright exceptions, which could obviate licensing requirements.

The issues identified above extend beyond the research proposed by the DSFSI (Marivate 2021). Many research institutions and scholars in South Africa and across the African continent have expressed frustration with the process and difficulties posed by copyright protection mechanisms in accessing data held by public bodies for purposes of NLP research, AI development and/or machine learning processes (Hlomani and Ncube 2023; Marivate 2021). Although the example considered is South African and is shaped by the South African legal context, the issues discussed are of broader relevance across Africa given the number of NLP projects undertaken across Africa in Kenya, Ghana, Uganda, Nigeria, etc. The copyright law in many African countries is similar in requiring the same eligibility criteria and providing for limitations and exceptions. Recognising the prevalence of the copyright challenges but also taking cognisance of the fact that it is not possible to cover the entirety of African countries, this paper also provides a Nigerian perspective and makes recommendations for guiding principles around best practices. In essence, this paper focuses on South African and Nigerian copyright law but the examples of NLP projects across Africa as highlighted above show that the issues discussed are continent wide.

The first part of this paper sets out the issues raised by data science research use of PSBs' data using the DSFSI Proposal as an illustration. The second part discusses these issues from the perspectives of South Africa's Copyright Act of 1978 as amended and

Nigeria's Copyright Act of 2022, drawing out their respective implications (particularly of the fair dealing provision) for data science research as contemplated by the DSFSI. In the third part, the paper acknowledges that South Africa's Parliament has passed an amended copyright statute – the Copyright Amendment Bill B13F-2017, which is currently before its Constitutional Court for a decision on its constitutionality. The South African Copyright Amendment Bill proposes a departure from fair dealing exception to fair use exception. In this regard, the third part explains the boundaries of what data science researchers may do with copyright-protected data within the fair dealing copyright exception in both South Africa and Nigeria (and the fair use exception, should it become law in South Africa) and also points out areas of uncertainties in the application of the law on copyright exception to data science research. The fourth part concludes.

The interface between data science research and copyright law in Africa

Under the copyright statutes in both South Africa section 23(1), and Nigeria (section 36), a prospective user of copyright-protected content must, where the use implicates any of the exclusive rights (for example, reproduction, adaptation, broadcasting or rebroadcasting, transmission in a diffusion service, publishing, etc.) granted by copyright law, procure the consent of the relevant copyright owner or risk copyright infringement liability. In essence, the legal authority of a copyright owner to grant or decline a licence for the use of its protected material as training data, comes from the copyright statute (sections 6 to 10 of the South African Copyright Act and sections 9 to 13 of the Nigerian Copyright Act), which grants it the exclusive right to reproduce, adapt, broadcast, etc. its literary, musical and artistic works, cinematograph films, sound recordings, broadcasts, etc. (Oira and Ndlovu 2018). However, for data represented by copyright-protected materials, both general and specific limitations and exceptions have been established (for example, in the EU's Articles 3 and 4 of the Directive on Copyright in the Digital Single Market 2019) that could exempt the text and data mining necessary for NLP research from copyright infringement liability (Craig 2024). These include provisions excluding certain kinds of ordinarily protectable subject-matter from copyright protection either because copyright has expired in the work or because the law explicitly or implicitly does not extend copyright protection to such works. In other instances, copyright exceptions such as fair dealing may apply to exclude ordinarily infringing activities from infringement liability and obviate the need to obtain a licence from the relevant copyright owners (Greenleaf and Bond 2013). Evident from experiences outside the African continent (Erickson et al. 2015), copyright limitations and exceptions including limitations as to scope of protectable subject-matter and exclusive rights can offer an enhanced access to data in African languages (Ncube and Rutenberg 2020; Okorie 2023). Exceptions offer the public access to copyright-protected works, which could in turn engender innovation and development. Put differently, although other legal considerations (attribution, source/endorsement, etc.) may still necessitate a licence or at least conditions/terms of use (Lee 2016), copyright limitations and exceptions have the potential to help data science researchers overcome the problem of data paucity (discussed above) and enable them to easily conduct NLP research that will benefit the country and the broader African society. However, the implementation and use of copyright exceptions require an understanding of their scope (Okorie 2023a). Further, the applicability and/

or application of copyright exceptions to data science research oftentimes depends on the entity undertaking the research, the nature of the research and the activities involved in the research. These are briefly described in this part.

The DSFSI proposal and the experiences of other African data science researchers

In early 2023, a number of data science researchers from South Africa, Kenya and Senegal were part of a panel convened at an academic conference held at the University of Pretoria, South Africa (P1 2023). These researchers shared their experiences on various research projects involving the collection and use of text, speech and audio-visual data ('language data') about African languages. One project involved the collection of language data on various African languages from an internationally known religious organisation – Jehovah's Witness (Agic and Vulic 2019). Another project (which is used illustratively in this paper) sought to collect a range of South African language data from South Africa's sole public service broadcaster, the South African Broadcasting Corporation (SABC 2016). The other project from Senegal sought to access data from academic publications in Africa (P1 2023).

These instances involved the collection of what is described here as 'copyright data' – materials which are used as data in the NLP context and which are or could be subject of copyright protection. DSFSI had, for several years before the January, 2023 conference, been seeking authorisation and consent from the SABC to use data in text, audio and video formats including transcripts therefrom from the SABC News website (collectively, 'SABC News Content') to create annotated datasets and to develop African NLP models to classify news content; translate news content from one language to another; perform named entities recognition (NER) on news content, and perform parts of speech (POS) tagging on news content to identify parts of speech (P1 2023). At the conference, a former board member of the SABC, who was a serving board member at the time the leader of the DSFSI research group approached the SABC for authorisation was asked his opinion as to why the SABC did not seem inclined to provide the requested authorisation, and he stated that (P1 2023):

> The SABC is one of the most commercially dependent public broadcasters on the planet. It is currently 80% dependent on commercial revenue. It only gets 3% of its revenue from the state … and in a situation like that with a massive public mandate and very little state support; what could be happening is pressure on the people sitting in that institution when they are approached by someone with genuine intentions like Vukosi start thinking 'well, hang on a second, am I passing up the opportunity of potential revenue for SABC down the line? Am I gonna get into trouble if I give a blank cheque?'. That's the first thing. And one can understand in a revenue constrained environment that they would be doing their job to see whether there is potential revenue for the SABC down the line. The second issue is on the legal side and I'm not a copyright lawyer and won't express an opinion but they were relying on the Copyright Act, on a section which escapes me now …

Following the January 2023 conference, the DSFSI prepared and shared with the SABC, a research proposal dated 5 April 2023 outlining the activities, methodology and expected outcomes of the NLP research it intended to undertake using the SABC News Content as training data (Annex 1). According to the DSFSI, the SABC expressed willingness to grant a non-exclusive, non-transferable, non-sub-licensable, royalty free licence to the research

group to use the news content for the research as stipulated in the Proposal on the condition that it (SABC) would retain ownership of all its intellectual property rights to the News Content, any translated content pursuant to the tools created/developed from the NLP research and any annotated datasets from the research, which are deemed capable of commercialisation. This illustration focuses on the SABC but it applies also to Nigeria's public service broadcasters – the Nigerian Television Authority (NTA) and the Federal Radio Corporation of Nigeria (FRCN), which have similar kind of repertoire.

In view of the foregoing assertion, unauthorised use of the copyright-protected materials as training data *could* amount to copyright infringement.

Copyright subsistence and ownership

As indicated in the copyright statutes in both South Africa and Nigeria and confirmed by case law, to determine whether infringement has taken place, two conditions must be satisfied: first, it is necessary to establish that in relation to a protectable subject matter, the alleged infringer's behaviour or the behaviour of a person acting through them, falls within the scope of any applicable exclusive right; second, such behaviour must be without the consent of the relevant copyright owner. Infringement, therefore, presupposes unlicensed or unauthorised copyright use i.e. use that falls within the scope of the exclusive rights attaching to the work in question. This is evident from South African cases such as *Haupt t/a Soft Copy v Brewers Marketing Intelligence (Pty) Ltd and Others* (1989*)*; *Jacana Education (Pty) Ltd v Frandsen Publishers (Pty) Ltd* (1998*)*; *Galago Publishers (Pty) Ltd and another v Erasmus* (1989*)* and Nigerian case law including *Married Media Ltd v Lawrence Akpa* (1990-1997*)*; *Plateau Publishing Ltd v Adophy* (1986*)*.

The SABC News Content to be deployed as corpora in the training and development of these NLP models are potentially copyright protected. These are materials in text, audio and video formats, which as broadcast contents could qualify as literary works, sound recording and cinematograph film respectively as seen from section 1(1) of the South African Copyright Act (1978) and s108 of the Nigerian Copyright Act (2022). Broadcast copyright in South Africa as it is defined and as the rights are structured in section 10 of the Copyright Act is essentially an ensemble of other works (Oira and Ndlovu 2018). The copyright statutes require that for a work to be protected, it must be original and fixed in a material form and must be authored by a 'qualified person' (i.e. by citizens, residents and juristic persons). While originality is not defined in the copyright statute, relevant case law in both jurisdictions such as *Haupt* (1989*)*; *Yemitan v Daily Times & Anor.* (1977-1989*)*; *Yeni Anikulapo-Kuti & Ors. V Iseli & Ors* (2003-2007*)*; etc. shows that it is to be assessed on a case-by-case basis and also that the leading standard is a 'sweat of the brow' standard requiring some independent thought and intellectual effort by the author (Geyer 2022). As materials commissioned and/or produced by the SABC and existing on the SABC News website, the SABC News Content are fixed in material form and it is highly likely that they satisfy the requirement of originality and also, authorship by a qualified person (Oira and Ndlovu 2018). The same conclusion is drawn in respect of content from Nigeria's PSBs in view of statutory provisions on originality (section 1(2)(a) of the Nigerian Copyright Act), qualified persons as author (section 5 of the Nigerian Copyright Act) and case law such as *Married Media Ltd v Lawrence Akpa*

(1990-1997); *Plateau Publishing Ltd v Adophy* (1986) where the eligibility criteria for copyright protections were set out in similar terms as South African case law.

As explained earlier, the process of training and developing NLP models may result in the creation and annotation of datasets. In terms of the Proposal as embodied in Annex 1 of this paper, the expected outcomes of the data science research are the development of various NLP models for categorisation, translations, NER tags, POS tags; and the release of the derivative data/models under a permissive licence for other researchers to be able to use. These envisaged outcomes raise the question of IP (especially copyright) subsistence (i.e. are the datasets to be considered protectable subject matter?), authorship and ownership of NLP models and labelled datasets. Neither the algorithms nor datasets are patentable inventions as they are excluded by s25(2) of the South African Patent Act. They also do not meet the criteria for trade mark protection or other recognised IPRs. This is also the case in Nigeria.

On the issue of whether the labelled datasets constitute protectable subject matter under copyright law, it must first be noted that while licensing is usually the go-to mechanism for permitting access to and use of labelled datasets, the datasets may not actually be copyright-protectable materials and hence appropriate objects for copyright licensing even if other legal and contractual frameworks can form the bases for a licence (Society of Composers, Authors and Music Publishers of Canada (SOCAN) v Bell 2012; Burrow 2021). However, where they are eligible for and attract copyright protection because they meet the eligibility criteria of originality and fixation, it raises not only the issue of copyright ownership of such materials but the rationale for limitations and exceptions. Even if the fair dealing exception did not apply, both section 2(3) of the South African Copyright Act and section 2(4) of the Nigerian Copyright Act are clear that the fact that the making of a work infringes on an existing work is not by itself a relevant consideration in the determination of the eligibility of that new work for copyright protection.

The labelled datasets would be in text form and, therefore, could be a collection of literary works (specifically, a database) within the meaning of the copyright statutes in both South Africa and Nigeria. Section 1 of the South African Copyright Act defines 'literary work' to include 'tables and compilations, including tables and compilations of data stored or embodied in a computer or a medium used in conjunction with a computer, but shall not include a computer program'. Section 108 of the Nigerian Copyright Act (2022) defines literary works along these lines but also includes computer programs. As already discussed earlier, the standard for assessing originality which is necessary for copyright subsistence, is said to be the 'sweat of the brow' test. This also applies to databases (Moleya 2020). This standard is admittedly low especially for databases given their informational nature and the conflict between the interests of database developers and those of the public in accessing information contained in databases (Moleya 2020). But, while there is merit in the argument that the sweat of the brow test is inappropriate for databases and that the protection of databases requires a high creativity-based standard instead, this paper focuses on merely acknowledging the possibility of and the basis for copyright subsistence in the labelled/annotated datasets as databases. This is without prejudice to other bases for a licence to be required for lawful use.

Following from the foregoing, the annotated datasets, if original, may qualify as a work to be protected under copyright law. Where that is the case, the author of such literary work (i.e. the dataset) would be the data science researcher who curated, created and/

or labelled the datasets as envisaged by the definition of 'author' in section 1(1) of the South African Copyright Act and also section 108 of the Nigerian Copyright Act. By virtue of section 21(1)(a) of the South African Copyright Act, the author is the first owner of copyright unless where the exceptions apply. This is similar to Nigeria's section 28(1) of the Copyright Act 2022. These exceptions per section 21(1)(b)-(d) of the South African Copyright Act (1978) and section 28(2) and (3) of the Nigerian Copyright Act (2022) relate to works created for publication in a newspaper, magazine or periodical; works created under a commission; works created in the course of one's employment. As such owners, the exclusive rights to reproduce, adapt, broadcast the work, etc. belong to them. DSFSI indicates in the Proposal, its intention to share the annotated datasets publicly under a permissible licence for other researchers to use. The intention to licence presupposes that there are some possible proprietary rights held over such datasets. To the extent that the annotated datasets are protected under copyright law, the licensing of the datasets is within the purview of their rights as copyright owners.

Where the datasets lack originality and cannot be protected by copyright, it may be difficult for the DSFSI (and other data scientists in similar situations) to maintain copyright control over their datasets. However, they may still use licences to relinquish control over their datasets.

Copyright protection and NLP research

Machine learning has been useful across many sectors such as fraud detection in the financial sector; health diagnosis in the field of medicine, understanding text for spam detection, answering questions, grouping documents and sentiment analysis, etc (Marivate 2021). Building and/or developing trained language models to perform these tasks requires a significant amount of data to be used as an input into the machine learning algorithm. For NLP, the training data is usually labelled or unlabelled text. The dominant approach to training and developing NLP models involves the identification and selection of the training data (i.e. media in written, audio and video formats). The data then undergoes pre-processing i.e. conversion into a format that can be read by machines or by the NLP tools. The pre-processed data will then be annotated/labelled. Annotation or labelling involves a human or an NLP tool reading the files and assigning appropriate labels to various segments of the data based on pre-defined instructions – statistical data, grammatical rules etc. Thereafter, the NLP model is 'trained'. This involves using a software programme (i.e. the training tool) that applies a machine learning algorithm (a test dataset) to the annotated data to make evaluations and/or analyse the annotated data to extract appropriate characteristics (Abebe et al. 2021; de Castilho et al. 2018; Marivate et al. 2020). As such, access to data for NLP research can be by identification, selection and repurposing of existing data or materials (so-called, text and data mining) as is the case with the SABC News Content or by crowdsourcing data donations as is the case with Mozilla Common Voice project or by paying for data contributions usually from funding to do so such as with the African Next Voices project, a one year pilot project funded by Bill and Melinda Gates Foundation aiming at data collection through the KenCorpus consortium of Universities. The purpose of the African Next Voices project is to support the establishment of a corpus of text and voice data for several local languages spoken in Kenya (Maseno University 2024). This paper focuses

on copyright issues arising from NLP research undertaken with existing copyright-protected data or materials.

It is possible that the activities involved in accessing existing data, annotating data and generally training and developing NLP models could implicate the exclusive rights of reproduction and adaptation. Section 1(1) of the South African Copyright Act and section 108 of the Nigerian Copyright Act does not provide an exhaustive definition of reproduction. In *Adenuga v Ilesanmi*, the Court of Appeal in Nigeria accepted that reproduction of a copyright-protected material involved making copies of such material. In *Media24 Books v Oxford University Press*, to 'reproduce' within the meaning of the South African Copyright Act was held to mean 'to copy'. In *Blind SA v Minister of Trade, Industry and Competition and Others*, the Constitutional Court of South Africa was called upon to inter alia determine whether the technologies and the activities involved in making accessible format copies of copyright-protected works for persons with visual disabilities amounted to reproduction and/or adaptation. The court noted that while 'a content-based distinction between reproduction and adaptation will not always be definitive', adaptation involves some 'interpretative engagement with the text so as to render its meaning'. Further, the court noted that with making copyright-protected materials accessible to wider audiences using technology, it may sometimes be a question of copying the work into another format and no more, and at other times, it may involve more than mere reproduction and require some translation, transformation that requires interpretation. The court concluded that a comprehensive and appropriate copyright exception would be one that speaks to both the right of reproduction and the right of adaptation. This paper agrees with the reasoning applied by the Constitutional Court of South Africa and submits that in the present case, the process of text and data mining, processing and annotating could go beyond reproduction to also involve some interpretative exercise as to include adaptation of the copyright data. However, even where the unauthorised copyright use of a protected work satisfies the infringement criteria indicated above, there would be no infringement liability when copyright limitations and exceptions apply.

In the light of the foregoing, a key interpretive issue is whether the research proposed by the DSFSI and similar research undertaken by other groups or researchers using existing copyright data from the SABC or any PSB or similar public body/source fell within the scope of copyright exceptions obviating the need for a licence or whether the circumstances of the proposed or actual use were such as to warrant a copyright licence. It is imperative that data science researchers engaged in developing and training NLP models have certainty as to the copyright status of language data used for such activities. The answer to the questions whether a data science researcher in a public research institution like the one represented by the DSFSI may be held liable for copyright infringement because they, for research purposes, used copyright-protected broadcast news content to train NLP models will inform the decisions and practices of data science researchers around the lawful reuse of existing materials in data science and AI development. Also, seeing as the copyright in the broadcast news content in the instant scenario is held by a public service broadcaster, the answer presented in this paper will further guide what approach public bodies should adopt in relation to data and/or copyright-protected materials generated as part of the discharge of their duties as such public bodies also considering the overarching constitutional rights of access to information and cultural participation.

It is against the context presented above, that data scientists and others making decisions and policies on data use and governance in South Africa and across Africa including those relating to the DSFSI's research as indicated in Annex 1, will undertake their decision-making.

Fair dealing and NLP research

Sections 12 to 19B of the South African Copyright Act and Sections 20 to 27 of the Nigerian Copyright Act provides for general and special exceptions from copyright protection where otherwise infringing activities would not be considered infringing.

In South Africa, section 12(1) of its Copyright Act refers to an exhaustive list of activities ('dealing'), which are to be considered within the parameters of fairness. Fair dealing for the purposes of research or private study, personal or private use, criticism or review, reporting current events in the case of literary or musical or artistic works section 12(1) and 15(4); broadcasts (section 18); published editions (section 19A) and for the purposes of criticism or review, and/or reporting current events in the case of cinematograph films (section 16(1), sound recordings (section 17) and computer programs (section 19B), would not be infringing. What is required there is that the activities alleged to be infringing are undertaken for any of the purposes referred to in that provision (i.e. private or personal use, research, criticism or review, reporting current events) and undertaken in a manner considered fair. While that provision does not indicate the meaning to be ascribed to 'fair dealing', case law offers guidance.

Part II of Nigeria's Copyright Act (sections 20 to 27) provides for general and special exceptions from copyright protection where otherwise infringing activities would not be considered infringing. Unlike South Africa which provides an exhaustive list of activities ('dealing'), which are to be considered as fair dealing purposes, Section 20(1) of the Nigerian Copyright Act offers an inexhaustive list of fair dealing purposes through its use of the expression 'such as' even while explicitly providing an indicative list of 18 fair dealing purposes. The use of the expression 'such as' within that provision connotes that it is possible for fair dealing purposes not explicitly listed to still be considered as a fair dealing purpose (Okorie 2023b). Within the explicitly listed fair dealing purpose is 'non-commercial research' (Section 20(1)(c) of the Nigerian Copyright Act 2022). This is a departure from the provisions of the now repealed Copyright Act of 2004 which permitted 'research' generally. Like its South African counterpart, the activities alleged to be infringing must be undertaken for a fair dealing purpose and undertaken in a manner considered fair. Unlike its South African counterpart, section 20(1) of Nigeria's Copyright Act indicates the meaning to be ascribed to 'fair dealing'. As will be seen in following paragraphs, this is similar to the case law guidance available in South Africa and related jurisdictions.

Exempted fair dealing activities

To reiterate, there are four exempted fair dealing activities in South Africa: research, private study or personal or private use, criticism or review and reporting current events. Nigeria has 18 listed exempted fair dealing activities with room for other activities not explicitly listed. Common to both jurisdictions are research, private study or personal

or private use, criticism or review and reporting current events, though Nigeria only exempts non-commercial research. Of these four common exempted fair dealing activities, only the 'criticism or review' and 'reporting current events' activities have come up for judicial interpretation in South Africa and Nigeria.

In *Moneyweb v Media24 and another,* a South African high court had the opportunity to consider the scope of the fair dealing exception for the purpose of reporting current events and its applicability in addressing issues of copyright infringement liability. Moneyweb and Media24 (via its Fin24 platform) are both online financial news publishers and business competitors. Moneyweb claimed that Media24 infringed its copyright by reproducing seven of its articles, or alternatively, had engaged in unlawful competition. While Media24 admitted reproducing a substantial parts of those articles, it argued that some of those reproduced articles were not original and even for the original works, its allegedly infringing reproduction was fair dealing under section 12(1)(c)(i) of the South African Copyright Act (i.e. reporting of current events). Media24 also argued that the articles were not copyright protected as they were news of the day under section 12(8)(a) of the South African Copyright Act which excludes news of the day from copyright protection.

Starting with the relevant fair dealing purpose (i.e. reporting of current events), the court considered that even though a current event need not be one that occurred on the day of the report, such event must be relatively close in time to the report. The court also indicated that the phrase 'reporting of current event' should be given its ordinary, wide meaning. It has also been held that the fair dealing purpose of reporting current event requires an 'element of currenthood' to be in the 'predominant or material of the work' (SABC v Vollenhoven and Apollis Independent CC 2016).

In *SABC v Via Vollenhoven,* SABC commissioned a documentary film titled *Project Spear* from Via Vollenhoven and Appollis Independent CC, an independent production company, around 2011. The production was governed by a Total Programme Commissioning (TPC) agreement which expressly granted the SABC full ownership of the copyright and editorial control over the final work. After Via Vollenhoven completed the film in 2012, the SABC withheld its broadcast, citing concerns about its content, which it believed could breach editorial policy or be defamatory. Although attempts were made to resolve the dispute, the film was ultimately never aired. In 2013, Via Vollenhoven arranged for limited private screenings, including at the Franschhoek Literary Festival and via the publication *Noseweek*, without SABC's consent. In response, the SABC sought an interdict to prevent further distribution or screening of the film, arguing that such use violated its contractual and statutory copyright. Via Vollenhoven argued inter alia, that it was entitled to retain a copy of the film and to exhibit it under fair dealing provisions of section 12 of the Copyright Act, especially for purposes such as criticism, review, or reporting current events. The court held that this was 'patently a sham'. In arriving at this conclusion, the court considered the evidence that the respondent had, in an interview with some radio stations, insisted that it intended to get the story contained in the work out to the public. While the court declined to formulate a specific meaning for fair dealing, the court further noted that the purpose of dealing must encompass a 'genuine purpose and not a pretext for a purpose which is not contemplated under fair dealing'.

In all, when it comes to the interpretation of the specific purpose of the dealing, the courts in South Africa appear consistent in holding that the context of the work in

question, and the facts existing at the time of dealing are important considerations. Because this relates to the specific purpose of the dealing, it is submitted that these considerations, i.e. the context of the work in question and the facts existing at the time of dealing will also be relevant when assessing the research and/or personal or private use purposes.

In the case of Nigeria, available case law on fair dealing as stated earlier has focused on the reporting of current events purpose under the now repealed Copyright Act of 2004. In *Peter Obe v Grapevine Communication Ltd,* where the plaintiff's photographs of the Nigerian Civil War and which were published in his book were used by the defendant in the inaugural edition of its newsmagazine without the plaintiff's authorisation, a Federal High Court in Lagos rejected the defence of fair dealing. The defendant's defence that its use of the photographs, even though unauthorised, constituted fair dealing for the purpose of reporting of current events failed because according to the court, the defendant did not acknowledge the plaintiff as the author of the photograph and the book and failed to acknowledge the title of the Plaintiff's work. Section 5 of the repealed Nigerian Copyright Act of 2004 which applied to the case and which is similar to section 20(1)(d) of the extant Copyright Act of 2022 recognises inter alia the defense of fair dealing for the reporting of current events, provided that the title of the work and its authorship are properly acknowledged. In essence, the title of the works and its authorship is a strict requirement to have a chance at successfully relying on fair dealing for the purpose of reporting current events.

The interpretation to be given to 'research' (of any kind) has not been considered judicially in South Africa and Nigeria. However, the meaning to be ascribed to 'research' has been discussed in decisions of courts outside both countries (CCH v Law Society 2004; SOCAN v Bell 2012; Oriakhogba 2023) as well as in scholarly literature (Appadurai 2006; Okorie 2023b; Von Kries and Winter 2015). These foreign decisions and scholarly authorities are not binding on Nigerian and South African courts. However, based on a plethora of cases, foreign decisions are generally of strong persuasive authority in both countries.

In *Ladoja v INEC*, Nigeria's Supreme Court held that the general principle of law is that decisions of courts outside Nigeria are of persuasive authority in Nigeria. There are however instances where a foreign decision is considered binding in Nigeria. These include instances where the principle adopted from a foreign decision has been repeatedly applied by Nigerian courts over the years. This exception was laid down by the Supreme Court of Nigeria in *Adetoun Oladeji (Nig) Ltd v Nigerian Breweries Plc*. The second exception is that Nigerian courts may resort to foreign decisions where there are no known Nigerian decisions on a principle of law. Thus, English authorities can be binding where the facts before a Nigerian Court are similar to the English case, and there are no known Nigerian decisions on the same set of facts. Thus, in *Omega Bank Plc. v. Govt. Ekiti State*, the Court of Appeal in Nigeria relied on the foreign precedent of *Derbyshire C.C. v. Times Newspapers* in reaching its decision. Based on the general rule and this second exception, it is submitted that Nigerian courts will likely rely on foreign decisions either persuasively or as binding, if the facts are similar.

For South Africa, foreign decisions are also persuasive and not binding authority and South African courts are not bound by the dicta of foreign tribunals. In *Sanderson v Attorney-General, Eastern Cape*, the court warned that 'the use of foreign precedent requires

circumspection and acknowledgment that transplants require careful management'. Generally, courts in South Africa have recourse to foreign decisions to identify the correct problem, or to identify it properly, using foreign decisions to see how foreign courts have solved the problem, what methodology has been used in resolving the problem, what the competing considerations have been, and whether any potential dangers were identified in the process (Ackermann 2006; Lawrence, S v Negal, S v Solberg 1997; MEC for Education KwaZulu-Natal v Pillay 2008; Rautenbach 2015).

In *CCH v Law Society,* the Canadian Supreme Court held that 'research' must be given a large and liberal interpretation in order to ensure that 'users' rights are not unduly constrained'. In summary, research in that case was construed to include lawyers carrying on the business of law for profit and accessing materials for the purpose of advising clients, giving opinions, arguing cases, preparing briefs and factums. The approach of giving a large and liberal interpretation to fair dealing purpose here is similar to the stipulation by the South African court in *Moneyweb* regarding the news reporting purpose.

In *SOCAN v Bell*, the Canadian Supreme Court construed research has been construed as an activity not limited to non-commercial or private contexts, while also holding that the nature of the research (i.e. whether for commercial or non-commercial purpose) is a relevant factor in weighing the fairness of the dealing. In that case, research was also construed to include library staff making copies of the requested cases, statutes, excerpts from legal texts and legal commentary. According to the court, research includes 'many activities that do not demand the establishment of new facts or conclusions' and 'can be piecemeal, informal, exploratory, or confirmatory' or be undertaken 'for no purpose except personal interest'. Consumers using music previews for the purpose of identifying which music to purchase were also considered to be conducting research.

Research also includes 'those processes of study, experiment, conceptualization, theory-testing and validation involved in the generation of new knowledge' (United Nations Educational, Scientific and Cultural Organization, Resolution 15).

Applying the guidance above to the circumstances presented by the DSFSI's Proposal and data science research generally, it is submitted that the activities contemplated in the Proposal and involving the use of the SABC News Content qualifies as 'research' and is a purpose permitted as fair dealing under section 12 of the South African Copyright Act and section 20 of the Nigerian Copyright Act. More particularly in the case of Nigeria, these activities would be non-commercial research as they would be undertaken by researchers in a public research university and not for a commercial purpose.

The 'fairness' of the dealing

Having established that the NLP research as contemplated above would be a permissible fair dealing purpose under South African and Nigerian copyright law, it becomes necessary to apply the second ambit of the fair dealing exception: whether the dealing (even for the permitted purposes) is fair.

Section 20(1)(d)(i) to (iv) of the Nigerian Copyright Act offers guidance as to some relevant criteria with which to assess fairness. It provides that in determining whether the use of a work in any particular case is fair dealing, the factors to be considered shall include the – (i) purpose and character of its usage, (ii) nature of the work, (iii) amount and substantiality of the portion used in relation to the work as a whole, and (iv) effect

of the use upon the potential market or value of the work. The interpretation and application of these factors have not been explored in case law so far. But, foreign case law from jurisdictions such as South Africa, the US and the UK on the application of these factors will be of strong persuasive authority in Nigeria. This is because section 107 of the US Copyright Act provides a similar set of factors for consideration in determining fairness of the dealing. In the case of South Africa and the UK, while their copyright statutes do not explicitly set out factors for consideration in determining fairness of the dealing, guidance is offered in case law with factors similar to those outlined in the Nigerian and US copyright statutes, albeit focused on fair dealing for the purpose of reporting current events and to some extent, for criticism or review. With regard to this, the courts in South Africa have, while cautioning against wholesale adoption of foreign jurisprudence on this, indicated the following factors as relevant: the nature of the medium in which the work have been published; whether the original work has already been published; the time lapse between the publication of the two works; the extent of the acknowledgement given to the original work; the nature and purpose of the use; the nature of the copyright work; the amount and sustainability of the use; the effect on the market and the value of the work; etc (Moneyweb (Pty) Limited v. Media 24 Limited and Another 2016; SABC v Via Vollenhoven 2016). These criteria are similar to those indicated in s20(1) of the Nigerian Copyright Act (Okorie 2023b).

Paraphrasing the application of these factors in *Moneyweb*, even where the dealing was for a purpose indicated in section 12(1) of the South African Copyright Act, such dealing will not be fair dealing where publication of the alleged infringing article was made within 1 d of publication of the original article; where the alleged infringer contributed little or nothing to what it had copied from the original article; or where what was copied was likely to be a substitute for the original article even if the original article was acknowledged as the source. In *SABC v Via Vollenhoven*, the court did not proceed with the interpretation and/or application of these factors because of its finding that the respondent's dealing did not fall within any of the purposes recognised under section 12 of the Act. This is also the case in Nigeria as the copyright statute in section 20 requires a consideration of both fair dealing purpose and fairness.

For Nigeria, there is explicit statutory provision on factors to be taken into consideration in assessing whether a dealing is fair. Those factors apply whether the dealing is for research or otherwise. These factors are similar to those that have been considered by courts in South Africa, Canada, the US and the UK. In the absence of case law from Nigeria, decisions from those courts applying the fairness factors to *research* will be of persuasive authority in Nigeria. This paper therefore relies on relevant foreign decisions in considering how the statutory listed fairness factors may be applied to NLP research in Nigeria.

For South Africa, reliance is also placed on relevant foreign decisions in considering what fairness factors may be applied to NLP research and how such factors should be applied. As stated earlier, these foreign decisions are of persuasive and guiding authority in South Africa. Fair dealing case law in South Africa suggests that once the purpose of dealing is within the purposes listed in the statute, what is needed to determine whether that purpose as expressed in the dealing is fair is an inexhaustive list of factors that do not necessarily apply cumulatively. The factors that are relevant and applicable therefore depends on the nature of the work and the context of the use (Shay 2014).

In essence, there is selectiveness in the factors that are considered relevant to a given context. As held by the court in *Moneyweb*, these factors are not exhaustive and one factor may be more important than the other (Moneyweb (Pty) Limited v. Media 24 Limited and Another 2016). It is noted that the proposed fair use provision in South Africa's Copyright Amendment Bill B13F-2017 offers a similar statutory listed fairness factors as Nigeria and the US and also allows room for the consideration of additional factors. This approach accords with *Moneyweb* (2016) in that the factors are not exhaustive.

In Canada, a dealing with musical works for research purposes was found to be fair dealing and permissible because: the real purpose of using the work was for research; there were 'reasonable safeguards' in place to ensure that the work is actually used for that purpose (SOCAN v Bell 2012; CCH v Law Society 2004); the character of the dealing was such that only single, temporary copies were distributed (SOCAN v Bell 2012; CCH v Law Society 2004); the amount or quantity of the work taken as part of the dealing is small when compared against the entire work taken from (SOCAN v Bell 2012; CCH v Law Society 2004); the use/dealing was the 'most practical, most economical and safest' way to achieve the 'ultimate purpose' of the dealing (SOCAN v Bell 2012; CCH v Law Society 2004), the nature of the work was such that it should be widely disseminated (not conflating or mistaking availability with dissemination) (SOCAN v Bell 2012), and the effect of the dealing on the original work was not adverse or competing (SOCAN v Bell 2012). In considering the purpose of the use, the relevant perspective is that of the person using the work and regard must be had to the 'real purpose or motive' behind using the work (SOCAN v Bell 2012).

What is evident from the above analysis is that the following factors are relevant criteria in considering/assessing fairness of the dealing for NLP research in both Nigeria and South Africa: *purpose and character of its usage; nature of the work; amount and substantiality of the portion used in relation to the work as a whole; and effect of the use upon the potential market or value of the work.* Even though these factors have not been judicially applied on fair dealing for the purpose of research in both countries, it is submitted that they are relevant factors (in varying degrees of importance) to consider in ascertaining whether a dealing for purposes of NLP research is fair. In the case of Nigeria particularly, section 20(1)(d) of the Copyright Act which presents the relevant factors refer to all fair dealing purposes and does not distinguish between various dealings or purposes. In the case of South Africa, while case law suggests that the factors relevant to a consideration of fairness may depend on the fair dealing purpose, case law aligns with the Nigerian approach that one factor may be more or less important than another, given the context of usage/dealing and, the list of factors is not exhaustive.

In both jurisdictions, however, the weight to be attached to any factor taken into consideration in determining the fairness of the dealing or use will differ depending on the context. In some cases, the effect of the use upon the potential market or value of the work and the purpose and character of the usage of the work may weigh more heavily than the nature of the work and/or amount and substantiality of the portion used in relation to the work as a whole. This was the position taken by a US court in a recent decision involving fair use in the use of copyright-protected material in AI innovation – *Thomson Reuters Enterprise Centre GmbH v. Ross Intelligence Inc.*, – where the court while reiterating that it 'must consider at least four fair-use factors' as listed in the copyright statute also noted that 'the first and fourth factors weigh most heavily in the

analysis'. These are the use's purpose and character, including whether it is commercial or nonprofit and how the use affected the copyrighted work's value or potential market. As noted in that decision, this was also the approach (i.e. different weighting to different factors) in *Authors Guild v. Google, Inc*.

Fair dealing purposes + fairness

In this section, this paper describes the meaning of each of the four fairness factors generally and in relation to NLP research. It also proposes the weight that should be attached to each factor based on relevant foreign case law and scholarly analysis. It is important to note that no single fairness factor is determinative. Instead, it is a question of conducting a holistic balancing test, weighing all four factors together to arrive at a conclusion.

Training NLP models involves a distinct process that fundamentally shapes fair dealing considerations. This process entails analysing massive amounts of textual data, where NLP researchers and AI developers routinely make digital copies of complete works to extract patterns, language structures, and factual information. This is a computational activity, distinct from traditional human reading or direct consumption of the copyright-protected material. A critical distinction in fair dealing analysis for NLP lies between the copyright-protected materials used as inputs for training and the outputs subsequently generated by the NLP or AI model (Rosati 2024). Courts might generally be more receptive to fair dealing arguments concerning the training process itself because the *process* of learning, rather than the *result* of learning, constitutes the 'use' of or the dealing with the copyright-protected material. This redefines what constitutes 'copying' and 'infringement' in the digital age, giving rise to concepts such as a 'fair learning' doctrine (Wei et al. 2025). This shift in focus from the final product to the underlying computational process is significant to a consideration of the fairness of the dealing for NLP research purposes. However, the fair dealing status of the outputs generated by the NLP model is a separate inquiry outside the scope of this paper. The following paragraphs offers an analysis of the different factors and explores why certain factors may deserve greater weight in the specific context of NLP research.

Purpose and character of the use

The purpose and character of the use are relevant considerations as they focus on why and how the work is used. When fair dealing is for the purpose of research, the central issue is whether the inquiry-based nature/purpose of research justifies the use of the original work bearing in mind the intended effect of research. In the instant case, the nature of the use of the SABC News Content is to be considered fair, being a research project as clearly set out in the Proposal and also to promote the use of South African languages which are currently low-resource languages. As discussed above, case law in South Africa and elsewhere shows that the real purpose of the dealing is a relevant factor and the assessment of this factor is to be made from the perspective of the user (SOCAN v Bell 2012; CCH v Law Society 2004). In the present case, the purpose of mining and scraping the works from the SABC News website is truly and primarily for research in the setting of a public research institution and to enable other NLP researchers working on South African languages to be able to do so. Furthermore, the distinction between commercial and non-commercial use also significantly impacts this factor.

Non-commercial, educational, and personal uses are generally favoured in fair dealing analysis especially when they avoid supplanting the market for original works (SOCAN v Bell 2012; CCH v Law Society 2004).

Regarding the character of the dealing, it is submitted that the fact that the NLP research process involves reproducing the SABC News Content in machine-readable format and then annotating them for machine learning tasks meant that copies made of the original works could not be used for the traditional purpose of music, other audio, text and/or video files (Margoni and Kretschmer 2022). In *SOCAN v Bell* (2012), it was accepted that streaming as opposed to downloads meant non-duplication or further dissemination by users. For NLP training, the argument could also be made that the character of the use is transformative because the system 'learns' functional patterns and extracts information to generate original output, which serves a fundamentally different purpose than the original expressive intent of the training data. In *Bartz and others v Anthropic*, the US District Court held that the AI system's ability to distil information from thousands of works to produce new text was 'quintessentially transformative' because the training on copyright-protected works by Anthropic was not replicate or supplant the works but to create something different.

An emerging area of consideration is the role of 'guardrails' and the intended output of the NLP model. In its non-binding report on whether the unauthorised use of copyright-protected materials to train generative AI systems is defensible as a fair use, the United States Copyright Office (the USCO) indicated that the implementation of 'guardrails' by AI model developers or deployers to prevent or minimise the creation of infringing outputs weighs in favour of a fair use argument. Such guardrails might include blocking certain prompts, employing training protocols designed to make infringing outputs less likely, or using internal system prompts to instruct a model not to generate copyrighted characters or styles. This suggests that the design and deployment of the AI system significantly influence the assessment of its transformative purpose. The implementation of these technical controls directly influences the 'purpose and character of the use' factor. This indicates a shift in legal analysis from merely scrutinising the intent of the use to also examining the technical design choices and operational controls embedded within the AI system. This intertwining of legal compliance with engineering practices creates a new dimension of responsibility for AI developers, where their technical decisions can directly support or undermine a fair dealing defence. With respect to the NLP research contemplated in this paper, in addition to the research being non-commercial research, the Proposal indicates that there are reasonable safeguards in the form of a permissible licence to ensure that the resulting datasets will only be used for research purposes (Annex 1).

In sum, the first fairness factor favours fair dealing for the SABC News Content being used for NLP research.

The nature of the work

This factor evaluates the characteristics of the original work itself and its publication status. Generally, factual works (e.g. news reports, scientific articles, non-fiction) and unpublished works may be more amenable to fair dealing. According to the Canadian Supreme Court in *CCH v Law Society of Canada*, the dealing may be more likely to be considered 'fair' if the work is unpublished since 'its reproduction with acknowledgement could lead to a wider public dissemination of the work'. The defendant in the case

easily met this factor as the works in question were essential to legal research and were subject to its Access Policy stating that the patron's purpose to access the works must be for research, private study, criticism, review or use in legal proceedings.

Applying this approach, it is submitted that the nature of the works as broadcast content in African languages is one which should be widely disseminated for the benefit of citizens in African countries. It is indisputable that the dissemination of works in national African languages is desirable and beneficial and while these works are easily available on the SABC News website or the website of African PSBs, it does not mean that such works are published, accessible and widely disseminated especially in the technology and computational environment. Unless these works are processed in machine readable format and for machine learning, the work will not be published or disseminated for NLP research purposes. Also, given that the works in current form are not easily used in the machine learning environment without processing, NLP activities are of immense benefit to promoting further dissemination.

The second factor points in favour of fair dealing for the NLP research.

Amount and substantiality of the portion used in relation to the work as a whole

This factor assesses the quantity of the original work used and whether the portion taken was significant or constituted the expressive, original elements/portions of the work. For NLP and AI training, developers often make digital copies of complete works because machine learning processes often require ingestion of entire works in order to be effective. The consideration of the amount and substantiality of use requires nuancing if it is to be relevant. Because of the nature of NLP research which requires access to the entire work before pre-processing and processing can determine which parts make it into the training data and the NLP model, a consideration of the amount and substantiality of the use in determining fairness of the dealing would be irrelevant and miss the point entirely if the quantity used is the focus without due qualification. As a first step, NLP research would always involve or, more accurately, require access to as opposed to actual or active use of a substantial, if not the entire portion of the work. Taking 'amount and substantiality' of use into consideration without discountenancing factual, unoriginal elements of the given works, would defeat the purpose of the research fair dealing (Margoni and Kretschmer 2022). Some scholars such as (Shay 2014) have argued that this would still remain a relevant consideration for any work but the weight to be attached to it would be quite low. Even if this factor were considered relevant in the case of NLP research, it is submitted that this use should be considered fair. In terms of quality, the dealing for the purposes of NLP research does not take the expressive elements of the works in their traditional context. In terms of quantity, this is also the case when the amount taken from *each* work is not significant when viewed against the entirety of the work (SOCAN v Bell 2012; CCH v Law Society 2004). As the court in *Moneyweb* noted, one must look at the dealing vis-à-vis each work (SOCAN v Bell 2012; CCH v Law Society 2004).

The third factor thus favours fair dealing for the NLP training copies to be made.

Effect of the use upon the potential market or value of the work

The effect of the dealing on the potential market for, or value of, the original requires a consideration of whether the unauthorised use causes economic harm to the market

for the original work or its potential market, including the market for permissions or licences. Such economic harm may be from market substitution where as noted by the UK court in *Ashdown v Telegraph Group Ltd*, the alleged fair dealing is in fact commercially competing with the proprietor's exploitation of the copyright work, a substitute for the probable purchase of authorised copies, and the like. On this factor with respect to the Canadian case of *CCH v Law Society of Canada*, D'Agostino observed that 'there was no evidence advanced to indicate that there was an effect on the publishers' market. Rather, the publishers continued to produce new reporter series and other legal publications during the period of the Great Library's request-based copying' (D'Agostino 2008). The analysis in case law from the US often distinguishes between 'market substitution' and 'market dilution'. For NLP research, market substitution occurs when the output from the NLP model directly replaces the original work. The *Thomson Reuters*' case, where the AI legal search tool was found to directly compete with Thomson Reuters' copyright-protected headnotes, is a good example. 'Market dilution' occurs when AI-generated outputs, even if not identical to specific works, compete in the same market by generating new works in the same style, genre, or category (US Copyright Office 2025). This can increase competition and potentially reduce the market value of the originals. A significant concern here relates to systems that can rapidly produce content that could compete with human-authored works. The availability of licensing and the 'permissions market' are also relevant. Courts consider whether affordable and readily available licensing options exist for the intended use (Authors Guild, Inc. v. Google Inc. 2015). For instance, Getty Images offers licences for AI training – Stability AI did not pay for this licence but rather used the images for training without permission. Given that the 'availability of permissions or licenses' is recognised as one of the potential values for copyright-protected works, the argument from the copyright owners of materials intended to be used as training data would be that NLP and AI data use can harm the potential market to license works for NLP or AI training. In fact, the statement from the SABC's former director quoted in this paper, alludes to the fact that the SABC sees possibility of generating revenue from licensing the SABC News Content for NLP training purposes. But, it is argued here that requiring a licence for non-commercial research and in essence, holding that such non-commercial NLP research imposes market harm and is not fair dealing, can be detrimental to academic research. In the case of the NLP research as proposed in Annex 1 and for data science research in public research institutions generally, it is submitted that because of the conversion of the News Content into annotated datasets for NLP, it can hardly be said that annotated datasets are in competition with the use and enjoyment of the individual broadcast content itself. And since the effect/outcome of the NLP research is to increase access to and dissemination of African languages – an outcome within the purview of the mandate of the SABC, NTA and many other national PSBs, it cannot be said the datasets have a negative impact on the works (SOCAN v Bell 2012).

Also, since the research outcomes (per the Proposal) is to automate the production of news and 'help individuals and organisations find news articles that match their interests using a search engine and produce more high-quality content in less time', it cannot be said that the research has a negative impact on the SABC News Content. In the instant case, the works are being used in a specific manner: factual and informational elements are taken, not to cut out the copyright owner's primary market for the work but to enable

uses in a setting where the copyright owner does not operate. Moreover, if the SABC or another national PSB were to operate in that setting, it cannot charge the public a fee for such services because its statutory mandate requires it to make those contents available and accessible to the public. Even though a licensing market could exist for the SABC or another national PSB in terms of the SABC News Content, it is not a licensing market that a public service broadcaster such as the SABC or another national PSB should exploit to the detriment of research, transformation and decolonisation.

The fourth factor thus favours fair dealing for the NLP training copies to be made.

Based on the foregoing analysis, the fair dealing factors could be ranked for NLP research using copyright-protected materials, from most to least relevant and weighty (Wei et al. 2025). It is submitted that the purpose and character of the use factor may be a paramount and most relevant factor given its emphasis on transformative use. The core argument for fair dealing in NLP is that the system learns patterns and generates new outputs, serving a different purpose than the original copyright-protected works and sometimes, even enhancing access to and better appreciation of the original copyright-protected works. The specific design and deployment of the NLP model, including guard-rails to prevent infringing outputs, directly influence this assessment. The purpose and character of the use factor may be followed closely by the effect on the potential market for the work in weight and relevance because the latter focus on the risk of market dilution or direct substitution for the original work. Because of the ability of NLP models to undertake tasks or generate content at speed and scale, they could pose a risk of market dilution or direct substitution which could undermine a fair dealing defence. The nature of the work factor is moderately relevant especially where NLP involves using factual and published works, which generally favour fair dealing. The amount and substantiality of the portion used factor may be less determinative for using copyright-protected works as NLP training inputs. This is because as explained above, NLP models frequently require ingestion of entire works for effective learning. While wholesale copying traditionally weighs against fair dealing, this negative weight is often mitigated if the 'purpose and character of the use' is strongly transformative.

In essence, while all four statutory factors are considered, the 'purpose and character of the use', particularly its transformative aspect, and the 'effect on the potential market for or value of the work' emerge as the most critical determinants for NLP training. A robust transformative purpose for the NLP learning process can significantly mitigate negative findings related to the amount of material used. However, any significant market harm can independently undermine an otherwise favourable fair dealing defence.

Other relevant factors

As stated earlier, under current South African copyright law, courts are allowed to apply any relevant factor in assessing the fairness of a dealing. Nigerian copyright law on the other hand mandates the consideration of the four factors discussed above while allowing that additional relevant factors may be considered. Accordingly, it is noted that factors such as alternatives to the dealing was considered relevant to the consideration of fair dealing for research purposes by the Canadian Supreme Court in *CCH v Law Society of Canada* and could equally be relevant in considering NLP research. The questions arising with respect to this factor include whether there was a non-copyright protected work available as an alternative and whether the use of the copyright-protected work

was not reasonably necessary to achieve the 'ultimate purpose. In this regard, the court in *CCH v Law Society of Canada* took the view that because twenty per cent of the defendants' patrons were outside Toronto and researchers were not allowed to borrow materials from the defendant's library there were no alternatives to the photocopying service and the need for copying was justified. In applying this factor to the NLP context of this paper, it is submitted that for NLP research in South African local languages, there are no sufficient and suitable non-copyright equivalent of the work that could have been used. Further, the processing of the content is necessary and crucial to achieve the ultimate purpose. Scraping these data, annotating them and using them to develop and train NLP models is the only practical way that such data researchers conduct research in NLP.

The extent of the acknowledgement given to the original work is a factor that has been considered relevant and even mandatory for fair dealing for purposes of reporting current events both in Nigeria and in South African case law (Moneyweb (Pty) Limited v. Media 24 Limited and Another 2016; Peter Obe v. Grapevine Communication Ltd 1997). For this factor, what is important is that the author and source of the used works are acknowledged as much as is practicable. This factor could also be relevant in the NLP environment to ensure transparency and accountability regarding the source of the training data. It is submitted that the acknowledgement proposed to be given to the SABC News Content will be sufficient and fair in both South Africa and Nigeria as the storage and management of the annotated data sets will indicate the SABC News website as the source of the original data sets and a link to the SABC News website will be provided.

The nature of the medium in which the works have been published is another factor that has been considered relevant for fair dealing for purposes of reporting current events in South African case law (Moneyweb (Pty) Limited v. Media 24 Limited and Another 2016). The nature of the medium in which the works have been published in the case of the SABC News Content is the internet, specifically on the website of the SABC. Given the open nature of the internet and the statutory role of the SABC and in the case of Nigeria, NTA and or FRCN, there should be no unfairness in accessing and using the works for purposes of NLP research. Given the NLP research context as indicated in the Proposal and also given that the relevant parties [researchers in a public research institution and the PSB] are public institutions whose mission is essentially to serve in the public interests (Akingbulu 2010; Bronstein and Katzew 2018), the factors considering whether the original work has already been published and the time lapse between the publication of the two works are not as relevant. Even if they were, the SABC News Content have been published.

In the light of the foregoing exploration of the four statutory factors and any other factors deemed relevant and weighing the results, it is submitted that, a data science researcher in a public research institution in South Africa or Nigeria such as the members of the DSFSI research group, using copyright-protected broadcast news content to train NLP models would be dealing fairly with such works and would not be infringing on copyright. The activities contemplated in the Proposal and involving the use of the SABC News Content qualifies as 'research' and involve dealing fairly with the News Content. Such use should be considered non-infringing and not require a copyright licence.

African NLP research and fair dealing/fair use

The focus of the preceding parts of this paper has been on NLP research conducted by data science researchers in the setting of a public research institution (i.e. non-commercial or not-for-profit research). Further, the interpretation has been on the basis of the current copyright statute in South Africa and Nigeria. These two considerations leave various questions unaddressed such as whether data scientists conducting research in environments outside public research institutions can claim the fair dealing exception and whether or not, in the case of South Africa, the interpretation proposed in the preceding parts of this paper would change (for better or worse) under the fair use exception proposed in its Copyright Amendment Bill B13F-2017. Clause 15 of the Bill proposes the deletion of the current fair dealing exception and in its place, the insertion of a fair use exception covering an open-ended list of permitted purposes and including an equally open-ended list of factors to be taken into consideration in determining whether a given use is fair. This is similar to what now obtains in Nigeria even though the Nigerian Copyright Act still uses the term, 'fair dealing'.

With respect to the proposed fair use exception, the fundamental differences when compared with the current fair dealing provision are that it provides a non-exhaustive list of fair use activities along with an equally non-exhaustive list of factors to be taken into consideration in determining whether or not a use is fair (Okorie 2023). The effect of these is that there would now be statutory basis in South Africa for considering not just research but a plethora of activities as fair use activities. Further, there would also be statutory basis for taking all relevant factors into consideration in determining fairness. In essence, the proposed fair use provision, by broadening the exempted fair use purposes and providing for a non-exhaustive, non-cumulative list of factors to considered in determining fairness, both retains and strengthens the interpretation proffered above for the use of copyright data in data science research in South Africa. However, the liberty to choose what factors to consider in deciding whether a dealing or use is fair will now be restricted to adding other relevant factors. The four factors expressly and explicitly listed in the Bill – (i)*the nature of the work in question; (ii)the amount and substantiality of the part of the work affected by the act in relation to the whole of the work; (iii)the purpose and character of the use, including whether – such use serves a purpose different from that of the work affected; and it is of a commercial nature or for non-profit research, library or educational purposes; and (iv)the substitution effect of the act upon the potential market for the work in question* – must be considered. As already stated above, these factors are quite similar to those listed in section 20(1)(d)(i)-(iv) of Nigeria's Copyright Act.

Further, even though the proposed fair use exception retains research (of any kind) as an explicitly listed fair use purpose (unlike Nigeria that only permits non-commercial research), the fact that it is enjoined to consider inter alia whether the purpose and character of a use is of a commercial nature or for non-profit research, library or educational purposes is likely to weigh more in favour of non-commercial data science research as described in this paper than for commercial data science research. Nigeria is explicitly for non-commercial research as a fair dealing purpose such that commercial research will likely not make it to the round of consideration of fairness (Oriakhogba and Olubiyi 2023).

Conclusion

Research in African NLP especially as represented by the Proposal of the DSFSI research group in Annex 1 requires data science researchers (and any person assessing data science research use of copyright works) to determine the scope of the fair dealing exceptions and how they apply to the use of copyright-protected works in the context of emerging technologies. The particular context of South Africa, Nigeria and African NLP also present an opportunity to reflect on how public institutions including PSBs should discharge their statutory mandates in the copyright environment.

Insofar as the fair dealing exception is concerned, case law in South Africa and Nigeria, while limited, aligns with foreign decisions particularly those on fair dealing in Canada and the UK where the specific dealing must first be one of the permitted purposes and such dealing must be fair. In this regard, it is required to consider the perspective of the user and the context of the use/dealing. The considerations for determining the fairness of the dealing are not fixed and vary depending on the nature of the work. Nigeria's copyright statute and South Africa's upcoming Copyright Amendment Bill also upholds this position. Of course, South Africa currently has leeway to pick and choose relevant fairness factors in the absence of an explicit statutory directive but a mandatory consideration of explicitly listed factors will apply should the Copyright Amendment Bill become law.

With specific regard to the question of whether data science research in the context of public research institutions and involving use of copyright data held by public bodies, is infringing, this should be answered in the negative: a data science researcher in a public research institution such as the members of the DSFSI research group, using copyright-protected broadcast news content to train NLP models would not be infringing on copyright in South Africa or Nigeria as the activities contemplated in the research and involving the use of copyright-protected content, qualify as 'research' and involve dealing fairly with the content. Such use is non-infringing and does not require a licence. Moreover, where copyright subsists in the outcomes of data science research, the relevant copyright owner (barring contractual overrides) is the person(s) who created or authored those outcomes.

To conclude, in terms of both South African and Nigerian law, it would be fair dealing (or fair use when South Africa's Copyright Amendment Bill becomes law) for data scientists in public research institutions to use copyright data as training data for NLP research. Based on the discussion in this paper, such data scientists should also consider the following aspects when using copyright-protected materials as training data for African NLP and data science research generally especially where copyright in those materials are held by a public body:

- the need to indicate and acknowledge the source of the training data;
- the need to safeguard the resulting annotated datasets and trained models also for research purposes by using licences that allow non-commercial research;
- the need to inform, clarify and/or caution users of their research outcomes and outputs as to the extent of accuracy and the limits of possible uses/reuses of the underlying data.

Acknowledgements

Special thanks to Professors Sean Flynn and Vukosi Marivate for their kind review and very helpful comments. All the views expressed herein, as well as any errors, are solely attributable to the author.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Arcadia, a Charitable Fund of Lisbet Rausing and Peter Baldwin through the PIJIP's Project on the Right to Research in International Copyright; and the University of Pretoria's Research Development Grant programme.

ORCID

Chijioke I. Okorie http://orcid.org/0000-0002-1794-4396

References

Abebe, Rediet, Kehinde Aruleba, Abeba Birhane, Sara Kingsley, George Obaido, Sekou L. Remy, and Swathi Sadagopan. 2021. "Narratives and Counternarratives on Data Sharing in Africa." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 329–341.

Ackermann, Laurie WH. 2006. "Constitutional Comparativism in South Africa." *South African Law Journal* 123 (3): 497–515.

Adenuga v Ilesanmi (1991) 5 NWLR (Pt. 189) 82.

Adetoun Oladeji (Nig) Ltd v Nigerian Breweries Plc (2007) LPELR-SC.91/2002

Agic, Željko, and Ivan Vulic. 2019. "JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages." *Association for Computational Linguistics* 3204–3210.

Akingbulu, Akin. 2010. "Public Broadcasting in Africa: Nigeria." In *African Minds*, 134.

Appadurai, Arjun. 2006. "The Right to Research." *Globalisation, Societies and Education* 4 (2): 167–177. https://doi.org/10.1080/14767720600750696

Ashdown v Telegraph Group Ltd [2001] EWCA Civ 1142 (CA)[106]

Authors Guild, Inc. v. Google Inc. 804 F.3d 202 (2nd Cir. 2015)

Babirye, Claire, Joyce Nakatumba-Nabende, Andrew Katumba, Ronald Ogwang, Jeremy Tusubira Francis, Jonathan Mukiibi, Medadi Ssentanda, Lilian D. Wanzare, and Davis David. 2022. "Building Text and Speech Datasets for Low Resourced Languages: A Case of Languages in East Africa." https://repository.maseno.ac.ke/handle/123456789/5278.

Bartz and others v. Anthropic PBC, Case No. 3:24-cv-05417.

Blind SA v. Minister of Trade, Industry and Competition and Others [2023] CCT 320/21 (2) BCLR 117 (CC).

Birhane, Abeba. 2020. "Algorithmic Colonization of Africa." *SCRIPTed* 17:389. https://doi.org/10.2966/scrip.170220.389

Braun, Mikio L., and Cheng Soon Ong. 2018. "Open Science in Machine Learning." In *Implementing Reproducible Research*, 343–365. New York: Chapman and Hall.

Broadcasting Act of 1999. (South Africa).

Bronstein and Katzew. 2018. "Safeguarding the South African Public Broadcaster: Governance, Civil Society and the SABC." *Journal of Media Law* 10(2): 244–272.

Burrow, Sheona. 2021. "The Law of Data Scraping: A Review of UK Law on Text and Data Mining." *Zenodo*. https://doi.org/10.5281/zenodo.4635759.

CCH Canadian Ltd. v. Law Society of Upper Canada [2004] 1 SCR 339

Copyright Act 2022 (Nigeria).

Copyright Act 98 of 1978 (South Africa).
Copyright Amendment Bill B13B–2017. (South Africa).
Craig, C. J. 2024. "The AI Copyright Trap." SSRN. https://doi.org/10.2139/ssrn.4905118.
D'Agostino, Gluseppina. 2008. "Healing Fair Dealing-A Comparative Copyright Analysis of Canada's Fair Dealing to UK Fair Dealing and US Fair Use." *McGill LJ* 53: 309–362.
de Castilho, R. E., G. Dore, T. Margoni, P. Labropoulou, and I. Gurevych. 2018. "A Legal Perspective on Training Models for Natural Language Processing." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan: European Language Resources Association (ELRA).
Erickson, Kris, Paul J. Heald, Fabian Homberg, Martin Kretschmer, and Dinusha Mendis. 2015. "Copyright and the Value of the Public Domain: An Empirical Assessment." *Intellectual Property Office Research Paper*: 15–16.
Galago Publishers (Pty) Ltd and another v. Erasmus [1989] (1) SA 276 (A)
Geyer, Sunelle. 2022. "Determining Originality in South African Copyright Law: Is It 'Or,' 'And,' or Something More?" *THRHR* 85: 176–195.
Greenleaf, Graham, and Catherine Bond. 2013. "Copyright: What Makes Up Australia's Public Domain" *Australian Intellectual Property Journal* 111–138.
Haupt t/a Soft Copy v. Brewers Marketing Intelligence (Pty) Ltd and Others [2006] 908 JOC (A).
Hlomani, Hanani, and Caroline B. Ncube. 2023. "Data Regulation in Africa: Free Flow of Data, Open Data Regimes and Cyber Security." https://publication.aercafricalibrary.org/server/api/core/bitstreams/22843761-f593-4859-8199-cfa750491c15/content.
Hu, Songbo, Abigail Oppong, Ebele Mogo, Charlotte Collins, Giulia Occhini, Anna Barford, and Anna Korhonen. 2025. "Natural Language Processing Technologies for Public Health in Africa: Scoping Review." *Journal of Medical Internet Research* 27:e68720. https://doi.org/10.2196/68720
Jacana Education (Pty) Ltd v. Frandsen Publishers (Pty) Ltd [1998] (2) SA 965 (SCA)
Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–6293.
Kruit, Benno. 2023. "Minimalist Entity Disambiguation for Mid-Resource Languages." In *Proceedings of the Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, 299–306.
Ladoja v. INEC. 2007. 12 NWLR (Pt. 1047) 119.
Lee, Jyh-An. 2016. "Licensing Open Government Data." *Hastings Business Law Journal* 13: 207–240.
Marcowitz-Bitton, Miriam. 2015. "Commercializing Public Sector Information." *Journal of Patent & Trademark Office Society* 97:412–441.
Margoni, Thomas, and Martin Kretschmer. 2022. "A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology." *GRUR International* 71 (8): 685–701. https://doi.org/10.1093/grurint/ikac054
Marivate, Vukosi. 2021. "Why African Natural Language Processing Now? A View from South Africa #AfricaNLP." In *Leap 4.0: African Perspectives on the Fourth Industrial Revolution*, 126.
Marivate, Vukosi, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. 2020. "Low Resource Language Dataset Creation, Curation and Classification: Setswana and Sepedi." arXiv Preprint arXiv:2004.13842. https://doi.org/10.48550/arXiv.2004.13842.
Married Media Ltd v. Lawrence Akpa [1990–1997] 3 IPLR 202
Moleya, Ndivhuwo Ishmel. 2020. "Evaluating the Copyright Protection of Databases in South Africa: A Comparative Analysis with the European Union." *South African Intellectual Property Law Journal* 8 (1): 56–79.
Maseno University. 2024. "African Next Voices workshop." https://www.maseno.ac.ke/africa-next-voices-workshop#:~:text=AFRICA%20NEXT%20VOICES%20is%20a,Kimathi%20(DeKut)%20and%20LDRI%3A
MEC for Education KwaZulu-Natal v Pillay 2008 1 SA 474 (CC)
Moneyweb (Pty) Limited v. Media 24 Limited and Another [2016] 3 All SA 193.

Moody, Glyn. 2023. "A "Blatant No" from a Copyright Holder Stops Vital Linguistic Research Work in Africa." *Walled Culture*. https://walledculture.org/a-blatant-no-from-a-copyright-holder-stops-vital-linguistic-research-work-in-africa/.

Ncube, Caroline, Lucienne Abrahams, and Titilayo Akinsanmi. 2014. "Effects of the South African IP Regime on Generating Value from Publicly Funded Research: An Exploratory Study of Two Universities." In *Innovation and Intellectual Property: Collaborative Dynamics in Africa*, 282–315.

Ncube, Caroline, and Isaac Rutenberg. 2020. "Intellectual Property and Fourth Industrial Revolution Technologies." In *Leap 4.0: African Perspectives on the Fourth Industrial Revolution*, edited by Z. Mazibuko-Makena, and E. Kraemer-Mbula, 393–416. Johannesburg: MISTRA.

Nekoto, Wilhelmina, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, et al. 2020. "Participatory Research for Low-Resourced Machine Translation: A Case Study in African Languages." arXiv preprint arXiv:2010.02353.

Nigerian Television Authority Act 24 of 1977.

Oira, Hezekiel, and Lonias Ndlovu. 2018. "The Dichotomy between Signal and Content as Basis of Broadcast Copyright: A Kenyan and South African Perspective." *Obiter* 39 (2): 399–429. https://doi.org/10.17159/obiter.v39i2.11371

Okorie, Chijioke. 2023a. "Beyond Intellectual Property Protection: Other AI IP Strategies for the African Context." In *Artificial Intelligence and the Law in Africa*, edited by Caroline Ncube, Desmond Oriakhogba, Tobias Schonwetter, and Isaac Rutenberg, 153–174. South Africa: LexisNexis.

Okorie, Chijioke. 2023b. "Fair Use or Fair Dealing in Africa: The South African Experience." In *Developments and Directions in Intellectual Property Law: 20 Years of the IPKat*, edited by Hayleigh Bosher, and Eleonora Rosati, 257–275. London: Oxford University Press.

Omega Bank Plc. v. Govt. Ekiti State (2007) 16 NWLR (Pt. 1061), 445.

Okorie, Chijioke. 2023. "Realising a 'Right' to Research in Nigeria and South Africa: The Role of the Executive Arm of Government." *Journal of Comparative Law in Africa* 10 (2): 141–173. https://doi.org/10.47348/JCLA/v10/i2a5

Oriakhogba, Desmond. 2023. *The Right to Research in Africa: Exploring the Copyright and Human Rights Interface*. Switzerland: Springer Nature.

Oriakhogba, Desmond, and Ifeoluwa Olubiyi. 2023. *Intellectual Property Law in Nigeria: Emerging Trends, Theories and Practice*. Benin City: Parclerd Press Limited.

P1 Computational Research. 2023. "Africa Examples, Right to Research in Africa Conference, Pretoria, January 23, 2023." *YouTube video*. Accessed 12 September 2023. https://www.youtube.com/watch?v=rZ-3MHcu1oA.

Peter Obe v. Grapevine Communication Ltd, case no. FHC/L/CS/1247/97 (unreported)

Plateau Publishing Ltd v. Adophy [1986] 4 NWLR (Pt. 34) 205

Rautenbach, Christa. 2015. "The South African Constitutional Court's Use of Foreign Precedent in Matters of Religion: Without Fear or Favour?" *Potchefstroom Electronic Law Journal (PELJ)* 18 (5): 1546–1570.

Rosati, Eleonora. 2023. "Copyright Reformed: The Narrative of Flexibility and Its Pitfalls in Policy and Legislative Initiatives (2011–2021)." *Asia Pacific Law Review* 31 (1): 33–54. https://doi.org/10.1080/10192557.2022.2117482

S v Lawrence, S v Negal, S v Solberg 1997 4 SA 1176 (CC).

Sanneh, Lamin. 2015. *Translating the Message: The Missionary Impact on Culture*. New York: Orbis Books.

Sanderson v Attorney-General, Eastern Cape 1998 2 SA 38 (CC)

Shay, R. M. 2014. "Exclusive Rights in News and the Application of Fair Dealing." *SA Mercantile Law Journal* 26 (3): 587–605.

Society of Composers, Authors and Music Publishers of Canada v. Bell Canada [2012] 2 SCR 326

South African Broadcasting Corporation SOC Ltd v. Vollenhoven and Appollis Independent CC and Others [2016] 4 All SA 623.

Thomson Reuters Enterprise Centre GmbH v. Ross Intelligence Inc, No. 1:20-CV-613-SB (2025)

United Nations Educational, Scientific and Cultural Organization. 2017. *Resolution 15, Annex II, Recommendation on Science and Scientific Researchers, adopted at the 39th session of the*

General Conference (October 30–November 14, 2017). https://en.unesco.org/themes/ethics-science-and-technology/recommendation_science.
Von Kries, Caroline, and Gerd Winter. 2015. "Defining Commercial and Non-commercial Research and Development under the Nagoya Protocol and in Other Contexts." In *Research and Development on Genetic Resources*, 60–74. London: Routledge.
Wei, Johnny Tian-Zheng, Maggie Wang, Ameya Godbole, Jonathan H. Choi, and Robin Jia. 2025. "Interrogating LLM Design Under a Fair Learning Doctrine." arXiv preprint arXiv:2502.16290.
Yemitan v. Daily Times & Anor. [1977–1989] 2 IPLR 141–156.
Yeni Anikulapo-Kuti & Ors. v. Iseli & Ors. [2003–2007] 5 IPLR 53–73.

Annex 1. Proposal by the Data Science for Social Impact (DSFSI) research group, University of Pretoria

3 Apr 2023

Improving News Categorization, Translation, Named Entity Recognition, and Part-of-Speech Tagging with Natural Language Processing Techniques.

Background

The South African news media, particularly SABC, has played an important political and social role in the two and a half decades since apartheid ended. With strong constitutional protections for free speech and a strong civil society, the SABC has helped create a culture of democratic debate while acting as a watchdog to hold political power accountable by looking into corruption and wrongdoing. People are getting more and more connected to the Internet, and there are more and more news articles available to them every minute and second. This has led to a problem called 'information overload' among Internet users. Hence, there is a need to automatically put news into the categories to reach more readers in a more preferred subject to meet their needs about social economics, the policies of the government, and enlightenment. Thus, every day, news organisations put out a huge amount of content, which makes it hard to keep up with high-quality content. Effective news categorization, translation, named entity recognition (NER), and part-of-speech (POS) annotation are all necessary for understanding and analysing news content. All of this will make meta-data more useful for publishers and easier to process automatically using natural language processing (NLP) and artificial intelligence technology to make tools that could be used to automatically model the inherent subjectivity in natural language and improve the overall quality of news content.

Objectives:

The main goal of this research proposal is to find out how well NLP techniques can improve news categorization, translation, natural language understanding (NER), and POS tagging. Specifically, the following are the contributions and objectives that will be pursued:

(1) To develop an NLP model to categorise news content based on topics such as politics, sports, and entertainment.
(2) To develop an NLP model to translate news content from one language to another.
(3) To develop an NLP model to perform NER on news content to identify named entities such as people, places, and organisations.
(4) To develop an NLP model to perform POS tagging on news content to identify parts of speech such as nouns, verbs, and adjectives.

Methodology

The proposed study will be conducted in four phases. In the first phase, a natural language processing (NLP) model will be made to classify news content based on the subjective nature of preference topics. The model will be trained using a dataset of news content sourced from the SABC integrated with a large language model (LLM) to identify a few representative stylistic elements that can be used to classify news content.

The training dataset could be transcripts from radio, television, and written media that the SABC produces. We would like to be able to access media in all formats: written, audio, video etc.

In the second phase, an NLP model will be developed to translate news content from one language to another. The model will be trained with a set of news articles written in two languages. This will help it figure out the patterns and structures of language translation. This may require the translation of SABC news content.

In the third phase, an NLP model will be developed to perform NER on news content to identify named entities such as people, places, and organisations. The model will be trained with a set of news articles that have been labelled with names of people, places, and things.

In the fourth phase, an NLP model will be developed to perform POS tagging on news content to identify parts of speech such as nouns, verbs, and adjectives. The model will be trained using a dataset of news content with annotated POS tags.

Expected outcomes

The expected outcomes of this research proposal are as follows:

- The development of an NLP model to categorise news content based on topics.
- The development of an NLP model to translate news content from one language to another.
- The development of an NLP model to perform NER on news content to identify named entities such as people, places, and organisations.
- The development of an NLP model to perform POS tagging on news content to identify parts of speech such as nouns, verbs, and adjectives.
- Release of the derivative data/models for: categorisation, translations, NER tags, POS tags under a permissive licence for other researchers to be able to use.

Conclusion

The goal of this research proposal is to find out how well NLP techniques with a large language model (LLM) can help improve news categorization, translation, natural language understanding (NER), and POS tagging. By using NLP and deep learning techniques to automate the production of news, it will help individuals and organisations find news articles that match their interests using a search engine and produce more high-quality content in less time. The results of the proposed study could be useful for news organisations that want to improve the quality of their content and keep their readers coming back. provide valuable insights for news corporations looking to enhance their content quality and maintain a loyal audience.