UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

INTERVENTIONS RECOMMENDATION: PROFESSIONALS' OBSERVATIONS ANALYSIS IN SPECIAL NEEDS EDUCATION

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS, MENCIÓN COMPUTACIÓN

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL EN COMPUTACIÓN

JAVIER LUCIANO MUÑOZ CARVAJAL

PROFESOR GUÍA:
FELIPE BRAVO MARQUEZ

MIEMBROS DE LA COMISIÓN:
ANDRES ABELIUK KIMELMAN
JOCELYN DUNSTAN ESCUDERO
ELIANA SCHEIHING GARCIA

SANTIAGO DE CHILE
2022

RESUMEN

El desarrollo de nuevas herramientas en los campos del Procesamiento del Lenguaje Natural y *Deep Learning* han permitido construir diferentes herramientas para ayudar a los estudiantes y profesores de la educación tradicional como parte de la Minería de Datos en la Educación. Sin embargo, existen problemas particulares presentes únicamente en la Educación Diferencial, como es el decidir las intervenciones más adecuadas para mejorar la calidad de vida y el rendimiento escolar de cada alumno.

En este problema los profesionales deciden estas intervenciones basándose en el diagnóstico de cada estudiante y distintas observaciones que dan médicos, fonoaudiólogos, psicológos y educadores diferenciales. Cada estudiante puede tener más de una intervención elegida y no todos los tipos de profesionales trabajan con cada estudiante, por lo que no siempre están presente sus observaciones. En esta tesis utilizamos diferentes enfoques y herramientas computacionales para resolver este problema para poder asistir a los profesionales en esta delicada tarea.

Comenzamos presentando un nuevo conjunto de datos, que contiene registros de estudiantes chilenos pertenecientes a la Educación Diferencial, el corpus SNEC. Mostramos la fuente de estos datos, junto con el proceso de su limpieza y anonimización. También presentamos un análisis exploratorio de las características de los datos y su distribución en determinados escenarios.

A continuación, presentamos el foco de esta investigación, experimentos diseñados y ejecutados con diferentes enfoques, uno de los cuales es, por ejemplo, el uso de *transformers* preentrenados de *Deep Learning* como parte de nuevas arquitecturas que utilizan un enfoque de aprendizaje multietiqueta multiinstancia para generar recomendaciones de intervenciones que ayuden a los profesionales de la Educación Especial a decidir las intervenciones más adecuadas para cada alumno.

Finalmente presentamos los resultados de estos experimentos y las conclusiones y reflexiones de este estudio. Junto con esto presentamos el marco de estudio para futuras investigaciones sobre el mismo problema.

Abstract

Development of new tools in the fields of Natural Language Processing and Deep Learning have made it possible to build different tools to help students and teachers in traditional education as part of the Educational Data Mining. However, there are particular problems present only in Special Need Education, such as deciding the most appropriate interventions to improve the quality of life and school performance of each student.

In this problem, Special Needs Education professionals decide the interventions based on the diagnosis of each student and different observations given by doctors, speech therapists, psychologists and special needs education teachers. Each student may have more than one intervention chosen and not all types of professionals work with each student, so their observations are not always present. In this thesis we use different approaches and computational tools to solve this problem in order to assist professionals in this sensitive task.

We begin by presenting a new dataset containing records of Chilean Special Needs Education Students, the SNEC corpus. We show the source of this data, together with the cleaning and anonymization process. We also present an exploratory analysis of the data features and their distribution in certain scenarios.

We then present the core of this research, several experiments designed and conducted with different approaches, one of which is, for example, using pre-trained Deep Learning transformers as part of new architectures that use a multi-instance multi-label learning approach to generate interventions recommendations to help Special Needs Education professionals decide the most appropriate interventions for each student.

Finally, we present the results of these experiments and the conclusions and reflections of this study. Along with this we present the framework for future research on the same problem.

A mi familia y amigos.

Saludos

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Technological advances in computational tools have made possible to develop diverse tools that help people to carry out specific tasks in many fields. One of these is the educational field where nowadays there are several tools to help students to obtain a better performance in different areas of their education, such as writing texts, evaluating their spelling, coherence and cohesion, or also recommendations of curricula changes or books based on their grades or behavior [1].

The development of the described tools could not be achieved without the huge progress in Natural Language Processing (NLP) [2], field that joins linguistics, computer science and artificial intelligence and allows machines to interpret free text instances written by humans. The tools that allowed this progress have been mainly pre-trained language modeling models, such as BERT [3]. These models are neural networks that have been trained with large datasets, such as all Wikipedia articles, and are able to learn from words and their context. These pre-trained models can be used in many tasks with smaller datasets by just little adjustment to learn about specific features.

However, there is a particular subfield of education that is probably the one that needs the most this kind of tools and there are not many of them, the Special Needs Education (SNE) field. In this research we propose to solve the problem of recommending interventions to SNE students knowing their diagnosis and professionals observations.

In Chile, SNE professionals write down the students' observations they have assisted in different documents in the first months of the school year that are then required by the Chilean ministry of education. At the end of the year they then write a new document containing the student progress, also indicating the interventions that helped and could have helped the student to improve his or her quality of life.

Currently, more and more SNE schools use different applications for a better management of the digitized students' documents, together with different records of the students, such as basic information and their diagnosis. This information has been mainly used to help professionals in diagnosing the students and applying interventions to help them, but not deciding these interventions. The choice of the most appropriate interventions is essential to improve the quality of life of each student since the professionals apply these interventions during the year and only one re-evaluation is made at the end of the year.

The professionals consider each student's information on an individual basis and their previous experience with other students to decide the best interventions. However, there are students with particularities observed by professionals that were not observed before in some schools but were observed in others, evaluating the performance of the interventions for these students. This information is no shared between the schools because of different reasons beyond the focus of this research.

With this in mind, we propose our problem as a technical NLP problem with multiple input types, as the professionals observations are free text instances and the students diagnosis can be coded as a number or a one-hot encoded vector. Furthermore, this problem can be faced with a multi-instance approach because the number of observations of each student is not always the same as not all the students need the support of the same type of professional. Besides these features, another challenge of this problem is that we can use a multi-label approach to model it since each student can have between 0 and "n" recommended interventions.

To solve this problem, we have built a new dataset of SNE students, collecting records from about 3,000 Chilean students enrolled in different SNE schools. The schools were previously consulted on the use of this data and with a commitment to anonymize the data before making it public. Each student has at least one observation type: SNE teacher, speech therapist, psychologist or doctor. Each student also has one diagnosis between 12 possible options and one or more associated interventions as the most appropriate for him or her, chosen from 14 options.

We then use shallow Machine Learning (ML) models with manually defined features and sparse n-grams representations of our text instances to explore this dataset and prove that we can use computational tools to make better interventions recommendation than just using naïve algorithms, such as just recommending the most common interventions. We use different input features combinations following a feature ablation methodology and observe which set features are the most useful.

Finally, we designed and conducted different end-to-end Deep Learning experiments using modern NLP models, which are based on neural network models previously trained with large datasets. These models are powerful representation learning tools since they also use the context of the text instances to make the representations.

We use pre-trained transformer model BETO [4] to make new representations of the professionals observations and use these representations in new neural network architectures with different settings that try to overcome the main challenges of the interventions recommendation problem. Besides, we also use similar architectures in a binary relevance scenario, decomposing the multi-label learning task into many independent binary learning tasks, to evaluate the performance of Deep Learning architectures in the prediction of each individual intervention.

1.1. Research Problem

The focus of this research is to study the validity of modeling the problem with Machine Learning techniques and to compare various approaches in order to establish which is more suitable for the interventions recommendation task.

Since we do not have previous results on this dataset, before conducting our Deep Learning results we performed an exploratory analysis to observe possible correlations between the different students attributes and several simple preliminary experiments to establish a performance baseline.

We try different Deep Learning (DL) tools to find a neural network architecture able to outperform ML models that use sparse n-grams representations combined with additional hand-coded features in shallow learning systems trying to solve many binary problems. As we mentioned before, one of the most useful tools of DL are the pre-trained models, but there are other tools not related to the NLP field that helped us to face this problem. For example, LSTM layers can be used to process a variable number of instances for a single input example, and using the bidirectional variant of this layer we can capture also possible correlations between the different types of professionals observations and the student diagnosis.

There are also additional challenges not related to the standard multi-instance multi-label learning (MIMLL) approach. First, the professionals observations do not have a length limit and the pre-trained models have a number of tokens limit, in the case of BERT this limit is 512 tokens. This can cause that if we want to use these models some observations should be truncated to be transformed into the embedding vector. However, there are different mechanism to get the representation of the entire document and even fine-tune these pre-trained models using the full content of the text features. For example, we can divide the text instances into smaller overlapped chunks, as shown in [5]. This way we can keep some context of the entire text in each chunk and then use a mechanism to generate a final representation, such as just calculate the average of each chunk representation, and use this as the entire text representation.

Another challenge present in this problem is that the set of labels (interventions in our particular problem) is not balanced, for example some labels are present in less than 30 % of the students and other are present in more than the 70 %. For the multi-label approach we used a non conventional loss function that can assign higher weights to the harder labels, that usually are the unbalanced ones. This loss function is the Focal Loss presented in [6].

1.2. Research Hypothesis

The hypothesis we postulate in this research is as follows:

"The problem of recommending interventions in special needs education can be successfully addressed using Machine Learning techniques. We also postulate that multi-instance multi-label learning approach can be a valuable solution for this problem."

1.3. Objectives

The main objective of this thesis is to try different ML architectures and evaluate their usefulness for the interventions recommendation problem. This objective includes to use different approaches, such as the multi-instance multi-label learning and the binary relevance approach, and also different methods and tools, such as using state-of-the-art pre-trained models as part of more complex architectures that are also capable of learning from possible correlations in labels and correlations between labels and instances.

A secondary objective of this research is to build an architecture that can generate appropriate interventions recommendations for SNE students based on their diagnosis and professionals' observations.

1.4. Methods

To achieve the main objectives of this thesis we divide our research in two parts. We first built the dataset, containing different students records with their diagnosis and professionals' observations that supported them used as the input of our classifiers and their interventions as the target of the predictions. Then, we conducted a data exploration analysis, observing the behavior of the student data, especially when a specific feature is present, such as, for example, associated interventions when there is a particular diagnosis or specific professionals observation types.

Secondly, we conduct several experiments, starting by designing experiments using a binary relevance approach to evaluate the performance of shallow Machine Learning models with sparse n-grams representations and hand-coded features in the most suitable interventions for each student recommendation problem. The second part of our experiments consisted of designing different end-to-end Deep Learning models to conduct new experiments to analyze the performance of neural network tools in the interventions recommendation problem with novel Deep Learning methods and tools.

1.4.1. Data and Exploratory Analysis

We first built the dataset collecting students' records from a web application and various digitized documents containing the professionals' observations and the interventions. We then cleaned the dataset and performed a data anonimization process, replacing sensitive data with special tokens.

Before conducting any experiment, we performed a statistical analysis of the different students data. We observed the percentage of interventions present in each diagnosis, the percentage of observation types present in each diagnosis and observation types in each intervention, first observing the individual behavior of the observations and then considering the different combinations of present observations. We also observed the behavior of the interventions when there was another intervention present, to analyze possible intervention correlations.

We then observed the most common words, bigrams and trigrams in the different observation types and interventions, also analyzing if the distribution changes depending on the students diagnosis. We also calculated the Pointwise Mutual Information (PMI) score, metric used in NLP to measure the co-occurrence of two words, of the bigrams to get an approach of the influence of some pair of words or tokens in the recommendation of each intervention.

1.4.2. Experiments

For the multi-label approach experiments of this research we use two main evaluation metrics. We use Jaccard similarity coefficient, metric used to measure similarity between finite sample sets by calculating the size of the intersection divided by the size of the union of the sample sets. We analyze the entire set prediction performance with this metric, since it allows us to measure the difference between the real set of labels and the predicted set of labels, but being more flexible than the standard accuracy, where if the predicted and real set of labels are different this is considered as a bad prediction even if it is just one label difference. We also use the F1 score to measure the prediction performance of each label, but considering the average of this measurement to make a general analysis. As most of our target labels are unbalanced classes (present in less than 15 % of the students or in more than 85 % of the students), we use the macro average of the above metrics to measure the performance in our multi-label experiments. In our individual label experiments we use the F1 score of the results for each experiment as our evaluation metric, since, as we previously mentioned, most of our target labels are unbalanced classes.

After the data exploration we conducted several experiments to evaluate the performance of shallow Machine Learning models in the interventions recommendation task and also learn from the results to then design more sophisticated experiments. Thus, we designed experiments using Machine Learning algorithms and techniques together with sparse n-grams representations of the professionals' observations to analyze the way that the different features of the students can help us to get interventions recommendations. In these experiments we use "feature ablation" method in order to achieve the mentioned objective. We also limited the dataset used in these experiments to only the training and validation dataset, since we should not learn from the results of the test set.

We then designed different Deep Learning experiments using pre-trained text transformers to get new representations of the professionals' observations in neural networks architectures. We used BETO [4] as text transformer to generate new observations representations in most of our Deep Learning experiments, since this transformer model has shown state-of-the-art performance in many tasks and also our main goal is to build a multi-instance multi-label learning architecture and not to find the best Deep Learning model that generates the best observations representations because addressing this problem is beyond the limits of our objectives.

One of the challenges using pre-trained transformers is that these models have input length limitations, BETO (and other BERT models) in particular, has a maximum of 512 tokens allowed for each input meaning that each text input should have around 500 words at most. As we show in Section 3.3, our observations have around 200 words, except for the special needs education teacher. This professional's observations contains around 500 and there are

many observations of these type of professional containing more than 500 words. Therefore, we considered using the fine-tuning techinque presented in [5] as a valuable approach to face the text length problem in our first Deep Learning experiments. We use this approach in our multi-instance multi-label approach and compare the performance using it to the performance obtained using simpler approaches, such as truncating the observations to certain number of tokens.

As the half of the interventions we are considering for our problem are unbalanced labels, the number of epochs needed for a model to learn which are the most relevant features to determine if a intervention should be recommended or not can be higher than the needed for balanced interventions. Since we are not using a large dataset, using a high number of epochs can cause the model overfits to these data and cannot be useful in unkown records. On the other hand, if we use a low number of epochs the model could not be able to learn how to generate unbalanced interventions. To deal with this problem we designed experiments using the Focal Loss loss function [6]. This loss function addresses the class imbalance by reshaping the standard cross entropy loss such that it down-weights the loss assigned to well-classified examples.

We also designed new architectures adding a BiLSTM layer to evaluate the performance when we use this layer over the embedding layer, similar to the way it is used in [7], using the word embeddings of the joined observations, and also when we use each observation type embedding as different sequence input for this layer.

Additionally, we designed experiments in a binary relevance scenario using similar architectures to the used in the above experiments. Thus, we can observe the performance of these models in a single label scenario and evaluate it individually.

1.5. Outline

This thesis is organized as follows. In Chapter 2 we present and describe the background of the scientific areas of this thesis, Educational Data Mining and Natural Language Processing. We also show state-of-the-art approaches in the different aspects of our technical problem, the use of Deep Learning architectures and methods in a multi-instance multi-label learning approach.

In Chapter 3 we state the practical problem of the interventions recommendation, giving details of the real world workflow of Special Needs Education students for a better understanding of our problem. We also present our data collecting, cleaning and anonymization process to build the SNEC corpus used in the experiments of our research. Finally, we also present the conducted exploratory analysis and its findings.

The conducted experiments description and results are presented in Chapter 4. We first describe the preliminary experiments using shallow ML methods and their results, followed by the Deep Learning experiments and results, first using the MIMLL approach and then in a binary relevance scenario.

Finally, in Chapter 5 we show the conclusions and contributions derived from our work.

In this chapter we also provide a future work methodology to develop further research.

Chapter 2

Background and Related Work

In this chapter we first present the background of the areas of knowledge on which this thesis is based: NLP and Educational Data Mining. We begin with a brief introduction to the reader in the area of Natural Language Processing (NLP) and Educational Data Mining. We then present different work related to the specific topics of our research: Multi-Instance Learning, Multi-Label Classification and then both topics together as a different problem. Finally, we address a brief discussion, arguing the usefulness of progress in the technical problem of using NLP techniques in a multi-instance multi-label learning problem and also the practical problem in the EDM area.

2.1. Scientific Disciplines

The challenges of this thesis are part of two well-defined scientific areas. First, one of the main challenges of our problem is to handle with the professionals' observations of each student, because these free text instances contain the more relevant information of each student. Thus, we need to know about different tools and techniques of the Natural Language Processing area to face this challenge. We introduce a review of the background of this discipline in the subsection 2.1.1.

Besides, we also need to know about the state-of-the-art of similar problems using computational tools to solve educational problems. These problems are handled in the Educational Data Mining field. Thus, we present a review of the background in the subsection 2.1.2.

2.1.1. Natural Language Processing

Natural Language Processing is a subfield of linguistics, computer science and artificial intelligence concerned with the interactions between computers and human language. The goal of this subfield is to design a computational tool capable of "understanding" the content of any document, including the contextual nuances of the language within them. This goal is pursued through progress on a variety of well-defined tasks that helps humans in different ways. Some of these tasks are:

- Sentiment Analysis: This task can be considered as a classification one, where each category represents a sentiment. Sentiment analysis provides a means to estimate the extent of a product or service and determine strategies to improve its quality [8].

- Question Answering: As its name suggests, the goal of this task is to provide relevant answers in response to questions proposed in natural language. This task is a specialized area in the field of Information Retrieval and presently is mainly used in the development of chatbots that provide virtual assistance in different industries [9].
- Language Modeling: This task consists of predicting the next word or character in a document using various statistical and probabilistic techniques to determine the probability of a given sequence of words occurring in the document. Currently, the progress in the Deep Learning field allow us to use different models or architectures designed to solve this task in other tasks, such as the mentioned above, using different Transfer Learning methods [10].

As we mention above, we can adapt different models designed for a particular task to another using different Transfer Learning techniques [11] and, in this way, obtain positive results for the target task. One of the most used techniques is to use models designed with a large number of parameters previously trained using a large amount of data [12]. Nowadays these models can be used with simple adaptations to get state-of-the-art performance in different tasks. These type of pre-trained models can generate word and document embeddings by also capturing information from the words context, generating better representations in most cases.

One of the most used pre-trained model in text classification tasks is BERT [3]. This model is a bidirectional encoder and its main feature is that it can capture information from the future and previous context of each word, producing better word and documents embeddings. Using this model, it is possible to obtain state-of-the-art performance across different problems by fine-tuning the model with just one additional output layer. However, it is important to remark, that the original version of BERT was trained on English documents, and our target task is in Spanish.

BETO [4] is a BERT-based model trained over Spanish free text instances obtained from the Wikipedia, movies subtitles and different public documents written in Spanish, that has shown state-of-the-art performance in different problems of Spanish text classification.

Today, there are different frameworks that facilitate the use of these pre-trained models. One of them is the Flair framework [13], which simplifies the use of these models and the construction of new architectures using these embeddings as inputs for new layers.

State of the art research have also shown that the use of additional layers over the pre-trained embeddings can improve the performance on NLP tasks. In this context, the authors of [7] use an additional LSTM layer over the embedding layer to learn summarization-specific features

2.1.2. Educational Data Mining

Educational Data Mining, as defined in [1], is a field of research that aims to answer educational questions and design strategies or applications to support students in their school development using computational tools and data-driven models. One of the problems this

field tries to solve is to support teachers by analyzing different student records, such as their grades, assistance or school behavior and providing feedback to then carry out strategies to help the students.

The authors of [14] build a software that uses learning information of the student (called "learning portfolio") to generate a list of recommended courses for each student, based on his or her preferences and performance in different areas. This was accomplished by defining a student model, extracting learning patterns to build a decision tree and finally generating an activity tree for each student.

In Special Needs Education there are also different tools that help teachers and other professionals to work with students. A.S. Drigas and R.-E. Ioannidou [15] show different tools and projects trying to solve particular problems present only in SNE, such as diagnosing SNE students using different A.I. learning methods. This work also presents different instruments that help professionals carry out the interventions chosen for each student.

2.2. Multi-Instance Multi-Label Learning

Considering that this work formulates the task of recommending interventions as a Multi-Instance Multi-Label Learning problem, we first discuss these two approaches separately as well as both together below.

2.2.1. Multi-Instance Learning

Multi-instance learning is a type of supervised learning where the training examples are labeled bags composed of instances instead of receiving the individually labeled instances. In single-class classification, a bag is labeled positive if there is at least one instance in it which is positive [16]. This approach is helpful when the instances are connected, like the professional's perceptions of a student in our identified problem. However, this might not hold in our problem domains, because the number of positive instances for a label might not be one and even need to certain instances to be positive. Hence, we could new approach such as a two-level approach as defined in [17], first exploring instance interactions and then structuring the instance space.

Therefore, the key challenge in MIL is to deal with the ambiguity of not knowing which of the instances in a positive bag are the actual positive examples and which ones are not. This approach is not easy to implement in neural network architectures since the learning model has to receive multiple instances as input. The model proposed in [18] uses a pre-trained model to compute instance-level representations for an input image (instance) in the bag and then use the set of representations as input for a convolutional neural network to produce multi-label predictions. This model achieve higher performance than models that generate feature vectors formed by simply combining instance level representations.

2.2.2. Multi-Label Classification

Many real world problems could be formulated as a problem of multi-label classification, since an image or a text input usually belongs to more than one conceptual class as shown

in [19]. In this approach each label is usually predicted independently from the others, so we cannot use any loss function to optimize a possible neural network architecture that solves our problem. The most used loss function for this approach is the cross-entropy loss function together with sigmoid activation function since sigmoid function gives independent probabilities for each class.

This approach also has been widely used to tackle image classification problems. However, a serious problem with existing approaches is that they are unable to exploit correlations between class labels, but the framework presented on [20] achieve this by modeling interactions between labels in an efficient manner with a correlated label propagation. The use of CNN models to solve multi-label classification problem on text has been most often used in cases where there are a large number of labels, but it is not our case. However, a close approach is used in [21] to classify text using a hidden bottleneck layer, allowing the architecture to learn better document representations and improve prediction accuracy.

Another approach to solve multi-label problems is to use classification chains. This approach is used in the framework presented on [22] and the model achieves competitive results, both in terms of predictive performance and time complexity. So far, this approach, for the best of our knowledge, has not been employed on neural networks architectures to solve multi-label classification problems, so we could perform various experiments to achieve better results.

It is also important to consider that in the real world, and in our particular problem, it is common to have an unbalanced distribution of labels, i.e., some high frequent labels and other infrequent ones, which may cause a learning model to fail in learning to correctly classify these labels in a few epochs or iterations. A new loss function is proposed in [6], the Focal Loss function, that can assign more weight to the hard-to-predict labels, which are usually the ones that are distributed as we mentioned before.

2.2.3. Multi-Instance Multi-Label Learning

The two approaches above give us the feeling our problem could be solved using both together. This mixed approach has been used in scene classification [23] because an image usually contains multiple patches, each of which can be described by a feature vector, and the image can belong to multiple categories since its semantics can be recognized in different ways. The authors of the mentioned research obtain positive results by using SVM algorithms, transforming the original multi-instance multi-label task into a single-instance multi-label one by converting each bag of instances example into a single instance example (i.e., mapping a bag of instances X_i into a single instance z_i using constructive clustering).

The use of Convolutional Neural Networks (CNN) to solve multi-instance multi-label learning problem has been more researched on the image classification problem. In [24] for example, the model outperform the state-of-the-art. Also, for the relation extraction task of NLP, in [25] the authors employ this approach to even outperform state-of-the-art of that moment by using neural networks. This could be the closest approach to solve a problem like ours.

2.3. Discussion

Recent advances in NLP, including pre-trained language models, can be useful in obtaining better observations representations than just using bag of n-grams representations, since modern text transformers also use the context of each word to produce the embeddings. This can be a valuable feature considering professionals' observations are free text instances describing different aspects of the student life and behavior.

We present a multi-instance multi-label learning approach suitable for our problem, even though most of these techniques are applied on the image classification field. Also, because of the non-uniform distribution of the interventions we are considering as labels, we think that using the previously mentioned Focal Loss function, can be a very useful tool to handle our problem.

Even though Educational Data Mining is a well-defined area that has made great progress in recent years and many schools are using computational tools developed in different research of this area, there are not as many computational tools for SNE as for traditional education. We also observed that, to the best of our knowledge, there are no computational tools to help professionals in choosing the best interventions for each student.

Thus, the problem studied in this thesis in addition to providing several technical challenges from the Machine Learning perspective, it is also novel and worth being studied from the educational point of view.

Chapter 3

Problem & Data

This chapter begins by stating the practical problem of the interventions recommendation, presented in Section 3.1. We then present the process of collecting, cleaning and anonymization of student records to build the dataset used in our experiments presented in Chapter 4, the SNEC corpus. The details of this process are shown in Section 3.2.

Finally, before conducting the core experiments of this thesis we perform an exploratory analysis on the SNEC corpus to observe the distribution, behavior and any possible interesting feature in this new dataset. The results of this exploratory analysis are shown in Section 3.3

3.1. The Interventions Recommendation Problem

Our problem lies in the field of Educational Data Mining, specifically in the area of Special Needs Education. It consists of automatically recommending interventions for students with different disabilities and behaviors, and its main goal is to help SNE teachers and professionals decide the best interventions for their students.

In Chile, Special Needs Education professionals write different documents that contain observations and relevant information of students they have worked with at the beginning of the school year. In these documents, the diagnosis of the SNE students is established from a set of pre-defined diagnoses. When students have more than one diagnosis they are diagnosed with 'Multiple cognitive deficit' and more details of the diagnoses are also settled down in the documents. Professionals also give details about the students behavior and relevant observed information together with interventions that can help students with their problems.

The professionals then decide the best interventions for each student by analyzing the information of the student's different documents and diagnosis. At the end of the year these professionals write a single document for each student, writing down the interventions that were useful for the student and also interventions that were not applied, but professionals think would have been useful had it been applied. Interventions that did not work are also written, but for this research we only use interventions that did work. The workflow is illustrated in Figure 3.1, showing the working process of the professionals with the students.

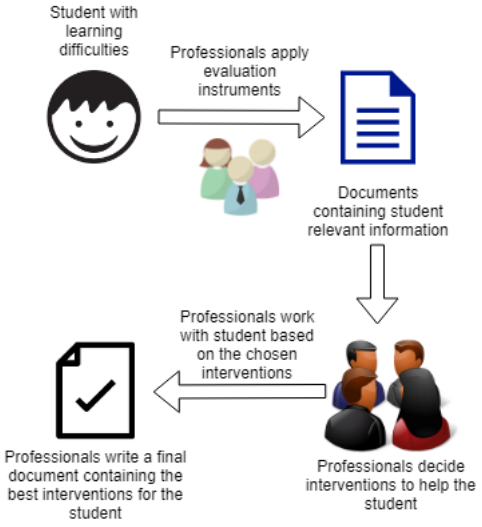

Figure 3.1: Students workflow.

3.2. Dataset Construction

We have built a new dataset for this research: the Special Needs Education Corpus (SNEC). We collected data from around 3,000 Chilean students enrolled in different SNE schools between 2018 and 2019, with the prior consent of the schools. These records were obtained from a web application used by schools to store students' basic data, such as their names, diagnosis, data of their relatives, among other data. The application is also used to store the results of different tests applied to the students together with documents written by professionals that support the students. In these documents the professionals write their observations about the students behavior, tests results and anything they consider relevant of the student.

Then, in a specific document named "Formulario Único de reevaluación" professionals write down the interventions that helped the student to improve their quality of life and school performance together with interventions that could have helped the student if were applied. In this research, we consider the diagnosis and the professionals observations as base input data and the interventions as the target data we want to predict for each student.

The students' diagnosis were obtained directly from the web application. For this research we have summarized the possible diagnosis of a student, grouping the same diagnoses but different severity into a single general diagnosis, for example students with the attention deficit disorder with symptoms of hyperactivity diagnosis and students without these symptoms were grouped into the attention deficit disorder, another example is the specific learning disorder group, that includes students with specific learning disorder in reading, writing and mathematical calculations tasks. This way, students have a diagnosis from 12 possible diagnoses and the number of students per diagnosis are shown in Table 3.1.

Diagnosis	# of students
Specific learning disorder	855
Specific language impairment	616
Intellectual disability	566
Borderline intellectual functioning	398
Attention deficit disorder	392
Autism spectrum disorder	111
Down's Syndrome	28
Hearing impairment - Hearing loss	27
Motor disorders	27
Multiple cognitive deficit	6
Global developmental delay	5
Visual impairment	4

Table 3.1: Number of students per diagnosis.

Professionals' observations are free text instances and were obtained from digitized documents required by Chilean laws for Special Needs schools. Depending on the document type there are observations of different professionals that have worked with the student. These professionals can be doctors, speech therapists, psychologists or Special Needs teachers. Each student can have 0 or more observations of each professional type, but must have at least one observation. This is due to the fact that depending on the student diagnosis certain professionals are required to support to the student from the beginning of the year, for example speech therapist professionals are the ones who work with the least number of diagnoses, since they tend to treat more specific problems. A student can also have more than one document containing observations from the same professional, being able to have more than one observation per professional type. For this work if the student has more than one observation per observation type, these are joined and used as a single observation. The number of students per observation type are shown in Table 3.2

Observation type	# of students
SNE Teacher Observations	2,560
Psychologist Observations	2,523
Doctor Observations	1,884
Speech Therapist Observations	525

Table 3.2: Number of students per observation type.

The best interventions for each student were obtained from another special document required by Chilean laws named "Formulario Único de reevaluación"[1]. This document is completed at the end of the school year and contains the applied interventions both those that

[1] https://especial.mineduc.cl/implementacion-dcto-supr-no170/formulario-unico/

worked and those that did not work as expected. This document also contains non-applied interventions that could have worked. For this study, we use interventions that worked and interventions that could have worked. Interventions are different actions that can help students to improve their quality of life and school performance by implementing treatments with certain professionals, adapting or changing some schools requirements, collaborative work with other students to promote the peer interaction, among other actions. Besides, interventions were grouped into larger groups to reduce the label space and get a better understanding of the interventions, for example, "Pedagogical support in subjects" intervention includes pedagogical support in any subject, such as maths, history and geography, biology or reading and writing in the native language. This way, students have 1 or more assigned "best interventions" from a set of 14 possible interventions.

Intervention	**# of students**
Access curricular adaptation	2,264
Involve the family in the process	2,048
Interdisciplinary support	1,441
Pedagogical support in subjects	1,314
Special Needs teacher support	1,311
Personal pedagogical support	1,240
Psychologist support	588
Speech therapist support	378
Peer tutoring	350
Objectives curricular adaptation	281
Occupational therapist support	153
General medical support	64
Neurological monitoring	63
Kinesiologist Support	32

Table 3.3: Number of students per intervention.

After collecting our data, we cleaned it by removing student records without any intervention or professional observation. This usually occurs when documents are not digitized but scanned versions of a physical document, being impossible to extract the data correctly and easily. In these cases we could not obtain the observations or interventions. We then anonymized our dataset by replacing the names of the students or their relatives in the observations with a special token, depending on who the name refers to, replacing the student name by "[ESTUDIANTE]" token or his or her relative's names by "[OTHER_NAME]" token. We also replaced schools and hospital names we found with "[ESCUELA]" and "[HOSPITAL]" tokens. This process was performed manually, since Spanish names can be easily confused with commonly used words, making it hard to use a simple find and replace process. We did not use any Machine Learning tool for this process because we should review the data anyway due to its sensitivity. However, we think it is necessary to make an exhaustive review of the anonimization process before making the dataset public.

A student is diagnosed with one of 12 different diagnoses and has different evaluations containing text observations from a doctor, a psychologist, and a special education teacher observations, also including speech therapist observations if the student requires it. Professionals then choose 1 or more interventions from 14 available. The problem ontology is illustrated in Figure 3.2.

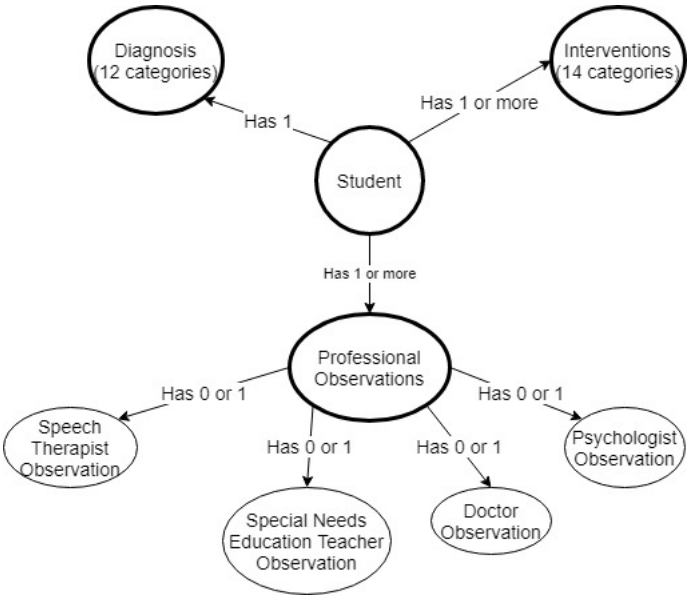

Figure 3.2: Problem ontology.

As we are dealing with a multi-label classification task, we employ the stratification method of [26] to generate training, validation and testing partitions dividing our data in a ratio of 3:1:1 respectively. This way, we use 1,836 students records for training and 612 for validation and testing.

3.3. Exploratory Analysis

We started our data exploration by observing the percentage of interventions per diagnosis. In this process we first observed different relations between the interventions and the diagnoses, for example, even though the "Involve the family in the process" intervention is the second most present in the students, it is the most common intervention per diagnosis, being present in all the diagnosis types with a large number of students on each one, while the most present intervention, "Access curricular adaptation", is not common in the non frequent diagnoses and "Objectives curricular adaptation" is present in greater quantity in students with these diagnoses. Furthermore, we observed that each diagnosis has a different interventions distribution and that this distribution is similar between frequent diagnoses while in non frequent diagnoses this distribution is different between them and also between all the interventions. Interventions distribution samples are shown in Figure 3.3 and all of them can be observed in Figure A.1.

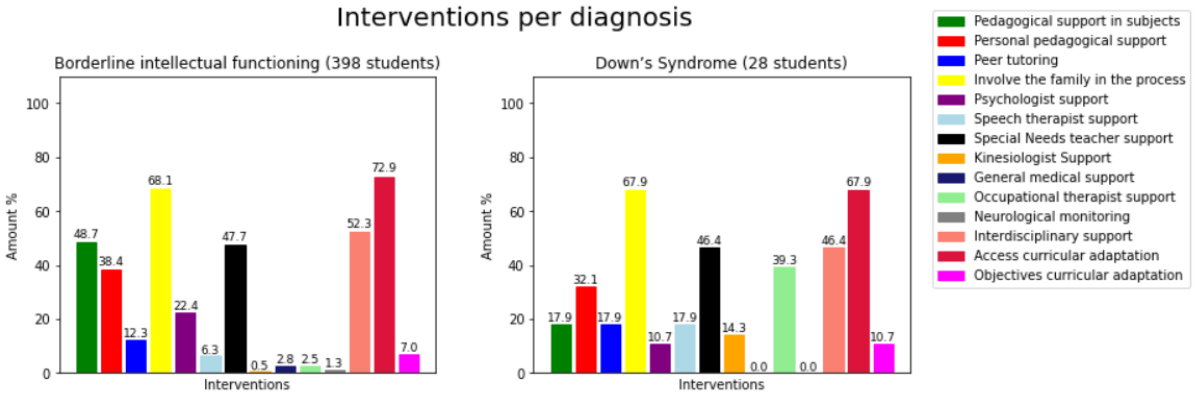

Figure 3.3: Interventions distribution for "Borderline intellectual functionning" and "Down's Syndrome" diagnoses.

We also analyzed the professionals observations distribution depending on the student diagnosis and observed that the psychologists and SNE teachers observations are present in most of the students, the medical observations are present in more than half of the students of any diagnosis and the speech therapist observations are present in a 20 % of the students of certain diagnosis and not present in the other diagnoses. Observation types presence distribution samples are shown in Figure 3.4 and all of them can be observed in Figure A.2.

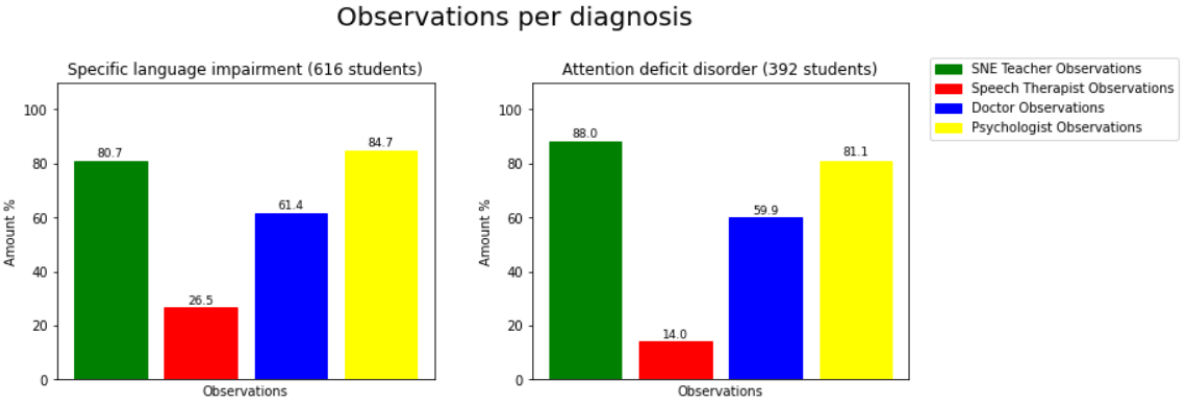

Figure 3.4: Observation types presence distribution for "Specific language impairment" and "Attention deficit disorder" diagnoses.

Besides, we analyzed the observations distribution depending on the interventions and observed that the distribution of the observations is similar to the distribution depending on the diagnosis. In addition, we observed that in the distribution of the set of present observations, and depending on the intervention, the most common set of present observations is to have the SNE teacher, psychologist and medical observation types, being present in the 50 % of the student with any intervention. The distribution of the other possible combinations of observations are different depending on the intervention analyzed. Observation types presence distribution for interventions samples are shown in Figure 3.5 and all of them can be observed in Figure A.3.

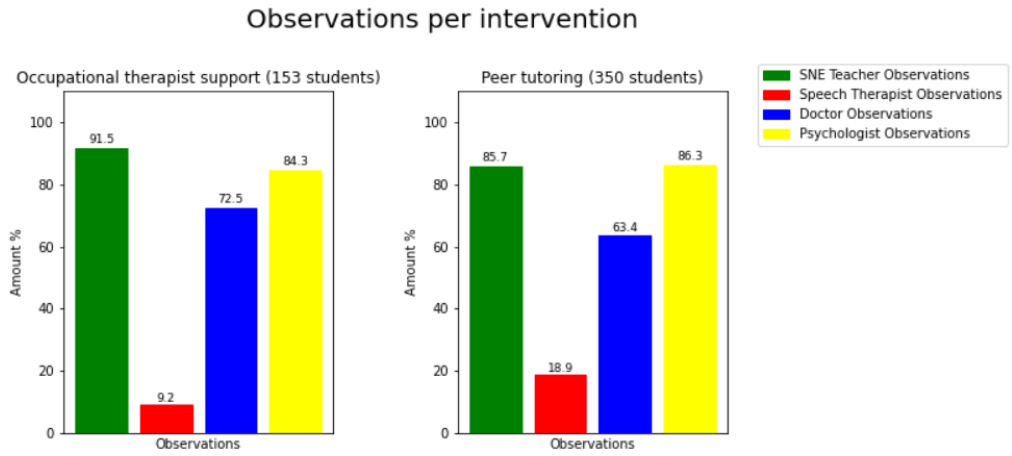

Figure 3.5: Observation types presence distribution for "Occupational therapist support" and "Peer tutoring" interventions.

We also observed the distribution of the interventions when a certain intervention is present and noted that the distribution is different depending on the observed intervention, but the most common interventions are also the most present interventions with different distributions and the rare interventions are the least present interventions. Interventions distribution for certain interventions samples are shown in Figure 3.6 and all of them can be observed in Figure A.4.

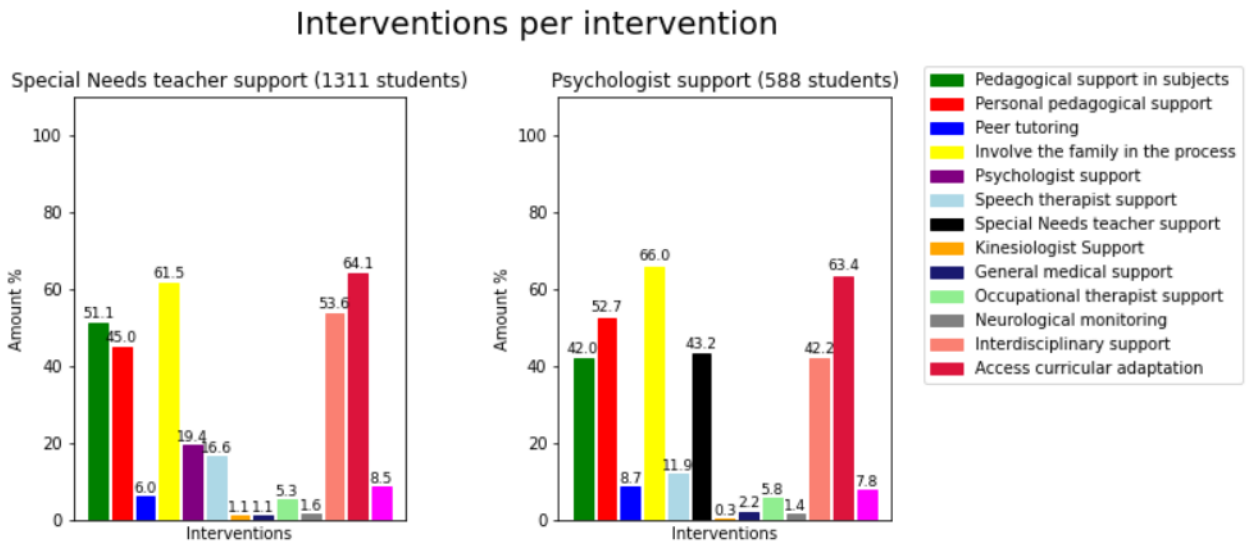

Figure 3.6: Interventions distribution when "Special Needs teacher support" and "Psychologist support" interventions are present respectively.

To conclude the data exploration, we analyzed the pointwise mutual information (PMI) score of different bigrams in the observations depending on each intervention. Thus, we observed that the top 10 highest scoring bigrams of each intervention are distinct from the others and also in some of them there are pairs that seem to be related to the objective of the intervention. For example, we observed that in the "Personal pedagogical support"

intervention the highest scoring bigrams are related to possible grammatical errors described by some professional, such as "/gue/ - /gui/", used by professionals to indicate difficulties on writing or pronouncing certain syllables. On the other hand, in "Involve the family in the process" intervention there are bigrams related to grammar errors, but also others related to social issues of the student, such as "empatizar - socializaciones", Spanish words for empathize and socialization.

Intervention	**Top #1**	**Top #2**	**Top #3**
Pedagogical support in subjects	apreciar-doblajes	empatizar-socializaciones	/avuelos-abuelos/
Personal pedagogical support	/gue/-/gui/	/b-/ll	/ge/-/gi/
Peer tutoring	adaptadas-alumnos/as	/g/-/j/	acompañada-corregida
Involve the family in the process	aprendizaje-esforzo	cantidades-incorporar	empatizar-socializaciones
Psychologist support	alumna-egresada	cantidades-incorporar	tolerancia-frustracion-lectura
Speech therapist support	aprendizaje-esforzo	expone-dramatiza	maneja-ordinalidad
Special Needs teacher support	/a/u/i/-copiarlas	/f/-/j/utilizar	/kl/-/tl/
Kinesiologist Support	envio-semanal	afeccion-pulmon	aerobica-disociacion
General medical support	afectan-significativamente	requiere-aprobacion	es-responsable
Occupational therapist support	abrigarse-desabrigarse	actos-serios	adentro/afuera-cerca/
Neurological monitoring	se-aprecian	accion-agregar	actitudes-descuidadas
Interdisciplinary support	empatizar-socializaciones	/b-/ll	/f/-/j/utilizar
Access curricular adaptation	apreciar-doblajes	aprendizaje-esforzo	perseverancia-compromiso
Objectives curricular adaptation	apreciar-doblajes	acortar-brecha	acotadas-pretende

Table 3.4: Top 3 words per intervention

Chapter 4

Interventions Recommendation Experiments

The core of this study is to evaluate and compare various Machine Learning approaches for our task at hand: recommending interventions to special needs students. In particular, we experiment with two types of approaches: shallow models, and Deep Learning models. We refer to "shallow" models as approaches based on manually designed features and logistic regression algorithm to build the classifier.

We first designed and conducted several preliminary experiments using shallow ML models to have a baseline reference and also to analyze the relevance of the students' data collected to build the dataset. Secondly, we use modern Deep Learning and NLP models and methods to design new experiments using pre-trained Deep Learning transformers.

This chapter is therefore organized as follows: in Section 4.1 we present our preliminary experiments using traditional ML models and methods followed by their results. Then, in Section 4.2 we first present the designed Deep Learning architectures and experiments using the MIMLL approach and their results, followed by the experiment details of the binary relevance scenario experiments and the results if these experiments. Finally, in Section 4.3 we discuss the results and findings of our experiments.

4.1. Preliminary Experiments

The objective of these experiments is first to use simple approaches and analyze if these approaches help us to get better interventions recommendations than naïve algorithms, such as just recommending the most common interventions to everyone, to validate the use of computational tools for this problem. Secondly, we also wanted to evaluate the relevance of each one of the students' features in order to generate the best interventions recommendation for each student.

4.1.1. Preliminary Classification Experiments

For these experiments we train logistic regression classifier models, provided by sklearn Python library, using a "L2" norm penalty, "liblinear" algorithm for the optimization problem and 10,000 iterations. We use the training partition to train the classifier and the validation

partition to evaluate the performance of these models, excluding the test partition in this stage of the experiments.

We consider different subsets of the feature space. The diagnosis of the students is one-hot encoded and is denoted with the letter 'D' in the experiments of Table 4.1 while the professionals observations are represented by a bag of n-grams representation, considering from 1-gram to 3-grams. The interventions are considered as binary features, showing the presence of each interventions and are denoted as 'AI' in Table 4.1 when all the interventions are used except for the predicted one and 'SI' when only one is used. We use a many-binary classification problems, considering each intervention as an individual problem to solve.

In this context, we first use the diagnosis of the student, his or her professionals' observations and also the interventions of the student (except for the one we are trying to predict) as input data for the model. We designed two versions of this experiment: first using the raw n-grams tokens as a single document, denoted as 'JOD' in Table 4.1. In the second version we use a special token as a prefix of the n-grams tokens to differentiate from which type of professional observation is each token. The second version of this experiment is used to observe possible differences in the information that each token can provide to the model depending on the professional that wrote the observation and is denoted as 'OST' in the results table.

We also designed experiments using only one of the other interventions as a binary attribute, keeping the OHE representation of the diagnosis and the observations sparse n-grams representation. In these experiments we also use both versions explained above. Additionally, we designed experiments using the OHE representation of the diagnosis and the sparse n-grams representation of professional observations as input data for new classifiers, without knowing any other intervention of the student.

We next designed experiments using the non-textual features, the OHE diagnosis and the interventions (excluding the one we are predicting on each experiment) as binary attributes, as input for the classification model for each intervention.

Our following designed experiments are focused on the professionals observations. Thus, in our next experiment we use only one type of professional observation representation, and OHE diagnosis' representation as input for the classifier. We use 'DO' to denote the doctor observations, 'PO' for psychologist observations, 'STO' for speech therapist observations and 'SNEO' for SNE teacher observations in Table 4.1.

Finally, the last experiments were designed considering only one of our features: diagnosis, observations or interventions, using also the special token to differentiate the observation types and the joined observations versions when we use the observations as features.

4.1.2. Preliminary Experiments Results

The preliminary experiments results are shown in Table 4.1. We first show the results of the dummy classifier using the naïve algorithm of predicting the majority class and then the results of the experiments explained in the previous section.

The experiment using all the features of a student, professionals' observations, diagnosis and other interventions, outperformed by a huge margin the results of the dummy classifier. Despite F1 score for each intervention were much better than the obtained with the dummy classifier, the Jaccard score obtained by analyzing the prediction of the whole set of interventions was not as superior as F1 score.

We also observed small individual differences between using the special token as prefix or not. We obtained slightly better results in three interventions using the token and worse results in the other. However, performance in the interventions in which the best results were obtained without using the tokens was better for a slightly greater margin than in those in which it was better to use the token.

The performance of the experiments using only one intervention as input together with the observations and diagnosis was slightly worse than that obtained in our previous experiment for every intervention by 0.01 on average in the F1 score measure.

Experiment names	**F1**	**Jaccard Score**
Baseline (Dummy classifier)	0.44	0.34
Only diagnosis (D)	0.47	0.38
All other interventions (AI)	0.51	0.39
Joined Observations Document (JOD)	0.64	0.44
Observations with special token (OST)	0.62	0.44
JOD + D	0.64	0.46
OST + D	0.62	0.44
DO + D	0.5	0.4
PO + D	0.55	0.4
STO + D	0.48	0.38
SNEO + D	0.63	0.43
AI + D	0.53	0.4
JOD + D + SI	0.64	0.46
OST + D + SI	0.63	0.45
JOD + D + AI	0.64	0.47
OST + D + AI	0.64	0.45

Table 4.1: Macro-averaged preliminary experiments results.

In the experiment using observations, diagnosis and only one intervention we also analyzed the weights the classifier assign to each feature for each intervention and observed that for certain interventions there are interventions used as binary attributes showing a high positive weight and in the same way, for other interventions there were interventions used as a binary attribute with a high negative weight as well, as shown in Table 4.2 and Table 4.3, showing the top positive and negative features using "Special Needs teacher support" intervention as a binary attribute in two different interventions.

The interventions used as binary attributes with a high negative or positive value were mainly most balanced interventions (present in more than 15 % of the students and less than 85 % of the students) and we did not observe a pattern on which interventions show correlation indications, since there were balanced and unbalanced interventions showing better performance when certain interventions were used as binary attributes.

Top #	**Feature**	**Weight**
1	Special Needs teacher support (Intervention attribute)	-1.12
2	'salud' (Spanish word for 'health')	-0.73
3	Specific language impairment (OHE diagnosis)	-0.70
4	'normas' (Spanish word for 'norms')	-0.49
5	'muestra' (Spanish word for 'show')	-0.49

Table 4.2: Top negative features for "General medical support" using "Special Needs teacher support" intervention as a binary feature.

Top #	**Feature**	**Weight**
1	Specific language impairment (OHE diagnosis)	2.28
2	Special Needs teacher support (Intervention attribute)	0.81
3	'palabra' (Spanish word for 'word')	0.80
4	'buena disposicion' (Spanish phrase for 'readiness')	0.79
5	'estructura' (Spanish word for 'structure')	0.78

Table 4.3: Top positive features for "Speech therapist support" using "Special Needs teacher support" intervention as a binary feature.

The performance using only professionals' observations and OHE diagnosis was equal to or worse than the obtained using also a single intervention as a binary attribute. We observed that in most interventions we obtained better performance using the joined professionals observations as a single document than using special tokens to differentiate the professional that wrote each observation. However, there are certain interventions, in both groups balanced and unbalanced interventions, showing better performance using these tokens, the same three interventions that followed this rule in the previous experiments.

The results obtained without using the professionals observations were worse than the obtained using any of the previous settings for all the interventions by 0.1 on average in the F1 score measure. These results indicate that the most relevant features of the students are the professionals' observations.

The performance using just one type of observation together with the OHE diagnosis was worse than in any of the previous experiments using textual features, except when using the SNE teacher observation, but we also observed that depending on the analyzed intervention, using certain types of observations we obtained better performance than using the other types, with the SNE teachers' observations being the most useful in most of the interventions.

The best results using only one attribute were obtained when using the joined professionals' observations, both in balanced and unbalanced interventions, while the worst results were obtained using only the doctor observations.

4.2. End-to-End Deep Learning Experiments

In this section we conduct several experiments to validate our multi-instance multi-label learning approach using different Deep Learning architectures and tools. We also conduct additional experiments using a many-binary label approach to evaluate the performance of Deep Learning architectures individually for each intervention. Thus, we first present the architectures and results of the MIMLL approach models experiments. Secondly, we describe the architectures used for the many-binary label approach and the results of the experiments. Finally, discussions of the results and findings of these experiments are presented.

4.2.1. MIMLL Experiments

We designed different architectures to address this approach, using Deep Learning tools to deal with the transformer limitations of the text inputs and the unbalanced distribution of the interventions.

The simplest designed architecture for this approach was using the concatenation of each students' observations as a single text input and using a different number of tokens limits to truncate the text to then use it to fine-tune BETO and build the classification model. We conducted this experiment using 200, 300, 400 and 500 as maximum tokens and a binary cross-entropy loss function to normalize the model outputs.

We also conducted experiments using the approach presented in [5] to fine-tune BETO with the full content of the observations. We used both approaches of the paper, first considering overlapped chunks of the joined observations as different inputs for the pre-trained transformer to fine-tune it, classifying these chunks independently and then classify the entire document using the fine-tuned model to obtain the document embedding. In the second approach we used the chunks embeddings produced by the transformer as input for an additional LSTM layer and use its output as input for the final classification layer.

Furthermore, we designed experiments using the Focal Loss function presented in [6] to assign higher weights to those labels that are more difficult to predict correctly, that usually are the unbalanced ones. Thus, we changed the traditional binary cross-entropy loss function used in the above experiments for the Focal Loss function. We also used a simple thresholding strategy, looking for the threshold value that maximizes the geometric mean score for each intervention in the validation dataset outputs.

In these experiments we did not use extremely unbalanced interventions, since interventions applied to less than 100 students seems not enough to let any architecture learn relevant features. We also truncated the observations to the maximum tokens limit allowed by BETO, but to avoid losing information we considered the observations as different inputs for the architecture, transformed by the same model.

Finally, the last experiments designed for this approach consisted in adding a BiLSTM layer following the embedding layer, based on the research presented in [7]. We built two architectures using the observations in different ways, in the first architecture we used the additional BiLSTM layer over the the embedding layer using the word embeddings of the joined observations document as the sequence input for the new layer. In the second architecture we used each observation type embedding as sequence inputs for the additional layer, using a sequence of 4 elements.

4.2.2. MIMLL Experiments Results

The results of our end-to-end Deep Learning experiments using the multi-instance multi-label learning approach are shown in table 4.4.

Experiment names	**F1**	**Jaccard Score**
max token 500 BETO diagnosis Single BETO	0.24	0.32
Overlapped pieces (independent chunks)	0.12	0.25
Overlapped pieces (sequence for LSTM layer)	0.21	0.27
Focal Loss + Thresholding (joined document)	0.34	0.27
Focal Loss + Thresholding (observations as different inputs)	0.38	0.3
LSTM over word embeddings	0.4	0.3
LSTM over observations embeddings	0.25	0.2

Table 4.4: Macro F1 score and Jaccard Score of MIMLL experiments

The performance using the truncated document of joined observations improved as we increased the tokens limit. However, the performance was far worse than that obtained in the preliminary experiments, obtaining less than half the macro F1 score of those experiments. These results indicate that the amount of content used of the observations to fine-tune the model and make the interventions recommendation directly affects the performance of the models.

The experiment using the overlapped chunks approach showed a training loss decrease with each epoch, while the validation loss and evaluation metrics did not, indicating a possible overfitting. Analyzing these results we observed that a possible cause of the poor performance in these experiments is the difficulty in predicting unbalanced interventions.

We also performed a brief analysis of the key part of our architectures, the BETO transformer of the above approaches architectures, since the performance of the neural network architectures built were not as expected, being unable to outperform simpler approaches using traditional ML methods. We observed the embeddings generated by fine-tuned BETO of different professionals' observations. As result of this analysis we noted the embeddings of most observations were the same and only a few were different from the others and between them. Thus, we concluded that the sensitiveness of BETO was lost in these experiments, possibly causing the bad performance of our architectures.

We observed that using the Focal Losss function the validation loss significantly decrea-

sed, but the evaluation metrics did not improve. Therefore, we decided to analyze the raw output of the model and observed that only the balanced interventions had a large range of values, while for unbalanced interventions the trained models predicted values similar to the distribution of the interventions, low values in rare interventions and high values in extremely common interventions.

The performance obtained using the thresholding strategy together with the Focal Loss function was better than any of the previous Deep Learning experiments in all our evaluation metrics, even almost doubling the macro F1 score obtained in some of the previous settings. Nevertheless, these results were not better than those obtained in the preliminary experiments.

Finally, using the additional BiLSTM over the embedding layer for the joined document approach we obtained a slightly better macro F1 score and Jaccard score and slightly worse results in micro F1 score. On the other hand, the performance using the second approach, considering the observation types as different inputs, was the worst performance of the Deep Learning experiments.

The findings of the experiments conducted so far show that the main challenge of using a multi-label approach to the interventions recommendation problem is to cope with unbalanced interventions. These labels affect the performance of the transformation models and produce insensitive embeddings of the observations. Therefore, we decided to employ the binary relevance approach (i.e., train independent single-label models for each label) instead of designing task-specific multi-label architectures.

This decision is motivated by the question that we still need to know whether Deep Learning methods are suitable in a single-label setting. In addition, we also want to find out whether neural network architectures can outperform the models in the preliminary experiments in this setting. We also hope that these experiments can provide new insights to design better multi-label architectures in the future, especially for unbalanced interventions.

4.2.3. Binary Relevance Experiments

In this section we describe the experiments conducted using a single label approach. The objective of designing additional experiments with this simpler approach is to analyze the performance of neural network architectures and tools in each intervention, being able to identify different behavior between the interventions that could help us in the multi-label approach.

In these experiments we do not consider students with any of the extremely rare interventions (interventions present in less than 3 % of students) since we do not have enough records for a Deep Learning classifier to learn relevant information for those interventions in the single label approach. Therefore, in these experiments we use 10 of the 14 possible interventions.

Thus, we first designed simple experiments, using the truncated document of joined observations and the OHE diagnosis as inputs for a new architecture of just two layers after the

input layer, first the embedding layer to transform the document followed by the classification output layer which receives the diagnosis and the embedding of the document as input. In addition, we designed another version of this architecture using the observation types as different inputs instead of joining them. With these settings we can observe the performance of the Deep Learning model for each intervention and also if simple approaches are enough to outperform the simple classifier model proposed in Section 4.1 in each intervention.

Additionally, we use different settings for the above experiments, changing the learning rate and batch size for the model. We evaluated two learning rates: $1e-5$ and $1e-6$, and three different batch sizes: 4, 8 and 16 elements for batch. This approach can help us to evaluate if using the most suitable learning rate and batch size for most intervention can help us in the multi-label approach, especially for predicting the most difficult interventions without worsening performance in the best performing interventions.

Finally, we also used the architecture with the additional BiLSTM layer over the embedding layer, used in the last experiment of the multi-instance multi-label approach experiments of Subsection 4.2.1. Thus, in the first setting we used the truncated joined observations document and the OHE diagnosis as inputs for the new architecture. The document was used as input for the embedding layer and the word embeddings of the document were used as sequence inputs for the BiLSTM layer. The output of this layer together with the OHE diagnosis of each student were used as input for the final classification layer to obtain the raw output.

For the second setting we use the observation types as different inputs and also the OHE diagnosis. Each observation was transformed using BETO and the document embedding of these observations were used as sequence inputs for the BiLSTM layer to finally use the output of this layer and the OHE diagnosis as input for the final classification layer.

4.2.4. Binary Relevance Experiments Results

The results of these experiments are shown in figure 4.1. We present a comparison of the F1 score obtained in each intervention first using the students' observations and diagnosis as input for the logistic regression classifier of preliminary experiments (PE AO + D, first column), then using the truncated document of joined observations and the OHE diagnosis (Joined O + D, second column), using observations as different inputs and OHE diagnosis (POE + D, third column), using the truncated document of joined observations and the OHE diagnosis and an additional BiLSTM layer over the embedding layer with the word embeddings as input (BoBETO WE + D, fourth column), and finally using the additional BiLSTM over the embedding layer but using the observations embeddings as inputs for the layer (BoBETO POE + D, last column).

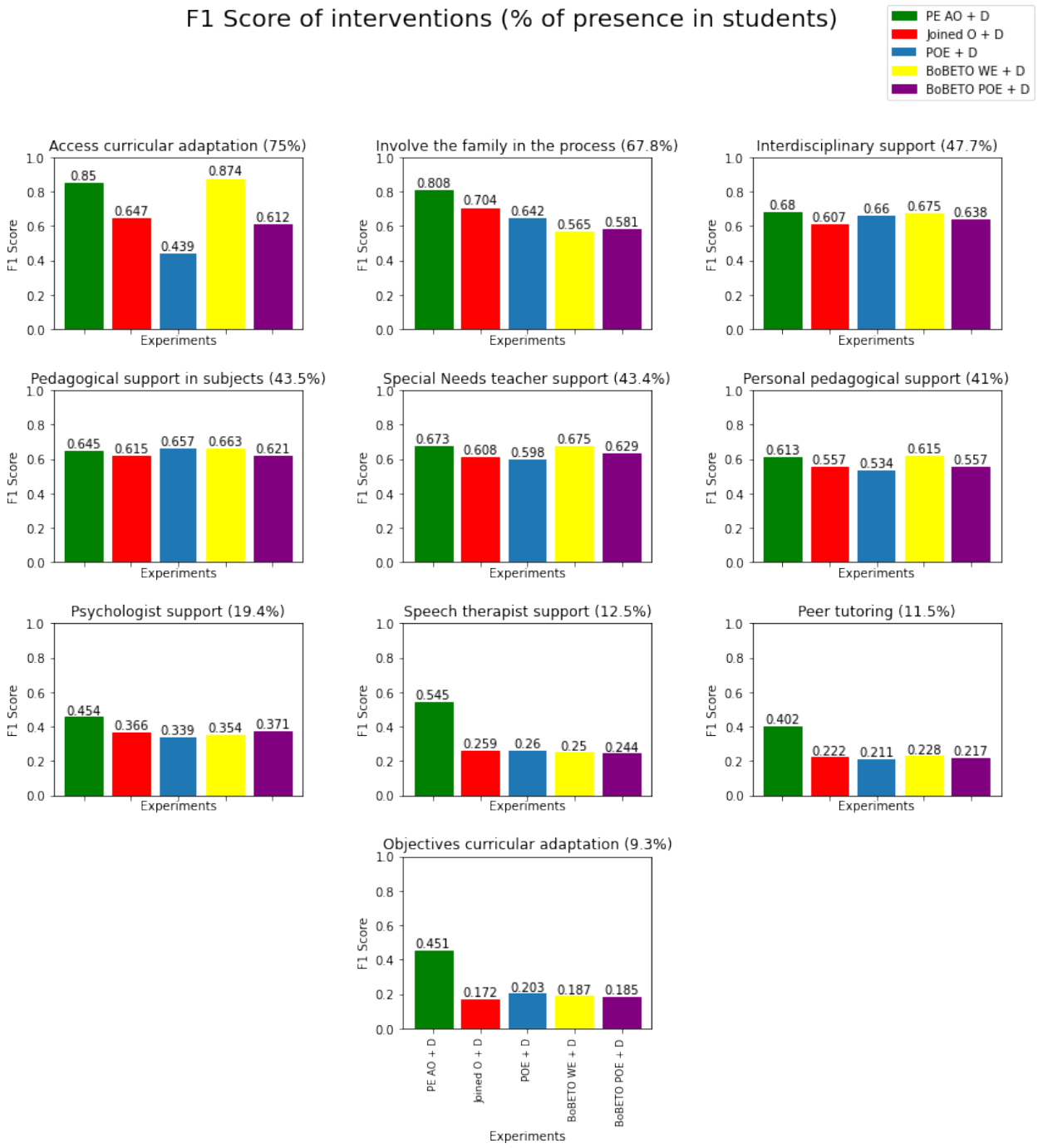

Figure 4.1: F1 score for each intervention using the students' observations and diagnosis as input for the logistic regression classifier of preliminary experiments and the designed end-to-end Deep Learning experiments. The graphs are shown in decreasing order by the number of students with the intervention present, with the "Access curricular adaptation" intervention being the most present intervention and the "Objectives curricular adaptation" the least.

The results of the first experiment, using the truncated observations, showed that performance of this architecture was better using the observation types as different input data than

using the joined document. The performance obtained using this setting was similar to the performance obtained in the preliminary experiments for the balanced interventions, even outperforming these experiments in "Pedagogical support in subjects" intervention. However, the performance for unbalanced interventions was much worse than in the preliminary experiments of Section 4.1.

We observed that using the observation types as different inputs were still better than using the joined document approach when we used different learning rates and batch sizes, with a significant difference in balanced interventions. However, the results for the unbalanced interventions showed that, even looking for the most suitable hyper parameters, the designed neural network architecture was not able to outperform preliminary experiments performance shown in Section 4.1 in these interventions. We also observed that using the joined observations approach the best batch size was of 8 elements, while using the different inputs approach it was better to use 16 elements per batch. Another finding was that, as expected, performance using $1e-6$ as the learning rate was better than using $1e-5$, but it was necessary to train the model for a larger number of epochs to converge.

The results of using the additional BiLSTM layer and considering the first approach, using the word embeddings as input for the new layer, were slightly better than those obtained without using the additional layer, and also were better than the preliminary experiments results in 4 interventions. However, the performance for unbalanced interventions of this architecture was still much worse than that obtained in preliminary experiments.

Nevertheless, using the different observation types embeddings as sequence inputs for the additional layer did not outperform the Deep Learning experiments that did not use the additional layer. We observed slightly better results in some interventions and slightly worse results in the other, but this architecture did not get the best results in any intervention.

4.3. Discussions

Preliminary experiments' results indicate that using Machine Learning tools we can outperform the majority class algorithm to recommend interventions, showing a better F1 score and Jaccard score in any scenario. Nevertheless, the interventions predictions of these models were different from the interventions written down in the documents by the professionals, obtaining a low Jaccard score using any settings.

The Deep Learning experiments' results show that the main challenge of this problem are the unbalanced interventions. These interventions causes that pre-trained transformers to lose sensitivity in generating representations of different observations. This lost of sensitivity affects directly the predictions of balanced interventions, causing the performance of predicting these interventions to be worse in a multi-label approach than in a single-label scenario.

On the other hand, thresholding techniques are a key feature to lead with unbalanced labels in the multi-label scenario. The Focal Loss function was also useful in preventing the transformer model from losing sensitivity, since we needed less epochs to learn from unbalanced interventions, thus avoiding overfitting.

However, any of the designed MIMLL architectures was able to outperform simple approaches using a binary relevance scenario. Since we observed that Deep Learning models can outperform simpler approaches in certain interventions, we think it is possible to outperform the simple approach using more complex models, using the full content of the observations, different embeddings and also additional layers to prevent loss of sensitivity of the pre-trained transformers.

The amount of available data can also make a significant difference in the learning capability of any Deep Learning models, thus we should also focus on collecting more data to increase the quantity and quality of the students records.

Chapter 5

Conclusions and Future Work

5.1. Conclusions

The first experiments results presented in Subsection 4.1.2 support the idea of using computational tools to help SNE professionals in deciding the best interventions for each student. The simple classifiers outperform the naïve algorithm of majority-class prediction in each intervention and in the multi-label approach.

Nevertheless, these classifiers did not perform sufficiently well in the multi-label approach considering the interventions written down by the professionals as the ideal set of interventions for each student, since the Jaccard and F1 score obtained in the test dataset were not good enough considering the context and sensitivity of our real world problem.

Despite the multi-label approach results, we observe reasonable results in certain interventions, supporting the use of computational tools for those interventions recommendation, while in the other interventions we also observe better performance than using a majority class algorithm. Additionally, we observed correlation between the interventions, diagnosis and observations of students, being this last source of information the most valuable for predicting interventions.

Although the designed end-to-end Deep Learning experiments presented in Subsection 4.2.1 results were much worse than expected, especially in the multi-label approach, we observed better performance in balanced interventions using more complex architectures and techniques than using the simple logistic regression classifier. Even though these results were not as expected, not allowing us to validate our hypothesis, we cannot reject it either, since we observed that using more complex architectures we can obtain better performance in our practical task.

Therefore, our experiments for the practical task of generating recommendations for interventions should be interpreted as initial experiments to solve this task and not as a practical solution to the problem. It is necessary to design more experiments not only with new models but also with new data sources of the students.

For the techincal problem of designing a neural network architecture using a pre-trained transformer together with other Deep Learning tools and methods in a MIMLL approach

that outperforms simple approaches we observed that the key challenge is preventing loss of sensitivity trying to predict harder or unbalanced labels. Focal loss function can be a valuable tool in dealing with this problem, since we use less epochs to learn from those labels, helping to avoid over-fitting and the sensitivity loss of the embedding model. A simple thresholding strategy can also help get a huge performance improvement in unbalanced dataset, because the final output of each label in the MIMLL approach can need a different threshold to consider the class as positive or negative.

The results of the Deep Learning experiments using the MIMLL approach and using the binary relevance approach, indicates that the models designed for the multi-label approach are affected by a large bias-variance problem. This can be deducted due to the large performance difference between the obtained by our Deep Learning models using the multi-label approach and the performance obtained with our shallow models. This performance difference is much lower when using the binary relevance approach, even outperforming the shallow models using the Deep Learning models.

Our end-to-end Deep Learning experiments results also indicate that more complex architectures, such as using additional BiLSTM layers over the embedding layer, can help us get a better performance in the MIMLL approach. Using these architectures we can prevent the sensitivity loss of the model and also allow the model to learn more about the words of the documents. This last feature can also be useful for predicting harder labels, since there are usually only specific words for these labels that indicate if the label is positive or not. Thus, it may be necessary to build more complex architectures for the MIMLL approach, architectures that exceeded our hardware and time limitations for this research.

Collecting more data is also an essential point to continue research in both areas of this research, the technical and the practical part. We believe it is necessary to have more data available from SNE students to develop further research and computational tools to help them and the professionals who support them.

5.2. Contributions

One of the main contributions of this research is the SNEC corpus, dataset containing information from more than 3,000 SNE students with different diagnosis, interventions and professionals' observations. However, an exhaustive anonimization process is still required, as there may be sensitive data of students that we did not notice previously.

Another contribution is the validation of computational tools for the interventions recommendation problem, since the results of our preliminary experiments using simple approaches outperform the naïve algorithm of majority class. Even though these results are not enough to consider these experiments as a solution for the interventions recommendation problem, we consider that the experiments presented in this research can be considered as a basis for future research on this problem.

In the technical side of our problem, the multi-instance multi-label learning problem using pre-trained models, despite any of our Deep Learning architectures obtained a better performance than our simple experiments using traditional Machine Learning techniques, we

learned from the results and highlight the use of particular techniques and methods, such as Focal Loss function and BiLSTM layers over the embedding layer.

We also observed that using pre-trained embedding models we can design simple Deep Learning architectures that can outperform traditional Machine Learning models in certain interventions, first showing that these models can be used in the interventions recommendation problem and also motivating us to develop further research to improve the results in these interventions and analyze the bad results in the other interventions.

5.3. Future Work

Finally, we propose a new methodology to develop further research in both areas of this research. First, we think it is essential to collect more students' data, especially from students with rare interventions to build better architectures and generate better recommendation of these interventions.

Considering the results of our multi-label experiments, we consider it necessary to first evaluate the following experiments in a binary relevance scenario, analyzing individual results of the architectures in each intervention before using them in a multi-label approach.

We observed better results with word-level embeddings and additional layers than using document embeddings for this problem. We could use character-level embeddings, such as the available in the Flair framework [13], together with the word or document embeddings and analyze the results. These architectures can be a valuable solution for this problem since we observed uncommon characters used by different professionals to indicate mispronunciation of words for example.

Next, another possible solution for this problem is to use different pre-trained models fine-tuned with a single observation type, so each one of these embeddings can learn from different context and generate different embeddings from a same phrase written down by different professionals if necessary. Nevertheless, these architectures would require a huge amount of memory to use more than one transformer model to generate the text representations.

Considering the results of using word embeddings instead of document embeddings and additional layers over this layer, a valuable solution for this problem using the MIMLL approach can be the use of word embeddings of each observation type instead of the document embedding of each type. These architectures would require a longer training time since the BiLSTM layer or any other kind of layer over the embedding layer will have several input data and new parameters to adjust.

Finally, we should not forget the main objective of this research, helping SNE professionals in deciding the most appropriate interventions for each student. Therefore, if our future models obtain better performance in this task, they should be validated with real world professionals. This validation can be crucial for a better understanding of the problem and to define better guidelines for future research.

We could also provide feedback about the text pieces the classifier considered relevant to

generate the recommendations in order to provide more confidence to the professionals. The Captum library [27] can be useful to provide interpretability of the decisions made by neural networks architectures.

Bibliography

[1] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601–618, 2010.

[2] N. Indurkhya and F. J. Damerau, *Handbook of natural language processing*, vol. 2. CRC Press, 2010.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[4] J. Cañete, G. Chaperon, R. Fuentes, and J. Pérez, "Spanish pre-trained BERT model and evaluation data," in *to appear in PML4DC at ICLR 2020*, 2020.

[5] R. Pappagari, P. Zelasko, J. Villalba, Y. Carmiel, and N. Dehak, "Hierarchical transformers for long document classification," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 838–844, IEEE, 2019.

[6] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

[7] Y. Liu, "Fine-tune bert for extractive summarization," *arXiv preprint arXiv:1903.10318*, 2019.

[8] R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach," *Journal of Informetrics*, vol. 3, no. 2, pp. 143–157, 2009.

[9] A. M. N. Allam and M. H. Haggag, "The question answering systems: A survey," *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, vol. 2, no. 3, 2012.

[10] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," *arXiv preprint arXiv:1602.02410*, 2016.

[11] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[12] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences*, pp. 1–26, 2020.

[13] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "Flair: An

easy-to-use framework for state-of-the-art nlp," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 54–59, 2019.

[14] W. Wang, J.-F. Weng, J.-M. Su, and S.-S. Tseng, "Learning portfolio analysis and mining in scorm compliant environment," in *34th Annual Frontiers in Education, 2004. FIE 2004.*, pp. T2C–17, IEEE, 2004.

[15] A. S. Drigas and R.-E. Ioannidou, "A review on artificial intelligence in special education," in *World Summit on Knowledge Society*, pp. 385–391, Springer, 2011.

[16] Z.-H. Zhou and M.-L. Zhang, "Neural networks for multi-instance learning," in *Proceedings of the International Conference on Intelligent Information Technology, Beijing, China*, pp. 455–459, 2002.

[17] N. Weidmann, E. Frank, and B. Pfahringer, "A two-level learning method for generalized multi-instance problems," in *European Conference on Machine Learning*, pp. 468–479, Springer, 2003.

[18] T. Zeng and S. Ji, "Deep convolutional neural networks for multi-instance multi-task learning," in *2015 IEEE International Conference on Data Mining*, pp. 579–588, IEEE, 2015.

[19] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.

[20] F. Kang, R. Jin, and R. Sukthankar, "Correlated label propagation with application to multi-label learning," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 1719–1726, IEEE, 2006.

[21] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 115–124, 2017.

[22] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine learning*, vol. 85, no. 3, p. 333, 2011.

[23] Z.-L. Zhang and M.-L. Zhang, "Multi-instance multi-label learning with application to scene classification," in *Advances in neural information processing systems*, pp. 1609–1616, 2007.

[24] L. Song, J. Liu, B. Qian, M. Sun, K. Yang, M. Sun, and S. Abbas, "A deep multi-modal cnn for multi-instance multi-label image classification," *IEEE Transactions on Image Processing*, vol. 27, no. 12, pp. 6025–6038, 2018.

[25] M. Surdeanu, J. Tibshirani, R. Nallapati, and C. D. Manning, "Multi-instance multi-label learning for relation extraction," in *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pp. 455–465, Association for Computational Linguistics, 2012.

[26] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 145–158, Springer, 2011.

[27] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Mel-

nikov, N. Kliushkina, C. Araya, S. Yan, *et al.*, "Captum: A unified and generic model interpretability library for pytorch," *arXiv preprint arXiv:2009.07896*, 2020.

ANNEXES

Annexed A

Complete Figures

A.1. Interventions Distribution for each Diagnosis

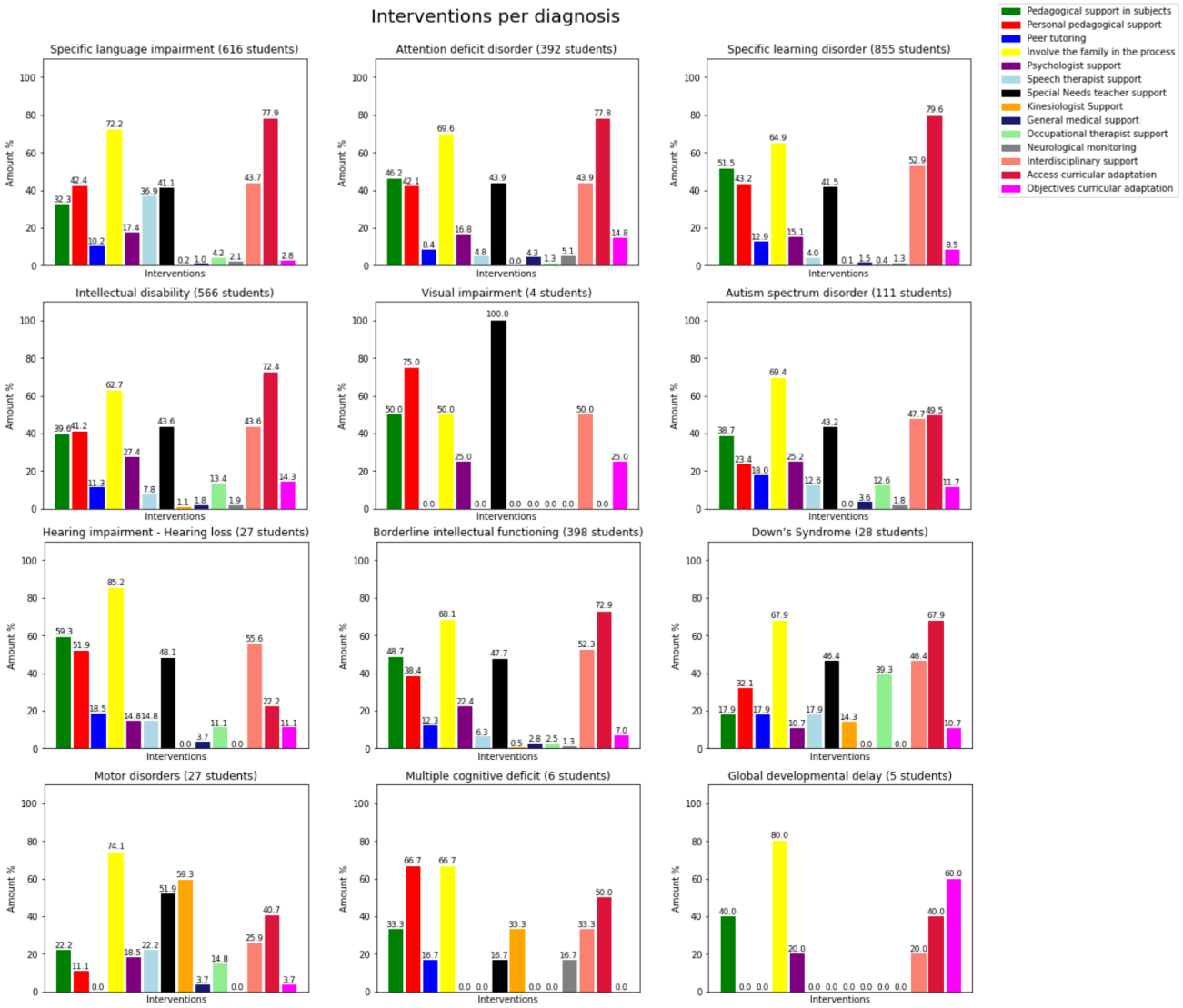

Figure A.1: Complete interventions distribution for each diagnosis.

A.2. Observation Types Presence Distribution for each Diagnosis

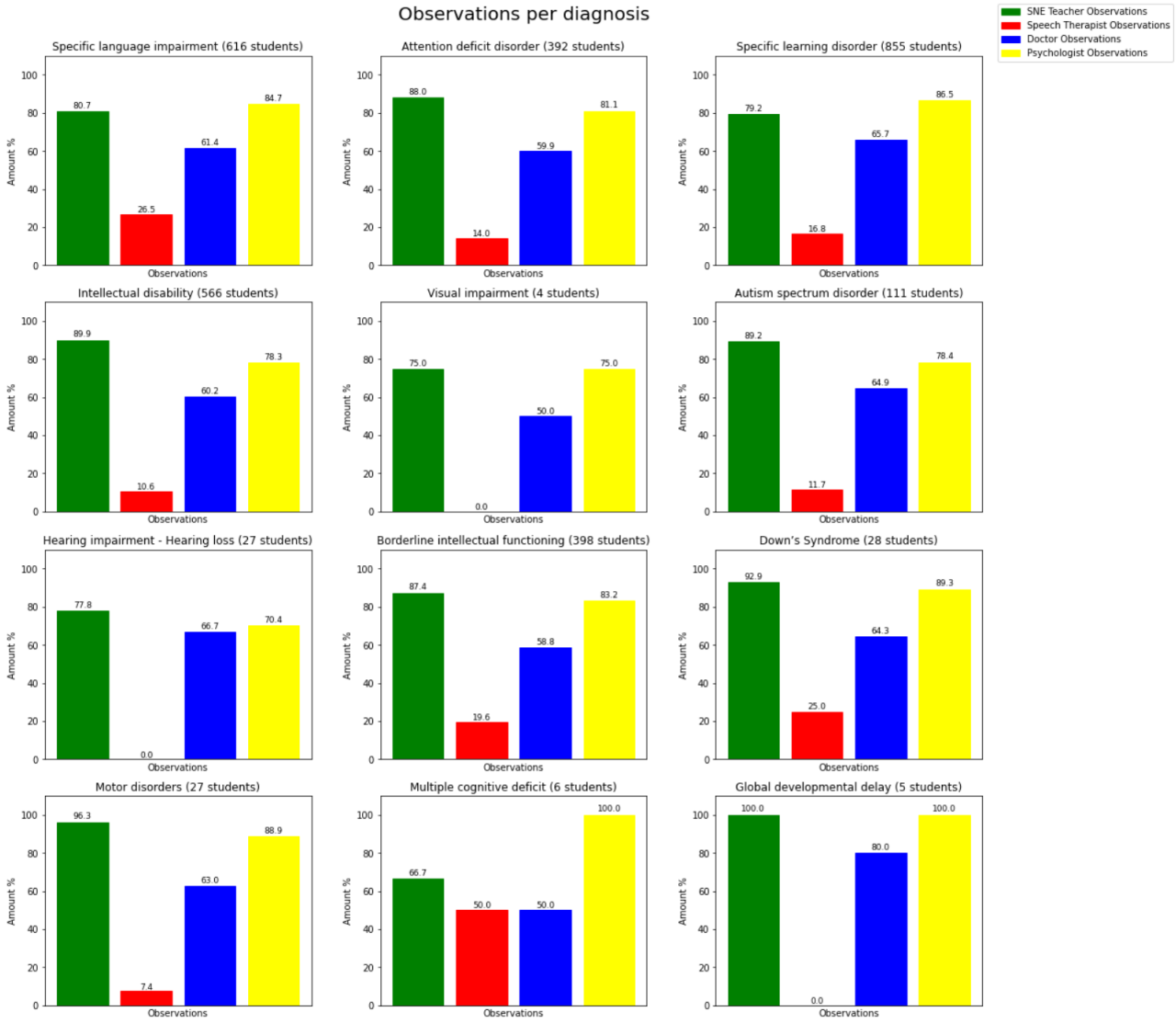

Figure A.2: Complete Observation types presence distribution for each diagnosis.

A.3. Observation Types Presence Distribution for each Intervention

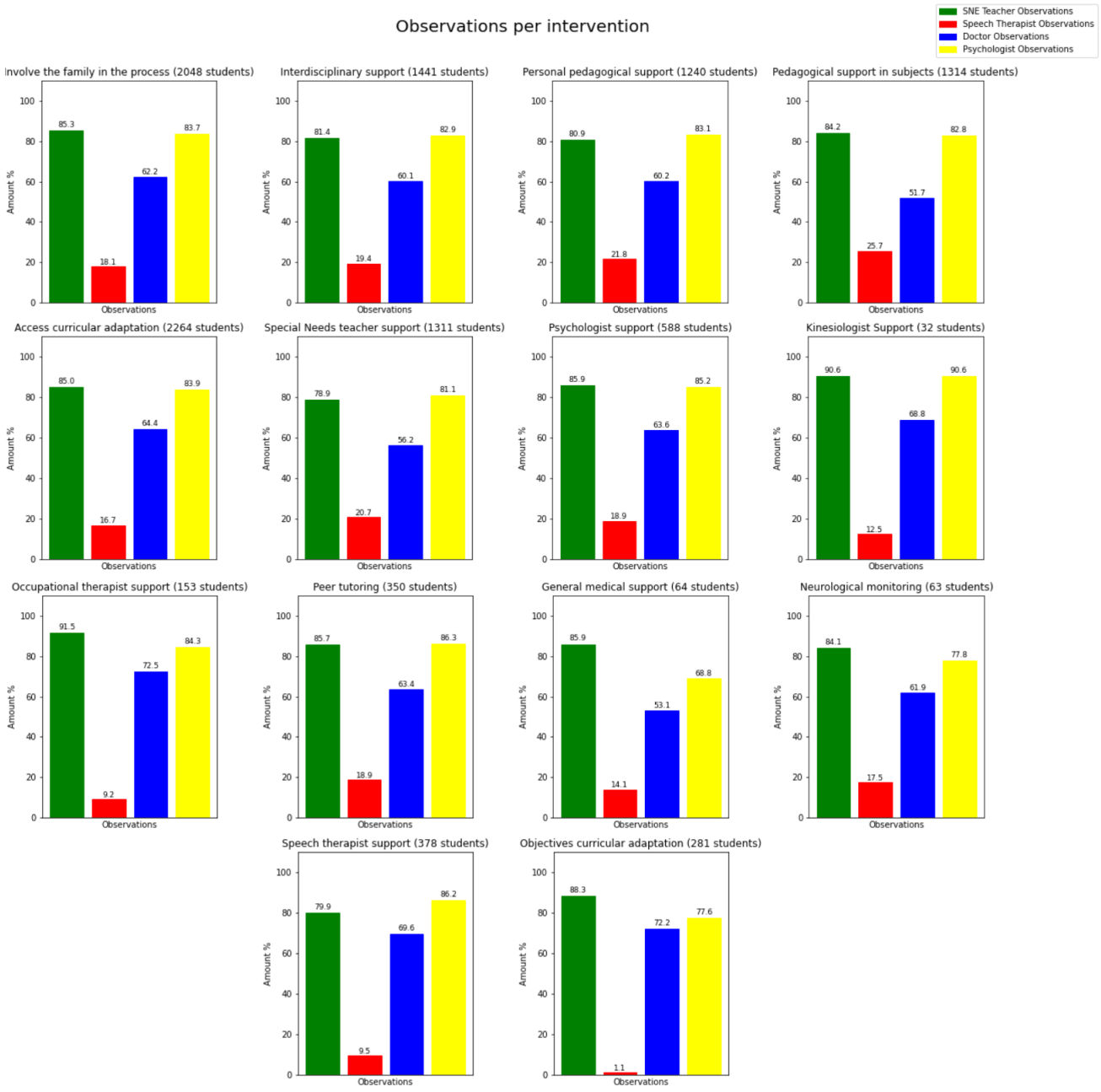

Figure A.3: Complete Observation types presence distribution for each intervention.

A.4. Interventions Distribution for each Intervention

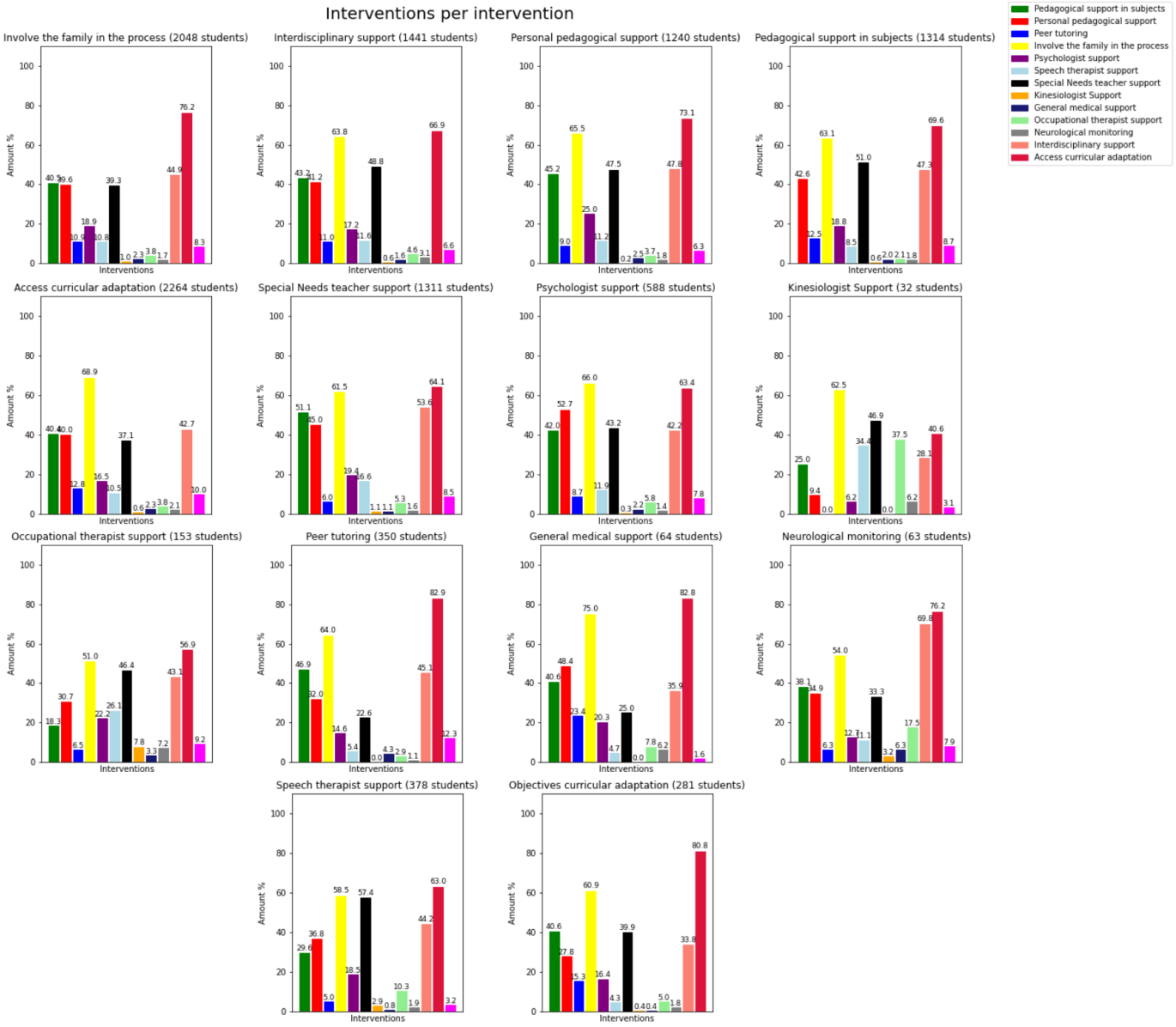

Figure A.4: Complete intervention distribution for each intervention.