ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Sentence Classification with Hierarchical Neural Networks for Rhetorical Sections Extraction

Ανδρέας Α. Νταργαράς

Επιβλέπων: Δρ. Χάρης Παπαγεωργίου, Ερευνητής Α' - Διευθυντής Ερευνών

ΑΘΗΝΑ

Μάιος 2021

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

INTERDEPARTMENTAL PROGRAM OF POSTGRADUATE STUDIES IN DATA SCIENCE AND INFORMATION TECHNOLOGIES

MSc THESIS

Sentence Classification with Hierarchical Neural Networks for Rhetorical Sections Extraction

Andreas A. Ntargaras

Supervisor: Dr. Haris Papageorgiou, Research Director

ATHENS

May 2021

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Sentence Classification with Hierarchical Neural Networks for Rhetorical Sections Extraction

Ανδρέας Α. Νταργαράς

Α.Μ.: DS2180012

ΕΠΙΒΛΕΠΩΝ: **Δρ. Χάρης Παπαγεωργίου**, Ερευνητής Α' - Διευθυντής Ερευνών
Ινστιτούτο Επεξεργασίας του Λόγου "ΙΕΛ" - ΑΘΗΝΑ, Ελλάδα

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Δρ. Μάρτιν Ρέτζκο**, Ειδικός Λειτουργικός Επιστήμονας Α',
Ερευνητικό Κέντρο Βιοϊατρικών Ερευνών "Αλέξανδρος Φλέμινγκ"
Δρ. Αλέξανδρος Δημόπουλος, Λέκτορας,
Σχολή Ναυτικών Δοκίμων

Μάιος 2021

Στους φίλους και την οικογένεια μου που βοήθησαν να προχωρήσω όλα αυτά τα χρόνια

MSc THESIS

Sentence Classification with Hierarchical Neural Networks for Rhetorical Sections Extraction

Andreas A. Ntargaras

S.N.: DS2180012

SUPERVISOR: **Dr. Haris Papageorgiou**, Research Director
Institute for Language and Speech Processing ”ILSP” - ATHENA, Greece

EXAMINATION COMMITTEE: **Dr. Martin Reczko**, Staff Research Scientist Professor Level
Biomedical Sciences Research Center ”Alexander Fleming”
Dr. Alexandros Dimopoulos, Lecturer
Hellenic Naval Academy

May 2021

To my friends and family that kept me going all these years

ΠΕΡΙΛΗΨΗ

Υπόβαθρο: Εκατομμύρια επιστημονικά άρθρα και επιστημονικές εργασίες δημοσιεύονται κάθε χρόνο, καθιστώντας την έρευνα για σχετική βιβλιογραφία όλο και πιο δύσκολη με κάθε μέρα που περνά. Ως εκ τούτου, οι σαφείς και ενημερωτικές περιλήψεις έχουν καταστεί απαραίτητο μέσο για να εντοπίζουν οι ερευνητές τις επιθυμητές πληροφορίες εγκαίρως και με αποτελεσματικό τρόπο. Πολλές περιλήψεις, ωστόσο, εξακολουθούν να στερούνται κοινών ρητορικών δομικών στοιχείων τα οποία θα βελτίωναν τους επικοινωνιακούς τους σκοπούς στο πλαίσιο του ακαδημαϊκού λόγου.

Στόχος: Στην παρούσα διατριβή στοχεύουμε να εξετάσουμε την αποτελεσματικότητα των μοντέλων ταξινόμησης προτάσεων για την εξαγωγή ρητορικών ενοτήτων σε περιλήψεις διαφορετικών τομέων και δομών και να δημιουργήσουμε ένα εργαλείο που αυτοματοποιεί αυτήν τη διαδικασία.

Μέθοδος: Τα μοντέλα ταξινόμησης προτάσεων που χρησιμοποιήθηκαν εδώ βασίστηκαν σε ένα ιεραρχικό νευρωνικό δίκτυο (HNN) που έχει εκπαιδευτεί σε τρία διαφορετικά σύνολα δεδομένων.

Αποτέλεσμα: Τα αποτελέσματά μας δείχνουν ότι τα μοντέλα μας επιβεβαιώνουν την ”state of the art” απόδοσή τους (SOTA) σε περιλήψεις του ίδιου επιστημονικού πεδίου με εκείνες που εκπαιδεύτηκαν, αλλά η διαπεδιακή ακρίβειά τους μειώνεται σημαντικά ειδικά όταν εφαρμόζονται σε μη κλασσικά δομημένες περιλήψεις.

Συμπέρασμα: Ένα ακριβές εργαλείο για την απόκτηση των ρητορικών τμημάτων των περιλήψεων μπορεί να αποτελέσει τη βάση για ένα μεγαλύτερο σύστημα που θα μπορεί να συνοψίζει τις πληροφορίες, βοηθώντας έτσι σε μεγάλο βαθμό την επιτάχυνση της διαδικασίας της βιβλιογραφικής έρευνας.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Επεξεργασία Φυσικής Γλώσσας

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Διαδοχική Ταξινόμηση Προτάσεων, Εξαγωγή Ρητορικών Ενοτήτων, Ιεραρχικά Νευρωνικά Δίκτυα

ABSTRACT

Background: Millions of scholarly articles and scientific papers are being published each year, making the search for relevant literature harder with each passing day. Clear and informative abstracts have therefore become an essential medium for researchers to locate their desired information in a timely and efficient manner. Many abstracts however, still lack common rhetorical structural elements that would improve their communicative purposes within the context of academic discourse.

Objective: In the present thesis we aim to review the efficacy of sentence classification models for rhetorical sections extraction on abstracts of different domains and structures and create a tool that automates this process.

Method: The sentence classification models used here were based on a hierarchical neural network (HNN) that has been trained on three different datasets.

Result: Our results show that our models manage to confirm their state of the art (SOTA) performance on abstracts of the same scientific field with the ones they were trained in, but their inter-domain accuracy drops significantly especially when applied to unordinarily structured abstracts.

Conclusion: An accurate tool for obtaining the rhetorical sections of abstracts can become the basis for a larger framework that could summarize information, helping tremendously to speed up the process of literature research.

SUBJECT AREA: Natural Language Processing

KEYWORDS: Sequential Sentence Classification, Rhetorical Sections Extraction, Hierarchical Neural Networks

ACKNOWLEDGEMENTS

Before continuing, I would like to take the time to thank everyone that helped me successfully conclude this thesis. First and foremost, I would like to express my deepest appreciation to my supervisor Haris Papageorgiou, whose valuable advice and insights helped shape the core of this thesis. His belief in my work from the start and his continuous support, both material and cognitive, are what drove this work to the realization of its full potential. Furthermore, I would like to give special thanks to the team at the Institute for Language and Speech Processing ”ILSP” - ATHENA, namely Aris Fergadis and Dimitris Pappas. Their helpful contributions, both theoretical and practical, not only helped me grasp the essence of the tools that were used throughout this thesis but also provided the solid foundation upon which the results of this thesis were extracted.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Problem Description and Motivation

In the course of this thesis we will be tackling the sentence classification problem, within the context of research article abstracts. What we are trying to achieve precisely, is to create artificial neural network (ANN) models that are able to correctly identify and categorize the sentences of a scientific abstract to a set of given labels. It is also important that these models retain the ability to generalize. To test this we will be using completely different scientific abstracts than the ones used to train them, therefore adding to this text classification problem the domain adaptation perspective. Finally, as a proof of concept, we will try to bring everything together in an Application Programming Interface (API) through the Flask module [1] in Python, which will include our trained models to an easy-to-use web-based tool for making predictions. This will serve as an attempt to identify its usefulness and limitations as a standalone tool and in consideration as part of a larger framework for information retrieval.

The main motivation behind our project could be identified as the easy and fast acquisition of required information for research. Every passing year the size of the scientific literature corpus is increasing at a rapid pace [2]. It is therefore becoming increasingly hard for researchers to find relevant literature, to help further their work, in a timely and efficient manner. This has shifted the focus towards abstracts, as their small and informative nature makes them ideal for quickly navigating through a lot of information. Not all abstracts however, have the necessary rhetorical structural elements that would improve their communicative purposes within the context of academic discourse. Consequently, an automatic tool that would identify those structural elements would be something extremely helpful and valuable as well as serve as the basis for a larger framework that would allow users of all scientific fields to easily locate their desired information and present it in a clear way.

Another, more subtle goal involves raising awareness in the researchers community about the importance of following a standardized scientific discourse. With abstracts becoming increasingly important, identification of the impeding factors for the accuracy of our model could provide important educational insights on future researchers regarding the steps they could take to improve the readability of their abstracts.

For the remainder of this chapter we will be providing some general background contextual information as well as a more formal definition of the problem through its individual tasks and describe our action plan. In the next chapter, we will discuss related approaches to this problem, while afterwards we will be focusing on the specifics of the methodology we followed to obtain our results. Finally, we will present these results and conclude by offering our comments and our closing remarks.

[1] `https://flask.palletsprojects.com/en/2.0.x/`

1.2 Background information

Deep learning (DL) is part of the broader research area of machine learning (ML), which is in turn considered part of artificial intelligence (AI) (Figure 1.1 [2]). Its applications use ANN models combined with an approach called deep contextual representations to solve problems in the fields of computer vision (CV), speech recognition (SR), natural language processing (NLP), precision medicine and many others ([3], [4], [5]). More specifically, DL is solving a central problem in representation learning, the fact that sometimes it is equally difficult to obtain a representation due to variation factors. It does so by using representations that are a combination of simpler representations, breaking down the more complex concepts into simpler ones [6].

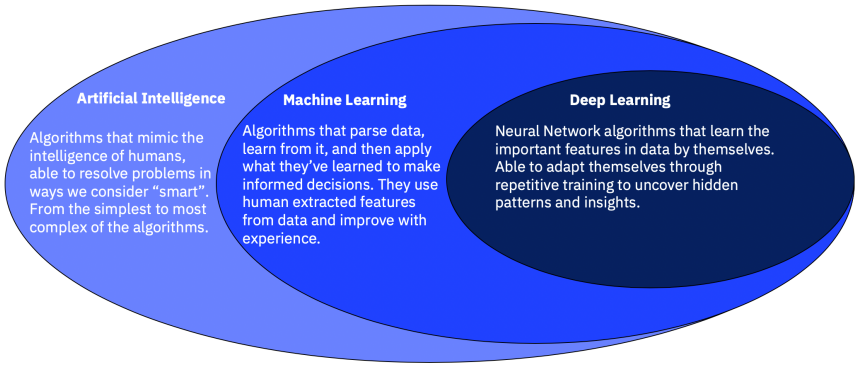


Figure 1.1: The relationship between AI, ML and DL.

As the main subject of this thesis falls within the field of NLP, we would like to provide at this point some more context by briefly describing its history and the tasks it addresses giving a little more detail. NLP has its roots in the 1950's [3] as a subfield of linguistics, computer science, and artificial intelligence [7]. It is concerned with the way machines perceive and generate language in written text form.

Early NLP models are known today as *Symbolic NLP*. They used rather pattern-based or rule-based approaches, such as hand-crafted rules or context-free grammars (CFG), for tasks like word-for-word translation which ultimately failed due to lack of coverage, ambiguity and the complexity of language in general. During late 1980's *Statistical NLP* emerged, with models that used probabilistic machine learning algorithms (which Chomsky's theoretical analysis had been skeptical about) and large annotated corpora to efficiently train them. One such model was the Markov-based n-gram language model, an approach that uses the previous words to predict the upcoming word, letter or phonemes in a sequence. These models are used in more practical applications such as auto-completion or spelling correction tasks. In an effort to address some of the shortcomings of these mod-

[2] Image taken from this IBM post: `https://ibm.co/3cG8pEQ`

[3] Alan Turing published an article in 1950, titled "Computing Machinery and Intelligence", proposing what is now known as the "Turing test" for measuring intelligence which involves the automated interpretation and generation of natural language

els Kneser and Ney [8] proposed a smoothing method to deal with unseen n-grams that would otherwise get a probability of 0.0.

With the dawn of the 21st century we moved forward to the current *Neural NLP* era. Conditional random fields (CRFs) [9], one of the most influential models for sequential classification, as well as the first large-scale modern feed-forward neural network for language modeling [10] were both proposed in 2001. During that time, bag-of-words models were also replaced by dense vector representations of words called word embeddings, which in 2013 were introduced by Mikolov et al. [11] approximating the objective function and making their training more efficient. Along with the word2vec implementation [12, 13], large scale training of word embeddings was made possible. One of the latest innovations in word embeddings is the projection of word embeddings from different languages into the same space enabling (zero-shot) learning [14, 15] opening a lot of applications for low-resource languages. From 2013-2014 until today, convolutional neural networks (CNNs, [16]) and more often recurrent neural networks (RNN, [17]) as well as (bidirectional) long short-term memory networks ((bi)-LSTM, [18, 19]) are still regarded as the go-to network architectures for state-of-the-art performance in NLP tasks, although the new Transformer architecture [20] is rapidly gaining ground on that regard [21]. Some of the more common NLP tasks include but are not limited to Text-to-Speech, Word Segmentation (Tokenization), Named Entity Recognition, Text Classification, Sentiment Analysis, Discourse Analysis, Text Summarization and many others.

1.3 Sequential Sentence Classification

Sentence classification is considered part of the broader task of text classification in NLP. Text classification is the task of assigning a sentence or document an appropriate category [22]. The categories depend on the chosen dataset and can range from topic to topic. Sequential text classification is the task where the whole sequence or context is taken into consideration for tagging the current token/ entity to be classified. Although it is one of the simpler tasks of the field, it is a precursor of many interesting applications such as spam detection or sentiment analysis. The true value of text classification models however comes from being able to perform unsupervised labeling of unstructured data, which accounts for the vast majority of the data found online. This holds true for scientific literature as well, so as we discussed in section 1.1, an accurate automated sentence classification tool would therefore be of high value.

As a sentence we will define the number of words between each consecutive period except the first. Although this is a simple enough definition for any human to understand, this can prove difficult for a sentence tokenizer as a lot of exceptions that we consider automatically need to be taught to the tokenizer in the form of rules (e.g. acronyms can have periods). Each sentence is assigned a label or as it is specifically called in the discourse analysis task a *rhetorical move*. As rhetorical move its considered a ”discoursal or rhetorical unit that performs a coherent communicative function in a written or spoken discourse” [23]. The moves proposed by Santos’ model in 1996 are illustrated in table 1.1. As we can see

common sections in abstracts include *BACKGROUND*, *OBJECTIVE*, *METHOD*, *RESULT* and *CONCLUSION*, which remains so until today. These sections are also presented in our datasets as well and are the exact labels we will be trying to predict.

Table 1.1: Framework for a five-move structure analysis of research article abstracts. [23]

Moves	Functions	Questions to ask
1. Situating the research	Setting the scene, topic generalization	What has been known about the field/topic of research?
2. Presenting the research	Setting the purpose of the study, research questions/hypotheses	What is the study about?
3. Describing the methodology	Describing the materials, subjects, variables, procedures	How was the research done?
4. Summarizing the findings	Reporting the main findings of the research	What did the research find?
5. Discussing the findings	Interpreting the results, giving recommendations, implications, applications	What do the results mean? So What?

1.4 Domain Adaptation

In a typical machine learning application, the test set is drawn from the same distribution as the train and validation sets. This would mean that you would have a manually annotated dataset for your specific task that you would split into three parts (typical ratios include 60% / 20% / 20%, 70% / 15% / 15% or 80% / 10% / 10% for train / validation / test) to use with your model. In some cases however, one might lack the amount of data required to fully train a model from the ground up for a particular task. In such a case, one could take a model trained in a source domain and transfer any relevant features from it to his task on another target domain.

From [24, 25] we get the definition of a domain $\mathcal{D}$, as a feature space $\mathcal{X} \subset R^d$ with a marginal probability distribution $P(\boldsymbol{X})$ or $\mathcal{D} = \{\mathcal{X},\ P(\boldsymbol{X})\}$. A task $\mathcal{T}$ is defined as a label space $\mathcal{Y}$ with the conditional probability distribution $P(\boldsymbol{Y}|\boldsymbol{X})$, so we can write $\mathcal{T} = \{\mathcal{Y},\ P(\boldsymbol{Y}|\boldsymbol{X})\}$. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are random variables. Now, a source domain $\mathcal{D}_S$ is defined as $\mathcal{D}_S = \{\mathcal{X}_S,\ P(\boldsymbol{X}_S)\}$ with a corresponding task $\mathcal{T}_S = \{\mathcal{Y}_S,\ P(\boldsymbol{Y}_S|\boldsymbol{X}_S)\}$. In a similar manner a target domain $\mathcal{D}_T$ is defined as $\mathcal{D}_T = \{\mathcal{X}_T,\ P(\boldsymbol{X}_T)\}$ with a corresponding task $\mathcal{T}_T = \{\mathcal{Y}_T,\ P(\boldsymbol{Y}_T|\boldsymbol{X}_T)\}$. Transfer learning is defined as the process of improving the outcome of task $\mathcal{T}_T$ by using information from $\mathcal{D}_S$ and $\mathcal{T}_S$ where $\mathcal{D}_S \neq \mathcal{D}_T$ and $\mathcal{T}_S \neq \mathcal{T}_T$. In the case where the source and target domains have the same feature space, $\mathcal{X}_S = \mathcal{X}_T$, we have a homogeneous transfer learning problem. Finally, the special case, where both domains share a common label space, $\mathcal{Y}_S = \mathcal{Y}_T$, is defined as domain adaptation.

Given the above formal definitions, our problem can be interpreted as a homogeneous domain adaptation problem in which a model trained on a source distribution is used in the context of a different yet related target distribution.

1.5 The Problem

Given the definitions in the previous sections, our problem is considered a complex mix of sentence classification and domain adaptation problems. Specifically, we are in the setting of the relaxed domain adaptation because the $\mathcal{Y}_S = \mathcal{Y}_T$ condition holds true. Our general action plan therefore, would be to train a deep learning model on one or more task specific datasets from a source domain and then test their efficacy on target domains.

The best and largest datasets are the PubMed RCT datasets, which refer to randomized controlled trials, curated by Dernoncourt and Lee [26]. The high level of curation of these datasets as well as their size, especially when compared to the older NICTA-PIBOSO dataset from Amini et al. [27], made them fitting for our needs. Also, having abstracts from a domain where we had more intimate knowledge increased their value for us and made these datasets a natural choice as our source domain. Finally, we opted to use some hand-made datasets of our own that also belonged to this domain. One was exactly like the PubMed RCT datasets, but at a larger scale of 1 million abstracts (400% increase), while the other had two more specialized labels pertaining to the PICO [4] framework used in Evidence Based Medicine for framing and answering of health care related questions [28, 29].

The choice of deep learning model to train with these datasets was equally straightforward, as we used the State-of-the-art (SOTA) sentence classification model on the aforementioned PubMed abstract datasets [5]. The model is called "*Hierarchical Sequential Labeling Network* (HSLN)" and was presented by D. Jin and P. Szolovits [1] in 2018. It is based on a hierarchical neural network (HNN) architecture which uses a CRF layer to account for the inter-dependence between consecutive labels along with a bi-LSTM layer atop the representation of each sentence to encode context and semantics from preceding and succeeding sentences for better label prediction.

As target domains we searched for ones that are as different as possible from our source domain while retaining the same label space. The energy and sociology domains satisfied these criteria so they served as our target domains of choice. However, to our knowledge there was no readily available datasets, therefore we needed to manually create them through data-mining open access repositories. Our repository of choice was the "Directory of Open Access Journals" archive [6], which is a community-curated online directory with good indexing of its vast array of high-quality, open access, peer-reviewed journals and articles.

Our final goal would be to construct a web-based API tool for automatic abstract sentence classification in both source and target domains, to use as a proof of concept for a larger framework of information retrieval.

[4] "PICO elements": Participants/Problem (P), Intervention (I), Comparison (C) and Outcome (O)

[5] `https://bit.ly/3xnG5PN`

[6] `https://doaj.org/`

2. RELATED WORK

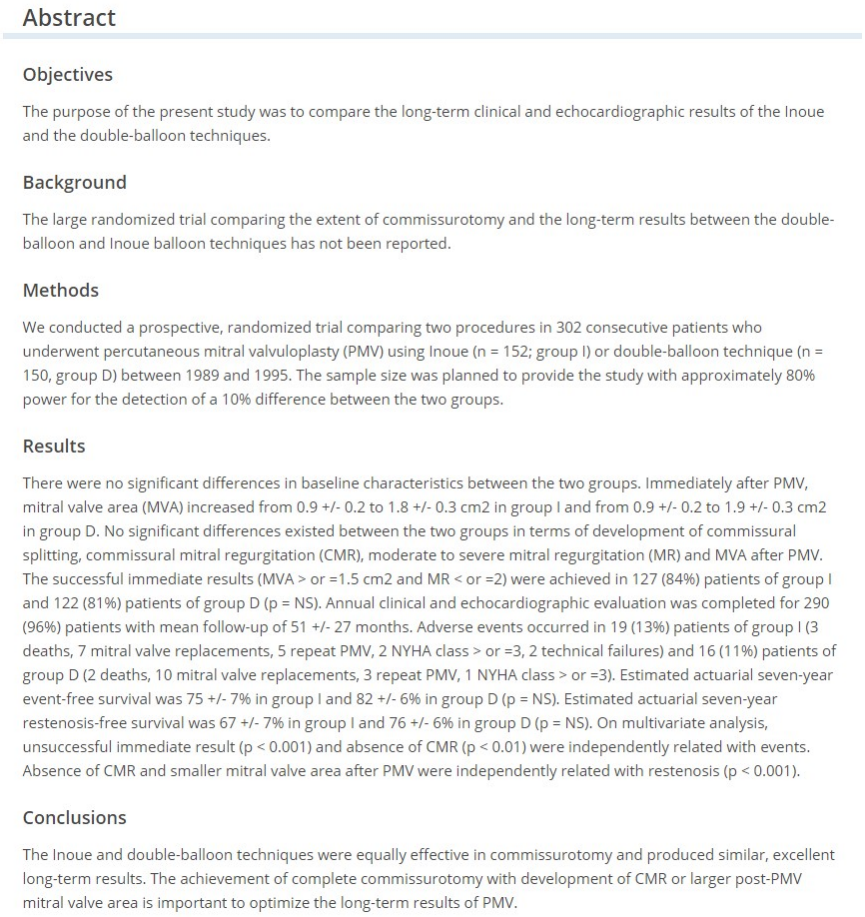
Before presenting in greater detail the model that we used, it is important to go over the other systems that were previously or are currently available for our task of sentence classification. This can help give more insight on the differences between them and our model of choice, by discussing some advantages and disadvantages of each system.

2.1 Different Models for the task

Through the years, a number of different models were used for the task of sequential sentence classification. In the pre-CNN and RNN era, the systems used were mainly based on either naive Bayes (NB), support vector machines (SVM), Hidden Markov Models or CRFs. One of the earliest successful attempts was reported in 2003 by McKnight and Srinivasan [30], who used an SVM to perform classification on both structured and unstructured abstracts utilizing a novel sentence location feature. Although it performed extensively better than traditional linear classifiers used previously, achieved an average F-score of 80% on structured and 74% on unstructured data. As ”structured”, the authors define the abstracts in which there where already structure labels present and as ”unstructured” the abstracts that were lacking those labels. An example of a structured abstract can be seen in figure 2.1.

A comparable effort using HMM was reported by Lin et al. [31] in 2006. Although their generative approach system of HMM coupled with Linear Discriminant Analysis (LDA) yielded similar results with McKnight and Srinivasan, the advantage of their approach is the linear complexity for training and testing their model compared to the quadratic complexity of SVMs. The use of an NB-based system proposed by Ruch et al [32] in 2007 showed more promise than the other two approaches, reporting an F-score of about 85% overall. Their biggest problems were the missclassification between the RESULTS and CONCLUSION moves, which was reported by other researchers as well and the very poor performance on unstructured data that fell bellow 70%. As unstructured data we mean abstracts that have no adherence to a formal format, with minimal or zero sensible links between the sentences. By the strict definition of the term, any abstract that is written in a single paragraph format and is not split into sections is considered unstructured. However, since such abstracts are still not widely adopted, for the purpose of this thesis we will consider as unstructured abstracts those that have a more narrative form, are often smaller and tend to be less informative about the contents of the paper they summarize.

The final, pre-deep learning, leap forward came in the form of CRFs [33] which saw an average of above 83% F-score in identifying PICO elements and above 94% on regular RCT moves. As it is evident from the above data, the latest models managed to achieve impressive scores in this given task however all of them lacked generalization properties. This is due to the fact that they rely heavily on carefully hand-crafted lexical, structural, statistical and sequential features and the fact that they are trained with smaller datasets.



Abstract

Objectives

The purpose of the present study was to compare the long-term clinical and echocardiographic results of the Inoue and the double-balloon techniques.

Background

The large randomized trial comparing the extent of commissurotomy and the long-term results between the double-balloon and Inoue balloon techniques has not been reported.

Methods

We conducted a prospective, randomized trial comparing two procedures in 302 consecutive patients who underwent percutaneous mitral valvuloplasty (PMV) using Inoue (n = 152; group I) or double-balloon technique (n = 150, group D) between 1989 and 1995. The sample size was planned to provide the study with approximately 80% power for the detection of a 10% difference between the two groups.

Results

There were no significant differences in baseline characteristics between the two groups. Immediately after PMV, mitral valve area (MVA) increased from 0.9 +/- 0.2 to 1.8 +/- 0.3 cm2 in group I and from 0.9 +/- 0.2 to 1.9 +/- 0.3 cm2 in group D. No significant differences existed between the two groups in terms of development of commissural splitting, commissural mitral regurgitation (CMR), moderate to severe mitral regurgitation (MR) and MVA after PMV. The successful immediate results (MVA > or =1.5 cm2 and MR < or =2) were achieved in 127 (84%) patients of group I and 122 (81%) patients of group D (p = NS). Annual clinical and echocardiographic evaluation was completed for 290 (96%) patients with mean follow-up of 51 +/- 27 months. Adverse events occurred in 19 (13%) patients of group I (3 deaths, 7 mitral valve replacements, 5 repeat PMV, 2 NYHA class > or =3, 2 technical failures) and 16 (11%) patients of group D (2 deaths, 10 mitral valve replacements, 3 repeat PMV, 1 NYHA class > or =3). Estimated actuarial seven-year event-free survival was 75 +/- 7% in group I and 82 +/- 6% in group D (p = NS). Estimated actuarial seven-year restenosis-free survival was 67 +/- 7% in group I and 76 +/- 6% in group D (p = NS). On multivariate analysis, unsuccessful immediate result (p < 0.001) and absence of CMR (p < 0.01) were independently related with events. Absence of CMR and smaller mitral valve area after PMV were independently related with restenosis (p < 0.001).

Conclusions

The Inoue and double-balloon techniques were equally effective in commissurotomy and produced similar, excellent long-term results. The achievement of complete commissurotomy with development of CMR or larger post-PMV mitral valve area is important to optimize the long-term results of PMV.

Figure 2.1: Example of structured abstract from the research article "Long-term clinical and echocardiographic outcome of percutaneous mitral valvuloplasty: randomized comparison of Inoue and double-balloon techniques.". (`https://europepmc.org/article/med/10636276`)

These problems were largely solved by the emergence of deep learning models, which removed the need for hand-picked features and use enormous datasets for their training. Many models based on CNNs and RNNs have achieved SOTA performances on various NLP tasks including our particular sentence classification task [34, 35, 36]. All these models consistently report results close to about 90% F-score and experience very good generalizability.What is more, depending on the architecture there are models that can utilize and operate at the smallest text unit i.e the character level as well as models, such as bi-LSTMs, which were capable of learning long-term dependencies and as such make very good candidates for our particular task.

Since their first introduction in 2017, transformer models [37] have taken the NLP field by storm improving the SOTA on various tasks including sentence classification mostly utilizing novel datasets tailored to their architecture such as SciCite [38]. The main advantage of a transformer is the fact that it can learn long-term dependencies regardless of distance

between elements and contrary to RNNs it has parallelization capabilities. A frequently used transformer model for NLP tasks is BERT [39] introduced in 2019 by Devlin J. et al.. Using its unique ability to pre-train deep bidirectional representations it can be fine-tuned with little effort to create new SOTA models. The current best performing model is OpenAI's GPT-3 [40]. It is the 3rd generation of the GPT model series and its full version has a capacity of 175 billion parameters, which is 10 times larger than the previously largest model Microsoft's Turing NLG [1].

Despite the above advantages BERT has some significant problems that later variations tried to solve and have mostly succeeded. First and foremost the fine-tuning and pre-training aspects are a bit inconsistent due to a lack of mark in the fine-tuning data. This has been partially solved by XLNet [41], which successfully addressed the problem of interdependence between mark words. The second problem lies with a corpus that is incomplete. Multiple variations such as ALBERT, RoBERTa and ERNIE [42, 43, 44] have been developed successfully addressing this issue. The last major problem is that the amount of calculations needed is too large leading to long training times. Although there are workarounds for this, by first calculating the attention to a low-dimentional space and then project to a high-dimensional space, it is certainly something not optimal.

In the biomedical field, the BioBERT [45] variation offers improvements in the NLP tasks of named entity recognition, relation extraction and question answering. In our particular task of sentence classification however, the HSLN model by Jin et al. has retained its SOTA performance and therefore was preferred over the alternatives. Another variation that is not specific to biomedical data but is used with scientific data in general is SciBert [46]. Although this variation achieved SOTA results on many different task in the computer science and multidomain fields of study, in the biomedical field it achieved better or similar results to BioBert on only the task of relation classification when tested on the ChemProt dataset [47] and of named entity recognition when tested on the BC5CDR [48] and the JNLPBA [49] datasets.

2.2 Different Datasets for the task

For our specific task of sentence classification on scientific abstracts, to our knowledge, there is no other dedicated datasets than the PubMed RCT and the NICTA-PIBOSO that we discussed in the previous chapter (section 1.5). If we move the focal point however from scientific abstract sentences in general to sentences with citation for intent classification, which is a related and very interesting task, there are two suitable datasets. The first of its kind was ACL-ARC [50], which is a collection of 10.920 (10.628 after cleaning) academic papers from the ACL Anthology. The newest and larger version, addressing multiple scientific domains is SciCite[38]. This dataset, which is five times larger than ACL-ARC, has solved not only the problem of domain-specificity but also the problem of sparse labels by combining a lot of the less frequent to more broad categories. This has greatly improved the generalizability of the dataset.

[1] `https://bit.ly/3zlXcC0`

3. METHODS

3.1 Model

As we discussed in section 1.5 of our introductory chapter, the deep learning model we chose for our problem was the HSLN model presented by Jin et al [1]. It was preferred over other models due to its SOTA performance on biomedical abstracts. In this section we will discuss in greater detail the model architecture. For convenience we will decompose the model into components based on the action they perform. These components are: the word embedding layer, the sentence encoding layer, the context enriching layer, and the label sequence optimization layer. A schema of the architecture of the model can be seen in Fig 3.1, along with indication of said components.

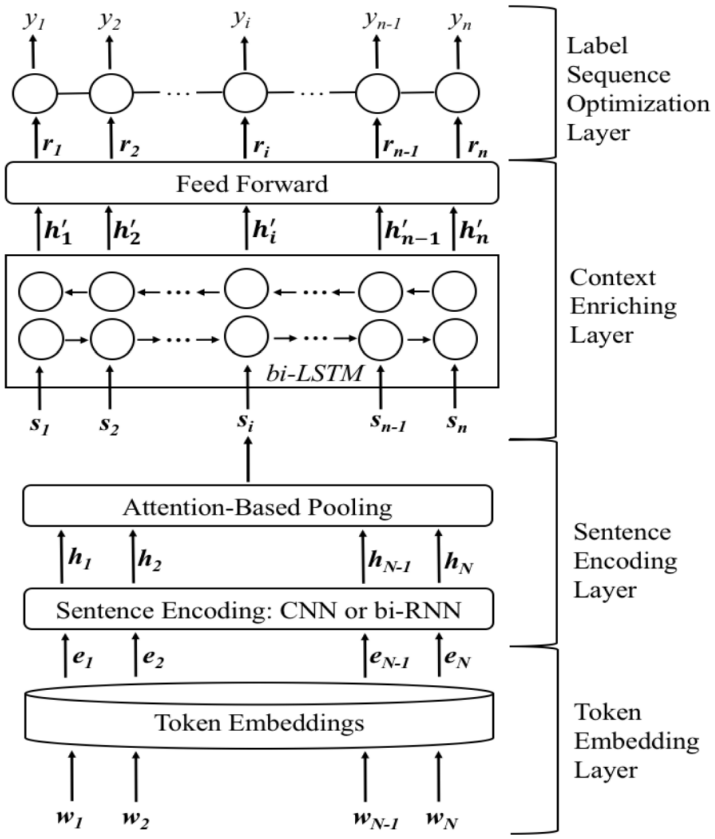


Figure 3.1: Model architecture as presented in Jin et al. [1]. w: original word; e: word embedding vector; h: sentence-level hidden state output by the bi-RNN or CNN layer; s: sentence representation vector; h_0 : abstract-level hidden state output by the bi-LSTM layer; r: sentence label probability vector; y: predicted sentence label.

The word embedding layer, takes a sentence of N words, $\mathbf{w} = [w_1\, w_2\, \cdots\, w_N]$ and maps each word to a vector of real values that serves as a representation of that word. Word representations are then encoded in an embedding matrix $W^{word} \in \mathbb{R}^{d^w \times |V|}$, where d^w is the dimension of the word vector and V is the vocabulary of the dataset. Each column j of the embedding matrix is the representation of j^{th} word in the vocabulary. Although a word embedding matrix can be randomly initialized, most models use pre-trained word embedding vectors such as word2vec, fastText and GloVe [51, 52, 53].

The sentence encoding layer accepts the embedding vector of all words in a sentence and produces the final encoding vector **s** of that sentence. This is achieved through the use of a bi-RNN or CNN layer that processes each embedding vector and outputs a sequence of hidden states $\mathbf{h}_{1:N}$ (each state corresponds to one word and N is the number of words). The final encoding vector results from attention-based pooling by reshaping the matrix S, coming from the use of the following equations, into a vector:

$$A = softmax(U_s \tanh(W_s H + \mathbf{b}_s)) \tag{3.1}$$

$$S = AH^T \tag{3.2}$$

where $H = [h_1\, h_2\, \cdots\, h_N] \in \mathbb{R}^{d^{hs} \times N}$, $W_s \in \mathbb{R}^{d^a \times d^{hs}}$ is a transformation matrix for soft alignment, $\mathbf{b}_s \in \mathbb{R}^{d^a}$ is a bias vector, $U_s \in \mathbb{R}^{r \times d^a}$ is a token-level content matrix for measuring the importance in the context of the whole sentence, softmax is performed along the second dimension of its input matrix and $A \in \mathbb{R}^{r \times N}$ is an attention matrix. Each row of U_s is a context vector $\mathbf{u_s} \in \mathbb{R}^{d^a}$ representing part of the semantics of the sentence. In our case we opted for the RNN variation of this layer as it has been shown to perform better than the CNN variant.

The context enriching layer takes the output of the previous layer for each sentence belonging to an abstract and enriches each vector with contextual information from nearby sentences. The sequence of sentence encoding vectors is input into a bi-LSTM layer to produce new hidden state vectors $\mathbf{h'}_{1:N}$ that will pass through a feed-forward NN with a single hidden layer to produce the probability vector $\boldsymbol{r} \in \mathbb{R}^l$ of the sentence. This vector depicts the probability of the sentence belonging to each label l.

Finally, the label sequence optimization layer uses the CRF algorithm to represent the dependencies between consecutive labels in order to boost overall model performance. The score of a label sequence is given by:

$$s(y_{1:n}) = \sum_{i=1}^{n} \boldsymbol{r}_i(y_i) + \sum_{i=2}^{n} T[y_{i-1}, y_i] \tag{3.3}$$

where T is a transition matrix whose $T[i, j]$ element corresponds to the probability of a labelled i token being followed by a labeled j token. The above score can be transformed into a probability by applying a softmax over the possible label sequences using the equation:

$$p(y_{1:n}) = \frac{e^{s(y_{1:n})}}{\sum_{\hat{y}_{1:n} \in Y} e^{s(\hat{y}_{1:n})}} \tag{3.4}$$

where Y represents the set of all possible label sequences.

3.2 Datasets

The datasets we used were comprised of abstracts, where each sentence is preceded by its annotated label reffering to one of the five moves of rhetorical structure and the PMID is used as an identification number. A typical example of an abstract is shown in table 3.1

Table 3.1: Example of typical abstract contained in the Pubmed 200K rct dataset (PMID: 16192451)

#	Gold label	Sentence
1	BACKGROUND	Iseganan , an antimicrobial peptide , is active against aerobic and anaerobic gram-positive and gram-negative bacteria as well as fungi and yeasts .
2	BACKGROUND	The drug has shown little resistance in vitro and to be safe and well tolerated in 800 patients with cancer treated for up to 6 wk .
3	OBJECTIVE	To determine the efficacy of iseganan for the prevention of ventilator-associated pneumonia (VAP) .
4	METHODS	Mechanically ventilated patients in the United States and Europe were randomized to oral topical iseganan or placebo (1:1) and treated six times per day while intubated for up to 14 d. Patients were eligible if randomized within 24 h of intubation and estimated to survive and remain mechanically ventilated for 48 h or more .
5	METHODS	The primary efficacy endpoint of the study was VAP measured among survivors at Day 14 .
6	RESULTS	A total of 709 patients were randomized and received at least one dose of study drug .
7	RESULTS	The two groups were comparable at baseline except iseganan-treated patients were , on average , 3 yr older .
8	RESULTS	The rate of VAP among survivors at Day 14 was 16 % (45/282) in patients treated with iseganan and 20 % (57/284) in those treated with placebo (p = 0.145) .
9	RESULTS	Mortality at Day 14 was 22.1 % (80/362) in the iseganan group compared with 18.2 % (63/347) in the placebo group (p = 0.206) .
10	RESULTS	No pattern of excess adverse events in the iseganan group compared with placebo was observed .
11	CONCLUSIONS	Iseganan is not effective in improving outcome in patients on prolonged mechanical ventilation .

The initial datasets available to us were the 20K and 200K PubMed RCT datasets (20K, 200K datasets) which are the largest published datasets for sequential sentence classification. They are based on the PubMed database. The sentences are categorized into one of the following five classes: *BACKGROUND*, *OBJECTIVE*, *METHOD*, *RESULT* and *CONCLUSION*.

Later we introduced two new and proprietary datasets, which we called 1M and 100K PubMed RCT datasets (1M, 100K datasets). These datasets where curated in-house and as their names suggest they are comprised of 1 million and 100 thousand abstracts in total respectively. Of those, the most interesting dataset is the 100K which includes two more labels belonging to the PICO framework of clinical discourse, namely *INTERVENTION* and *POPULATION*. Our goal with the inclusion of these two datasets is the inspection of any changes in the models accuracy compared to the other two standard datasets. All four datasets would serve as our source domain.

Finally, as target domains we chose abstracts pertaining to the energy and sociology fields from the ”Directory of Open Access Journals” archive. Our intention was to use abstracts that were as different as possible from the biomedical abstracts, in both written format and scientific discourse used. With that in mind, we utilized their API and we downloaded 919 energy and 605 sociology abstracts to create our initial datasets. By pre-processing the downloaded abstracts, we cleared the datasets of any non-English abstract translations that they had attached as well as disposed of any bad abstracts resulting in 860 energy and 290 sociology final abstracts. Then, we randomly picked 20% of the abstracts we downloaded to create two new test sets. This percentage was regarded as adequate for our needs, due to the fact that we would be annotating the abstracts in-house, therefore we needed a small yet representative test set that would serve as proof of concept. The quality of the abstracts within the test sets was not up to par with the highly curated source domain datasets, which can be seen from the amount of abstracts removed during preprocessing, however in their current form they would serve as invaluable real world samples to test the efficacy of our models.

In the last parts of this section, we will present some statistics regarding the datasets described above. Table 3.2 outlines the basic statistics regarding the datasets we used, such as the number of classes and the sizes of the vocabulary for each set. The development and test set sizes are kept small, especially for the 100K and 200K datasets, however we did not want to change the ratios provided by Dernoncourt and Lee [26]. In table 3.3 we present the statistics regarding the number of sentences per abstract and the average sentence length for each dataset (counted in whitespace tokens). As it is evident from the table, sentence length is comparable between all datasets whereas the abstract length is almost twice as long in the source domain compared to the target domains. This justifies our choice of very differently formatted abstracts. Last but not least, table 3.4 presents the statistics regarding the number of times each label appears in the train set for each dataset (in the energy and sociology dataset it is the test set since there was no train set). As we can see there is heavy class imbalance, with as much as 5-fold increase in some cases, which will influence our choice of the model accuracy metric.

Table 3.2: Statistics for the datasets used. Vocabulary measured in word tokens, Train - Development - Test sets in abstract tokens

Dataset	Classes	Vocabulary	Train	Development	Test
20K	5	46 k	15 k	2.5 k	2.5 k
200K	5	168 k	190 k	2.5 k	2.5 k
1M	5	613 k	1 m	100 k	100 k
100K	7	154 k	95 k	2.5 k	2.5 k

Table 3.3: Sentence (in word tokens) and abstract (in sentence tokens) lengths per dataset

	20K	200K	1M	100K	Energy	Sociology
Mean sentence length	26.34	26.23	20.75	20.82	25.04	26.19
Min sentence length	1	1	1	1	1	1
Max sentence length	296	338	72	125	136	81
Average doc length	12	11.6	11.3	11.57	7.44	6.76
Min doc length	4	3	5	5	1	2
Max doc length	31	51	30	30	25	23

Table 3.4: Number of sentences in train set for each label per dataset

Labels	20K	200K	1M	100K
BACKGROUND	18402	196689	1449519	114759
OBJECTIVE	13838	186601	790496	85001
METHOD	59281	722586	2872361	264659
RESULT	57953	766271	4262478	399754
CONCLUSION	27168	339714	1911280	183610
INTERVENTION	-	-	-	15622
POPULATION	-	-	-	36148

3.3 Experiments

The final section of this chapter is dedicated to describing our experimental process. In the pre-train part of the experiment we followed some basic pre-processing steps to prepare the datasets for the coming tests. At first, we performed a check of whether the train set of each dataset contained abstracts in common with the development and test sets, eliminating any duplicates from the train sets. This would help us avoid overfitting and miss-evaluating the performance of our models. Then, we annotated 20% of the downloaded energy and sociology abstracts, using two highly specialized and experienced annotators. Using the Python module ”Scikit-Learn” [1], we were able to calculate the inter-annotator agreement using Cohen’s kappa score [54] close to 0.7, which was deemed as adequate for the needs of this thesis.

[1] `https://scikit-learn.org/stable/index.html`

For model training we used a system equipped with an NVIDIA Titan X GPU (Pascal architecture, 12 GB of memory), which is one of the best GPUs regarding NLP tasks. Our batch size was 40, while the rest of the parameters were kept in their default values proposed by Jin et al. [1]. This meant that the sizes of the RNN layer for the sentence encoding layer and the bi-LSTM for the context enriching layer were 200 and the drop penalty and dropout were 0.01 and 0.5 respectively. The utilized Adam optimizer had an initial learning rate of 0.003, which decayed by 0.9 per epoch. Early stopping was set for a 5 epoch window with no improvement. All models took a long time to complete training, with time increasing proportionally with the train set size. One problem we faced during training was in regard to the 1M dataset which proved to be too large to train with this specific model as the TensorFlow [2] graph exceeded the 2GB hard limit. To circumvent this hurdle we rebuild the model in Keras [3]. In order to be confident about the switch we compared the results of the two versions on the 20K and 200K datasets and found that they were similar enough so that we could use them interchangeably (20K: TensorFlow wF1 = 92.20 - Keras wF1 = 92.03, 200K: TensorFlow wF1 = 94.07 - Keras wF1 = 93.97). Each pre-trained model was named after the dataset used to train it, i.e. the model trained with the 200K dataset is the 200K model. For accuracy we used the weighted-F1 score, because as we saw in table 3.4 there is heavy class imbalance, so it would serve as the optimal accuracy metric.

As our second goal was to prove that a larger framework for information retrieval is possible with these models as its base, we sought to create a Flask-based API which would run the models in the background and serve to the user a list of all the abstracts he uploaded with the models annotation for each sentence. To do so, we combined the Flask python module, along with CSS [4] and HTML [5] code to create a minimalistic yet functional API to serve as the proof of concept we required. Our tool requires from the user to chose the model he would like to use for the inference based on the dataset it was trained and to input the abstract or abstracts he would like to analyze either as singleton or in bulk. Then, the API runs at inference mode using the loaded model and returns the abstract with the predicted labels at the start of each sentence. Although the API could be deployed online, e.g. on a cloud service, for the purposes of this thesis we set it up to run on localhost. More information on the APIs structure and operation can be found in appendix A.

[2] `https://www.tensorflow.org/`
[3] `https://keras.io/`
[4] `https://www.w3.org/TR/CSS/`
[5] `https://html.spec.whatwg.org/`

4. RESULTS

4.1 Source domain results

After each successful model train, we compile all the different accuracy measures into a single comprehensive report table. Each of those tables contains results in terms of precision (P), recall (R) and F-measure (F1) for each label and overall accuracy values, given in percentages, obtained by a model based on a specific test set. The formulas for calculating those values are given in appendix B. Accompanying these tables there are also confusion matrices. Their rows correspond to predicted labels while columns correspond to true labels, giving us an indication in absolute numbers of how well the model classified each label. Our measure of choice for determining the accuracy of the models is the weighted F1-score. The reason for this is the heavy class imbalance that was evident in table 3.4. Due to this we needed a metric that would take into account the total number of instances of each class.

4.1.1 Per model intra-dataset test set results

In the following tables 4.1 - 4.4, we record the results for each of our models given their own test set.

Table 4.1: Consolidated results for the 20K-Model

20K-Model results on 20K test set				
	P	R	F1	Sup
B	73.20	86.09	79.12	3077
C	97.59	96.59	97.09	4571
M	95.37	97.29	96.32	9884
O	78.69	57.61	66.52	2333
R	96.35	95.67	96.01	9713
Total				
acc			92.35	29578
macro	88.24	86.65	87.01	29578
weighted	92.41	92.35	92.20	29578

(a) These are the results given by our model,when trained on the 20K dataset and evaluated on its own test set. Presented as percentages are the values for precision (P), recall (R) and F-measure (F1), while the last column gives the Support (Sup). B stands for Background, C for Conclusions, M for Methods, O for Objective and R for Results.

20K-Model Confusion Matrix					
	B	C	M	O	R
B	2649	4	67	347	10
C	1	4415	10	3	142
M	49	9	9616	14	196
O	918	0	67	1344	4
R	2	96	323	0	9292

(b) This is the confusion matrix given by our model, trained on the 20K dataset. B stands for Background, C for Conclusions, M for Methods, O for Objective and R for Results.

Table 4.2: Consolidated results for the 200K-Model

200K-Model results on 200K test set				
	P	R	F1	Sup
B	79.50	82.73	81.08	2663
C	97.82	96.43	97.12	4426
M	96.52	97.51	97.01	9751
O	81.08	76.44	78.69	2377
R	96.92	96.84	96.88	10276
Total				
acc			94.08	29493
macro	90.37	89.99	90.16	29493
weighted	94.07	94.08	94.07	29493

(a) These are the results given by our model, when trained on the 200K dataset and evaluated on its own test set. Presented as percentages are the values for precision (P), recall (R) and F-measure (F1), while the last column gives the Support (Sup). B stands for Background, C for Conclusions, M for Methods, O for Objective and R for Results.

200K-Model Confusion Matrix					
	B	C	M	O	R
B	2203	3	44	410	3
C	21	4268	11	0	126
M	33	10	9508	14	186
O	509	0	50	1817	1
R	5	82	238	0	9951

(b) This is the confusion matrix given by our model, trained on the 200K dataset. B stands for Background, C for Conclusions, M for Methods, O for Objective and R for Results.

Table 4.3: Consolidated results for the 1M-Model

1M-Model results on 1M test set				
	P	R	F1	Sup
B	74.93	89.53	81.58	105064
C	96.62	96.28	96.45	189007
M	96.61	95.85	96.23	279152
O	86.04	64.78	73.91	86725
R	96.14	97.01	96.57	417100
Total				
acc			93.26	1077048
macro	90.07	88.69	88.95	1077048
weighted	93.46	93.26	93.18	1077048

(a) These are the results given by our model, when trained on the 1M dataset and evaluated on its own test set. Presented as percentages are the values for precision (P), recall (R) and F-measure (F1), while the last column gives the Support (Sup). B stands for Background, C for Conclusions, M for Methods, O for Objective and R for Results.

1M-Model Confusion Matrix					
	B	C	M	O	R
B	94063	34	1695	8695	577
C	128	181972	86	40	6781
M	2438	120	267580	373	8641
O	28434	33	1821	56183	254
R	477	6171	5797	10	404645

(b) This is the confusion matrix given by our model, trained on the 1M dataset. B stands for Background, C for Conclusions, M for Methods, O for Objective and R for Results.

Table 4.4: Consolidated results for the 100K-Model

100K-Model results on 100K test set

	P	R	F1	Sup
B	76.43	81.19	78.74	691
C	96.50	96.11	96.30	4881
I	83.20	75.59	79.21	1815
M	91.62	92.89	92.25	5389
O	95.38	92.65	93.99	2720
P	89.64	83.19	86.30	3569
R	93.63	96.98	95.28	11263
Total				
acc			92.46	30328
macro	89.49	88.37	87.87	30328
weighted	92.41	92.46	92.39	30328

(a) These are the results given by our model, when trained on the 100K dataset and evaluated on its own test set. Presented as percentages are the values for precision (P), recall (R) and F-measure (F1), while the last column gives the Support (Sup). B stands for Background, C for Conclusions, I for Intervention, M for Methods, O for Objective, P for Population and R for Results.

100K-Model Confusion Matrix

	B	C	I	M	O	P	R
B	561	0	0	6	113	8	3
C	1	4691	3	0	0	2	184
I	0	1	1372	214	0	91	137
M	13	0	128	5006	8	143	91
O	151	0	3	33	2520	11	2
P	7	8	94	164	1	2969	326
R	1	161	49	41	0	88	10923

(b) This is the confusion matrix given by our model, trained on the 100K dataset. B stands for Background, C for Conclusions, I for Intervention, M for Methods, O for Objective, P for Population and R for Results.

Considering the accuracy report tables (4.1a - 4.4a), we will at first examine the overall model accuracies using the weighted F1-score as discussed earlier. By comparing the F1-scores between our 20K and 200K model results and their equivalents presented by Jin et al. [1], we see that the percentage difference between their reported accuracy and ours is about 0.4% for the 20K (our model: 92.20, Jin et al.: 92.6) and 0.2% for the 200K (our model: 94.07, Jin et al.: 93.90) dataset respectively. Given the stochastic nature of the Adam optimizer and the fact that we don't know the specific hardware setup they used for training their model, we can safely assume that our results are comparable which means that our training process has worked as intended. This is an important confirmation, as we needed to be certain of the validity of our actions before moving to our novel datasets for training and comparison. Viewing the F1 scores of all our models side by side, it shows that having a larger dataset leads to better accuracy scores in general (20K w-F1: 92.20, 100K w-F1: 92.39, 200K w-F1: 94.07). However, having too large of a dataset can spoil the accuracy (1M w-F1: 93.18) in favor of better generalization. Another interesting finding that arose from looking at the precision and recall values for each label per different model, is that all the models seem to be very capable of accurately classifying sentences belonging to the class of *CONCLUSION*, while they are having some trouble distinguishing between *RESULT* and *METHOD* and more trouble distinguishing between *OBJECTIVE* and *BACKGROUND*. This can be also confirmed by the confusion matrices, where we see nearly a third to a fifth of the background sentences (depending on the model) are mislabeled as *OBJECTIVE*. This phenomenon can be attributed to a number of reasons. One major problem could be the presence of imbalanced classes

in the training dataset, like *INTERVENTION* and *POPULATION* which have an order of magnitude lower representation compared to the other classes. Another problem could be due to the discoursal and structural similarities between sentences belonging to some classes. One such case is illustrated on table 4.5, where we see the results of the 20K model on one of its own test sets abstract. We can see that sentence 3 has been miss-classified as *BACKGROUND* instead of *OBJECTIVE*. A possible cause to this, that is a separate problem as well, might be the golden-true labels themselves. As seen on table 4.6 sometimes the sentence annotators have trouble themselves correctly identifying the golden-true labels as is the case in the first three sentences which should have been labeled as *BACKGROUND*.

Table 4.5: Example for label confusion

#	Predicted label	Gold label	Sentence
1	BACKGROUND	BACKGROUND	Recurrent aphthous stomatitis (RAS) is the most common oral mucosal disease.
2	BACKGROUND	BACKGROUND	However , the available therapies for RAS only relieve symptoms and do not provide a cure.
3	BACKGROUND	OBJECTIVE	This study assessed the response to treatment with levamisole and low-dose prednisolone drug combination in patients with RAS.
4	METHODS	METHODS	Fifty RAS subjects were enrolled in the single-blind randomized placebo-controlled trial.
5	METHODS	METHODS	Study medications were administered thrice daily for 3 consecutive days/week for 3 consecutive weeks.
6	METHODS	METHODS	Patients in Group 1 received placebo , Group 2 received levamisole (50 mg) and Group 3 received levamisole (50 mg) and low-dose prednisolone (5 mg).
7	METHODS	METHODS	Patients were followed up for 60 days after treatment.
8	METHODS	METHODS	Response to treatment was assessed using the following clinical parameters : pain due to ulcers , number of ulcers/episode , size of ulcers , duration of ulcers , and frequency of ulcers (episodes/month).
9	METHODS	METHODS	MannWhitney U-test.
10	RESULTS	RESULTS	A statistically significant improvement was noted in all parameters except for the size of ulcers in patients treated with levamisole alone and with combination of levamisole and low-dose prednisolone.
11	RESULTS	RESULTS	There was no statistically significant improvement in the placebo group.
12	RESULTS	RESULTS	Both active groups had significantly better improvement when compared to placebo group , while there was no significant difference between the two active groups.
13	CONCLUSIONS	CONCLUSIONS	Levamisole alone and combination of levamisole and low-dose prednisolone are effective modes of therapy for RAS.

Table 4.6: Example for bad golden-true label annotation

#	Gold label	Sentence
1	OBJECTIVE	Depressive disorders are one of the leading components of the global burden of disease with a prevalence of up to 14 % in the general population.
2	OBJECTIVE	Numerous studies have demonstrated that pharmacotherapy combined with non-pharmacological measures offer the best treatment approach.
3	OBJECTIVE	Psycho-education as an intervention has been studied mostly in disorders such as schizophrenia and dementia , less so in depressive disorders.
4	OBJECTIVE	The present study aimed to assess the impact of psycho-education of patients and their caregivers on the outcome of depression.
5	METHODS	A total of 80 eligible depressed subjects were recruited and randomised into 2 groups.
6	METHODS	The study group involved an eligible family member and all were offered individual structured psycho-educational modules.
7	METHODS	Another group (controls) received routine counselling.
8	METHODS	The subjects in both groups also received routine pharmacotherapy and counselling from the treating clinician and were assessed at baseline , 2 , 4 , 8 , and 12 weeks using the Hamilton Depression Rating Scale (HDRS) , Global Assessment of Functioning (GAF) , and Psychological General Well-Being Index (PGWBI).
9	METHODS	Results from both groups were compared using statistical methods including Chi-square test , Fisher 's exact test , Student 's t test , Pearson 's correlation coefficient , as well as univariate and multiple regression analyses.
10	RESULTS	Baseline socio-demographic and assessment measures were similar in both groups.
11	RESULTS	The study group had consistent improvement in terms of outcome measures with HDRS , GAF , and PGWBI scores showing respective mean change of -15.08 , 22 , and 60 over 12 weeks.
12	RESULTS	The comparable respective changes in the controls were -8.77 , 18.1 , and 43.25.
13	CONCLUSIONS	Structured psycho-education combined with pharmacotherapy is an effective intervention for people with depressive disorders.
14	CONCLUSIONS	Psycho-education optimises the pharmacological treatment of depression in terms of faster recovery , reduction in severity of depression , and improvement in subjective wellbeing and social functioning.

4.1.2 Transition matrices

To better understand how the ”Label Sequence Optimization Layer” described in section 3.1 is encoding the possible sequence of the labels in each model, we have generated transition matrices, which are given in tables 4.7 - 4.10. They reveal the transition probabilities between subsequent labels, with the rows representing the label of the previous sentence and the columns representing the label of the current sentence. Looking at the various tables for the 5-class models we see that the starting sentence is interchangeable between *BACKGROUND* and *OBJECTIVE*. Then come the sentences of the *METHOD* label followed by *RESULT* and finally the *CONCLUSION*. Concerning the 100K model, we see that the exclusive classes *POPULATION* and *INTERVENTION* are placed in this order between *METHOD* and *RESULT* labels. All of these results are consistent with our expectations. They also explain the difficulty all the models have in distinguishing between the *BACKGROUND* and *OBJECTIVE* labels as well as the *METHOD* and *RESULT* labels. This is indicative of the negatives this rigid form of sequence encoding has despite an overall improvement of the results.

Table 4.7: Transition matrix for 20K model

5 classes	BACKGROUND	METHOD	CONCLUSION	RESULT	OBJECTIVE	END
START	0.49	0.01	0.00	0.00	0.50	0.00
BACKGROUND	0.52	0.25	0.00	0.02	0.08	0.12
METHOD	0.00	0.75	0.01	0.23	0.00	0.00
CONCLUSION	0.10	0.01	0.45	0.00	0.00	0.44
RESULT	0.00	0.00	0.24	0.76	0.00	0.00
OBJECTIVE	0.02	0.65	0.00	0.01	0.32	0.00

Table 4.8: Transition matrix for 200K model

5 classes	BACKGROUND	METHOD	CONCLUSION	RESULT	OBJECTIVE	END
START	0.42	0.01	0.00	0.00	0.57	0.00
BACKGROUND	0.51	0.29	0.00	0.02	0.12	0.06
METHOD	0.00	0.74	0.01	0.25	0.00	0.00
CONCLUSION	0.03	0.00	0.44	0.00	0.00	0.52
RESULT	0.00	0.00	0.24	0.76	0.00	0.00
OBJECTIVE	0.02	0.67	0.00	0.01	0.29	0.00

Table 4.9: Transition matrix for 1M model

5 classes	BACKGROUND	METHOD	CONCLUSION	RESULT	OBJECTIVE	END
START	0.57	0.00	0.00	0.00	0.42	0.00
BACKGROUND	0.60	0.29	0.00	0.03	0.07	0.00
METHOD	0.00	0.67	0.00	0.33	0.00	0.00
CONCLUSION	0.00	0.00	0.48	0.00	0.00	0.52
RESULT	0.00	0.00	0.23	0.77	0.00	0.00
OBJECTIVE	0.01	0.65	0.00	0.01	0.33	0.00

Table 4.10: Transition matrix for 100K model

7 classes	BACKGROUND	METHOD	CONCLUSION	RESULT	OBJECTIVE	INTERVENTION	POPULATION	END
START	0.48	0.00	0.00	0.00	0.52	0.00	0.00	0.00
BACKGROUND	0.60	0.26	0.00	0.03	0.10	0.00	0.01	0.00
METHOD	0.00	0.64	0.00	0.27	0.00	0.00	0.08	0.00
CONCLUSION	0.00	0.00	0.49	0.00	0.00	0.00	0.00	0.51
RESULT	0.00	0.00	0.23	0.76	0.00	0.00	0.00	0.00
OBJECTIVE	0.01	0.66	0.00	0.01	0.29	0.00	0.04	0.00
INTERVENTION	0.00	0.01	0.01	0.61	0.00	0.37	0.00	0.00
POPULATION	0.00	0.20	0.01	0.25	0.00	0.25	0.30	0.00

4.1.3 Per model inter-dataset test set results

As a final step, before moving to the different domain datasets, we tested the accuracy of the models when they are evaluated on a different test set. This would give us a good sense on the robustness of each model generalization capabilities, by using data that are closely related to their own before moving to completely different data. Table 4.11 presents the results of these tests. We have included the original papers results as a reference wherever it was provided to ascertain that the values are very close to ours. Due to having the two extra classes the 100K model test set could not be used to test the 5-class models, however the opposite was feasible.

As can be seen from the table, the 200K model seems to be the best overall. This is the second indication that there exists an optimal size for the training dataset in order to be as accurate as possible for sentence classification within a specific domain. This robustness however tends to spoil the results when one tries to generalize the model on different domains. Having a bigger or more diverse training set will prove to be more important in those situations.

Table 4.11: Consolidated accuracy results for all the models

Consolidated Results						
Trained On	**Tested On**	**Dev Size**	**Test Size**	**Dev score (F1)**	**Test Score (F1)**	**Jin et al. Score (F1)**
1M	1M	100000	100000	92.96	**93.18**	-
	200K		2500		91.65	-
	20K		2500		90.34	-
200K	1M	2500	100000	94.23	91.36	-
	200K		2500		**94.07**	93.9
	20K		2500		93.22	-
20K	1M	2500	100000	92.47	88.51	-
	200K		2500		91.66	-
	20K		2500		**92.20**	92.6
100K	1M	2500	100000	92.47	85.86	-
	200K		2500		82.91	-
	20K		2500		83.55	-
	100K		2500		**92.39**	-

Another interesting observation is the very bad accuracy results exhibited by the 100K model when evaluated on the other models' test sets. Although this might seem alarming at first glance, it was to be expected considering the fact that it incorporates two more labels in prediction that the golden-true labels of the other test sets do not account for. To illustrate this point more clearly we used an abstract that is part of the test set of the 200K dataset, and we run it individually through our API with the 100K model to obtain the results shown in table 4.12. As expected, the predicted labels are almost identical to the golden-true labels, however on top of the regular *BACKGROUND* - *OBJECTIVE* confusion the model has "incorrectly" identified sentence 8 as *POPULATION*. Upon closer inspection one can identify that this is actually correct, however since this is an abstract from the 200K dataset it could never have had this golden-true label in the first place making these cross evaluations less impactful.

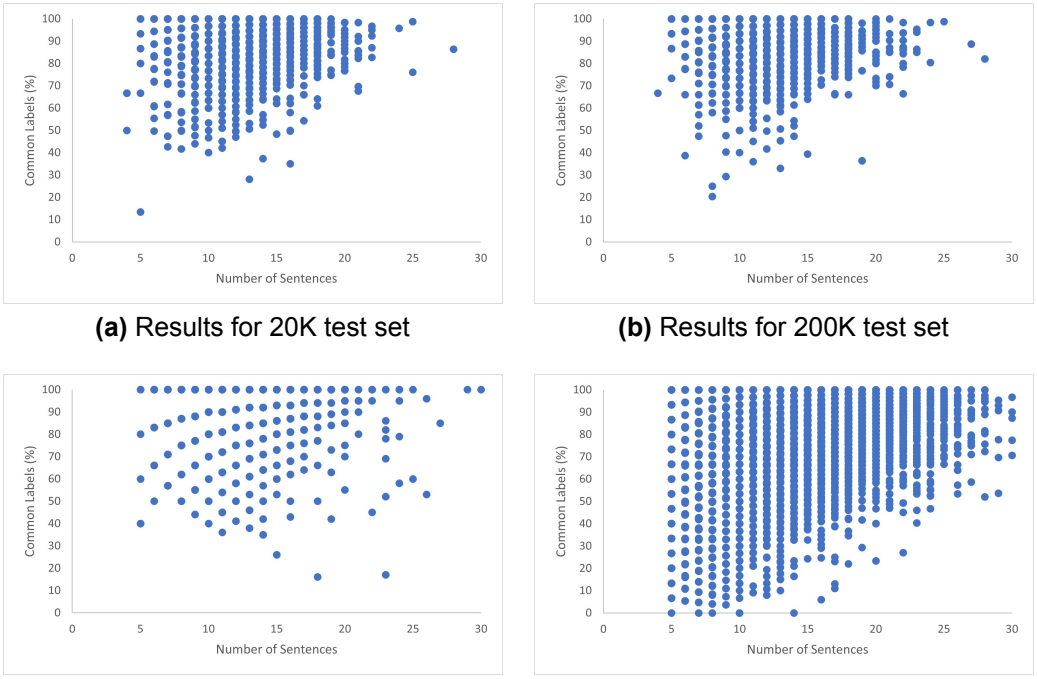
Table 4.12: Example for label changes based on the model used

#	Predicted label	Gold label	Sentence
1	BACKGROUND	BACKGROUND	The efficacy and safety of regular-strength beclomethasone dipropionate MDI prescribed within its recommended dosing range of 2 to 5 puffs three to four times daily has been well established in more than 25 years of worldwide use.
2	BACKGROUND	BACKGROUND	A more concentrated formulation delivering 84 microg per puff was developed to provide for a more convenient twice-daily dosing regimen.
3	BACKGROUND	OBJECTIVE	This randomized , single-blinded , positive and placebo-controlled , parallel-group , multiple-dose bioactivity study was conducted to assess the potential of a new beclomethasone dipropionate 84 microg double-strength metered-dose inhaler (Vanceril 84 microg Double Strength Inhalation Aerosol/Key) to cause hypothalamic-pituitary-adrenocortical axis suppression.
4	METHODS	METHODS	Beclomethasone dipropionate double-strength 84 microg was compared with beclomethasone dipropionate regular-strength 42 microg , orally administered prednisone , and placebo inhaler after 36 consecutive days of administration in adults with moderate asthma.
5	METHODS	METHODS	Beclomethasone dipropionate double-strength was administered as 5 puffs BID and beclomethasone dipropionate regular-strength was administered as 10 puffs BID for the same total daily dose of 840 microg of beclomethasone dipropionate.
6	METHODS	METHODS	Oral prednisone was administered by mouth at 10 mg once a day.
7	METHODS	METHODS	The potential for hypothalamic-pituitary-adrenocortical axis suppression was evaluated by an adrenocorticotropic hormone (ACTH) stimulation test using cosyntropin 250 microg in 500 mL normal saline infused over six hours on the 36th day of treatment.
8	POPULATION	METHODS	Sixty-four patients completed this study.
9	RESULTS	RESULTS	No clinically significant post-study findings were observed from physical examination , electrocardiogram , or clinical laboratory evaluation for any treatment group.
10	RESULTS	RESULTS	No serious or unexpected adverse events were reported.
11	RESULTS	RESULTS	On the 36th day of treatment , there was a significant (P <.01) difference in the plasma cortisol concentration response to cosyntropin stimulation between the prednisone and placebo treatment groups at the sixth hour of infusion.
12	RESULTS	RESULTS	There was no significant difference in the plasma cortisol concentration response to cosyntropin stimulation between the beclomethasone dipropionate double-strength and beclomethasone dipropionate regular-strength treatment groups and the placebo group.
13	RESULTS	RESULTS	In addition , comparison of the response between the beclomethasone dipropionate double-strength and beclomethasone dipropionate regular-strength groups showed no significant difference.
14	CONCLUSIONS	CONCLUSIONS	Beclomethasone dipropionate , administered either via a double-strength (84 microg/puff) or regular-strength (42 microg/puff) inhaler dosed at 840 microg/day showed no evidence of hypothalamic-pituitary-adrenocortical axis suppression in adults with moderate asthma.

4.1.4 Per test set common label results

In an effort to extract deeper insights about the test sets, we have generated scatter plots indicating the average percentage of common labels per abstract sentence length for each test set. The graphs are presented in Figure 4.1.

The first observation to be made is that the more sentences an abstract has the better the chances of each model achieving a higher accuracy in identifying the correct label. This holds true for all test sets, but it is more evident in the results for 1M test set (Fig 4.1c). A possible explanation for this behaviour could be that a higher number of sentences ensures a better abstract format, which the models can more easily recognize. Another interesting observation is that the results for the 20K, 200K and 100K test sets (Figs 4.1a - 4.1c) seem to indicate a more robust performance overall of each model compared to



(a) Results for 20K test set

(b) Results for 200K test set

(c) Results for 100K test set

(d) Results for 1M test set

Figure 4.1: Scatter plots indicating the percentage of average common results between golden-true and predicted labels across all models based on number of abstract sentences for each test set.

the results for the 1M test set (Figs 4.1d). This is evident in the fact that for the first three test sets the average accuracy of the models rarely drops below 40%, while on the 1M test set there can be seen cases where the accuracy is close to 0%. In an effort to explain this behaviour one can assume that one factor contributing to its appearance could be the volume of this laters test set. Having so many abstract could inevitably lead to the appearance of such bad results. As, however, an accuracy of 0% is rather extreme one should consider the possibility of those specific abstracts having bad golden-true labels. Although such a phenomenon is rare, it is not entirely impossible. Finally, having terrible abstract format is another valid consideration.

4.2 Target domain results

For the target domain tests we will report both the model accuracies as well as our other results for the annotated abstracts that we randomly selected from the sum of processed abstracts for each category.

4.2.1 Energy and Sociology domain per model results

Considering again the accuracy report tables (4.19a - 4.24a) first, we will focus on the weighted F1 scores to determine the overall model accuracies when faced with the novel abstracts.

As it is evident the accuracy of all the models has significantly dropped compared to the accuracies they all exhibited with the test sets of the source domain datasets. Although the results are not optimal, due to the fact that the number of abstracts considered is relatively small, we can observe from the tables that the models have a relatively good classification capacity. The best performing models for both domains seem to be the 100K and 1M, which were indicated as the most robust from table 4.11. By observing the confusion matrices one can identify a possible problem. Due to an apparent absence of easily identifiable abstract format traditionally good classified labels such as *CONCLUSION* and *METHOD* exhibit a larger than usual mislabeling. This was expected to some degree as real world data are vastly different to carefully curated data, however since we want to see if this tool could serve as part of a larger information summarization framework we will attempt to discover all the factors that contribute to this result and discuss them in depth.

The first and most important problem that contributes to the poorer sentence classification is the absence of a standardized scientific discourse that is similar between the biomedical abstracts used for training the models and the ones coming from the energy and sociology domain. In cases where the abstract had a format closer to the one expected by the ”Label Sequence Optimization Layer” the models performed best, to the point of even achieving 100% accuracy. However, in abstracts where this was not the case, the CRF tried to enforce the sequence it had encoded, forcibly ”discovering” incorrect labels to try and mimic the sequence of labels it was expecting. One such case was the energy abstract seen in table 4.13, where the results were spoiled by a combination of the non-standard discourse and the ambiguity in the way the sentences are tied which the models fail to recognise. This is clearly evident in sentences no. 3 and 4 where the wording used is more related to *BACKGROUND* but the meaning of the sentences is clearly towards *CONCLUSION*.

The results also showed that another problem which affects the model stability is the number of abstract sentences. From the scatter plots presented in figure 4.2, we can see that abstracts which are 10 - 15 sentences long exhibit more stable results. This is expected as it has also been previously observed in figure 4.1 and can be clearly observed in tables 4.14 - 4.15. More sentences generally relate to a better and more complete abstract format, in turn leading to better classification as shown by the example in table 4.17. How-

Table 4.13: Negative example of energy abstract with non-standard discourse

#	Sentences	Gold label	100K model Predicted label	1M model Predicted label	200K model Predicted label
1	"This article states basics of the "green economy" concept and the role of transferring from industrial to ecologically responsible development of renewable power sources.",	OBJECTIVE	OBJECTIVE	OBJECTIVE	OBJECTIVE
2	"The main aim of this article is to determine principle regularities that stipulate and restrict abilities of the BRICS countries to use the renewable energy sector in order to replace high-carbon consumption in economic and social sector.",	OBJECTIVE	OBJECTIVE	OBJECTIVE	OBJECTIVE
3	"Basic conclusions of this work are the following: 1) The aggregate of economic, social, ecological, and energetic problems faced by the modern civilization are directly interrelated.",	CONCLUSION	BACKGROUND	METHOD	RESULT
4	"That is why in order to preserve the environment and natural resources for future generations, it is necessary to refuse from resources of unsustainable and high-carbon areas of development.",	CONCLUSION	RESULT	RESULT	BACKGROUND
5	"2) The concept of green economy lies in the fact that needs of the humankind must be rationalized, above all, in the power context, through ensuring a reasonable refusal from using economically destructive carbons by replacing them with renewable power sources.",	CONCLUSION	RESULT	RESULT	RESULT
6	"3) A lot of European countries made the power breakthrough simultaneously developing traditional and renewable energetic.",	CONCLUSION	RESULT	RESULT	RESULT
7	"However, it is impossible to make the same conclusion in relation to the BRICS countries.",	CONCLUSION	RESULT	RESULT	RESULT
8	"In the BRICS coalition only one country – China - can be acknowledged as a leader in using the renewable energy sector.",	CONCLUSION	CONCLUSION	RESULT	CONCLUSION
9	"However, at the same time this country is an "anti-leader" in polluting the environment.",	CONCLUSION	RESULT	RESULT	CONCLUSION
10	"4) At the present time economy of the BRICS countries cannot be yet acknowledged as green.",	CONCLUSION	CONCLUSION	CONCLUSION	CONCLUSION
11	"However, along with this, India, China, RSA, Russia, and Brazil have a considerable natural, climatic, and geographical potential for efficient use of benefits of the renewable energy sector.",	CONCLUSION	CONCLUSION	CONCLUSION	CONCLUSION
12	"In the future it will allow them to transfer from the industrial and unsustainable vector of development to ecologically responsible development."	CONCLUSION	CONCLUSION	CONCLUSION	CONCLUSION

ever, as we see from table 4.18 all models retain their accuracy when the sentences are clearly worded with a format in mind. Along with those domain-specific problems, the models are still burdened by the problems they were exhibiting before. Namely they still exhibit for the most part *METHOD* - *RESULT* as well as *BACKGROUND* - *OBJECTIVE* label confusion, which is further amplified by the first problem.

Lastly, the final problem that contributes to errors in the results has to do with the sentence splitter used. Although the tool of choice in our case is a proprietary one that has superior performance compared to other generic choices (like the NLTK tokenizer), there are some edge cases (ex. word abbreviations and others) that are incorrectly causing it to decide that the end of the sentence has been reached as seen in table 4.16. Although this is generally a serious problem, especially for an automated tool, it can be somewhat avoided by requiring a more specific input format from the user. As such, it has lower severity factor than the others.

Table 4.14: Negative example of energy abstract

#	Sentences	Gold label	100K model Predicted label	1M model Predicted label	200K model Predicted label
1	This paper presents the issues of electromagnetic interactions in a four-circuit and dual-voltage power line.	OBJECTIVE	OBJECTIVE	OBJECTIVE	CONCLUSION
2	Such solutions are increasingly used in practice due to difficulties in land acquisition for the construction of new power lines.	BACKGROUND	OBJECTIVE	CONCLUSION	CONCLUSION
3	Lines of this type, however, have some disadvantages, incl.,	BACKGROUND	OBJECTIVE	CONCLUSION	CONCLUSION
4	the electromagnetic interactions between the circuits and voltages induced as their consequence.	BACKGROUND	OBJECTIVE	CONCLUSION	CONCLUSION
5	These issues are considered in relation to an existing four-circuit, 110 kV and 15 kV line.	METHOD	OBJECTIVE	CONCLUSION	CONCLUSION
6	Results of the studies of the interaction effects in a real system, and an analysis of selected ways to reduce the voltage induced in 15 kV line circuits are presented.	CONCLUSION	CONCLUSION	CONCLUSION	CONCLUSION

Table 4.15: Negative example of sociology abstract

#	Sentences	Gold label	100K model Predicted label	1M model Predicted label	200K model Predicted label
1	January 24, 2012, may not go down as a particularly noteworthy day overall, but for the growing sustainable food systems field it marked an important milestone.	BACKGROUND	BACKGROUND	BACKGROUND	BACKGROUND
2	On this day, the Community Food Security Coalition's venerable COMFOOD listserv (http://www.foodsecurity.org/list.html) announced it was separating job announcements that were routinely posted on the list into a new listserv, COMFOOD JOBS.The emergence of a dedicated vehicle for posting jobs in sustainable food systems is a coming-of-age event for our field.	OBJECTIVE	BACKGROUND	BACKGROUND	BACKGROUND
3	In the few short months since the listserv came online, over 400 jobs and related posts have gone on the list.	RESULT	BACKGROUND	BACKGROUND	BACKGROUND
4	The diversity of job titles, geographies, and education and experience requirements is extraordinary.	CONCLUSION	BACKGROUND	BACKGROUND	BACKGROUND
5	Consider that in just June of this year job announcements have been made for positions ranging from a driver for a mobile livestock program in Taos, New Mexico, to a healthy food access expert in California, to a business manager for a New York–based food systems consulting firm.	CONCLUSION	OBJECTIVE	CONCLUSION	BACKGROUND
6	Farms, businesses, and community-based organizations are looking for everything from interns to experienced experts in creating the new food system....	CONCLUSION	CONCLUSION	CONCLUSION	METHOD

Table 4.16: Negative example of energy abstract with bad splitting

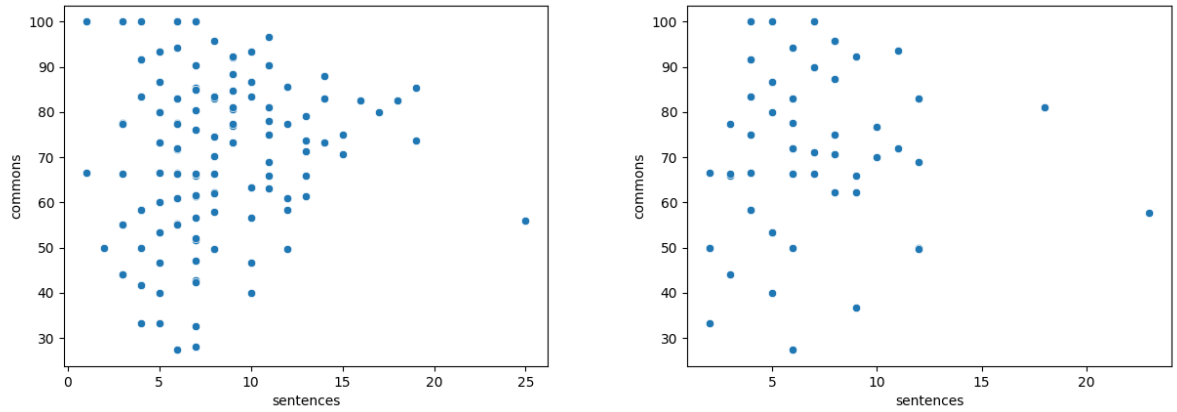
#	Sentences	Gold label	100K model Predicted label	1M model Predicted label	200K model Predicted label
1	"The desire to reduce carbon emissions due to transportation sources has led over the past decade to the development of new propulsion technologies, focused on vehicle electrification (including hybrid, plug-in hybrid and battery electric vehicles).",	BACKGROUND	BACKGROUND	BACKGROUND	BACKGROUND
2	"These propulsion technologies, along with advances in telecommunication and computing power, have the potential of making passenger and commercial vehicles more energy efficient and environment friendly.",	BACKGROUND	BACKGROUND	BACKGROUND	BACKGROUND
3	"In particular, energy management algorithms are an integral part of plug-in vehicles and are very important for achieving the performance benefits.",	BACKGROUND	BACKGROUND	BACKGROUND	BACKGROUND
4	"The optimal performance of energy management algorithms depends strongly on the ability to forecast energy demand from the vehicle.",	BACKGROUND	BACKGROUND	BACKGROUND	BACKGROUND
5	"Information available about environment (temperature, humidity, wind, road grade, etc.)",	BACKGROUND	METHOD	RESULT	METHOD
6	"and traffic (traffic density, traffic lights, etc.",	BACKGROUND	METHOD	METHOD	METHOD
7	"), is very important in operating a vehicle at optimal efficiency.",	BACKGROUND	RESULT	RESULT	RESULT
8	"This article outlines some current technologies that can help achieving this optimum efficiency goal.",	CONCLUSION	CONCLUSION	CONCLUSION	CONCLUSION
9	"In addition to information available from telematic and geographical information systems, knowledge of projected vehicle charging demand on the power grid is necessary to build an intelligent energy management controller for future plug-in hybrid and electric vehicles.",	CONCLUSION	CONCLUSION	CONCLUSION	CONCLUSION
10	"The impact of charging millions of vehicles from the power grid could be significant, in the form of increased loading of power plants, transmission and distribution lines, emissions and economics (information are given and discussed for the US case).",	CONCLUSION	CONCLUSION	CONCLUSION	CONCLUSION
11	"Therefore, this effect should be considered in an intelligent way by controlling/scheduling the charging through a communication based distributed control."	CONCLUSION	CONCLUSION	CONCLUSION	CONCLUSION

Table 4.17: Positive example of sociology abstract with optimal number of sentences

#	Sentences	Gold label	100K model Predicted label	1M model Predicted label	200K model Predicted label
1	"Minas Jaya is one of the villages adjacent to Sultan Syarif Hasyim Forest Park (Tahura SSH).",	BACKGROUND	BACKGROUND	BACKGROUND	BACKGROUND
2	"Tahura SSH is one of the conservation areas in Riau Province which is currently in critical condition due to forest encroachment, illegal logging, forest fires and illegal land conversion.",	BACKGROUND	BACKGROUND	BACKGROUND	BACKGROUND
3	"In order to restore it requires an approach that combines conservation and community empowerment.",	BACKGROUND	BACKGROUND	BACKGROUND	BACKGROUND
4	"One is the concept of Conservation Village.",	BACKGROUND	BACKGROUND	BACKGROUND	BACKGROUND
5	"The initial stage in Conservation Village development needs to be a priority class map of conservation.",	BACKGROUND	BACKGROUND	BACKGROUND	BACKGROUND
6	"The problems faced by partners to produce such maps require special expertise in spatial planning, mapping, and understanding of conservation village concepts.",	BACKGROUND	BACKGROUND	BACKGROUND	BACKGROUND
7	"Based on the mapping identification showing that data of the conservation priority area 1, were identified it consisted of 243.92 hectares, then the conservation priority area 2 consisted of 257.87 hectares.",	RESULT	RESULT	RESULT	RESULT
8	"Further, the conservation priority area 3 also identified and consisted of 504.28 hectares, moreover the conservation level 4 conservation area around 1,868.57 hectares, and conservation priorities 5 identified around 1,082.79 hectares.",	RESULT	RESULT	RESULT	RESULT
9	"Conservation program directives were linked to each priority of conservation classes.",	RESULT	RESULT	RESULT	RESULT
10	"It generally includes a good forest covers protection activities, enrich the land with tree crops.",	CONCLUSION	CONCLUSION	CONCLUSION	CONCLUSION
11	"Furthermore, critical land rehabilitation with agroforestry patterns has the choice of species and proportion of annual crops and trees adapted to the degree of land criticality and gradient."	CONCLUSION	CONCLUSION	CONCLUSION	CONCLUSION

Table 4.18: Positive example of small energy abstract with good format

#	Sentences	Gold label	100K model Predicted label	1M model Predicted label	200K model Predicted label
1	This study investigates the energy consumption-growth nexus in Algeria.",	OBJECTIVE	OBJECTIVE	OBJECTIVE	OBJECTIVE
2	"The causal relationship between the logarithm of per capita energy consumption (LPCEC) and the logarithm of per capita GDP (LPCGDP) during the 1965-2008 period is examined using the threshold cointegration and Granger causality tests.",	METHOD	METHOD	METHOD	METHOD
3	"The estimation results indicate that the LPCEC and LPCGDP for Algeria are non cointegrated and that there is a uni-directional causality running from LPCGDP to LPCEC, but not vice versa.",	RESULT	RESULT	RESULT	RESULT
4	"The research results strongly support the neoclassical perspective that energy consumption is not a limiting factor to economic growth in Algeria.",	CONCLUSION	CONCLUSION	CONCLUSION	CONCLUSION
5	"Accordingly, an important policy implication resulting from this analysis is that government can pursue the conservation energy policies that aim at curtailing energy use for environmental friendly development purposes without creating severe effects on economic growth.",	CONCLUSION	CONCLUSION	CONCLUSION	CONCLUSION
6	"The energy should be efficiently allocated into more productive sectors of the economy."	CONCLUSION	CONCLUSION	CONCLUSION	CONCLUSION



(a) Energy domain results.

(b) Sociology domain results.

Figure 4.2: Scatter plot indicating the percentage of average common results between golden-true and predicted labels across all models based on number of abstract sentences.

Table 4.19: Consolidated results for the 200K-Model on the Energy abstracts

200K-Model results on the Energy abstracts				
	P	R	F1	Sup
B	64.94	88.42	74.88	354
C	67.80	74.74	71.10	293
M	72.14	65.02	68.40	223
O	81.82	45.00	58.06	140
R	72.96	53.16	61.51	269
acc			69.04	1279
macro	71.93	65.27	66.79	1279
weighted	70.38	69.04	68.23	1279

(a) These are the results given by our model, when trained on the 200K dataset and evaluated on the Energy abstracts. Presented in percentage are the values for precision (P), recall (R) and F-measure (F1). B stands for Background, C for Conclusions, M for Methods, O for Objective and R for Results.

200K-Model Energy Confusion Matrix					
	B	C	M	O	R
B	313	13	11	6	11
C	48	219	8	2	16
M	35	14	145	6	23
O	56	8	10	63	3
R	30	69	27	0	143

(b) This is the confusion matrix given by our model, trained on the 200K dataset. B stands for Background, C for Conclusions, M for Methods, O for Objective and R for Results.

Table 4.20: Consolidated results for the 1M-Model on the Energy abstracts

1M-Model results on the Energy abstracts				
	P	R	F1	Sup
B	77.25	82.49	79.78	354
C	80.28	79.18	79.73	293
M	77.98	58.74	67.01	223
O	85.19	49.29	62.44	140
R	62.81	84.76	72.15	269
acc			74.43	1279
macro	76.70	70.89	72.22	1279
weighted	75.90	74.43	74.04	1279

(a) These are the results given by our model, when trained on the 1M dataset and evaluated on the Energy abstracts. Presented in percentage are the values for precision (P), recall (R) and F-measure (F1). B stands for Background, C for Conclusions, M for Methods, O for Objective and R for Results.

1M-Model Energy Confusion Matrix					
	B	C	M	O	R
B	292	24	10	9	19
C	1	232	2	0	58
M	29	11	131	3	49
O	45	8	9	69	9
R	11	14	16	0	228

(b) This is the confusion matrix given by our model, trained on the 1M dataset. B stands for Background, C for Conclusions, M for Methods, O for Objective and R for Results.

Table 4.21: Consolidated results for the 100K-Model on the Energy abstracts

100K-Model results on the Energy abstracts				
	P	R	F1	Sup
B	78.09	78.53	78.31	354
C	78.85	83.96	81.32	293
I	00.00	00.00	00.00	0
M	79.82	40.81	54.01	223
O	57.78	74.29	65.00	140
P	00.00	00.00	00.00	0
R	68.30	77.70	72.70	269
acc			72.56	1279
macro	51.83	50.75	50.19	1279
weighted	74.28	72.56	72.12	1279

(a) These are the results given by our model, when trained on the 100K dataset and evaluated on the Energy abstracts. Presented in percentage are the values for precision (P), recall (R) and F-measure (F1). B stands for Background, C for Conclusions, I for Intervention, M for Methods, O for Objective, P for Population and R for Results.

100K-Model Energy Confusion Matrix							
	B	C	I	M	O	P	R
B	278	16	1	15	38	1	5
C	6	246	0	0	7	0	34
I	0	0	0	0	0	0	0
M	30	13	4	91	27	5	53
O	22	8	0	1	104	0	5
P	0	0	0	0	0	0	0
R	20	29	0	7	4	0	209

(b) This is the confusion matrix given by our model, trained on the 100K dataset. B stands for Background, C for Conclusions, I for Intervention, M for Methods, O for Objective, P for Population and R for Results.

Table 4.22: Consolidated results for the 200K-Model on the Sociology abstracts

200K-Model results on the Sociology abstracts				
	P	R	F1	Sup
B	54.61	78.30	64.34	106
C	77.89	71.84	74.75	103
M	66.67	66.67	66.67	60
O	73.53	51.02	60.24	49
R	70.59	48.65	57.60	74
acc			65.82	392
macro	68.66	63.30	64.72	392
weighted	67.95	65.82	65.65	392

(a) These are the results given by our model, when trained on the 200K dataset and evaluated on the Sociology abstracts. Presented in percentage are the values for precision (P), recall (R) and F-measure (F1). B stands for Background, C for Conclusions, M for Methods, O for Objective and R for Results.

200K-Model Sociology Confusion Matrix					
	B	C	M	O	R
B	83	4	4	4	11
C	20	74	6	1	2
M	13	2	40	3	2
O	20	2	2	25	0
R	16	13	8	1	36

(b) This is the confusion matrix given by our model, trained on the 200K dataset. B stands for Background, C for Conclusions, M for Methods, O for Objective and R for Results.

Table 4.23: Consolidated results for the 1M-Model on the Sociology abstracts

1M-Model results on the Sociology abstracts				
	P	R	F1	Sup
B	82.41	83.96	83.18	106
C	86.52	74.76	80.21	103
M	75.00	70.00	72.41	60
O	66.67	73.47	69.90	49
R	69.41	79.73	74.21	74
acc			77.30	392
macro	76.00	76.38	75.98	392
weighted	77.93	77.30	77.40	392

(a) These are the results given by our model, when trained on the 1M dataset and evaluated on the Sociology abstracts. Presented in percentage are the values for precision (P), recall (R) and F-measure (F1). B stands for Background, C for Conclusions, M for Methods, O for Objective and R for Results.

1M-Model Sociology Confusion Matrix					
	B	C	M	O	R
B	89	4	1	10	2
C	2	77	2	2	20
M	5	3	42	6	4
O	9	2	2	36	0
R	3	3	9	0	59

(b) This is the confusion matrix given by our model, trained on the 1M dataset. B stands for Background, C for Conclusions, M for Methods, O for Objective and R for Results.

Table 4.24: Consolidated results for the 100K-Model on the Sociology abstracts

100K-Model results on the Sociology abstracts				
	P	R	F1	Sup
B	76.24	72.64	74.40	106
C	89.13	79.61	84.10	103
I	00.00	00.00	00.00	0
M	70.21	55.00	61.68	60
O	50.65	79.59	61.90	49
P	00.00	00.00	00.00	0
R	77.78	66.22	71.53	74
acc			71.43	392
macro	52.00	50.44	50.52	392
weighted	75.80	71.43	72.90	392

(a) These are the results given by our model, when trained on the 100K dataset and evaluated on the Sociology abstracts. Presented in percentage are the values for precision (P), recall (R) and F-measure (F1). B stands for Background, C for Conclusions, I for Intervention, M for Methods, O for Objective, P for Population and R for Results.

100K-Model Sociology Confusion Matrix							
	B	C	I	M	O	P	R
B	77	3	0	4	17	2	3
C	4	82	0	2	4	2	9
I	0	0	0	0	0	0	0
M	5	3	4	33	11	2	2
O	8	1	0	1	39	0	0
P	0	0	0	0	0	0	0
R	7	3	1	7	6	1	49

(b) This is the confusion matrix given by our model, trained on the 100K dataset. B stands for Background, C for Conclusions, I for Intervention, M for Methods, O for Objective, P for Population and R for Results.

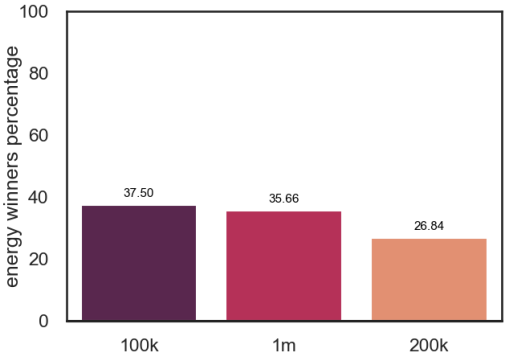
4.2.2 Best performing model per domain

In the previous section we saw that the more robust 100K and 1M models achieved better results in these novel domains. In this section we will elaborate more in depth about which model performs best and why.

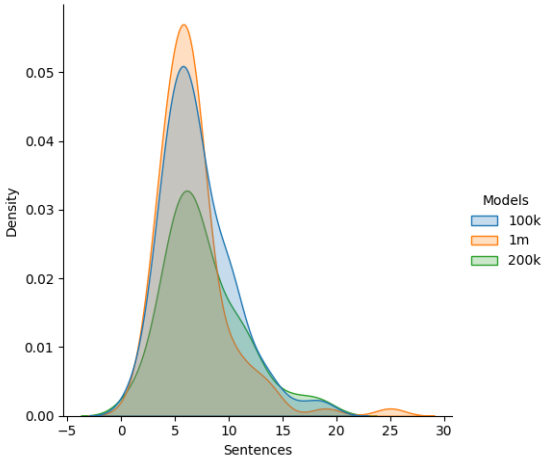
For each abstract we kept track of the model that correctly identified the most common labels. In the case of ties we considered both models as winners. We present the results in the two bar plots in figures 4.3a and 4.4a, given as percentages for better interpretability. These plots confirm the accuracy results from the tables in the previous section and give us some new insights. A very interesting observation is that the 100K model performed very well, even managing to come up ahead in the energy domain, even though the golden labels were lacking the two specific PICO labels exclusive to th 100K dataset, i.e. *POPULATION* and *INTERVENTION*. This could possibly be attributed to the fact that the ”Label Sequence Optimization Layer” of the 100K model is more flexible when it comes to the possible label sequence than the other models due to the two extra labels. The 1M model however is the clear winner in the sociology test set as it managed to achieve a percentage of common labels that is 26.83% and 41.47% increased compared to the 100K and 200K models respectively.

To further clarify as well as justify these results we drew a Kernel Density Estimation (KDE) plot for each domain, which are presented in 4.3b and 4.4b, where we illustrate the model wins based on the number of sentences in the abstracts. Unlike the scatter plots in figure 4.2a, here we distinguish each models performance instead of combining all the results together. The interesting finding in the energy graph, that could help explain bar plot results, is that the 100K model and to some smaller extent the 200K model have better accuracy than the 1M model for abstracts containing 10-15 sentences. As this is the range for the most stable and good results,it is apparent why the 100K model performed better even though overall it had a lower accuracy result than the 1M model.

From the above results it is evident that to increase the accuracy of the results one would need to diversify and enrich the training dataset as much as possible, which is something that holds true for most deep learning models, as well as try to fine tune the ”Label Sequence Optimization Layer” to account for a more diverse discourse.

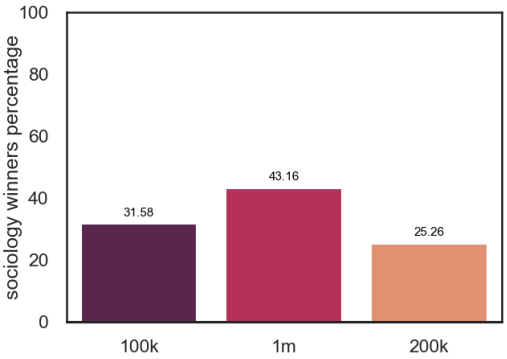


(a) Bar plot indicating the number of wins per model on the Energy dataset.

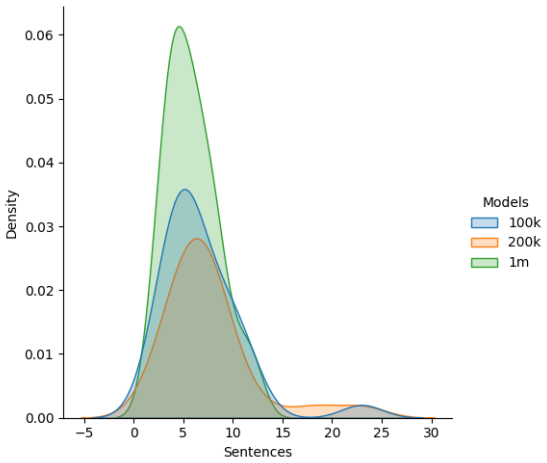


(b) KDE graph of model wins vs the number of sentences in the abstract.

Figure 4.3: Energy domain results



(a) Bar plot indicating the number of wins per model on the Sociology dataset.



(b) KDE graph of model wins vs the number of sentences in the abstract.

Figure 4.4: Sociology domain results

5. CONCLUSIONS

During the course of this thesis our goal was twofold. The primary goal was to explore the capabilities of a neural network model in correctly classifying abstract sentences of a specific research field as well as their ability to generalize to other fields. The secondary, yet equally important, goal was to create a web-based tool, which would encompass all the models we have created, to identify its potential both as a standalone application and as a basis to a larger information retrieval framework.

To achieve these goals, we started with deploying the current SOTA model in Sentence Classification of PubMed publications, which is based on a Hierarchical Neural Network architecture consisting of 4 layers for word embedding, sentence encoding, context enriching and label sequence optimization. Along with the two bench-marking datasets, PubMed 20K and 200K, we trained the network with two novel proprietary datasets both of which were curated with generalization in mind. One was the PubMed 1M which had a larger amount of data and the other was the 100K which had seven classes instead of the usual five, adding *"POPULATION"* and *"INTERVENTION"* as two clinical discourse specific classes referring to the PICO framework. The potency of the models was tested on a wealth of abstracts of both the source and target domains, the second of which were curated by us.

Based on our results, we can conclude that all the models had similar performance when it came to their own datasets with only the 200K and 1M models achieving somewhat higher results. As we saw from table 4.11, all the 5-class models retained most of their accuracy when bench-marked with the test sets of other datasets, while the 7-class model was the least successful mainly due to the other test sets not having the corresponding extra labels in their annotated abstracts. In the inter-domain results however we see that, although these abstracts lack the labels as well, the 100K model is tied with the 1M for highest accuracy. This is indicative of the generalization capability a better encoding of the label sequencing layer offers to the robustness of a model considerably smaller than the other best performer. Finally, based on the scatter plots in Fig 4.1 we can conclude that the larger an abstract was the better the models could identify labels of its sentences. This is most probably due to the label sequence optimization layer achieving better results due to having more sentences to work with. Another factor however is the fact that researchers who write larger abstracts have a better format structure in their mind, therefore making it easier for the model to identify it and proceed to the labeling of the sentences.

After extracting and analyzing all the results while using the API we have gathered a few of the potential hurdles one must overcome to integrate this web-based tool, which was created for the purposes of the current thesis, to a larger framework for information retrieval. These can be categorized into two broad categories, the datasets used for training and abstracts in general and the software itself.

In the first category, lack of a standardized scientific discourse across all fields is one, if not the biggest, problem that has detrimental effects on the capabilities of any such potential tool. Educating accordingly researchers of various fields on the advantages of

adhering to such a standardized discourse, would not only benefit their readership but also help promote their work through specialized automated frameworks. This can also have the added effect of reaching a point where the number of sentences in an abstract will not factor in the effectiveness of that abstract relaying information successfully. An easy way of implementing something similar would be to standardize the creation of structured abstracts. This way, by having a specific format with clear and intuitive rules (e.g Background - Objective - Method - Results - Conclusion) the readability and indexing capabilities of the abstracts by automated tools will increase by a fair amount. The other big problem in this category is the absence of a more generalized training dataset. The two proprietary datasets we used, proved that having a larger or a highly curated dataset can lead to increased model robustness to the point of even solving the problem of lack of a standardized discourse. This venture however, would require a lot of highly skilled people annotating and curating such a dataset.

For the second category, regarding the software, all the models exhibited some form of label confusion problems as can be seen from our confusion matrices. One reason behind this could something inherent in the way the models word embedding and sentence encoding layers work, as the problem mostly appears in sentences that have similar semantics. However another, greater problem could be the main source behind this. This problem could be attributed to errors the annotators make during initial labeling of the abstracts that are used for the training of the models. It is highly probable that the reason the models mix certain labels is because the annotators themselves were mixing those labels, therefore transferring this error to the models. This noise in annotation can only be addressed by educating annotators upon a robust set of rules for correctly identifying and labelling each potential sentence. Another problem in this category pertains to abstract tokenization in sentences. If the tool is to be truly automated and easy to use, the end user needs to be able to simply paste an abstract and get their results. Consequently, a highly efficient sentence tokenizer would be required to be included in the software. The proprietary one used as part of the API created for this thesis serves as an excellent starting point for any such solution. Another excellent consideration would be the ScispaCy [55] developed by the Allen Institute for Artificial Intelligence (AI2), which is a custom tokenizer trained on biomedical data specifically, enhancing spaCy's rule based tokenizer.

To further build upon the work that has been presented here, apart from the suggestions we listed above for the individual problems one can try a number of other things to improve the performance. One such solution could involve the use of some form of ensemble model combining various different models together to improve the overall accuracy. Another course of action could involve the retraining of all or some layers of the model, with new application specific data. This would help the model perform better on abstracts that have no related discourse to the ones used to initially train it. Lastly, one could try and merge similar labels wherever possible to help the model avoid label confusions, however this might require the design and implementation of a whole new rhetorical structure schema.

As a final remark, we could claim that the major contributions of this thesis are with regard to the two proprietary datasets and the API tool. The datasets showed that a solid network architecture can exhibit good generalizability with the right training set. Also, within the

limited time of conducting this thesis, we demonstrate that a basic yet fully functional user interface can be achieved for automatically generating the sentence labels for one or multiple abstracts. Supporting this with a web scrapping tool, e.g. using the Beautiful Soup [1] module from python to build such a tool, while adding more functionality after acquiring the sentence labels such as identification of specific words pertaining to identification of who did what in the abstract could lead to a framework that could really help researchers in their work. Although a fully optimized model and such a complicated framework were beyond the scope of our current work, our results can serve as a strong proof of concept for the further exploration of this idea in a way that would contribute to the more complex problems of text summarization and information extraction.

[1] https://www.crummy.com/software/BeautifulSoup/

APPENDIX A. THE FLASK API

An Application Programming Interface is a software intermediary that allows two applications to communicate with each other. In our case the we wanted to create a front end user interface that would run our models in the background in inference mode and then serve the results to the user. For this reason we employed the Flask Python module, which is a lightweight framework designed for quick and easy API builds that can be scaled up to complex applications.

A.1 App structure

The Python version required in the virtual environment for running the API is 3.7.7 while the rest of the required packages are included in a *requirements.txt* file for easy and fast deployment in any local or cloud-based server. A collapsed diagram of the file structure of the API can be seen on Fig A.1. All files except the ones pertaining to the models, those included to the *data*, *model* and *results* directories have been written or curated by us. More specifically, the *hsln_api.py* file is the main app file that contains the Flask-related code and serves as the link between the trained models and the front-End, whose code is contained in the *templates* directory. The *static* directory contains the supplementary CSS styling files for the pages displayed, along with the Java Script files for displaying the final results and the images used throughout the API. File *my_sentence_splitting.py* contains the proprietary code that is used in the API to tokenize abstracts into their sentences.

A.2 Front-End structure

The Front-End part of the tool consists of six HTML files as shown in Fig A.2. The *index.html* and the *abstract_input.html* files are the core pages that all API users will encounter and they have the outline for the home page - model selection and the abstract input forms respectively. Depending on the input the model runs in inference mode and the user is redirected to one of the three *results_display_{}.html* pages where he can view and download his results. Finally, the *show_graphs.html* page displays a scatter plot such as the ones presented in Fig 4.1 or Fig 4.2.

A.3 How to use

The flowchart diagram in Fig A.3 displays a rough estimate of the way the API is operating. As described in the previous subsection A.2 the first page the API user encounters is the home page where he will be prompted to select a model out of the pre-trained models available in order to label his chosen abstract. A depiction of the home page can be seen on figure A.4.

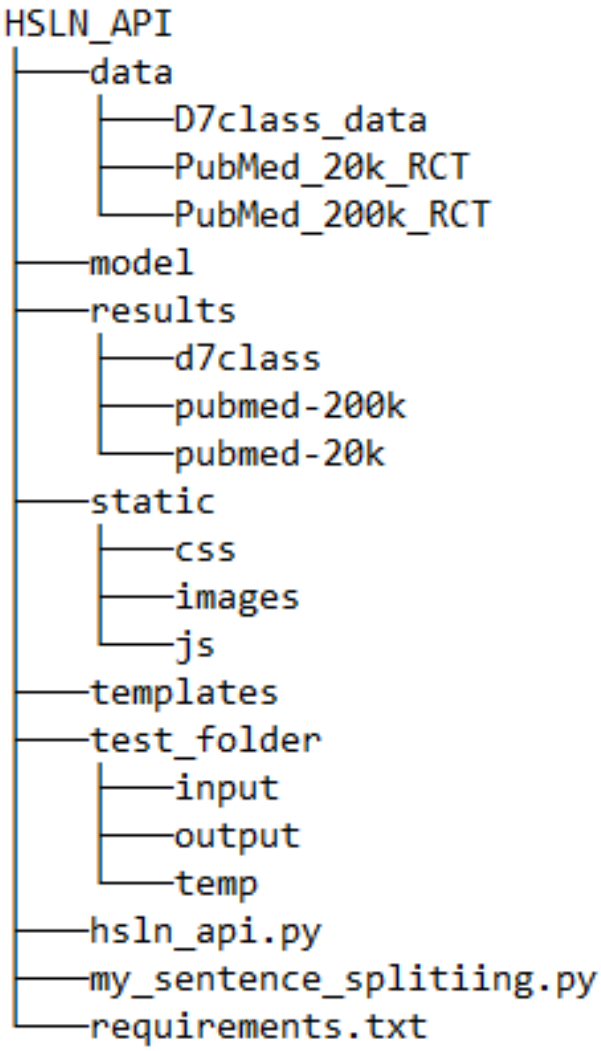


Figure A.1: File structure of the API

After the model selection has been made the user is redirected to the page seen in Fig A.5. There, he has three possible input methods. The leftmost method can be used to paste the text from a single abstract that the user would like to be labeled. This is the easiest, fastest and most straightforward method of input for users that want to try one or two abstracts for labeling. The middle method allows the user to input one or more abstracts for labeling using the json format. The abstracts can be as many as the user wants and they should be in the form of ”abstract name”:”abstract text” within the file. Otherwise this method is the same as the previous one, except it is best used when one wants to label many abstracts to avoid unnecessary back and forth. The rightmost method is reserved for when the user would like to test the results of a model against an already annotated abstract or abstracts. This is why in the end the user, apart from the usual download button he is prompted with a second ”Show Graphs” button that displays the scatter plot. A possible results page can be seen in Fig A.6.

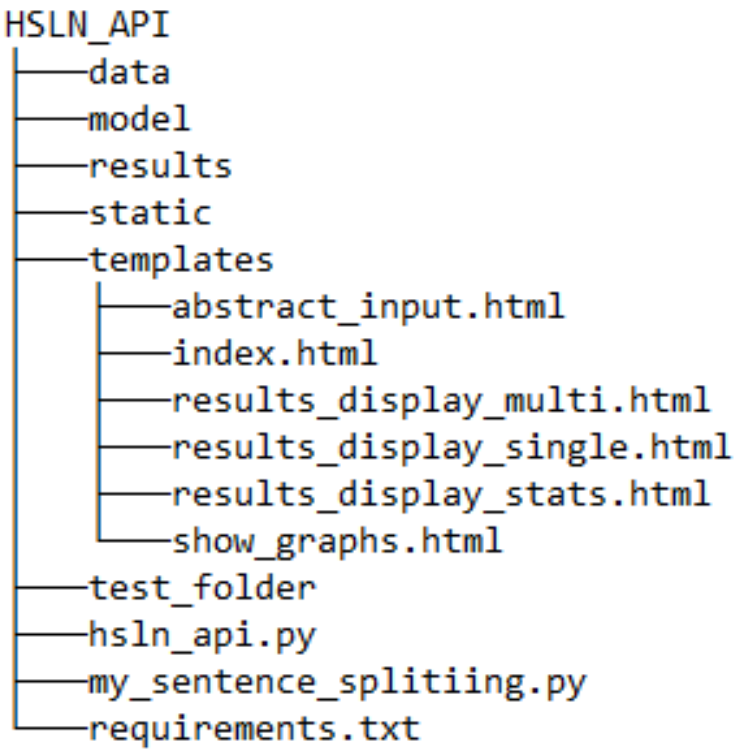


Figure A.2: File structure of the Front-End

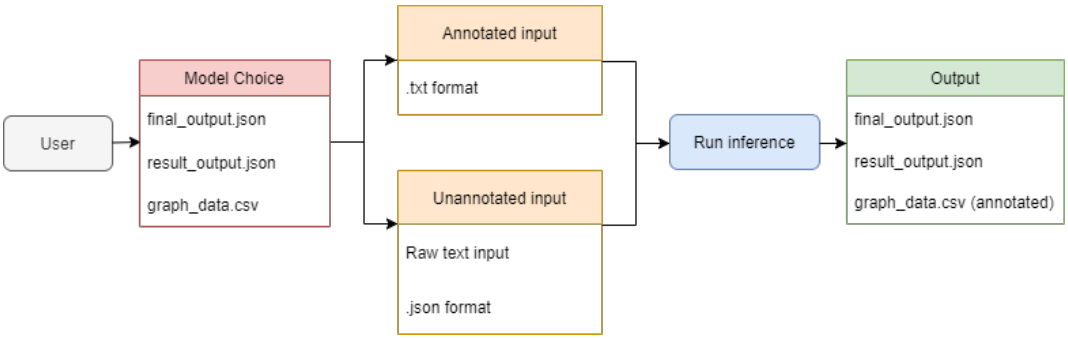


Figure A.3: Flowchart of the functionality of the API tool

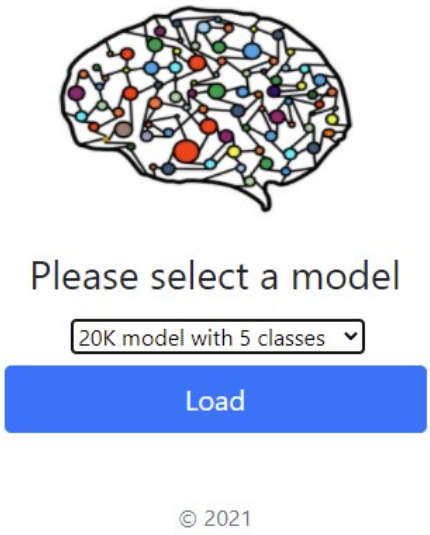


Figure A.4: Home page of our proprietary API

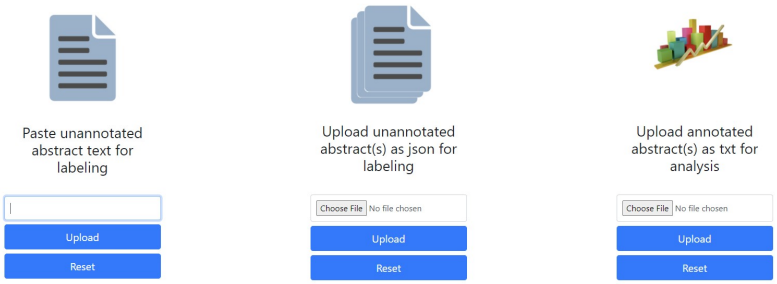


Figure A.5: Abstract input page of our proprietary API

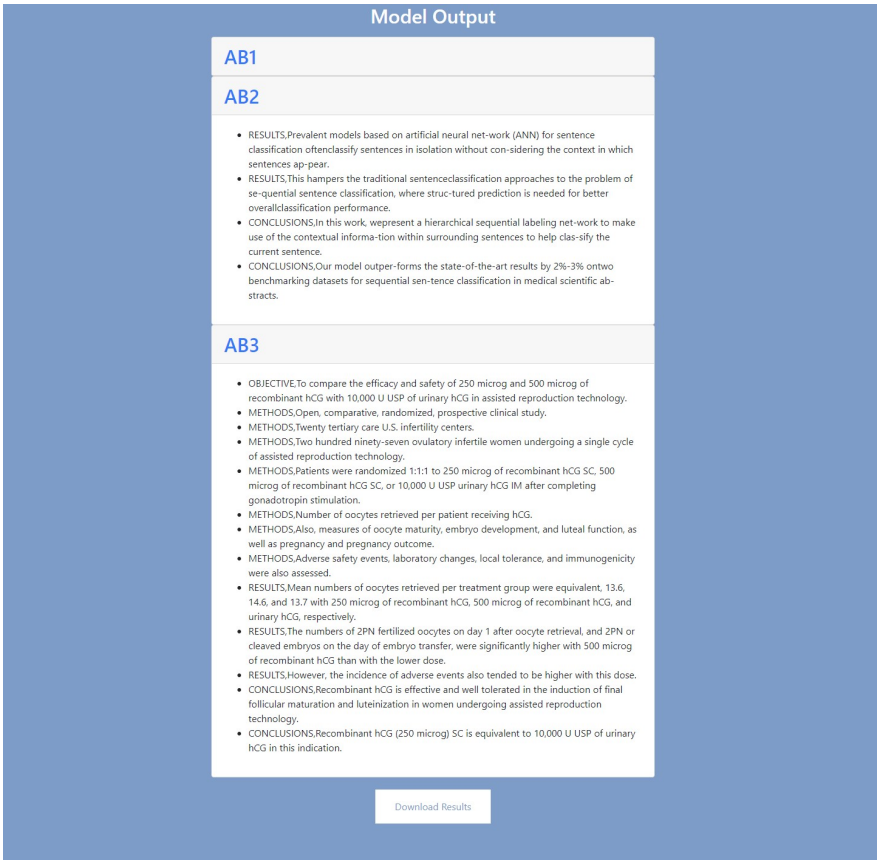


Figure A.6: Output page of our proprietary API

APPENDIX B. FORMULAS FOR THE RESULTING ACCURACY TABLES IN CHAPTER 4

$$Precision = \frac{TP_c}{TP_c + FP_c} \tag{B.1}$$

$$Recall = \frac{TP_c}{TP_c + FN_c} \tag{B.2}$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{B.3}$$

$$= \frac{TP_c}{TP_c + \frac{1}{2}(FP_c + FN_c)} \tag{B.4}$$

where TP_c = True positives of class c, FP_c = False positives of class c and FN_c = False negatives of class c

$$Acc = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C} Sup_c} \tag{B.5}$$

$$Macro = \frac{\sum_{c=1}^{C} F1_c}{C} \tag{B.6}$$

$$Weighted = \frac{\sum_{c=1}^{C} F1_c \cdot Sup_c}{\sum_{c=1}^{C} Sup_c} \tag{B.7}$$

where Sup_c = Support for each class c and C = total number of classes c

ABBREVIATIONS - ACRONYMS

ANN	Artificial Neural Network
API	Application Programming Interface
DL	Deep Learing
ML	Machine Learning
AI	Artificial Intelligence
CV	Computer Vision
SR	Speech Recognition
NLP	Natural Language Processing
CFG	Context-Free Grammar
CRF	Conditional Random Field
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
(bi)-LSTM	(bidirectional) Long Short-Term Memory
RCT	Randomized Controlled Trial
SOTA	State-of-the-Art
HSLN	Hierarchical Sequential Labeling Network
HNN	Hierarchical Neural Network
NB	Naive Bayes
SVM	Support Vector Machines
HMM	Hidden Markov Model
LDA	Linear Discriminant Analysis
GPT-3	Generative Pre-trained Transformer 3
20K (dataset)	PubMed 20K RCT dataset
200K (dataset)	PubMed 20K RCT dataset
1M (dataset)	PubMed 1 million abstracts dataset
100K (dataset)	PubMed 7-class 100K abstracts dataset

20K model	Model trained on the 20K dataset
200K model	Model trained on the 200K dataset
1M model	Model trained on the 1M dataset
100K model	Model trained on the 100K dataset
KDE plot	Kernel Density Estimation plot

REFERENCES

[1] D. Jin and P. Szolovits, "Hierarchical neural networks for sequential sentence classification in medical scientific abstracts," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (Brussels, Belgium), pp. 3100–3109, Association for Computational Linguistics, Oct.-Nov. 2018.

[2] M. Ware and M. Mabe, *The STM report: An overview of scientific and scholarly journal publishing.* 03 2015.

[3] K. Vougas, T. Sakellaropoulos, A. Kotsinas, G.-R. P. Foukas, A. Ntargaras, F. Koinis, A. Polyzos, V. Myrianthopoulos, H. Zhou, S. Narang, V. Georgoulias, L. Alexopoulos, I. Aifantis, P. A. Townsend, P. Sfikakis, R. Fitzgerald, D. Thanos, J. Bartek, R. Petty, A. Tsirigos, and V. G. Gorgoulis, "Machine learning and data mining frameworks for predicting drug response in cancer: An overview and a novel in silico screening process based on association rule mining," *Pharmacology & Therapeutics*, vol. 203, p. 107395, 2019.

[4] T. Thireou and M. Reczko, "Bidirectional Long Short-Term Memory Networks for Predicting the Subcellular Localization of Eukaryotic Proteins," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 4, pp. 441–446, Aug 2007.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[7] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," *Journal of the American Medical Informatics Association*, vol. 18, pp. 544–551, 09 2011.

[8] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependences in stochastic language modelling," *Computer Speech & Language*, vol. 8, no. 1, pp. 1–38, 1994.

[9] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001.

[10] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," in *J. Mach. Learn. Res.*, 2000.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR*, 2013.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv*, Jan 2013.

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *arXiv*, Oct 2013.

[14] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word Translation Without Parallel Data," *arXiv*, Oct 2017.

[15] M. Artetxe, G. Labaka, and E. Agirre, "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Melbourne, Australia), pp. 789–798, Association for Computational Linguistics, July 2018.

[16] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A Convolutional Neural Network for Modelling Sentences," *arXiv*, Apr 2014.

[17] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, pp. 179–211, Apr 1990.

[18] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, pp. 1735–1780, Nov 1997.

[19] A. Graves, N. Jaitly, and A. rahman Mohamed, “Hybrid speech recognition with deep bidirectional lstm,” *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 273–278, 2013.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” *arXiv*, Jun 2017.

[21] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (Online), pp. 38–45, Association for Computational Linguistics, Oct. 2020.

[22] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. 12 2020. `https://web.stanford.edu/~jurafsky/slp3/`.

[23] K. Doró, “The Rhetoric Structure of Research Article Abstracts in English Studies Journals,” *Prague Journal of English Studies*, vol. 2, pp. 119–139, Feb 2014.

[24] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[25] K. R. Weiss, T. M. Khoshgoftaar, and D. Wang, “A survey of transfer learning,” *Journal of Big Data*, vol. 3, pp. 1–40, 2016.

[26] F. Dernoncourt and J. Y. Lee, “PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts,” *arXiv*, Oct 2017.

[27] I. Amini, D. Martinez, and D. Molla, “Overview of the ALTA 2012 shared task,” in *Proceedings of the Australasian Language Technology Association Workshop 2012*, (Dunedin, New Zealand), pp. 124–129, Dec. 2012.

[28] W. Richardson, M. C. Wilson, J. Nishikawa, and R. Hayward, “The well-built clinical question: a key to evidence-based decisions.,” *ACP journal club*, vol. 123 3, pp. A12–3, 1995.

[29] X. Huang, J. Lin, and D. Demner-Fushman, “Evaluation of PICO as a knowledge representation for clinical questions,” *AMIA Annu. Symp. Proc.*, vol. 2006, no. 359-63., p. ;, 2006.

[30] L. McKnight and P. Srinivasan, “Categorization of Sentence Types in Medical Abstracts,” *AMIA Annu. Symp. Proc.*, vol. 2003, p. 440, 2003.

[31] J. Lin, D. Karakos, D. Demner-Fushman, and S. Khudanpur, “Generative content models for structural analysis of medical abstracts,” in *Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, (New York, New York), pp. 65–72, Association for Computational Linguistics, June 2006.

[32] P. Ruch, C. Boyer, C. Chichester, I. Tbahriti, A. Geissbühler, P. Fabry, J. Gobeill, V. Pillet, D. Rebholz-Schuhmann, C. Lovis, and A.-L. Veuthey, “Using argumentation to extract key sentences from biomedical abstracts,” *Int. J. Med. Inf.*, vol. 76, pp. 195–200, Feb 2007.

[33] G. Y. Chung, “Sentence retrieval for abstracts of randomized controlled trials,” *BMC Med. Inf. Decis. Making*, vol. 9, pp. 1–13, Dec 2009.

[34] Y. Kim, “Convolutional neural networks for sentence classification,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 1746–1751, Association for Computational Linguistics, Oct. 2014.

[35] R. Zhang, H. Lee, and D. R. Radev, “Dependency sensitive convolutional neural networks for modeling sentences and documents,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (San Diego, California), pp. 1512–1521, Association for Computational Linguistics, June 2016.

[36] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, "Very deep convolutional networks for text classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, (Valencia, Spain), pp. 1107–1116, Association for Computational Linguistics, Apr. 2017.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[38] A. Cohan, W. Ammar, M. van Zuylen, and F. Cady, "Structural Scaffolds for Citation Intent Classification in Scientific Publications," *arXiv*, Apr 2019.

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

[40] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," *arXiv*, May 2020.

[41] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[42] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *International Conference on Learning Representations*, 2020.

[43] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Ro{bert}a: A robustly optimized {bert} pretraining approach," 2020.

[44] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "ERNIE: Enhanced language representation with informative entities," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 1441–1451, Association for Computational Linguistics, July 2019.

[45] 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, pp. 1234–1240, 09 2019.

[46] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," *arXiv*, Mar 2019.

[47] J. Kringelum, S. K. Kjaerulff, S. Brunak, O. Lund, T. I. Oprea, and O. Taboureau, "ChemProt-3.0: a global chemical biology diseases mapping," *Database (Oxford).*, vol. 2016, p. bav123., Feb 2016.

[48] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu, "BioCreative V CDR task corpus: a resource for chemical disease relation extraction," *Database (Oxford).*, vol. 2016, 2016.

[49] N. Collier and J.-D. Kim, "Introduction to the bio-entity recognition task at JNLPBA," in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, (Geneva, Switzerland), pp. 73–78, COLING, Aug. 28th and 29th 2004.

[50] S. Bird, R. Dale, B. J. Dorr, B. Gibson, M. T. Joseph, M.-Y. Kan, D. Lee, B. Powley, D. R. Radev, and Y. F. Tan, "The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics," in *Proc. of the 6th International Conference on Language Resources and Evaluation Conference (LREC'08)*, pp. 1755–1759, 2008.

[51] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” *arXiv*, Oct 2013.

[52] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of Tricks for Efficient Text Classification,” *arXiv*, Jul 2016.

[53] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 1532–1543, Association for Computational Linguistics, Oct. 2014.

[54] J. Cohen, “A Coefficient of Agreement for Nominal Scales,” *Educational and Psychological Measurement*, vol. 20, pp. 37–46, Apr 1960.

[55] M. Neumann, D. King, I. Beltagy, and W. Ammar, “ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing,” in *Proceedings of the 18th BioNLP Workshop and Shared Task*, (Florence, Italy), pp. 319–327, Association for Computational Linguistics, Aug. 2019.