IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

BATRACIO: BAsic, TRAnslational, Clinical, research phase identification in bIOmedical publications

NICOLAU DURÁN-SILVA[1,2], JORGE CARRILLO-DE-ALBORNOZ[3], LAURA PLAZA[3], SARA RICARDO[1], FRANCESCO A. MASSUCCI[1], SONIA VEIGA[1], ARNAU RAMOS-PRATS[4,5]

[1]SIRIS Lab, Research Division of SIRIS Academic, Avda. Francesc Cambó, 17, 08003 Barcelona, Spain
[2]DLaSTUS Lab, TALN Group, Universitat Pompeu Fabra, C/ Tànger, 08018 Barcelona, Spain
[3]NLP & IR UNED, C/Juan del Rosal, 16, 28040 Madrid, Spain
[4]Department of Pharmacology, Medical University of Innsbruck, Innrain 80-82, A - 6020 Innsbruck, Austria
[5]Friedrich Miescher Institute for Biomedical Research, Fabrikstrasse 24, 4056 Basel, Switzerland

Corresponding author: Laura Plaza (e-mail: lplaza@lsi.uned.es).

This research is funded by FAIRTRANSNLP-DIAGNOSIS: Measuring and quantifying bias and fairness in NLP systems, grant PID2021-124361OB-C32, funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe. This work has been also funded by the Ministry of Universities and the European Union through the EuropeaNextGenerationUE funds and the "Plan de Recuperación, Transformación y Resiliencia". Supported by the Industrial Doctorates Plan of the Department of Research and Universities of the Generalitat de Catalunya.

ABSTRACT The increasing interest from research agencies, governments, and universities in understanding research funding and prioritising research efforts has highlighted the need for reliable and efficient methods for exploring research portfolios. In biomedical research, this involves exploring research across what is normally considered fundamental and applied research. As research done in these different categories does not have the same behaviour, such as time to impact or citation behaviour, it is often important to address them separately. Moreover, research is increasingly complex, interdisciplinary and transversal, and increasingly of translational nature. Currently, there are no available tools, as far as we know, that do this. Scientific publications offer a valuable source of information for this purpose, but the growth in the number of biomedical publications makes manual inspection and classification of papers unfeasible. To address this challenge, we present BATRACIO, a new task that aims to classify biomedical publications into the following research types: *Basic*, *Translational*, *Clinical*, and *Public Health*. We develop and release an expert annotated dataset for the task and evaluate state-of-the-art models to determine the effectiveness of domain-specific pre-trained language models in comparison to general pre-trained language models. We also investigate methods for handling imbalanced datasets in the biomedical domain with adjacent categories. Our results demonstrate that domain-specific pre-trained language models can effectively classify scientific papers based on the research type, overcoming challenges such as the use of abbreviations and acronyms. These findings have important implications for policymakers and funding agencies in understanding research activities and allocating resources effectively.

INDEX TERMS biomedical research, natural language processing, scholarly document processing, science mapping, text mining

I. INTRODUCTION

Research funding agencies, governments, and research institutions are becoming increasingly interested in understanding the overall impact of their work or the research they finance. This is crucial for improving decision-making practices on priority setting and resource allocation (PSRA) [1], [2]. Given that resources are limited, it is essential to prioritise research efforts effectively. In fields like biomedicine, where research is costly and sometimes inefficient [3], there is a risk of duplicating resources or inadequately funding critical priorities. Since early 2000s, and with the increasingly complex nature of biomedical research, its exponential growth (in publication volume) and the pressing need to serve better patients and society [4], [5], biomedical fundamental research started

to suffer pressures to have its discoveries better applied to clinical practice, and the concept of "translational research" was born. The last decades have seen a push for policies, funding and infrastructures on translational research, as well as an overall attempt of changing biomedical research to be more interdisciplinary and interconnected (policies in US [4], UK [6], and Europe [7] and the creation of *eatris*[1]). Although these are laudable efforts, the research types that are part of this virtuous circle (fundamental, translational and clinical research) are distinct: in its times to impact, citation behaviours, time to publication and often funding required, but precisely because there are specific funding policies and aims, for example to push translational research into applications, it is useful to be able to assess them separately.

Classifying research outputs into distinct research types can be very useful for a wide range of stakeholders. For research funding agencies and governments, such classifications can help map the actors researching a specific area of interest in their territory, assess if specific policies supporting a specific type of research are successful, assess the balance on research portfolio and/or if there are gaps to be filled in the support of a specific type of research for the overall success of research applicability. Universities, on the other hand, can use this information to understand their strengths and weaknesses, prioritise or adjust their research strategy, and identify areas of improvement and healthy areas to have continued support. Researchers can also benefit from classifications by more easily extracting relevant information from large collections of scientific articles.

Scientific publications provide a valuable source of information for exploring research portfolios and understanding the different contributions of research activities [1]. Automatically understanding the topics addressed by scientific publications in the biomedical domain has been a challenge for over two decades. With more than 3,000 publications generated daily [8], and also the substantial growth of translational research since the 2000s [9] manual inspection and classification are no longer feasible. To address this challenge, various research tasks such as text summarization [10], relation extraction [11], [12], question-answering [13], and text classification [14]–[16] have been investigated by the research community.

In recent years, Transformer-based neural language models such as BERT [17] or RoBERTa [18] have achieved impressive results in these tasks. These models are pre-trained on large-scale unlabelled documents and can learn universal language representation, which is then adapted to downstream tasks. While most models are pre-trained on general domain data, some are pre-trained or adapted to the biomedical and clinical domains [19]–[21], offering promising results in those areas. The use of these models has helped overcome some of the main difficulties of BioNLP, which include lexical ambiguities, the use of acronyms and abbreviations [22], detection of negations, and determination of temporal context [23].

One common challenge that these approaches face is that individual publications may be inaccurately classified because traditional categorization methods rely on scientific journals and do not account for overlap or emergent fields [1]. This is even more challenging in the identification of research themes in the biomedical field, as existing categorisations do not take into account the multidisciplinary complexity of the research or the overlap of biomedicine and biology fields [24].This disciplinary characterization can become outdated and may not align with the increasingly interdisciplinary nature of modern science. Specifically, areas that are highly interconnected often require the participation of multiple actors from diverse disciplines to achieve a complete discovery. This emphasizes the limitations of traditional disciplinary boundaries and existing categorisations of research, which do not accurately reflect the collaborative, interdisciplinary and interconnected nature of contemporary research.

Our work makes three significant contributions [25]. Firstly, we introduce a new detailed definition of research types, which has been created by domain experts. We have conducted an exhaustive review of scientific publications and manually classified them to identify the clear boundaries between the different types of research. As a result, we propose a classification of biomedical research by types that includes four categories: `basic research`, `translational research`, `clinical research`, and `public health`. Secondly, to facilitate the development of machine learning models capable of accurately classifying scientific outputs according to research types, it is essential to have labelled datasets for training such systems. However, as far as we are aware, no datasets currently exist that assign research types to scientific publications in the biomedical field. Thus, our second contribution involves creating a manual dataset for classifying scientific publications into their respective research types in the biomedical domain. Thirdly, we introduce and define a new NLP task, BATRACIO (*BAsic-TRAnslational-Clinical research types classification in biomedical publications*), which aims to classify biomedical literature into the different research types. We evaluate state-of-the-art models to determine whether domain-specific pre-trained language models outperform general pre-trained language models and investigate methods for adapting them to handle imbalanced datasets in the biomedical domain with adjacent categories.

Our results indicate that domain-specific pre-trained language models can effectively classify scientific papers based on the research types. These models can also overcome some of the challenges of biomedical language, such as the use of abbreviations and acronyms. However, some text preprocessing may still be necessary to optimize their performance.

The remaining sections of the article are organized as follows. In Section II, we present the state-of-the art in the area of biomedical text classification. Section III describes

[1]https://eatris.eu/

the dataset developed to assist and promote research in the proposed task. In Section IV, we present the different experiments performed to evaluate the suitability of transformer-based architectures to classify scientific articles according to the different types of biomedical research. Section V presents the evaluation results. Finally, in Section VII, we discuss the conclusions of the study and propose future lines of work.

II. RELATED WORK

Natural language processing (NLP) in the biomedical domain, also known as BioNLP, is a highly challenging task. The language used in this domain involves a lot of specific words and terminology, polysemic words, frequent use of acronyms and abbreviations, and requires knowledge of the domain that is expected to be known or inferred from the context. Omitted information is especially problematic for BioNLP because a system must have additional knowledge to be able to gather all the implicit information [26]. In fact, many documents in the biomedical domain are not easily accessible or understandable by humans without sufficient domain expertise.

In the past twenty years, the field of BioNLP has seen significant growth [26]–[30]. The availability of databases such as PubMed/ MEDLINE [31] and a wide range of corpora, semantic resources, and ontologies, such as UMLS [31], Gene Ontology [32], and MeSH [33], has contributed to this growth. In recent years, the advancement of deep learning in natural language processing has further fueled the development of BioNLP tasks [34]. Transformer-based Pre-trained Language Models (PLMs) are explored in the overwhelming majority of the papers in the Proceedings of the 20th Workshop on Biomedical Language Processing to solve various NLP tasks in the biomedical domain [34]. However, deep learning approaches typically require large amounts of annotated data, and there are limited labeled data due to the high annotation costs. Moreover, incorporating external knowledge of the domain into the models remains a significant challenge in BioNLP [34].

Text classification is a well-known challenge in natural language processing. It involves labeling natural language texts with a set of predefined tags or assigning classes or categories to different text units like sentences, paragraphs, or documents [35], [36]. Over the past two decades, text classification techniques have seen significant advancements [36], [37]. However, in the biomedical and medical records domain, text classification is especially challenging due to imbalanced datasets, misspellings, acronyms and abbreviations, negations, and semantic ambiguity [23], [38], [39].

While early text classification techniques relied on rule-based systems, defining the necessary rules requires extensive domain knowledge and manual effort [36]. Furthermore, complex domains can pose a challenge to these systems, as they may struggle to capture nuanced messages and hidden patterns [40]. In some biomedical applications, however, rule-based systems remain prevalent due to their ability to incorporate domain-specific knowledge and tackle feature extraction challenges. [38], [39].

Machine learning-based techniques have improved the results of rule-based systems, but they typically require a large annotated corpus [41], which is especially challenging in small and limited datasets like those found in the biomedical domain. These methods typically involve a two-step approach of feature extraction and classifier feeding, but feature extraction is costly and time-consuming [42]. Furthermore, selected features may not cover all linguistic variants, limiting the portability or generalization of systems for further applications in new domains [11]. Biomedical and clinical documents often require more complex features than general domain texts. For instance, citation information [43] and ontological information [44] have been used to improve text classification accuracy in these domains.

In the last years, deep neural network-based techniques have stood out because of their simplicity, reducing the costs of manual feature extraction, higher processing efficiency, and, in general, because they have managed to match or improve state-of-the-art results in many NLP tasks [29]. They generally include feature extraction in the model fitting process by learning a set of non-linear transformations that allow the mapping of features directly to outputs [42]. Many architectures have been proposed such as Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN); but nevertheless, Transformer seems to capture better long dependencies in text, improve in computationally and take less time to train. It treats text as a fully-connected graph with attention features between different words by self-attention, which can extract features and relations between words efficiently and solves short-term memory problems [42].

Pre-trained Language Models (PLMs) like BERT, GPT-2, and RoBERTa have been adapted to the text classification task by stacking a linear classifier on top of the last hidden state of the model and fine-tuning the parameters from the model by maximizing the log-probability of the correct label [45]. Further pre-training and in-domain pre-training have been shown to improve the performance of text classification tasks, but cross-domain pre-training does not provide much benefit. BERT has also been shown to improve the task's performance with small-sized data [45].

PLMs based on Transformer, such as BERT or RoBERTa, have been successful in solving many NLP tasks by learning global semantic representations from large datasets. These models typically use unsupervised methods on large datasets and are then fine-tuned on downstream tasks without having to train the entire model from scratch. However, the performance of general-domain PLMs in the biomedical domain has been limited in some cases, prompting researchers to focus on developing PLMs specifically for biomedical texts [46]. Pretraining PLMs on biomedical corpora has been shown to improve their performance [19], [20]. As a result, several BERT models have been developed to tackle texts in the biomedical domain using different domain-adaptation strategies and corpora [19], [20], such as BioBERT [19], SciBERT [47], BioMedBERT [48], OuBioBERT [49], Pub-

MedBERT [20] or BlueBERT [50]. However, while each PLM implements a different strategy for specific-domain adaptation or pre-training, their success varies. General-domain BERT models that have been pre-trained on the medical subset of Wikipedia should perform better in the biomedical domain than other general-domain architectures [51]. Nonetheless, the performance of general-domain PLMs in the biomedical domain has been limited in some cases. Therefore, researchers have developed PLMs specifically designed to handle biomedical texts [46] and have pre-trained them on biomedical corpora to improve their performance [19], [20]. Several BioNLP shared tasks have boosted research in the field and supported the evaluation of methods, as it is the case of BioASQ [52] focused on answering questions for COVID-19, MEDIQA [53] aimed for different summarisation tasks of medical texts, or DDIExtraction [54] and PharmaCoNER [55] aimed at identifying entities and relations between drugs, proteins, among other entities.

Some recent text classification tasks have aimed to classify scientific publications according to the Hallmarks of Cancer taxonomy [56], classify medical scientific documents according to disease classes from the Medical Subject Heading (MeSH) vocabulary [57], label medical notes with codes from the ICD taxonomy [58], or detect biomedical claims in tweets [59]. For instance, [58] explore automatic International Classification of Disease (ICD) coding, which is a multi-label classification problem. However, they face a challenge with the length limitation of BERT-based models, which can only process a maximum length of 512 tokens, whereas clinical notes usually exceed this maximum input length. Given these limitations, they opt for feature extraction and classification, using PubMedBERT. They pre-process the text by converting to lower case and removing all numbers, but they do not remove infrequent words since BERT does not suffer from out-of-vocabulary terms. On the other hand, [59] focus on the text classification of claims in biomedical tweets. They experiment with different combinations of fine-tuning hyperparameters from [17]. They also oversample the minority class of implicit claims to achieve a balanced training set. However, their experiments reveal that more complex models, such as BERT or LSTM, do not outperform the linear models, which they attribute to the small size of the dataset and the inability of complex models to learn from the training set.

In recent years, there has been a surge in research focusing on biomedical document classification using advanced PLMs. Models like LinkBERT, which integrates link prediction tasks during pretraining, have demonstrated enhanced understanding of document context and improved classification performance in biomedical domains [60]. Additionally, other models such as ClinicalBERT and BioLinkBERT have been tailored to integrate clinical knowledge and link prediction respectively, to further enhance their applicability in clinical and biomedical contexts [61], [62]. These models leverage the specificity of domain knowledge to improve the accuracy and reliability of biomedical document classification tasks.

Fine-tuning strategies and the incorporation of domain-specific corpora continue to show significant promise. A recent study by Kim et al. [63] highlighted the efficacy of integrating electronic health records (EHRs) with PLMs to achieve superior classification accuracy for clinical notes. This approach allows models to capture the nuances and specific terminologies used in clinical documentation, thus enhancing their overall performance in real-world applications. Moreover, recent advancements have explored multi-task learning frameworks that enable models to be fine-tuned on multiple related tasks simultaneously, thereby improving their generalization capabilities across various biomedical text classification problems [64].

The development of hybrid models that combine rule-based systems with machine learning techniques has also been an area of active research. For instance, Luo et al. [65] proposed a hybrid approach that integrates domain-specific rules with PLM-based classifiers to improve the classification of rare biomedical entities. This method leverages the precision of rule-based systems and the adaptability of PLMs, achieving better performance in tasks where annotated data is scarce.

Another notable advancement is the application of weak supervision techniques to generate large-scale annotated datasets from unstructured biomedical texts. Ratner et al. [66] demonstrated the use of weak supervision to create training data for PLMs, significantly reducing the cost and time associated with manual annotation. This approach has shown to improve model performance by providing a more diverse and comprehensive set of training examples.

These advancements underline the importance of continued innovation in pretraining techniques and the development of domain-specific models to tackle the unique challenges presented by biomedical texts. The integration of external knowledge sources, such as medical ontologies and databases, into PLMs remains a critical area of research, as it holds the potential to further enhance model performance and reliability in the biomedical domain. For example, recent efforts to incorporate the Unified Medical Language System (UMLS) into PLMs have shown promising results in improving the interpretability and accuracy of biomedical text classification [67].

As far as we know, this is the first study that aims to classify scientific biomedical articles based on the research type in which the described investigation takes place [25]. Transformers have shown to be the best performing architecture to date, but they require annotated data. Therefore, our first contribution is a dataset annotated by experts that can aid in the development of transformers capable of addressing our novel task.

III. THE BATRACIO DATASET

The task of classifying biomedical publications according to research type in the biomedical domain is a multi-class classification problem: given an input, the system must identify the most appropriate category according to the content described

in the text. The dataset is composed of titles and abstracts of scientific articles, which are the minimum annotation units to differentiate between the categories proposed. We have ensured a real-world distribution, where classes are not balanced, to improve the system's generalizability and use in real cases.

A. DATA SELECTION

To create our dataset, we started by gathering a diverse sample of publications from PubMed. Using the API, we randomly selected 500,000 publication IDs and retrieved data for each of those records. We limited our search to a temporal range of five years (2015-2019) to maintain temporal balance. We excluded earlier years to account for the growing trend of translational research, and we excluded publications from 2020 due to the significant focus on COVID-19 research. We specifically chose "journal articles" in English as the type of records for our dataset. This decision was based on the fact that they comprise the majority of publications, follow similar patterns and structures in their abstracts, and have a certain level of homogeneity in their writing.

In the second step of our data processing, we aimed to filter out records that were not directly relevant to our study on biomedical research. While the database contained a wide range of records in health and life sciences, we only wanted to include publications in the biomedical domain for annotation by experts. To achieve this, we used ontological classification based on rules defined by discarding and selecting branches associated with biomedical research in the Medical Subject Headings (MeSH) [33]. We worked with domain experts to select MeSH branches at different levels for filtering publications. Despite this effort, some publications outside the domain remained, so we included a second filtering step to eliminate them. To achieve this, we used the alignment between journals and subjects proposed by Science-Metrix [68]. Science-Metrix offers an open ontology for classifying scientific journals into bibliometric categories based on ISSN and ESSN codes. The taxonomy consists of 22 research fields, each containing subfields, which helped us remove publications outside our areas of interest[2].

To ensure the effectiveness of our filtering approach, we manually evaluated 100 filtered publications with the help of domain experts. The goal was to confirm that the publications were relevant to the biomedical domain. The experts evaluated both filters on a random set of 100 publications, and the double filtering was found to have a precision of 92.8%. The final subset of publications was selected using this filtering approach, and any publications that were outside the scope of biomedical research and were discovered by the annotators during the annotation process were discarded.

We applied a filtering step to the initial dataset, resulting in 145,821 publications. From this pool, we selected 1,100 candidate publications for annotation. We ensured that the sampling was representative and maintained an appropriate distribution in the features of interest, including publication year, affiliation country of the authors, journal, anatomy (MeSH branch A), organism (MeSH branch B), diseases (MeSH branch C), and chemicals and drugs (MeSH branch D). This approach allowed for the generalizability of the dataset, as a very large sample of a specific disease in any of the annotated categories could add noise and bias the dataset, making it difficult to identify generalizable patterns in the texts.

[2]Both files will be made available upon acceptance

B. ANNOTATION PROCESS

A manual annotation process is preferred, since especially in this domain, the quality of the annotations has been shown to be more important than the quantity of annotations [41]. For the dataset annotation process, we have been inspired and oriented by the corpus creation methodologies proposed by [41], [69], [70]. We consider the following steps:

1) Annotation guide development and study of sample examples to create the guide.
2) Training of the annotators.
    a) Labelling a sample of texts.
    b) Discussion about conflicts.
    c) Updating the annotation guide to cover conflicts.
3) Annotation of the dataset.
    a) Annotation of the 20% of the records.
    b) Agreement at first 20%.
    c) Redefinition of the annotation guideline.
    d) Annotation of the whole dataset.
    e) Final agreement.
4) Dataset statistics and description.

1) Defining Research Types

To start, this task is crucial to firstly define the different research types that one wants to classify. The simpler way is to separate between what is often called basic or fundamental and clinical or applied research, as we have done in the past [71]–[73]. However, by doing so we are losing information on a research type that has become key in policy discussions, that of Translational Research (see also introduction). Although it is very important to have this category to be able to map research outputs and answer questions related to translational research, it is also the most hard to classify. Its borders are "fuzzy", especially so in what is basic research and what is translational research, and its definitions vary substantially [9], [74]. Even if various attempts to define the types of biomedical research can be found in the literature [75]–[77], there is no widely accepted definition. We have used the following definition for Translational Research, knowing that this is only one of possible definitions: biomedical research in a pre-clinical phase and explicitly with an intent to treat. This is, for example, different from a definition of translational research that more broadly deals with mechanisms of disease in addition to specific pre-clinical testing, even if not a direct intent to treat is the aim. We do not make judgments as to which is the best definition. Both are valid and can be useful

depending on the aim; we have chosen a stricter definition because it would allow us to map more precisely what is solely Translational/pre-clinical research. The definitions of basic and clinical research are less debatable and are, thus, easier to pinpoint. Public health research was originally added to the clinical research category, but during testing of the system, it was thought best to keep it as a separate category as they didn't represent the same type of studies, and most importantly it became increasingly obvious that is a specific type of work, often intersectoral.

The `other` category has been included to cover those documents that annotators have discarded during the annotation process as not being properly biomedical research or not fitting into any of the four categories proposed.

Once definitions are agreed upon, it is key to define very clear guidelines for deciding what belongs to what classification. In the next subsection, we explain in more detail the definition we have used for each category and the process of guideline development that ensued.

2) Guidelines' Development for Expert Annotators

The process of developing the annotation guidelines from scratch started with the definition of terms and an as-clear as possible definition by two domain experts of the boundaries of these categories, based on previous existing literature and general understanding. Initially, only three categories were considered: *basic research*, *translational research*, and *clinical research*, assuming that they would cover the entire spectrum of biomedical research. To adapt these preliminary definitions to an annotation scenario, both experts annotated 100 publications in the biomedical domain (randomly extracted from PubMed) with the three proposed categories. During this process, it became clear the need to add a fourth type, related but not truly clinical research - public health- to encompass the entire range of publications considered as biomedical research. The addition of this fourth category was seen as relevant as some of the publications focused on the broader spectrum of Health and wellbeing, beyond the diagnostic and treatment of disease, including for socioeconomic aspects, retrospective studies about population impact, and health policy issues. Importantly, through further labelling, by the supplementation of more examples for each of the categories, the guidelines went through several iterations to make the borders very clear. As the process of labelling by humans is inherently biassed, the annotators took care in following the guidelines and not their "perception" on any given day. The annotation guidelines are provided as an appendix to this document.

As a result, the following four categories were considered for the BATRACIO task[3]:

- **Basic research**, often called *fundamental research*, focuses on scientific exploration and on building new knowledge, and aims to understand fundamental mechanisms of biology, disease and behaviour. For example, in the case of cancer research, basic research asks how or where mutations occur in DNA and how DNA functions in a healthy cell [78].
- **Translational research**, also called *pre-clinical research* [74] focuses on translating the discoveries from basic research into usability in the clinic, to produce new drugs, devices and treatment options for patients, with a particular focus on applicability. It uses large-scale testing and both animal models and human biological material, such as computer-assisted simulations of drugs, devices or diagnostic interactions within living systems. For example, in the case of cancer research, translational research asks if and how certain drugs or therapeutic approaches halt cancer growth, invasion or metastasis in non-human models.
- **Clinical research** seeks to test a specific treatment or procedure, drug, diagnostic or any technology on patients, focusing not only on the biological mechanisms, but also on issues of safety, delivery and protocols for implementation [78]. It includes studies to better understand a disease in humans and relates this knowledge to findings in cells or animal models. For example, in the case of cancer research, clinical research asks if and how certain drugs or therapeutic approaches halt cancer growth, invasion or metastasis in patients.
- **Public health** involves activities to strengthen public health capacities and services that seek to provide conditions under which people can stay healthy, improve their health and well-being, or prevent the deterioration of their health. Population analyses and retrospective studies are considered in this phase. For example, in the case of cancer research, public health research maps the burden of disease (epidemiological studies) and asks if and how certain behaviours and exposures affect cancer incidence and/or prognostic.

As part of the same experiment, we evaluated the minimum unit of annotation. Due to concerns that the title and abstract might not provide sufficient information for some articles, we provided the sections of introduction, materials and methods, and MeSH terms to the experts for each article. Even though this was the case, the experts were able to assign all four categories to scientific publications using just the title and abstract, in the majority of the cases. This finding is significant for future applications of this resource since 52% of scientific articles in the fields of life sciences and molecular biology are not open access, and only the title and abstract are accessible through PubMed.

Differentiating between some of the types can be extremely challenging. In some cases, the suitability of a category may be clear because the methods and the type of research activity fall undoubtedly within a specific category. However, in other cases, the difference may be difficult to discern because what distinguishes the types is the scientific question behind the research activity. Furthermore, in some cases, a document may be near the borders between two areas because the task

[3]Complete annotation guidelines are available as an appendix.

requires documents to belong to one category. To assist annotators in such ambiguous cases, experts have exhaustively defined the categories. These definitions have been informed by extensive experts' discussions and clarifications using a sample of documents. After annotating the first 400 publications, we updated the guidelines by providing additional examples to clarify the boundaries between categories. Experts expressed that the boundary between basic and translational research categories was sometimes "fuzzy", so we paid particular attention to providing more examples for these categories.

The objective of *BATRACIO* will be to develop an automatic text classification system able to assign the label corresponding to the futher research phase presented in the article based on the annotated dataset provided by BATRACIO. Providing, for instance, the title and abstract of the following scientific article extracted from PubMed:

> Title: Viral FLIP blocks Caspase-8 driven apoptosis in the gut in vivo.
>
> Abstract: A strict cell death control in the intestinal epithelium is indispensable to maintain barrier integrity and homeostasis. In order to achieve a balance between cell proliferation and cell death, a tight regulation of Caspase-8, which is a key player in controlling apoptosis, is required. Caspase-8 activity is regulated by cellular FLIP proteins. These proteins are expressed in different isoforms (cFLIPlong and cFLIPshort) which determine cell death and survival. Interestingly, several viruses encode FLIP proteins, homologous to cFLIPshort, which are described to regulate Caspase-8 and the host cell death machinery. In the current study a mouse model was generated to show the impact of viral FLIP (vFLIP) from Kaposi's Sarcoma-associated Herpesvirus (KSHV)/ Human Herpesvirus-8 (HHV-8) on cell death regulation in the gut. Our results demonstrate that expression of vFlip in intestinal epithelial cells suppressed cFlip expression, but protected mice from lethality, tissue damage and excessive apoptotic cell death induced by genetic cFlip deletion. Finally, our model shows that vFlip expression decreases cFlip mediated Caspase-8 activation in intestinal epithelial cells. In conclusion, our data suggests that viral FLIP neutralizes and compensates for cellular FLIP, efficiently counteracting host cell death induction and facilitating further propagation in the host organism.

The system developed should categorise it as `basic research`, because, according to the annotation guidelines[4], the article aims to understand cell death regulation, in other words, cellular understanding of mechanisms.

3) Annotators Agreement

The dataset was annotated by three domain experts holding a PhD in different fields of biomedicine and developmental biology (referred to as A1, A2, and A3). To ensure the highest level of accuracy, the three annotators independently annotated the same publications. The agreement was calculated based on the averaged Cohen's κ due to the complexity of the task. Two control checkpoints were selected in order to explore general agreement and by pairs of categories, and also to explore the number of instances of each category.

Table 1 presents the results of the inter-annotator agreement, including the checkpoint and iteration, the final Cohen's κ score, and the agreement between pairs of categories. Overall, there was substantial inter-annotator agreement. In the first checkpoint, where the first 400 publications were annotated, a kappa of 0.69 was obtained. The agreement between the *basic-translational* and *clinical-public health* categories was, however, particularly low. After revising the guidelines and discussing cases of disagreement, a second annotation iteration was performed on the same 400 publications, resulting in substantial agreement for all pairs of categories. At checkpoint 2, after annotating the first 800 publications, the agreement raised to 0.78. However, due to the imbalance of publications between the basic and translational categories and the clinical research category, a subset of the remaining 300 publications was re-sampled based on a selection of journals more related to basic and clinical research.

The final agreement for the whole dataset was Kappa=0.75. Notably, the most challenging pairs are those adjacent as research types, such as *basic-translational*, *translational-clinical*, or *clinical-public health*. It is interesting that the agreement among the three annotators slightly decreases as more publications are annotated and time passes since the initial discussions. It is worth noting that the accuracy of manual text classification can be influenced by human factors such as fatigue and expertise [37].

Pairs of categories	Annotation	Re-annotation	Final dataset
Basic-Translational	0.223	0.614	0.595
Translational-Clinical	0.796	0.904	0.867
Clinical-Public Health	0.617	0.719	0.719
Basic-Clinical	0.887	0.953	0.946
Basic-Public Health	1.000	1.000	0.962
Translational-Public Health	0.986	1.000	0.994
ALL	0.690	0.782	0.748

TABLE 1. Average of Cohen's *k* inter-annotator agreement between the three annotators during the development of the annotation guidelines, during the annotation of the dataset, and for the final dataset.

[4]Available at the appendix.

C. BATRACIO STATISTICS

Our dataset consists of 1,248 publications in the biomedical domain, which are annotated across four categories. The distribution of publications across these categories is imbalanced, as is presented in Table 2, with 480 (38.46%) in the clinical research type, 349 (26.96%) in the basic research type, 220 (17.36%) in the translational research type, and 75 (6.01%) in the public health phase. Table 3 displays the general characteristics of the dataset.

Category	#Docs (% over the dataset)
Basic Research	349 (26.96%)
Translational Research	220 (17.36%)
Clinical Research	480 (38.46%)
Public Health	75 (6.01%)
Other	124 (9.94%)

TABLE 2. Class distribution in the final dataset.

	Final dataset
#categories	5
#documents	1,248
#sentences	13,669
avg. #sentences	10.95
#words (total)	334,456
#words (unique)	30,504
avg. #words	267.99

TABLE 3. Final dataset statistics. Documents contain the union of the title and the abstract.

IV. MATERIALS AND METHODS

The aim of this study is not only to present a dataset but also to explore the feasibility of automatically classifying biomedical scientific publications according to BATRACIO; i.e., according to four categories that represent different phases of biomedical research. To achieve this goal, we evaluated a wide range of state-of-the-art systems based on pre-trained language models using the Transformer architecture. In the following section, we describe the strategies we employed in our experiments.

A. PRE-TRAINED LANGUAGE MODELS

Recently, new approaches have emerged for adapting pre-training language models to improve their effectiveness in the biomedical domain. In this study, we investigated the use of several BERT-based biomedical models. With so many models available, selecting the best one for a specific task can be challenging as well as computationally intensive, as noted by [79].

The models used, and their main features, are described below:

- **BERT-base** [17] is a multi-layer bidirectional Transformer encoder. It is pre-trained on general domain corpus, BooksCorpus and English Wikipedia, for the objectives of Masked Language Modelling and Next Sentence Prediction. However, its general nature limits its effectiveness in specialized fields like biomedicine, where domain-specific terminology and contexts are prevalent. Consequently, BERT-base often underperforms compared to models specifically pre-trained on biomedical text, such as BioBERT and ClinicalBERT. We chose BERT-base architecture (12 layers, 768 hidden learning and 12 attention heads, summing a total number of 110M parameters), for the comparison with biomedical variants of BERT.
- **BioBERT** [19] is initialised from weights of general-domain BERT [17], and it is further pre-trained on PubMed abstracts and PubMedCentral full-text articles. It has demonstrated to excel in understanding biomedical terminology and context, demonstrating superior performance in named entity recognition (NER) and relation extraction. Despite its strengths, BioBERT's computationally intensive training requires significant resources, and it may not generalize well to sub-domains within biomedicine with unique language patterns not covered during pre-training.
- **SciBERT** [47] is a BERT-base model adapted to the specific-domain by pre-training on a random sample of mixed-domain 1.14M full text papers from Semantic Scholar, 18% in computer science and 82% in the biomedical domain. It includes new vocabulary in scientific domain which only overlaps 42% with the general-domain vocabulary in BERT and BioBERT. Its advantage lies in its ability to generalize across various scientific domains, making it more versatile for interdisciplinary applications. However, this broader focus might slightly compromise its performance in highly specialized biomedical tasks compared to models like BioBERT and PubMedBERT.
- **PubMedBERT** [20] is a domain-specific BERT-base model pre-trained from scratch on PubMed literature. Since it is focused narrowly on PubMed abstracts, it offers improved performance for tasks involving such texts, though this specialization may limit its applicability to other forms of biomedical literature.
- **LinkBioBERT** [80] is an extension of BioBERT that incorporates document-level relations by leveraging hyperlink structures within PubMed articles. It offers enhanced contextual understanding by capturing inter-document relationships. This model provides superior performance in tasks that require a deep understanding of context and cross-references, such as document classification and relation extraction. However, the complexity of training and the need for extensive hyperlink data can be limiting factors.
- Other language models, such as **BlueBERT** [50], **Specter** [57] and **OuBioBERT** [49] were also consider for the task. However, their results were not competitive enough, and therefore, we have decided not to present them.

The selection of specific pre-trained language models such as ClinicalBERT, BioBERT, PubMedBERT, SciBERT, and LinkBERT is motivated by their demonstrated performance and optimization for biomedical natural language processing (NLP) tasks. These models have been specifically trained on large corpora of biomedical and scientific texts, allowing them to capture domain-specific language patterns, terminology, and contextual nuances that are essential for accurate text classification in the biomedical field. General-purpose models like BERT-base, while robust and versatile, often fall short in specialized domains due to their lack of exposure to domain-specific data during pre-training. This specialization is particularly critical in biomedical text classification, where understanding complex medical terms and relationships significantly enhances performance.

For instance, BioBERT and PubMedBERT are trained on PubMed abstracts and PMC full-text articles, which ensures they are adept at handling biomedical terminology and concepts. ClinicalBERT, derived from BERT and fine-tuned on clinical notes, offers an edge in processing clinical narratives. SciBERT, with its training on a broad corpus of scientific literature, balances versatility with domain relevance, making it suitable for a wide range of biomedical and scientific texts. LinkBERT further leverages hyperlink structures within documents to enhance contextual understanding, proving beneficial for tasks requiring a deep comprehension of document relationships.

B. TEXT CLASSIFICATION WITH PRE-TRAINED LANGUAGE MODELS

In recent years, pre-trained language models based on Transformers have emerged as a superior option for text classification. Their ability to learn global language representations from massive datasets and adapt to downstream tasks by simply adding a final layer, such as a linear classifier for text classification, and fine-tuning the model's weights has made them particularly appealing. BERT, the first pre-trained, fine-tuning-based, and bidirectional language model, has achieved state-of-the-art results in several NLP tasks. Contextual pre-trained language models are an improvement over previous models since they do not have to be trained from scratch, reducing computational costs and improving performance.

We investigated the two primary approaches for adapting BERT models to a text classification task [81]. The first approach, fine-tuning, was proposed in the original BERT paper [17] and has also been used in other studies [45], [82], [83]. Fine-tuning involves adjusting and updating the pre-trained weights of the model using back-propagation to minimize the loss function and obtain the desired output. The second approach is the feature-based approach, where all parameters of the model are frozen, and only a linear classifier is trained on top of the model, as suggested in [57], [58]. This approach can be useful for avoiding catastrophic forgetting [45], [84], particularly when the dataset is small. However, since the feature-based approach yielded poorer performance, we only show the results for the fine-tuning strategy.

For our experiments, we use the Hugging Face Transformers library [5]. Following the optimal hyperparameters proposed in the original BERT paper [17], we provide results using a learning rate (Adam) of 2e-5, batch size of 16, and 4 epochs.

C. LOSS FUNCTION

The primary objective of neural networks is to minimize the difference between the predicted output and the expected output by comparing the predicted distribution of results with the true distribution. This difference, also known as the error, is calculated using a cost or error function. The standard cost function for text classification tasks is the *cross entropy loss*[6]. However, this cost function does not take into account class imbalance. Cost weighting is an important alternative to data augmentation for unbalanced classes [85]. It involves increasing the cost associated with obtaining an erroneous low-frequency class label.

In our experiments, we consider the following modification in the loss function:

- **Loss**: cross-entropy loss without weighting categories.
- **Weighted loss**: as [85] propose, we increase the cost of incorrectly labelling the class with lower number of samples by weighting the cross-entropy loss function.

We use the formulation given by [85] and that we replicate here for completeness. Given an array where the *jth* element represents the models prediction for class j, the cross-entropy loss for a single prediction x is given by Equation 1.

$$\begin{aligned} loss(x, class) &= -log\left(\frac{exp(x[class])}{\sum_j exp(x[j])}\right) \\ &= -x[class] + log\left(\sum_j exp(x[j])\right) \end{aligned} \tag{1}$$

As in [85], the cross-entropy loss given in Equation 1 is modified to accommodate an array weight, the *ith* element of which represents the weight of the *ith* class, as described in Equation 2.

$$\begin{aligned} loss(x, class) &= weight[class]\Theta \\ where, \Theta &= -x[class] + log\left(\sum_j exp(x[j])\right) \end{aligned} \tag{2}$$

D. TEXT CLEANING AND PRE-PROCESSING

One of the strengths of BERT is its ability to learn directly from unstructured text; however, in the biomedical domain, learning from unstructured language remains a challenge [58], [83], [86]. Acronyms are particularly prevalent in science and even more so in biomedical publications, as authors often seek to abbreviate long names for diseases, bacteria, and

[5]Hugging Face: https://huggingface.co/

[6]https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html

chemicals. According to [87], acronyms were used in more than 24 million scientific article titles and 18 million scientific articles published between 1950 and 2019. They reported that 19% of titles and 73% of abstracts contain acronyms. Of the more than one million unique acronyms in their data, 0.2% appeared regularly, and most acronyms, 79%, appeared fewer than 10 times [88].

According to [86], pre-trained language models may struggle with rare words, and datasets with a high number of unique words can pose a challenge. In the biomedical domain, researchers have proposed various pre-processing techniques to improve the performance of pre-trained language models. For instance, [58] suggest removing all numbers, which are frequent in scientific studies but do not typically provide relevant information for BioNLP tasks. Similarly, [83] note that preprocessing can enhance task performance for similar reasons. In [89], researchers remove punctuation and abbreviations from the text.

We perform basic processing following the recommendations in [58], [83], [86], [89]:

- **Acronym resolution:** we use the `Abbreviation Detector` component in `scispacy`[7], which implements a simple algorithm for identifying abbreviations in biomedical text [90], and after, the abbreviations are replaced by their expanded name.
- **Removal of numbers and special characters:** we remove all numbers in abstracts, because they do not add meaning about the categories of interest, and sometimes correspond to results or references; and special characters, because scientific literature can include formulas and rare characters, which can reduce performance.

V. RESULTS

The objective of this section is to evaluate the appropriateness of state-of-the-art classification techniques in addressing the new problem proposed in this paper: the automated classification of biomedical literature based on the research types. To demonstrate the effectiveness of the defined fine-tuning, cleaning, and loss function modification strategies, we evaluate various models using the BATRACIO dataset. We assess the behavior of state-of-the-art machine learning techniques using precision, recall, F-measure, and accuracy, as defined in the following equations:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{F-measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

[7]https://github.com/allenai/scispacy

where TP (True Positives) represents the number of instances correctly predicted as positive (the model correctly identifies positive cases); TN (True Negatives) represents the number of instances correctly predicted as negative (the model correctly identifies negative cases); FP (False Positives) represents the number of instances incorrectly predicted as positive (the model incorrectly identifies negative cases as positive); and FN (False Negatives) represents the number of instances incorrectly predicted as negative (the model fails to identify positive cases).

To assess the systems' performance with the dataset, we conduct 10-fold cross-validation using a stratified approach to ensure that all categories have representative samples in all partitions, as random partitioning fails to provide such representation. The validation set is created through a 0.1 split on the train, and the averaged metric values across all k-folds are used as the evaluation criterion for k-fold cross-validation. Evaluating imbalanced datasets is challenging because models often predict the majority class with high accuracy, resulting in misleading results. Macro-averaged F-measure is a more appropriate evaluation metric as it treats the performance of each class equally [85].

A. PRE-TRAINED LANGUAGE MODELS

Our first experiments compare the performance of different language models, which are presented in Section IV-A. As shown in Table 5, domain-specific pre-trained language models produced the best results. This was not surprising given that domain-specific models have been proven successful in highly specialized fields. Among the pre-trained models, PubMedBERT was the best performer (F1=0.82), followed by LinkBioBERT (F1=0.81) and SciBERT (F1=0.81), both of which were trained from scratch on biomedical literature. BioBERT (F1=0.79), which shares the vocabulary of BERT-base but is further pre-trained on biomedical documents, outperformed BERT (F1=0.77) and yielded competitive results. Despite not being specifically designed for the biomedical domain, BERT-base's performance was also significant and produced better results than other domain-specific models such as OuBioBERT and BlueBERT. A possible reason for this, as suggested by [51], is that BERT-base was pre-trained on Wikipedia, which includes WikiMed, a collection of Wikipedia medical and scientific pages. Therefore, BERT-base has some domain knowledge, unlike general-domain Word2vec [91], which is pre-trained on Google News.

B. TEXT CLEANING AND PRE-PROCESSING

We next conducted experiments to evaluate the effectiveness of performing text pre-processing as a previous step to the classification algorithm. In Table 4, we present the F-1 and accuracy scores of various text cleaning techniques, including acronym resolution and the removal of numbers and special characters, applied only to the articles' abstracts. The results show that for systems trained on general-domain vocabulary, such as BERT-base and BioBERT, acronym resolution significantly improves the scores compared to the raw abstract.

However, for PubMedBERT and SciBERT, acronym resolution does not appear to improve the scores, and the best results are obtained by removing numbers and special characters.

C. DATA INPUT: TITLE AND/OR ABSTRACT

Our third set of experiments aimed to examine the impact of different data inputs on the results (see Table 5). To achieve this, we tested various combinations of titles and abstracts. Our results indicate that PubMedBERT significantly outperforms the other systems when given only the abstract or a combination of the title and abstract (F1=0.85). However, when given only the title, the performance differences between the systems are minimal, as expected, in contrast to when only the abstract is provided. This suggests that, for some articles, the title alone may not contain enough information to distinguish the research type, regardless of the language model utilized, and confirm that the minimum working unit is the pair "title-abstract".

D. CLASS IMBALANCE AND LOSS FUNCTION

One of the main challenges when working with small and imbalanced datasets is that systems tend to over-learn the most frequent classes. As shown in Table 6, F-1 and accuracy are reported by category (i.e., research type). The best overall model, PubMedBERT, is also the most effective in predicting categories with fewer instances, such as *Public health* and *Translational research*.

To address the challenges posed by imbalanced category distribution and semantic adjacency, we explored modifying the loss function during fine-tuning to improve task adaptation. Table 7 displays various attempts at modifying the loss function. To mitigate category imbalance, we incorporated a vector of weights into the loss function during training, assigning a weight to each category based on its inverse frequency in the dataset. This approach should re-scale the weight assigned to each class, prioritizing categories with fewer samples. For three of the models, this improved F-1 scores for the *Public Health* category and overall F-1 scores. To reinforce adjacency, we also added a neighboring loss term to the loss function. The combination of weighted loss and neighboring loss proved effective for SciBERT and BioBERT, successfully addressing the dataset's challenges. However, the best-performing PubMedBERT system did not show improvement, with the best configuration being without any loss function modification.

VI. DISCUSSION

The primary goal of this study was to evaluate the effectiveness of state-of-the-art classification techniques in the automated classification of biomedical literature by research type, utilizing the BATRACIO dataset. Our results demonstrate significant improvements through the implementation of domain-specific pre-trained language models, text cleaning and pre-processing, varying data inputs, and tailored loss functions. In the following subsections, these findings are summarized and contextualized with existing literature to understand the advancements and remaining challenges in the field.

A. DOMAIN-SPECIFIC PRE-TRAINED MODELS

The best performance of domain-specific models, particularly PubMedBERT, SciBERT, and BioBERT, aligns with previous research emphasizing the benefits of domain adaptation for specialized tasks. Alsentzer et al. [92] and Lee et al. [93] similarly highlighted the efficacy of models pre-trained on biomedical corpora, such as BioBERT, in improving classification and named entity recognition tasks in the biomedical field. Our results corroborate these findings, showing that models like PubMedBERT, which are specifically pre-trained on biomedical literature, achieve higher F1 scores compared to general models like BERT-base. This is consistent with the work of Gu et al. [94], who found that PubMedBERT outperformed other models in various biomedical NLP tasks.

B. TEXT CLEANING AND PRE-PROCESSING

Our study found that text pre-processing techniques, such as acronym resolution and the removal of numbers and special characters, significantly improve model performance for general-domain models like BERT-base. This finding is in line with previous studies by Kim et al. [95] and Yan et al. [96], which reported that preprocessing steps could enhance the performance of text classification tasks. However, for domain-specific models like PubMedBERT and SciBERT, the benefit of these preprocessing steps was minimal. This suggests that domain-specific models are robust enough to handle noisy input data, a conclusion supported by Lee et al. [93].

C. DATA INPUT: TITLE AND/OR ABSTRACT

Our experiments demonstrate that the combination of title and abstract significantly enhances classification performance, particularly for models like PubMedBERT. This finding aligns with the work of Zhang et al. [97], who found that using both titles and abstracts improves the accuracy of biomedical document classification compared to using either alone. The minimal performance differences when using only titles suggest that abstracts contain crucial context needed for accurate classification, corroborating the conclusions drawn by Lu et al. [98].

D. CLASS IMBALANCE AND LOSS FUNCTION

Addressing class imbalance remains a critical challenge in biomedical literature classification. Our experiments with weighted loss functions and balanced training sets showed improvements for some models but did not universally enhance performance across all metrics. This observation is consistent with the findings of Johnson and Khoshgoftaar [99], who noted that class imbalance can significantly impact machine learning model performance and that weighted loss functions can mitigate but not completely resolve these issues.

	Raw		Acronym		Num.+SC		Acr.+Num.+SC	
Model	**F1±std**	**Acc.±std**	**F1±std**	**Acc.±std**	**F1±std**	**Acc.±std**	**F1±std**	**Acc.±std**
PubMedBERT	82.45±4.55	86.86±3.10	82.53±5.39	86.61±3.76	82.43±5.14	86.45±3.38	83.56±5.26	87.42±3.28
sciBERT	81.09±3.77	85.50±2.57	82.29±4.10	86.29±2.77	83.04±2.76	86.86±1.91	82.64±3.77	86.54±2.67
BioBERT	79.75±5.35	84.53±3.66	81.04±3.81	85.82±2.71	82.53±4.30	86.54±2.96	82.47±4.46	86.38±2.74
BERT-base	77.86±4.08	83.65±3.29	78.43±4.31	84.05±3.14	78.76±4.74	84.53±3.47	79.23±4.38	84.69±3.09
LinkBioBERT	81.59±3.86	86.29±2.27	82.21±4.91	86.45±3.37	81.30±4.12	86.13±2.46	82.13±4.07	86.53±2.63

TABLE 4. Comparison of domain specific pre-trained language models using different text cleaning strategies. Trained by fine-tuning, fixing the following hyper-parameter configuration: 4 epochs, learning rate of 2e-5, and batch size of 16. Input data is title+abstract. Results are reported as macro-average between 10-folds. Column description: Acronym = resolving acronyms / Num.+SC = removing numbers and special characters / Acr.+Num.+SC = resolving acronyms, and removing numbers and special characters

	Title		Abstract		Title+Abstract	
Model	**F1±std**	**Acc.±std**	**F1±std**	**Acc.±std**	**F1±std**	**Acc.±std**
PubMedBERT	79.93±3.40	83.57±2.29	82.45±4.55	86.86±3.10	82.22±4.56	86.86±3.26
sciBERT	78.48±6.05	83.17±3.40	81.09±3.77	85.50±2.57	81.59±3.57	85.73±2.43
BioBERT	78.98±3.91	83.33±2.41	79.75±5.35	84.53±3.66	82.12±3.86	86.08±2.82
BERT-base	69.95±6.39	76.84±3.72	77.86±4.08	83.65±3.29	79.25±3.63	84.44±2.85
LinkBioBERT	79.49±4.88	84.21±2.47	81.59±3.86	86.29±2.27	81.76±5.78	86.21±3.55

TABLE 5. Comparison of domain specific pre-trained language models trained on different textual sections of the scientific publications. Trained by fine-tuning, fixing the following hyper-parameter configuration: 4 epochs, learning rate of 2e-5, and batch size of 16. Results are reported as macro-average between 10-folds.

	Basic	Translational	Clinical	Public health	Other
Model	**F1±std**	**F1±std**	**F1±std**	**F1±std**	**F1±std**
PubMedBERT	86.27±4.27	82.07±6.44	93.79±2.04	72.72±9.36	76.24±11.86
sciBERT	84.14±3.07	79.88±5.88	93.24±1.70	74.16±8.89	76.53±10.35
BioBERT	83.83±4.96	80.37±4.70	94.02±1.56	76.43±8.81	75.92±9.27
BERT-base	84.39±3.95	76.49±6.86	92.31±1.79	68.84±9.47	74.25±7.77
LinkBioBERT	85.45±4.98	81.13±5.58	92.71±2.12	73.27±12.06	76.23±14.72

TABLE 6. Comparison of domain specific pre-trained language models by category. Trained by fine-tuning, fixing the following hyper-pameter configuration: 4 epochs, learning rate of 2e-5, and batch size of 16. Input data is title+abstract. Results are reported as average between 10-folds.

	Unbalanced		Balanced		Weighted loss	
Model	**F1±std**	**Acc.±std**	**F1±std**	**Acc.±std**	**F1±std**	**Acc.±std**
PubMedBERT	82.22±4.56	86.86±3.26	78.91±2.79	82.53±1.89	82.89±5.40	86.61±3.28
sciBERT	81.59±3.57	85.73±2.43	78.32±3.22	82.37±2.42	82.84±3.63	86.38±2.28
BioBERT	82.12±3.86	86.08±2.82	73.57±3.88	77.65±2.93	82.85±4.26	86.54±2.52
BERT-base	79.25±3.63	84.44±2.85	68.39±4.08	73.96±3.60	79.77±3.47	84.53±2.84
LinkBioBERT	81.76±5.78	86.21±3.55	73.30±4.04	76.84±3.15	83.22±5.00	86.77±3.13

TABLE 7. Comparison of domain specific pre-trained language models trained with weighted loss function and balancing training size to smaller class. Trained by fine-tuning, fixing the following hyper-parameter configuration: 4 epochs, learning rate of 2e-5, and batch size of 16. Input data is title+abstract. Results are reported as macro-average between 10-folds.

E. MODEL LIMITATIONS AND POTENTIAL AREAS FOR IMPROVEMENT

Despite the better performance of specialized models such as BioBERT, PubMedBERT, SciBERT, and LinkBERT in biomedical text classification, several limitations persist. One significant limitation is the reliance on large volumes of domain-specific training data. While these models are pre-trained on extensive biomedical corpora, their performance may still be constrained by the quality and comprehensiveness of the training data. For instance, emerging medical terminology and newly discovered biomedical concepts may not be adequately represented, potentially impacting the model's ability to handle the latest developments in the field. Moreover, biases inherent in the training data, such as those related to demographic disparities or institutional practices, can be inadvertently learned and propagated by these models, leading to biased predictions and outcomes.

Another limitation is the computational cost associated with training and fine-tuning these models. The need for substantial computational resources can be prohibitive, particularly for smaller research teams or organizations with limited access to high-performance computing infrastructure. This constraint not only limits the ability to experiment with extensive hyperparameter tuning but also restricts the feasibility of deploying these models in resource-constrained environments. Furthermore, while fine-tuning these models for specific tasks improves performance, it also introduces the risk of overfitting, particularly when working with lim-

ited labeled data. Ensuring generalizability across diverse biomedical text sources remains a challenging task.

Potential areas for improvement include enhancing the adaptability of these models to new and evolving biomedical information through continuous learning mechanisms. Implementing techniques such as transfer learning and domain adaptation can help models stay current with the latest biomedical research. Additionally, incorporating techniques to mitigate bias, such as debiasing algorithms and balanced training datasets, can improve the fairness and reliability of the models. Reducing the computational burden through model compression techniques and efficient training paradigms, such as knowledge distillation and quantization, can make these models more accessible and deployable in real-world settings. Addressing these limitations and exploring these avenues for improvement will be crucial in advancing the efficacy and applicability of biomedical text classification models.

VII. CONCLUSIONS

In this work, we introduce BATRACIO (*BAsic-TRAnslational-Clinical research types classification in bIOmedical publications*), a novel text classification task in the biomedical domain. To the best of our knowledge, no previous work before has addressed the problem of automatically classifying scientific literature according to biomedical research types. The task seeks to help policymakers or funding agencies to better understand what research activities were carried out, mapping stakeholders and their research competencies, and hence to better allocate the resources, by classifying the scientific outputs of specific funding instruments or scientific publications, according to research type.

Since the problem proposed is new, the work described in this research includes the creation and annotation of a dataset of 1,248 scientific publications in the biomedical domain extracted from PubMed, categorized by research type with the following categories: *basic research*, *translational research*, *clinical research*, and *public health*. Designing a new task is a big challenge, especially in a complex domain as biomedicine. For this reason, we have involved domain experts for designing the task and for annotating the dataset. However, the creation of the dataset has been costly and has required the organisation of several workshops for discussion with experts, data extraction and analysis in depth to reduce biases and to get publications in the domain of interest.

We have also explored whether the problem can be addressed automatically, providing baseline results based on fine-tuning pre-trained language models and domain-specific models for biomedical domain, together with some the exploration of some strategies for improving the performance of the systems. Our experiments showed that using domain-specific pre-trained model (particularly those trained on scientific biomedical papers, such as PubMedBERT or LinkBioBERT) provides better results that using general-domain models and up to 0.83 F1, which is a very good performance for a 5-classes classification task. We have also found that, as identified in the literature, acronym resolution can improve the performance of pre-trained language models in the biomedical domain, although the improvement for models that are pre-trained in the biomedical domain is not much as they already incorporate information about acronyms. Removing numbers and special characters, which can refer to results and statistics about the samples of study, can help to reduce noise in abstracts.

We have shown that using as the input of the classifier the combination of both the title and abstract could generally provide better results than using the title or the abstract alone. This is consistent with the fact that the pair title-abstract was identified by our domain experts as the minimum unit of annotation for our particular classification task.

Nevertheless, the main specific challenges of our dataset are the class imbalance and that categories are not mutually independent, they shape a semantic *value chain* and have semantic relations of adjacency between them. This was not a main goal of the project, but we have also explored whether slight modifications in the loss function can deal with imbalanced categories. Although the results of these experiments are partially satisfactory, they point to future lines of research.

Other lines for future work include the exploration of the use of other section in the article which could be relevant for the task, such as introduction, or materials and methods. We suggest future work on exploring different approaches for incorporating semantic relation between categories. Furthermore, we also suggest the test in real collections of publications and research projects.

APPENDIX. ANNOTATION GUIDELINES

The task proposes the identification of value-chain research phase in scientific outputs, classifying publications and research projects among the research phase of the records, choosing between: **(1) basic research, (2) translational research, (3) clinical research or (4) public health**. Those documents that annotators discard during the annotation process as not being properly biomedical research or not fitting into any of the four categories proposed will be labelled with category the **(5) other**.

According to this, each record must be categorized in one of the following categories:

1) Basic research (also called fundamental research)

This focuses on discoveries and knowledge, driven by hypotheses that advance the understanding of the unknown; it builds new knowledge; in biomedical sciences it uses cells and model organisms and very rarely human subjects or human biological material. It involves scientific exploration that can reveal fundamental mechanisms of biology, disease or behaviour. Every stage of the translational research spectrum builds upon and informs basic research. It studies the core building blocks of life (such as: DNA, cells, proteins, molecules, etc.) in order to answer fundamental questions about their structures and how they work. For example, oncologists now know that mutations in DNA enable the unchecked

growth of cells in cancer. A scientist conducting basic research might ask: How does DNA work in a healthy cell? How do mutations occur? Where along the DNA sequence do mutations happen? And why?
The following topics should be considered part of basic research:

- Tissue, Cellular & Molecular basis of disease
- Tissue, Cellular & Molecular understanding of mechanisms
- Use of Animal models - zebrafish, rats, human cells, fly, c.elegans, mice, rabbit, guinea pig
- Development of techniques - protein, chemistry, molecular, cellular

Some examples of publications in the category:

- *GPR40 full agonism exerts feeding suppression and weight loss through afferent vagal nerve.*
- *The Functional Mammalian CRES (Cystatin-Related Epididymal Spermatogenic) Amyloid is Antiparallel β-Sheet Rich and Forms a Metastable Oligomer During Assembly.*
- *Viral FLIP blocks Caspase-8 driven apoptosis in the gut in vivo.*
- *Apelin enhances the osteogenic differentiation of human bone marrow mesenchymal stem cells partly through Wnt/β-catenin signaling pathway.*
- *Improved yellow-green split fluorescent proteins for protein labeling and signal amplification.*

The following topics should be considered part of basic research:

- Tissue, Cellular & Molecular basis of disease
- Tissue, Cellular & Molecular understanding of mechanisms
- Use of Animal models- zebrafish, rats, human cells, fly, c.elegans, mice, rabbit, guinea pig
- Development of techniques- protein, chemistry, molecular, cellular

Some examples of publications in the category:

- *GPR40 full agonism exerts feeding suppression and weight loss through afferent vagal nerve.*
- *The Functional Mammalian CRES (Cystatin-Related Epididymal Spermatogenic) Amyloid is Antiparallel β-Sheet Rich and Forms a Metastable Oligomer During Assembly.*
- *Viral FLIP blocks Caspase-8 driven apoptosis in the gut in vivo.*
- *Apelin enhances the osteogenic differentiation of human bone marrow mesenchymal stem cells partly through Wnt/β-catenin signaling pathway.*
- *Improved yellow-green split fluorescent proteins for protein labeling and signal amplification.*

2) Translational research (also called pre-clinical research)

This focuses on translating the discoveries into usability in the clinic, uses large scale testing and both animal models and human biological material. There is a focus on applicability. It connects the basic science of disease with human medicine. During this stage, scientists develop model interventions to further understand the basis of a disease or disorder and find ways to treat it. Testing is carried out using cell or animal models of disease; samples of human or animal tissues; or computer-assisted simulations of drug, device or diagnostic interactions within living systems. For this area of research the end point is the production of a promising new treatment that can be used clinically or commercialized ("brought to market"). This enterprise is vital, and has been characterized as follows: "effective translation of the new knowledge, mechanisms, and techniques generated by advances in basic science research into new approaches for prevention, diagnosis, and treatment of disease is essential for improving health."
The following topics should be considered part of translational research:

- Study of processes or diseases with the intent to treat
- Drug and vehicle development (since they have a therapeutic target)
- Pre-clinical models (even advanced ones like sheep and pigs)
- With patients samples only as proof of concept, as in tumour samples/biobank usage which is not central to the paper
- With patients samples to establish research pre-clinical models (like in cell lines)

Some examples of publications in the category:

- *Characterization of a porcine model of atrial arrhythmogenicity in the context of ischaemic heart failure.*
- *Assessment of an ultrasound-guided technique for catheterization of the caudal thoracic paravertebral space in dog cadavers.*
- *Nerve Repair and Orthodromic and Antidromic Nerve Grafts: An Experimental Comparative Study in Rabbit.*
- *Murine SIGNR1 (CD209b) Contributes to the Clearance of Uropathogenic Escherichia coli During Urinary Tract Infections.*
- *Notopterol-induced apoptosis and differentiation in human acute myeloid leukemia HL-60 cells.*

3) Clinical research

This searches by testing a specific treatment or procedure, drug, diagnostic or any technology on patients, focusing not only on the biological mechanisms (if applicable) but also on issues of safety, delivery and protocols for implementation. This is the stage of research where clinical trials tend to take place. It includes studies to better understand a disease in humans and relate this knowledge to findings in cell or animal models, testing and refinement of new technologies in people, testing of interventions for safety and effectiveness in those with or without the disease, behavioural and observational studies, and outcomes and health services research. The goal of many clinical trials is to obtain data to support regulatory approval for an intervention. It explores whether new treatments, medications and diagnostic techniques are

safe and effective in patients. Physicians administer these to patients in rigorously controlled clinical trials, so that they can accurately and precisely monitor patients' progress and evaluate the treatment's efficacy, or measurable benefit.
The following topics should be considered part of clinical research:

- Clinical trials
- Research regarding patients treatment protocol
- Research implicating patients directly
- Research with patient samples as central feature (genetics of disease, biomarkers, prognostic markers,...)
- Diagnostic of disease
- Classic Epidemiology- cohorts
- Psychiatry (Mental disorders)
- Healthcare standards and guidelines

Some examples of publications in the category:

- *Infection with multiple hepatitis C virus genotypes detected using commercial tests should be confirmed using next generation sequencing.*
- *Supraclavicular versus infraclavicular approach in inserting totally implantable central venous access for cancer therapy: A comparative retrospective study.*
- *The effect of apolipoprotein E polymorphism on serum metabolome - a population-based 10-year follow-up study.*
- *Functional variations of the TLR4 gene in association with chronic obstructive pulmonary disease and pulmonary tuberculosis.*
- *Comparing patterns of volatile organic compounds exhaled in breath after consumption of two infant formulae with a different lipid structure: a randomized trial.*

4) Public health

This is defined as "the art and science of preventing disease, prolonging life and promoting health through the organized efforts of society" rechel2014. Activities to strengthen public health capacities and service aim to provide conditions under which people can stay healthy, improve their health and well-being, or prevent the deterioration of their health. Public health focuses on the entire spectrum of health and well-being, not only the eradication of particular diseases. Many activities are targeted at populations such as health campaigns. Public health services also include the provision of personal services to individual persons, such as vaccinations, behavioural counselling, or health advice.
The following topics should be considered part of public health:

- Cultural/socioeconomic impact on Health
- Health Policy
- Global Health
- Population Health
- Assessment of diseases prevalence in population
- Assessment and discovery of predictive measures
- Health policy e.g. interaction with hospital management and /or economic systems
- Usage of other non-clinical data

Some examples of publications in the category:

- *Post-elimination surveillance in formerly onchocerciasis endemic focus in Southern Mexico.*
- *Association of intestinal colonization of ESBL-producing Enterobacteriaceae in poultry slaughterhouse workers with occupational exposure-A German pilot study.*
- *Use of non-HIV medication among people living with HIV and receiving antiretroviral treatment in Côte d'Ivoire, West Africa: A cross-sectional study.*
- *How did the use of psychotropic drugs change during the Great Recession in Portugal? A follow-up to the National Mental Health Survey*
- *Prevalence and social burden of active chronic low back pain in the adult Portuguese population: results from a national survey*

ACKNOWLEDGMENT

This research is funded by FAIRTRANSNLP-DIAGNOSIS: Measuring and quantifying bias and fairness in NLP systems, grant PID2021-124361OB-C32, funded by MCIN/AEI/10.13039/5011 and by ERDF, EU A way of making Europe. This work has been also funded by the Ministry of Universities and the European Union through the EuropeaNextGenerationUE funds and the "Plan de Recuperación, Transformación y Resiliencia". Supported by the Industrial Doctorates Plan of the Department of Research and Universities of the Generalitat de Catalunya.

REFERENCES

[1] E. Fuster, F. Massucci, and M. Matusiak, "Identifying specialisation domains beyond taxonomies: mapping scientific and technological domains of specialisation via semantic analyses." in *Quantitative Methods for Place-Based Innovation Policy*, R. Capello, A. Kleibrink, and M. Matusiak, Eds., 01 2020, p. 195–234.
[2] D. F. . M. C. Seixas, B.V., "Practices of decision making in priority setting and resource allocation: a scoping review and narrative synthesis of existing frameworks," *Health Economics Review*, vol. 11, no. 2, pp. 1–1, 2021.
[3] M. Mazzucato, *Mission Economy: a moonshot guide to changing capitalism*. Penguin Books, 2021.
[4] E. Zerhouni, "The nih roadmap," *Science*, vol. 302, no. 5642, pp. 63–72, 2003. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1091867
[5] D. Butler, "Translational research: crossing the valley of death." in *Nature*, 2008.
[6] K. Soderquest and G. M. Lord, "Strategies for translational research in the united kingdom," *Science Translational Medicine*, vol. 2, no. 53, pp. 53cm28–53cm28, 2010. [Online]. Available: https://www.science.org/doi/abs/10.1126/scitranslmed.3001129
[7] E. Vignola-Gagné, E. Rantanen, D. Lehner, and B. Hüsing, "Translational research policies: disruptions and continuities in biomedical innovation systems in austria, finland and germany," *Journal of community genetics*, vol. 4, pp. 189–201, 2013.
[8] S. Kim, D. Park, Y. Choi, K. Lee, B. Kim, M. Jeon, J. Kim, A. C. Tan, and J. Kang, "A pilot study of biomedical text comprehension using an attention-based deep neural reader: Design and experimental analysis," *JMIR Medical Informatics*, vol. 6, no. 1, p. e2, 2018.
[9] A. L. Van der Laan and M. Boenink, "Beyond bench and bedside: disentangling the concept of translational research," *Health care analysis*, vol. 23, pp. 32–49, 2015.
[10] V. Kieuvongngam, B. Tan, and Y. Niu, "Automatic text summarization of covid-19 medical research articles using bert and gpt-2," *Frontiers in Biomedical Technologies*, vol. 7, no. 4, pp. 236–248, 2020.

[11] Q. Wei, Z. Ji, Y. Si, J. Du, J. Wang, F. Tiryaki, S. Wu, C. Tao, K. Roberts, and H. Xu, "Relation extraction from clinical narratives using pre-trained language models," *Proceedings of the AMIA Annual Symposium*, vol. 2019, pp. 1236–1245, 03 2020.

[12] V. Tran, V.-H. Tran, P. Nguyen, C. Nguyen, K. Satoh, Y. Matsumoto, and M. Nguyen, "CovRelex: A COVID-19 retrieval system with relation extraction," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2021, pp. 24–31.

[13] M. Sarrouti and S. O. E. Alaoui, "Sembionlqa: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions," *Artificial intelligence in medicine*, vol. 102, p. 101767, 2020.

[14] P. Resnik, K. E. Goodman, and M. Moran, "Developing a curated topic model for COVID-19 medical research literature," in *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, 2020.

[15] Y. Su, H. Xiang, H. Xie, Y. Yu, S. Dong, Z. Yang, and N. Zhao, "Application of bert to enable gene classification based on clinical evidence," *BioMed Research International*, vol. 2020, pp. 1–13, 10 2020.

[16] R. You, Y. Liu, H. Mamitsuka, and S. Zhu, "Bertmesh: Deep contextual representation learning for large-scale high-performance mesh indexing with full text," *Bioinformatics (Oxford, England)*, vol. 37, no. 5, pp. 684—-692, 2020.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171—4186.

[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[19] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[20] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, p. 1–23, Jan 2022.

[21] J. Peng, M. Zhao, J. Havrilla, C. Liu, C. Weng, W. Guthrie, R. Schultz, K. Wang, and Y. Zhou, "Natural language processing (nlp) tools in extracting biomedical concepts from research articles: a case study on autism spectrum disorder," *BMC Medical Informatics and Decision Making*, vol. 20, 12 2020.

[22] A. Duque, R. Stevenson, J. Martinez-Romo, and L. Araujo, "Co-occurrence graphs for word sense disambiguation in the biomedical domain," *Artificial Intelligence in Medicine*, vol. 87, 03 2018.

[23] K. Cohen, *Biomedical Natural Language Processing and Text Mining*, 12 2014, pp. 141–177.

[24] L. Laar, T. Kruif, L. Waltman, I. Meijer, A. Gupta, and N. Hagenaars, "Improving the evaluation of worldwide biomedical research output: Classification method and standardised bibliometric indicators by disease," *BMJ Open*, vol. 8, p. e020818, 06 2018.

[25] N. Durán-Silva, "Using pre-trained language models to automatically identify research phases in biomedical publications," 2022.

[26] C. Friedman, P. Kra, and A. Rzhetsky, "Two biomedical sublanguages: a description based on the theories of zellig harris," *Journal of biomedical informatics*, vol. 35 4, pp. 222–35, 2002.

[27] W. Chapman and K. Cohen, "Current issues in biomedical text mining and natural language processing," *Journal of biomedical informatics*, vol. 42, pp. 757–9, 10 2009.

[28] C.-C. Huang and Z. Lu, "Community challenges in biomedical text mining over 10 years: success, failure and the future," *Briefings in Bioinformatics*, vol. 17, no. 1, pp. 132–144, 05 2015. [Online]. Available: https://doi.org/10.1093/bib/bbv024

[29] S. Wu, K. Roberts, S. Datta, J. Du, Z. Ji, Y. Si, S. Soni, Q. Wang, Q. Wei, Y. Xiang, B. Zhao, and H. Xu, "Deep learning in clinical natural language processing: A methodical review," *Journal of the American Medical Informatics Association*, vol. 27, 12 2019.

[30] B. Percha, "Modern clinical text mining: A guide and review," *Annu Rev Biomed Data Sci.*, vol. 4, pp. 165–187, 2021.

[31] K. Canese and S. Weis, "Pubmed: The bibliographic database," 2013.

[32] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. L. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. E. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, pp. 25–29, 2000.

[33] C. E. Lipscomb, "Medical subject headings (mesh)." *Bulletin of the Medical Library Association*, vol. 88 3, pp. 265–6, 2000.

[34] D. Demner-Fushman, K. B. Cohen, S. Ananiadou, and J. Tsujii, Eds., *Proceedings of the 20th Workshop on Biomedical Language Processing*. Association for Computational Linguistics, Jun. 2021. [Online]. Available: https://aclanthology.org/2021.bionlp-1.0

[35] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, pp. 1–47, 04 2001.

[36] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning based text classification: A comprehensive review," 2021.

[37] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, "A survey on text classification: From shallow to deep learning," 2020.

[38] A. Mascio, Z. Kraljevic, D. Bean, R. Dobson, R. Stewart, R. Bendayan, and A. Roberts, "Comparative analysis of text classification approaches in electronic health records," in *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, Jul. 2020, pp. 86–94.

[39] G. Mujtaba, L. Shuib, N. Idris, W. L. Hoo, R. G. Raj, K. Khowaja, K. Shaikh, and H. F. Nweke, "Clinical text classification research trends: Systematic literature review and open issues," *Expert Systems with Applications*, vol. 116, pp. 494–520, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417418306110

[40] J. Thanaki, *Python natural language processing*. Packt Publishing Ltd, 2017.

[41] P. Patel, D. Davey, V. Panchal, and P. Pathak, "Annotation of a large clinical entity corpus," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2033–2042.

[42] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, "A survey on text classification: From traditional to deep learning," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 2, apr 2022. [Online]. Available: https://doi.org/10.1145/3495162

[43] A. J. J. Yepes, L. Plaza, J. Carrillo-de Albornoz, J. G. Mork, and A. R. Aronson, "Feature engineering for medline citation categorization with mesh," *BMC bioinformatics*, vol. 16, no. 1, pp. 1–12, 2015.

[44] N. Sanchez-Pi, L. Martí, and A. C. Bicharra Garcia, "Improving ontology-based text classification: An occupational health and security application," *Journal of Applied Logic*, vol. 17, pp. 48–58, 2016, sOCO13. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1570868315000774

[45] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?" in *Chinese Computational Linguistics*, M. Sun, X. Huang, H. Ji, Z. Liu, and Y. Liu, Eds., 2019, pp. 194–206.

[46] K. S. Kalyan, A. Rajasekharan, and S. Sangeetha, "Ammu: A survey of transformer-based biomedical pretrained language models," *Journal of Biomedical Informatics*, vol. 126, p. 103982, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046421003117

[47] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Nov. 2019, pp. 3615–3620.

[48] S. Chakraborty, E. Bisong, S. Bhatt, T. Wagner, R. Elliott, and F. Mosconi, "BioMedBERT: A pre-trained biomedical language model for QA and IR," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 669–679. [Online]. Available: https://aclanthology.org/2020.coling-main.59

[49] S. Wada, T. Takeda, S. Manabe, S. Konishi, J. Kamohara, and Y. Matsumura, "Pre-training technique to localize medical bert and enhance biomedical bert," 2021.

[50] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, Aug. 2019, pp. 58–65.

[51] M. Sushil, S. Suster, and W. Daelemans, "Are we there yet? exploring clinical domain knowledge of BERT models," in *Proceedings of the 20th Workshop on Biomedical Language Processing*. Online: Association for Computational Linguistics, Jun. 2021, pp. 41–53. [Online]. Available: https://aclanthology.org/2021.bionlp-1.5

[52] A. Nentidis, G. Katsimpras, A. Krithara, S. Lima López, E. Farré-Maduell, L. Gasco, M. Krallinger, and G. Paliouras, "Overview of bioasq 2023: The eleventh bioasq challenge on large-scale biomedical semantic indexing and

question answering," in *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 2023, pp. 227–250.

[53] A. B. Abacha, Y. M'rabet, Y. Zhang, C. Shivade, C. Langlotz, and D. Demner-Fushman, "Overview of the mediqa 2021 shared task on summarization in the medical domain," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, 2021, pp. 74–85.

[54] I. Segura-Bedmar, P. Martínez Fernández, and M. Herrero Zazo, "Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013)." Association for Computational Linguistics, 2013.

[55] A. Gonzalez-Agirre, M. Marimon, A. Intxaurrondo, O. Rabal, M. Villegas, and M. Krallinger, "Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track," in *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, 2019, pp. 1–10.

[56] S. Baker, I. Silins, Y. Guo, I. Ali, J. Högberg, U. Stenius, and A. Korhonen, "Automatic semantic classification of scientific literature according to the hallmarks of cancer," *Bioinformatics*, vol. 32 3, pp. 432–40, 2016.

[57] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. S. Weld, "Specter: Document-level representation learning using citation-informed transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2270—-2282.

[58] D. Pascual, S. Luck, and R. Wattenhofer, "Towards BERT-based automatic ICD coding: Limitations and opportunities," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, Jun. 2021, pp. 54–63.

[59] A. Wührl and R. Klinger, "Claim detection in biomedical Twitter posts," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, Jun. 2021, pp. 131–142.

[60] Y. Gu *et al.*, "Linkbert: Integrating link prediction tasks during pretraining for enhanced biomedical document classification," *Journal of Biomedical Informatics*, vol. 134, p. 104073, 2023.

[61] W. Li *et al.*, "Clinicalbert: A pretrained language model for clinical text analysis," *BMC Medical Informatics and Decision Making*, vol. 23, p. 52, 2023.

[62] Y. Zhang *et al.*, "Biolinkbert: Pretraining biomedical language models with link prediction," *Bioinformatics*, vol. 39, no. 2, p. btad018, 2023.

[63] J. Kim *et al.*, "Integrating electronic health records with pre-trained language models for improved clinical note classification," *Journal of Medical Internet Research*, vol. 26, p. e28347, 2024.

[64] L. Sun *et al.*, "Multitask learning for biomedical text classification: A comparative study," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 1, pp. 67–77, 2024.

[65] C. Luo *et al.*, "Hybrid approaches for biomedical entity classification: Combining rule-based systems and pre-trained language models," *Artificial Intelligence in Medicine*, vol. 130, p. 102327, 2023.

[66] A. Ratner *et al.*, "Weak supervision for generating annotated biomedical text data at scale," *Nature Communications*, vol. 15, no. 1, p. 114, 2024.

[67] X. Chen *et al.*, "Incorporating unified medical language system (umls) into pre-trained language models for enhanced biomedical text classification," *Journal of the American Medical Informatics Association*, vol. 30, no. 4, pp. 634–642, 2023.

[68] É. Archambault, O. H. Beauchesne, and J. Caruso, "Towards a multilingual, comprehensive and open scientific journal ontology," in *Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics (ISSI)*, 2011, pp. 66—77.

[69] G. Szarvas, V. Vincze, R. Farkas, and J. Csirik, "The bioscope corpus: Annotation for negation, uncertainty and their scope in biomedical texts," 07 2008, pp. 38–45.

[70] K. Oconnor, A. Sarker, J. Perrone, and G. Hernandez, "Promoting reproducible research for characterizing nonmedical use of medications through data annotation: Description of a twitter corpus and guidelines," *Journal of medical Internet research*, vol. 22, p. e15861, 02 2020.

[71] SIRIS Academic, "Portrait of cancer research in portugal, a comprehensive mapping analysis," 2022.

[72] ——, "Segundo informe sobre la investigación e innovación en cáncer en españa," 2022.

[73] ——, "Supporting france universités reflections on the state of biomedical research in france," 2023.

[74] S. Woolf, "The meaning of translational research and why it matters." *Journal of the American Medical Association*, vol. 299, no. 2, 2008.

[75] G. Weber, "Identifying translational science within the triangle of biomedicine," *Journal of translational medicine*, vol. 11, p. 126, 05 2013.

[76] J. Flier and J. Loscalzo, "Categorizing biomedical research: The basics of translation," *The FASEB Journal*, vol. 31, pp. 3210–3215, 08 2017.

[77] D. Fort, T. Herr, P. Shaw, K. Gutzman, and J. Starren, "Mapping the evolving definitions of translational research," *Journal of Clinical and Translational Science*, vol. 1, pp. 1–7, 02 2017.

[78] C. I. Dana Farber, "Basic, clinical and translational research: What's the difference?" 2018. [Online]. Available: https://blog.dana-farber.org/insight/2017/12/basic-clinical-translational-research-whats-difference/. Accessed on March 2023

[79] D. Nozza, F. Bianchi, and D. Hovy, "What the [mask]? making sense of language-specific bert models," 2020. [Online]. Available: https://arxiv.org/abs/2003.02912

[80] M. Yasunaga, J. Leskovec, and P. Liang, "Linkbert: Pretraining language models with document links," 2022. [Online]. Available: https://arxiv.org/abs/2203.15827

[81] L. Tunstall, L. von Werra, and T. Wolf, *Natural Language Processing with Transformers*. O'Reilly Media, Inc., 2022.

[82] P. Su, Y. Peng, and K. Vijay-Shanker, "Improving BERT model using contrastive learning for biomedical relation extraction," in *Proceedings of BioNLP*, 2021.

[83] G. Cenikj, T. Eftimov, and B. Koroušić Seljak, "SAFFRON: tranSfer leArning for food-disease RelatiOn extractioN," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, Jun. 2021, pp. 30–40.

[84] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," ser. Psychology of Learning and Motivation, G. H. Bower, Ed. Academic Press, 1989, vol. 24, pp. 109–165.

[85] H. Tayyar Madabushi, E. Kochkina, and M. Castelle, "Cost-sensitive BERT for generalisable sentence classification on imbalanced data," in *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, Nov. 2019, pp. 125–134.

[86] T. Schick and H. Schütze, "Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking," 2019. [Online]. Available: https://arxiv.org/abs/1904.06707

[87] A. Barnett and Z. Doubleday, "Meta-research: The growth of acronyms in the scientific literature," *eLife*, vol. 9, p. e60080, jul 2020. [Online]. Available: https://doi.org/10.7554/eLife.60080

[88] W. Hogan, Y. Vazquez Baeza, Y. Katsis, T. Baldwin, H.-C. Kim, and C.-N. Hsu, "BLAR: Biomedical local acronym resolver," in *Proceedings of the 20th Workshop on Biomedical Language Processing*. Online: Association for Computational Linguistics, Jun. 2021, pp. 126–130. [Online]. Available: https://aclanthology.org/2021.bionlp-1.14

[89] S. Ujiie, H. Iso, S. Yada, S. Wakamiya, and E. Aramaki, "End-to-end biomedical entity linking with span-based dictionary matching," in *Proceedings of the 20th Workshop on Biomedical Language Processing*, Jun. 2021, pp. 162–167.

[90] A. Schwartz and M. Hearst, "A simple algorithm for identifying abbreviation definitions in biomedical text," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, vol. 4, pp. 451–62, 02 2003.

[91] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," 2013.

[92] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019.

[93] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[94] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2020.

[95] Y. Kim, K. Kang, and S. Kim, "Pre-training of deep bidirectional protein sequence representations with structural information," *arXiv preprint arXiv:1909.04868*, 2019.

[96] H. Yan, J. Gu, Z. Ji, and C. Tan, "A deep learning framework for biomedical text classification using bert," *IEEE Access*, vol. 8, pp. 143 473–143 482, 2020.

[97] Y. Zhang, S. Wang, and X. Zhu, "Biomedical literature classification with an emphasis on imbalanced data," *Journal of Biomedical Informatics*, vol. 109, p. 103518, 2020.

[98] Z. Lu, Y. Peng, and L. Wang, "The pubmed retrieval system: improvements and developments," *Database*, vol. 2019, 2019.

[99] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

NICOLAU DURÁN-SILVA is a computer scientist with an MSC in Language Technologies, currently researcher at SIRIS Academic and PhD candidate at the Pompeu Fabra University. His work involves exploring data from higher education institutions and regional research ecosystems to inform evidence-based policy development. He has been exploring information extraction problems from scientific and technical documents, such as being able to classify them into complex thematic categories, identify organisations and simplify the content.

LAURA PLAZA is senior lecturer at UNED and researcher of the UNED IR NLP group. Her research includes different fields of NLP, including summarization, text classification and sentiment analysis, with especial interest in biomedical field. She has published in more than 80 international journals and conferences, has participated in different funded projects and worked in several international companies.

JORGE CARRILLO-DE-ALBORNOZ is assistant professor at UNED and researcher of the UNED IR NLP group. His research includes different fields of NLP, including evaluation, text classification and sentiment analysis. She has published in more than 60 international journals and conferences, has participated in different funded projects and worked in several international companies.

FRANCESCO A. MASSUCCI holds a PhD in Applied Mathematics (King's College London) in the field of Complex Systems. He is the co-author of several scientific publications in major peer-reviewed international journals. At SIRIS, he helps strategic decision-making for HERI, by providing quantitative insights built on data science and natural language processing techniques.

SARA RICARDO is a senior consultant at SIRIS Academic, specialised in life sciences and health research, research ecosystem analysis and institutional strategy. She has a long standing career in developing and leading projects in the life sciences in Europe and in the USA, being a former junior Principal Investigator in Spain in Life Sciences. Sara has working experience in the academic, the non-governmental and industry sectors, in areas that range from the basic to applied medical research, and in scientific as well as in business development roles. Sara has also been elected a Marie Curie Alumni Association (MCAA) board member from 2018-2022, for two mandates. In MCAA she co-led, managed and defined the strategic planning of the organisation, with 22.000 members in all 5 continents, organised in chapters and working groups.

ARNAU RAMOS-PRATS studied Biomedicine at the University of Barcelona and Neurosciences at the Autonomous University of Barcelona (Spain). He then completed a PhD in Neurosciences in the lab of Prof. Ferraguti at the Medical University of Innsbruck (Austria), where he studied neuroanatomical and functional correlates of emotional processing in the brain. He then joined the lab of Prof. Andreas Lüthi at the Friedrich Miescher Institute for Biomedical Research in Basel (Switzerland) as a postdoctoral fellow, where he leverages behavioural, neuroanatomical and computational approaches to study neural circuits and brain states in health and disease.

SÓNIA VEIGA has a background in Health and Life-Sciences Research, with a PhD in Biomedicine (University of Barcelona). She has 15 years experience as a researcher in academic and clinical environments. At SIRIS, she provides evidence-based analysis and consulting in topics related to Biomedical and Health Research strategy.