Interventions Recommendation: Professionals' Observations Analysis in Special Needs Education

Javier Muñoz[1], **Felipe Bravo-Marquez**[1,2]
[1]Department of Computer Science, University of Chile
[2]Millennium Institute for Foundational Research on Data, IMFD-Chile
jmunoz@dcc.uchile.cl, fbravo@dcc.uchile.cl,

Abstract

This paper presents a new task in educational NLP, recommending professional interventions for Special Needs Education (SNE) students using NLP techniques. The task is formulated as a multi-label classification problem in which each training example is formed by the student's diagnosis along with various free text observations made by teachers and professionals, and the target classes correspond to a set of interventions recommended based on that information. Using the previously mentioned structure, we build the Special Needs Education Corpus (SNEC), a new corpus of over 3,000 Chilean special needs students. We also train several machine learning models using different settings and feature representations of our data. Our results indicate that textual features are the most useful in terms of classification performance and that other non-textual features, such as diagnosis and other chosen interventions, are also beneficial. We also observed a positive effect of representing text inputs with a dense BERT-based representation over using sparse n-grams and non-contextual word embeddings. Our corpus and source code are available at https://github.com/dccuchile/SNEC.

1 Introduction

One consequence of the adoption of digital technologies in schools is the wide availability of data records associated with their enrolled students. These records have been successfully employed in many ways to help students improve their school performance. For example, Romero and Ventura (2010) present different projects that use student records, such as their grades in different subjects, to help them in different ways, for example recommending activities or books that can help them to improve their grades.

These investigations focus on, for example, correction of grammatical errors, automated writing evaluation, automated content evaluation, vocabulary analysis, among others. These tasks use free text pieces written by students as inputs to develop different kinds of outputs, allowing students to improve their grammar together with reading and writing skills (Costa-jussà and Alfonseca, 2019).

Another important task is to analyze free text observations and reports made by teachers and other professionals (e.g., physiologists, speech therapists) that have worked with the students to decide certain strategies or interventions that can help students improve their school performance and social skills.

This task is especially important in "Special Needs Education" (SNE), where students with different disabilities, such as motor disorders, Down's Syndrome, specific language impairment, among others, may need particular interventions, such as access curricular adaptation, involve the family in the educational process, interdisciplinary support, among others (Olakanmi et al., 2020).

As the decisions made by professionals working in SNE usually are more critical than the decisions made in traditional education, teachers and professionals are not always sure about their decisions (Podell and Soodak, 1993). For this reason, different tools have been developed to help SNE professionals diagnosing Special Needs students and performing the chosen interventions with these students (Drigas and Ioannidou, 2011).

In this work we study the problem of recommending the best interventions to students in Special Needs Education considering their diagnosis, professionals' observations and even other applied interventions, formulating this problem as a multi-label text classification task with additional data. We explore different input sets and text transformations that allow a machine learning system to make intervention predictions for each student.

To achieve this, we collected data from differ-

ent Chilean Special Needs schools to build a new the Special Needs Education Corpus (SNEC). This dataset was built with records from around 3,000 students with different diagnoses, Spanish-written observations by professionals and the best interventions for each of them.

Each student has one of 12 possible diagnoses and has one or more recommended interventions, chosen from a set of 14 possible interventions. The professionals that give their observations of students are in most cases doctors, psychologists and special education teachers, but speech therapists can also observe if the student is required to work with these professionals.

The content of the observations are free text instances and vary according to the professional who issues them. We have therefore designed different experiments to analyze whether these observations should be differentiated or use a simplification such as joining the observations and working with a single joined document and see if it is enough to get correct interventions prediction. As each student might need more than one intervention, we formulate this problem as a multi-label text classification task with additional attributes, such as the diagnosis and other chosen interventions.

We use different sparse bag of n-grams text representations and dense neural text transformers models, such as BERT (Devlin et al., 2019), to analyze which text representation approach is best suited to our problem.

The focus of this paper is therefore to introduce a new task in the Educational Data Mining field, specifically in SNE field using Natural Language Processing (NLP) techniques, together with a our new SNEC corpus with Spanish-written professionals' observations of students with different disabilities and results of several experiments using different approaches and text transformations to predict the best interventions for each one of these students using machine learning techniques.

The remainder of this paper is structured as follows. In Section 2 we present related work to our practical problem, recommending interventions in Special Needs Education, and the technical problem, use different text representations for observations written by professionals and different input sets to make intervention predictions. In Section 3 we propose our task and show the dataset origin and details of each attribute. In Section 4 we present our experiment settings and the evaluation metrics used to analyze the results of the experiments. The details and results of the experiments are shown and discussed in Section 5. Finally, in Section 6 we present our conclusions and future work.

2 Related Work

This work can be viewed from multiple perspectives: from a domain point of view it lies in the fields of Educational Data Mining (EDM) and SNE and from a technical point view we have a text classification problem. In this section we present different tools used both in traditional educational data mining and particularly in SNE and then we also present text classification methods that are relevant to this work.

Educational Data Mining is a specific field that uses computational approaches to analyze educational data in order to study educational questions, as defined in (Romero and Ventura, 2010). Our problem shares features with many tasks defined in the above paper, such as providing feedback for supporting instructors (where the objective is to provide feedback to support course teachers in decision making) and recommendations for the students (where the objective is to make recommendations directly to the students analyzing their personalized activities).

In Special Needs Education there are also different tools that help teachers and other professionals to work with students. Drigas and Ioannidou (2011) 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.

Our study also analyzes the technical problem of the interventions recommendation - the text classification problem when we use the free text instances, such as the professionals' observations. A relatively recent architecture that has yielded state-of-the-art results in many NLP tasks is the Transformer (Cer et al., 2018), which can be used for training large language models from unalabeled corpora (e.g., BERT (Devlin et al., 2019)). These pre-trained models can be later fine-tuned to almost any target task at hand with successful results. Although most of these large pre-trained models are only available for the English language, several language-specific versions have been made available in recent years, such as BETO (Cañete et al.,

2020), the Spanish version of BERT.

3 Problem Formulation

In this section we present the task studied in this work together with its context and our new corpus.

Our task 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 and in case 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.

With these document 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 use just the interventions that did work.

We built a new Special Needs Education Corpus (SNEC) from the records from around 3,000 students with different disabilities from Chilean SNE schools enrolled between 2018 and 2019. Each student record consists in his or her diagnosis, the observations of the professionals that have worked with each student at the beginning of the year and the interventions that professionals chose as the best for this student (applied or not) at the end of the scholar year.

The students' diagnosis were obtained from a web application that helps teachers store basic student information, such as his or her name and diagnosis, and students' digitized documents.

This web application is used to track the students' progress, both in school performance and evolution of their diagnosis. For this work we summarized the possible diagnosis of a student, grouping the same diagnoses but different severity into a single general diagnosis. This way, students have a diagnosis from 12 possible diagnoses. The number of students per diagnosis are shown in Table 1.

Diagnoses	Number 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 1: Number of students per diagnosis.

The 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.

The 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. 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 observation 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 that have each observation type are shown in Table 2.

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

Table 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 worked and those that did

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

not. This document also contains not applied interventions that could have worked. For this study, interventions were grouped into larger groups to reduce the label space and get a better understanding of the interventions. This way, students have 1 or more assigned "best interventions" from a set of 14 possible interventions as shown in Table 3.

Interventions	Number 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: 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.

The student data can thus be represented as shown in Figure 1. The student is diagnosed with one of 12 different diagnoses and has different evaluations containing free 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.

Based on this, we formulate the task of predicting interventions for a given student as a multi-label classification problem, in which all the free text observations and the student's diagnosis are used as input features.

4 Methods

As we are dealing with a multi-label classification task, we employ the stratification method of (Sechidis et al., 2011) to generate training, validation and testing partitions dividing our data in a ratio of 3:1:1 respectively. For this research we do not use the testing partition as we plan to build more complex architectures in future research. This way, we use 1836 students data for training and 612 for validation.

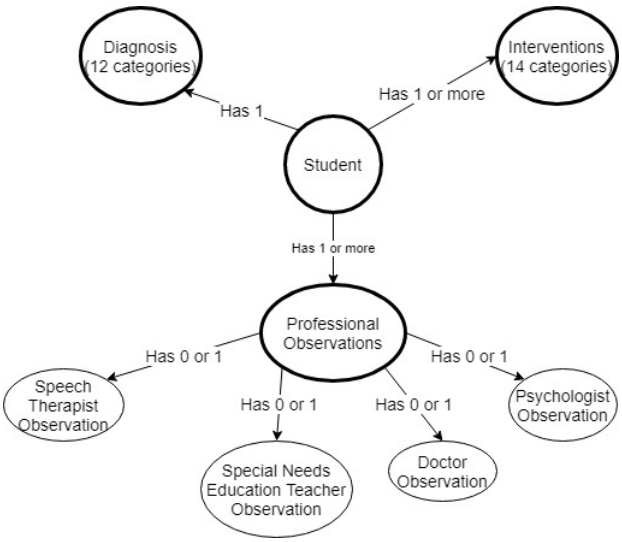

Figure 1: Problem ontology.

We then define general settings and evaluation metrics for all our designed experiments. Our experiments are applied to each intervention, considering a single intervention as output to analyze single intervention behaviors. As behavior of the predictions and possible correlations could be different in well-balanced and unbalanced interventions, we group interventions into two sets: unbalanced interventions (present in less than 15% of the students or in more than 85% of the students) and balanced interventions (not unbalanced interventions).

We consider a dummy majority vote classifier as the baseline. As this study explores the possible correlation between the diagnosis, professional observations or interventions of students and tries different text transformations for the professionals' observations, we use a 'simple' logistic regression classifier with the same settings for all the experiments, only changing the input data (data selection and transformations) for each intervention.

The diagnosis of each student is represented using a One-Hot Encoding (OHE) representation, adding 12 different binary attributes representing each one of them. Interventions are also represented as binary attributes, both when used as input attributes and as output.

In the "simple" text representation experiments we use a sparse bag of n-grams representation of

the professional observations, considering from 1 to 3 words as tokens, using unigrams, bigrams and trigrams this way. We also use a simple approach for cases where a student had more than one observation of the same observation type. In these cases we just join the observations and use it as a single observation.

We also design two versions for experiments in which we use the bag of n-grams representation of the observations, first considering the joined different types of observations of each student as a single text input and another version adding a special token to each word of the observations, using a different token depending on the professional that wrote the observation.

Furthermore, we design two versions for experiments using dense neural representations of the observations. As adding a special token to words will cause pre-trained dense transformers generate inaccurate representations, we use each observation type as a different text input and use their representations as different features as input for our classifier.

For this study we use the following metrics: accuracy, Cohen's kappa score (Vieira et al., 2010), F1 score average of positive and negative class and Area Under the Receiver Operating Characteristic Curve (ROC AUC) (Flach, 2016) score of each experiment for each intervention. In this work we report the macro-average of all the interventions.

5 Experiments

For our experiments we train logistic regression models considering different sub-sets of the feature space and also using different representation approaches for our textual features. We also have two groups of experiments: the first one consisted in using sparse n-gram representation of our text inputs, using the bag of n-grams mentioned above, while for experiments of the second group we use dense neural representations of the observations.

5.1 Sparse N-gram Representations Experiments

For these experiments we use a feature ablation technique to design our experiments. Thus, for our first experiment we use the OHE representation of students' diagnosis, bag of n-grams representation of the professionals' observations and all the interventions, except the one we are trying to predict, as binary attributes, using 0 if the intervention is not present and 1 if it is. We train the logistic regression classifier with these features and then we evaluate the performance on our test dataset. This experiment outperformed the dummy classifier mentioned in Section 4 in all the interventions using our both versions, adding special tokens and joining the observations. We also observe that the performance of this experiment is different depending on the intervention we are trying to predict and if we use or do not use the special token. However, this experiment is not realistic since we can try to predict the interventions using a pre-defined order but we cannot know every other predicted intervention always.

Next, we design new 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. The performance of these experiments was worse than the obtained in our previous experiment for all the interventions. We analyze the weights the classifier assign to each feature and observe that for certain interventions there are interventions used as binary attributes showing a high positive weight and in the same way, for other interventions some interventions used as a binary attribute with a high negative weight as well, as shown in Table 4 and Table 5, showing the top positive and negative features using 'Special Needs teacher support' intervention as a binary attribute. The interventions with a high negative or positive value were mainly balanced interventions and the interventions that showed a correlation with another intervention were unbalanced and balanced interventions.

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: Top positive features for 'Apoyo fonoaudiólogo(a)' intervention in experiment using all observations joined + OHE diagnosis + presence of 'Special Needs teacher support' intervention as binary feature.

We then remove the interventions as binary attributes and work with OHE representation of the diagnosis and the sparse n-grams representation of professional observations. The performance using these experiment settings are equal to or worse than the obtained using the above settings, with a single intervention as a binary attribute. Depending on the

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 5: Top negative features for 'General medical support' intervention in experiment using all observations joined + OHE diagnosis + presence of 'Special Needs teacher support' intervention as binary feature.

analyzed intervention, the results using the special token approach are better than using the joined observations. However, performance is better using the joined observations approach in most cases.

Our next experiment was using all the interventions, except the one we are trying to predict, as binary attributes, and OHE representation of the diagnosis of students as input for the classifier. The performance of this experiment is worse than the one obtained using the above setting.

We then use only one type of professional observation representation and OHE diagnosis' representation as input for our classifier. Using this setting, the performance is worse than in any of the previous experiments, but we observe that depending on the analyzed intervention, using certain type of observations we obtain better performance than using the other observations. This occurs using the SNE teacher observations in most interventions.

Our lasts experiments use only one of our features: diagnosis, observations or interventions, using also the special token and the joined observations versions when we use the observations as features. The best results using only one attribute are obtained using the professionals' observations, both in balanced and unbalanced interventions.

The results of our sparse n-grams experiments are summarized in Table 6. We present the macro-averaged results of all the interventions and the results of experiments using a single intervention were also summarized. However, as we mentioned before, results are different depending on the analyzed intervention and the intervention we use as binary attribute.

These results show that the observations of the professionals are the most relevant features for predicting interventions using our settings, being SNE teacher observations the most relevant in most cases. We also observe than depending on the analyzed intervention, using a special token to differentiate each type of observation we obtain a better performance than joining the observations and using them as a single text instance. In addition, we observe that diagnosis and interventions are not enough to get better performance than using a majority class algorithm by themselves, but are useful attributes for making better predictions when using the professionals observations. The interventions and diagnosis also presented different weights in our models depending on the analyzed intervention, showing correlation between certain interventions and between interventions and diagnoses.

5.2 Dense Neural Representations Experiments

The focus of these experiments is to use different dense neural representations of the observations as input for our classifier and compare their performance with the previous experiments. This way, in a preliminary experiment we use our own pre-trained embeddings built from our domain data, but the performance using this model to produce representations of the professionals' observations were as bad as the obtained in our sparse n-grams experiments using only the OHE representation of the diagnosis, even using the one-hot encoded diagnosis together with the word embedding representation.

We then use pre-trained frozen BETO transformation of the observations as input together with the one-hot encoded diagnosis. We use our two versions, first joining all the observations and using this joined document as a single text instance, and second encoding each observation and using each type as different features according to the professional who issued them. We also split the observations or joined observations that had more tokens than the allowed by the BETO transformer (512) into smaller pieces and then averaging these representations. This experiment, using observations representation as different features, outperforms all our previous experiments for balanced interventions, except for our first experiment of sparse n-gram representations, where we used all our available data. However, the performance for unbalanced interventions using this approach is worse than the obtained in the experiments using the bag of n-grams representation of the observations.

We also design experiments using a sentence embedding representation, using the cross lingual pre-trained model presented in (Conneau et al., 2020)

Experiment names	Acc	Kappa	F1	AUC
Baseline (Dummy classifier)	0.79	0	0.44	0.5
Only diagnosis	0.8	0.02	0.47	0.51
All other interventions	0.8	0.06	0.51	0.53
All joined observations	0.81	0.27	0.64*	0.62
Observations with special token	0.8	0.25	0.62	0.62
All joined observations + diagnosis	0.81	0.29	0.64	0.63*
Observations with special token + diagnosis	0.81	0.25	0.62	0.62
Doctor observations only + diagnosis	0.79	0.05	0.5	0.52
Psychologist observations only + diagnosis	0.8	0.11	0.55	0.55
Speech therapist observations only + diagnosis	0.8	0.03	0.48	0.51
Special Needs Education teacher observations only + diagnosis	0.81	0.26	0.63	0.61
All other interventions + diagnosis	0.81	0.1	0.53	0.54
All joined observations + diagnosis + 'Single Intervention'	0.81	0.29*	0.64*	0.63*
Observations with special tokens + diagnosis + 'Single Intervention'	0.81	0.26	0.63	0.62
All joined observations + diagnosis + all other interventions	0.82*	0.29*	0.64*	0.63*
Observations with special tokens + diagnosis + all other interventions	0.82*	0.28	0.64*	0.63*

Table 6: Macro-averaged sparse n-grams experiments results of all interventions. Scores with an asterisk correspond to the best macro-average results of each metric.

to make new representations of the professionals observations, but the results are not better than the ones obtained with BETO transformer representations used in the experiment above for any intervention.

Our last experiment consists of using the word embedding representation in the observations from the pre-trained model presented in (Bojanowski et al., 2017) with the "Spanish Billion Word Corpus"[2] embeddings, together with the one-hot encoded diagnosis. The performance using this representation is similar to that obtained with our sparse n-grams experiments where we use OHE representation of the diagnosis only, did not perform further experiments with this approach.

The summarized results of these experiments are presented in Table 7. As in Table 6, we present the macro-average results of all the interventions.

Using these representations we observe that text transformers like BETO also help to get better predictions, while static word embedding representations, where the context of the words is not really taken into account, are not as useful as transformers. We also observe that splitting our text instances into smaller pieces to use the BETO representation the performance of the classifier decrease as more splits are required.

6 Conclusions and Future Work

In this paper we have presented a new task, recommend interventions to professionals to work with Special Needs students using machine learning techniques. We also released SNEC, a new corpus with more than 3,000 students records containing their diagnosis, free text written observations and the best interventions for each of them. Finally, we showed the results of experiments using different text representations.

We observed that the professionals' observations of a student can be used to predict the best interventions for him or her. Besides, the diagnosis and other chosen interventions showed to be valuable in predicting some interventions, but not enough to improve on the predictions made using observations alone.

We also observed that contextualized neural representations of the observations, such as BERT or sentence embeddings, are useful to predict balanced interventions, but not for the unbalanced ones. It is important to remark, though, that, in this study the weights obtained from these pre-trained models were left frozen. The literature suggests that fine-tuning these weights to the task at hand can lead to significant improvements (Cañete et al., 2020).

As a general conclusion we can say that no single approach to represent observations consistently outperformed the others for all the interventions.

We believe that the models trained in this study can be useful to develop new tools to help teachers in deciding which interventions are the best for each student knowing his or her diagnosis, professionals' observations and other possible interventions.

Our next steps are to analyze our results further and find out the limitations of our approach.

We believe that there is plenty of room to im-

[2] http://crscardellino.github.io/SBWCE/

Experiment names	Acc	Kappa	F1	AUC
Baseline (Dummy classifier)	0.79	0	0.44	0.5
All observations with new word embedding	0.8	0.02	0.47	0.51
All observations (new word embedding) + diagnosis	0.8	0.04	0.49	0.52
All observations with BETO (average of word representation)	0.82*	0.16	0.57	0.57
All observations (BETO-average representation) + diagnosis	0.82*	0.16	0.58	0.57
Different observations with BETO (average)	0.82*	0.25*	0.62*	0.61
Different observations (BETO-average) + diagnosis	0.82*	0.25*	0.62*	0.62*
All observations (xlm-r-bert sentence embedding)	0.81	0.17	0.57	0.57
All observations (xlm-r-bert sentence embedding) + diagnosis	0.81	0.19	0.59	0.58
Different observations (xlm-r-bert sentence embedding)	0.81	0.19	0.6	0.59
Different observations (xlm-r-bert sentence embedding) + diagnosis	0.81	0.22	0.61	0.6
All observations (fasttext-sbwc) + diagnosis	0.8	0.03	0.48	0.51
Different observations (fasttext-sbwc) + diagnosis	0.82*	0.09	0.53	0.54

Table 7: Macro-averaged dense neural representations experiments results of all interventions. Scores with an asterisk correspond to the best macro-average results of each metric.

prove our results by designing an architecture more tailored to our problem.

For future work, we plan to design an end-to-end neural network architecture for this task and to fine-tune the pre-trained representations employed in this study.

Moreover, we did not experiment with approaches that can more efficiently exploit the multi-label nature of our problem i.e., the correlation between certain interventions. In this direction, we plan to experiment with multi-label loss functions such as the Hamming loss.

Acknowledgments

Felipe Bravo-Marquez was funded by ANID FONDECYT grant 11200290, U-Inicia VID Project UI-004/20 and ANID - Millennium Science Initiative Program - Code ICN17_002.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *to appear in PML4DC at ICLR 2020*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R Costa-jussà and Enrique Alfonseca. 2019. Proceedings of the 57th annual meeting of the association for computational linguistics: System demonstrations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Athanasios S Drigas and Rodi-Eleni Ioannidou. 2011. A review on artificial intelligence in special education. In *World Summit on Knowledge Society*, pages 385–391. Springer.

Peter A Flach. 2016. Roc analysis. In *Encyclopedia of Machine Learning and Data Mining*, pages 1–8. Springer.

Oluwabunmi Adewoyin Olakanmi, Gokce Akcayir, Oluwbukola Mayowa Ishola, and Carrie Demmans Epp. 2020. Using technology in special education: current practices and trends. *Educational Technology Research and Development*, 68(4):1711–1738.

David M Podell and Leslie C Soodak. 1993. Teacher efficacy and bias in special education referrals. *The Journal of educational research*, 86(4):247–253.

Cristóbal Romero and Sebastián Ventura. 2010. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 145–158. Springer.

Susana M Vieira, Uzay Kaymak, and João MC Sousa. 2010. Cohen's kappa coefficient as a performance measure for feature selection. In *International Conference on Fuzzy Systems*, pages 1–8. IEEE.