KU LEUVEN

FACULTY OF ENGINEERING SCIENCE

Learning What to Learn: Generating Language Lessons using BERT

João Pedro Olinto Dossena

Thesis submitted for the degree of Master of Science in Engineering: Computer Science, option Artificial Intelligence

Supervisor:
Prof. dr. Bettina Berendt

Assessors:
Ing. Pieter Delobelle
Prof. dr. G. Marra

Assistant-supervisor:
Ing. Pieter Delobelle

Academic year 2022 – 2023

Preface

I would like to thank everybody who kept me busy the last year, especially to my promoter and to my supervisor. Additionally, I would like to thank Marleen Vanderheiden and Elke Gilin, from the CLT and ILT, respectively, for taking the time to talk to me and give me insights about the thesis. I would like to give a special thanks to my family and friends, especially to my mother and to Daniel Šparemblek, for incentivizing and for helping me. I would also like to thank the jury for reading the text.

João Pedro Olinto Dossena

Contents

Abstract

Gathering material for language learning and building a course demands resources, such as time, effort and money. The goal of this thesis is to investigate whether Natural Language Processing techniques, particularly Transformer models (Vaswani et al. [2017]), such as BERT (Devlin et al. [2019]), can be used to partially automate the process of creating language learning material. More specifically, text difficulty and semantical topic are important features with which to categorize learning material, and the goal is to find these features automatically. The Portuguese language is taken as a case study for these goals.

Firstly, a BERTimbau (Souza et al. [2020]) model is fine-tuned in the COPLE2 corpus (Mendes et al. [2016]) for the task of classifying text difficulty in Portuguese using the CEFR scale (Europarat [2020]), and then is compared to an existing GPT-2 (Radford et al. [2019]) model in the same task (Santos et al. [2021]). The result is that the model extensively trained in Portuguese (BERTimbau) significantly overperforms the GPT-2 model, which was converted from English to Portuguese with significantly less training data.

Secondly, a dataset of over 273 thousand bilingual sentences in English and Portuguese from Tatoeba (Tatoeba Association [2023]) is used for Topic Modeling using BERTopic (Grootendorst [2022]). BERTopic then extracts semantical topics from the dataset, and three topic models with different amounts of topics are the result. The topics of these models are then compared to the semantical topics of three major language learning applications - Duolingo (Duolingo, Inc), Babbel (Babbel, GmbH) and Memrise (Memrise Limited). Since BERTopic allows to find the most similar topic to a certain search term, one can find the most similar extracted topics to the ones from the apps, and their percentage of semantic similarity. The three different sized topic models are compared, showing that the models with more topics have more semantically similar topics to the app topics, but at the expense of topic size. In the end, by using the methods of this thesis, one can have a dataset of text classified by difficulty and by semantical topics. These two attributes could then be used by a course creator or by an independent student who is mining language learning material.

List of Figures and Tables

List of Figures

List of Tables

List of Abbreviations and Symbols

Abbreviations

NLP	Natural Language Processing
LM	Language Model/Modeling
SLM	Statistical Language Model/Modeling
NLM	Neural Language Model/Modeling
PLM	Pre-trained Language Model/Modeling
LLM	Large Language Model/Modeling
ANN	Artificial Neural Network
LSTM	Long Short-Term Memory
biLSTM	Bidirectional LSTM
GPT	Generative Pre-trained Transformer
SOTA	State-of-the-art
CEFR	Common European Framework of Reference for Languages: Learning, Teaching, Assessment
BERT	Bidirectional Encoder Representations from Transformers
L2	Second Language
MLM	Masked Language Modeling
NSP	Next Sentence Prediction
STS	Sentence Textual Similarity
RTE	Recognizing Textual Entailment
NLI	Natural Language Inference
NER	Named Entity Recognition
AES	Automatic Essay Scoring

Symbols

1

Chapter 1

Introduction

Language learning apps have a carefully designed path of content, a skill tree, that prioritizes what students will learn. For example, in Duolingo's (Duolingo, Inc) Dutch course, Unit 1 teaches basic phrases; Unit 2 teaches how to form negative phrases; Unit 3 talks about food and animals, and so on. The app incrementally introduces complex sentence structures and vocabulary as the student progresses through the tree.

Manually crafting these skill trees is cumbersome and needs effort from experts with domain knowledge. Moreover, hiring such experts to construct skill trees for languages with few speakers might not be economically feasible. However, recent advances in NLP, particularly the development of transformer models, allow for machines to achieve performance levels that are comparable to those of humans in certain language tasks, such as text classification, which could alleviate these pain points. As a result, the application of these recent methods could potentially automate the process of creating skill trees, effectively addressing the issue of manual construction.

1.1 Goal

The goal of this thesis is, therefore, to investigate the possibility of automatically generating such trees by leveraging NLP models.

To facilitate the generation of skill trees, a prerequisite is the acquisition of a dataset of phrases in the target language, alongside the development of an automated classification algorithm that can rank their difficulty levels. This classification process would enable the app to then present phrases to the student in an ordered and progressive manner. Additionally, it might be useful to be able to group sentences by topic, or by a certain grammatical feature, such as tense. From phrases classified by difficulty and by topic an algorithm can be devised to generate said skill trees.

In order to give a proof-of-concept of this automatic process of creating language learning paths, a specific language and specific methods are chosen. The experimental section of the thesis aims to implement this feasibility study, and can be broken down in two parts: *sentence classification*, and *crafting skill trees*.

1.1.1 Part 1: Sentence Classification

For the first part, sentence difficulty classification will be performed. In order to make things more tangible, a language to make such classification has to be chosen, and this language is Portuguese. Similarly, as sentence difficulty has to be measured in some concrete form, the CEFR scale (Europarat [2020]) will be employed, since it is widely used in characterizing language proficiency in European languages. By leveraging transfer learning, a language model trained on data in Portuguese can be fine-tuned to classify text difficulty with CEFR levels. Since some monolingual language models have shown better performance than multilingual models in certain NLP tasks, a monolingual model pretrained fully in Portuguese will be used for this experimentation part of the thesis. This model will be fine-tuned to the goal task of classifying sentences by CEFR levels on a corpus annotated with such CEFR scale. Thus the first research question this thesis aims to answer is:

RQ1 - *"What impact does a monolingual language model have on classifying the CEFR level of phrases in Portuguese?"*

1.1.2 Part 2: Crafting Skill Trees

In the second phase of the experiments, sentences will be classified according to their corresponding difficulty levels and a skill tree must be constructed using these phrases. In essence, an algorithm must be devised to determine the sequence in which the sentences are presented to the student, in order to provide guidance and facilitate learning. The simplest algorithm is to order phrases by CEFR level, and then additional enhancements can be made. An example of further improvement is filtering or grouping phrases by topic (such as "groceries") or grammatical feature (such as the future tense). As a result, the student can be exposed to sentences in their target language by topic of interest and in a progressively challenging order. Hence the second research question this thesis aims to answer is:

RQ2 - *"How can a language skill tree be automatically built from CEFR-labeled sentences?"*

The goal of the thesis will hence be achieved by attaining the goals of *Part 1* and *Part 2*.

1.2 Structure

The rest of this thesis is outlined in this section. Chapter 2 contains the background work. In this chapter the reader can find brief explanations to essential concepts upon which the thesis expands. Chapter 3 presents work related to the goals of the thesis. Chapter 4 contains a more detailed explanation of the experiments' approach, and their results. The main conclusion from the experiments as well as future work are contained in the last chapter (5).

Chapter 2

Background

Understanding the goals and experiments of this thesis presupposes some background knowledge. The first research question (*"What impact does a monolingual language model have on classifying the CEFR level of phrases in Portuguese?"*) presupposes that the reader has some knowledge about classification tasks, language models, and the CEFR scale, for example. Additionally, the second research question (*"How can a language skill tree be automatically built from CEFR-labeled sentences?"*) presupposes that the reader has some familiarity with language learning apps, and topic modeling, for example. In this sense, this chapter will briefly provide a context for the reader to understand the subsequent chapters, without getting into fine details that are not relevant for understanding the topics at hand.

2.1 Language Models

To put it briefly, language models aim to model the probability distribution of sequences, or, in other words, they assign a probability for a given sentence to occur (Deoras et al. [2011], Zhao et al. [2023]). There are different kinds of models that differ in the technique employed for such probabilistic estimation: statistical language models, neural language models, pretrained language models, and large language models, to name a few (Zhao et al. [2023]).

Statistical language models (SLMs), are a traditional language modeling technique that uses statistical estimation methods (Rosenfeld [2000], Zhao et al. [2023]). One major example of SLMs is the *n-gram* model, which models sequences of linguistic units as a Markov process with a context of fixed length $n - 1$ (Rosenfeld [2000], Zhao et al. [2023], Deoras et al. [2011]). In other words, the probability of the next word to be predicted depends on the $n - 1$ previous words. As a consequence, it has been argued that these models have a pitfall in capturing long-distance dependencies between words, with a distance longer than n words (Deoras et al. [2011], Chomsky [2009]). In addition to that problem, language modeling using statistical methods does not capture syntactic and semantic meaning of words, and is also subject to the curse of dimensionality, in which estimating probability distributions with a relatively large context can lead to evermore increasing computation required (Zhao

et al. [2023], Bengio et al. [2000], Naseem et al. [2021]). Despite its disadvantages, these statistical language models have brought advancements to the field of NLP (Zhao et al. [2023], Rosenfeld [2000]).

In order to address the pitfalls of SLMs, some work has proposed using artificial neural networks to perform language modeling, from here on called neural language modeling (NLM). One prominent advancement is using an ANN to learn a *distributed representation* vector for words and sentences (Bengio et al. [2000], Zhao et al. [2023], Naseem et al. [2021]). In this manner, the model can also take into consideration a kind of semantic similarity between words by leveraging the similarity between continuous real valued vectors in $\mathbb{R}^m$ (Bengio et al. [2000], Mikolov et al. [2013], Zhao et al. [2023]). Representation learning or feature learning is thus the type of method which learns vector representations that can be used for downstream tasks such as classification (Naseem et al. [2021]), while a *word embedding* is the feature learned for each word (Bengio et al. [2013]). Since the performance of the models is greatly affected by it, learning a useful representation is of utmost importance (Bengio et al. [2013], Naseem et al. [2021]).

In this context, *ELMo* is an attempt at learning useful representations while capturing syntactical and semantical meaning and polysemy (Peters et al. [2018]). *Embeddings from Language Models*, or, in short, *ELMo*, is a feature learning method. It uses a bidirectional language model (more specifically a biLSTM), which is pretrained on a large corpus in order to learn rich context-sensitive word embeddings (Peters et al. [2018], Zhao et al. [2023]). After pretraining this LM on unlabeled data, it is then fine-tuned on for certain downstream NLP tasks. Fine-tuning is the act of training a model (by updating its weights) for a specific task using many labeled examples (Brown et al. [2020]). This "pretrain and fine-tune" paradigm for language modeling (PLM) was also used for models with different architectures, such as GPT-2 (Radford et al. [2019]) and BART (Lewis et al. [2019]), and for more efficient training methods, such as RoBERTa (Liu et al. [2019]) (Zhao et al. [2023]).

When a PLM is sufficiently scaled (in dataset size and quality, amount of computation and number of parameters), it is called a large language model (LLM) (Zhao et al. [2023], Wei et al. [2022], Kaplan et al. [2020], Hoffmann et al. [2022]). LLMs are distinct to PLMs in the sense that they have shown emergent abilities, which are defined as abilities present in large models but not present in smaller models Zhao et al. [2023], Wei et al. [2022]. Models' abilities are considered emergent when, past a certain model size threshold, their performance substantially surpasses random performance (Zhao et al. [2023], Wei et al. [2022]). For example, GPT-3 (Brown et al. [2020]), a 175B parameter model, showcases emergent abilities when compared to smaller models in in-context learning (ICL). In short, in-context learning differs from fine-tuning in the following way: the latter goes through supervised, labeled training specific for a given task, and updates its parameters; and the former learns by analogy by just being given examples during inference time, without parameter updates (Dong et al. [2023], Brown et al. [2020]). Further, within ICL, one can categorize three approaches: few-shot, one-shot and zero-shot learning (Brown et al. [2020]). For a certain task, few-shot ICL consists of prompting the LM K input-output examples of the task, followed by an incomplete (input) example, which the

model should complete by understanding the patterns in the given demonstration (Brown et al. [2020], Dong et al. [2023], Wei et al. [2022]). Similarly, one-shot ICL is simply few-shot ICL where $K = 1$, while zero-shot ICL has a description of the task instead of input-output examples (Brown et al. [2020]).

In conclusion, language models assign probabilities to sequences of words, and the methods for that have evolved over time: from SLMs, to NLMs; from PLMs to LLMs (Zhao et al. [2023], Deoras et al. [2011]). SLMs use statistical techniques relying on recent context (Zhao et al. [2023], Rosenfeld [2000]), while NLMs have been used in representation learning for NLP – learning good word representation vectors (Zhao et al. [2023], Bengio et al. [2000], Naseem et al. [2021], Mikolov et al. [2013]). Subsequently, PLMs such as ELMo have established the paradigm of pretraining on unlabeled data and fine-tuning on labeled data for a specialized downstream task, which was adopted by later models (Zhao et al. [2023], Peters et al. [2018], Radford et al. [2019], Lewis et al. [2019], Liu et al. [2019]). Finally, scaling such PLMs into LLMs led to emergent abilities, such as in-context learning (Zhao et al. [2023], Wei et al. [2022], Brown et al. [2020]). Some of the models mentioned in this section, such as the GPTs (Radford et al. [2019], Brown et al. [2020]), BART (Lewis et al. [2019]) and RoBERTa (Liu et al. [2019]) utilize the *Transformer* architecture (Vaswani et al. [2017]), which will be further explained in the next section.

2.2 Transformers

The *Transformer* is a neural network architecture which initially improved the state-of-the-art on two machine translation tasks while having a smaller training cost than the competing models (Vaswani et al. [2017]). Besides using an encoder-decoder framework, its distinctive feature is that it uses the *(multi-head self-)attention mechanism*, instead of the recurrence used in well established recurrent neural network (RNN) models (Vaswani et al. [2017]). RNNs can be slow to train because they must process their inputs sequentially, while transformers can have their training parallelized because of the aforementioned mechanism (Vaswani et al. [2017], Niu et al. [2021]).

2.2.1 The attention mechanism

The main idea behind computational attention is to mimic its biological counterpart: when presented with a lot of information, humans focus on a more important fraction of it in order to process it with limited resources (Niu et al. [2021]). To put it simply, the attention mechanism gives a numerical score to how similar/compatible each vector is to other vectors, i.e. how important each word is in relation to all other words in the sequence (Bahdanau et al. [2016], Vaswani et al. [2017]). By performing optimized matrix multiplications, an attention head computes a context-aware attention matrix representation for the sequence. Consequently, multi-head attention allows multiple heads to compute such matrices in parallel, which are then aggregated into a final result, allowing for efficient training (Vaswani et al. [2017]). For illustration purposes, figure 2.1 shows the attention value between words in two

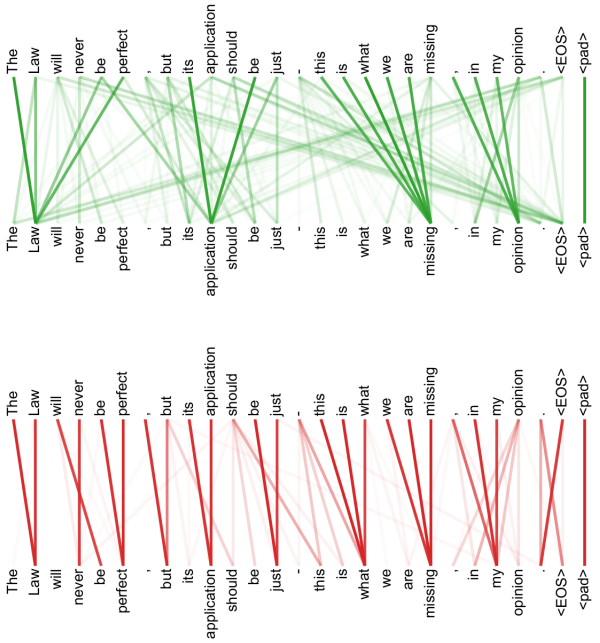

Figure 2.1: Attention values example for two different heads at layer 5 of the encoder part, taken from Vaswani et al. [2017]

different attention heads: the more intense the color in the connection is, the higher the attention value is between the words, and vice-versa (Vaswani et al. [2017]).

2.2.2 The encoder-decoder architecture

Besides the self-attention mechanism, the transformer architecture has an encoder-decoder structure (Vaswani et al. [2017], Lin et al. [2022]). In a nutshell, the encoder part receives as input a sequence of word embeddings $(x_1, ..., x_n)$, and outputs an intermediate representation $z = (z_1, ...z_n)$; while the decoder receives z and outputs a sequence $(y_1, ..., y_m)$ (Vaswani et al. [2017]). In figure 2.2, we can see the overall architecture of the transformer, with the encoder part on the left and the decoder part on the right (Vaswani et al. [2017]).

The encoder receives the sequence embeddings with positional encoding in order to capture word position information (Vaswani et al. [2017], Lin et al. [2022]). This is then fed into a multi-head attention block, and passes through layer normalization with a residual connection (Vaswani et al. [2017], Lin et al. [2022]), respectively proposed as solutions for faster training (Ba et al. [2016], Lin et al. [2022]) and to deal with accuracy degradation in deep networks (He et al. [2016], Lin et al. [2022]). The result goes into the position-wise feed-forward network module, which ensures the position information of the input, and then goes through another normalization with a residual connection step (Vaswani et al. [2017], Niu et al. [2021]). The encoder is formed by $N = 6$ stacked layers as described above, and a decoder is formed by

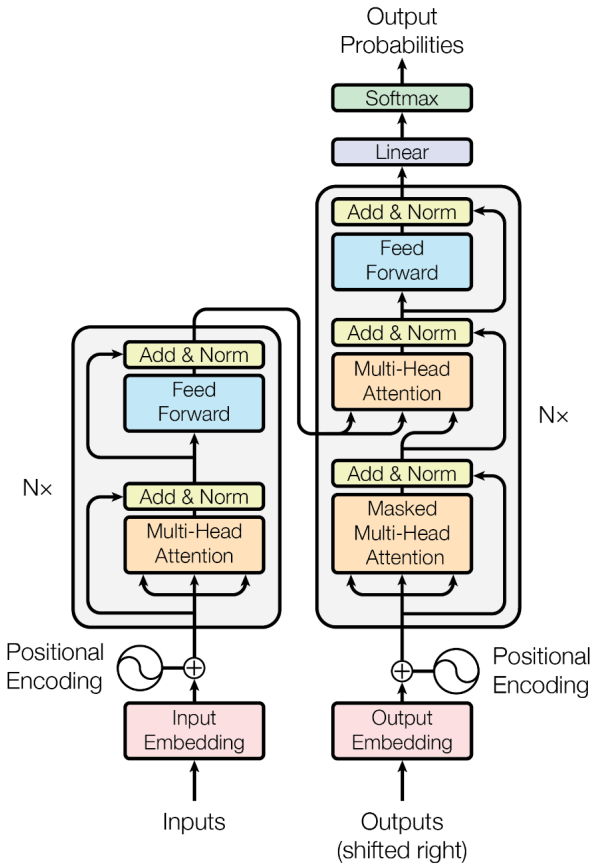

FIGURE 2.2: The Transformer architecture, taken from Vaswani et al. [2017]

$N = 6$ stacked layers that work in a similar fashion (Vaswani et al. [2017]). An important distinction in the decoder layers is the *masked multi-head attention* block, responsible for ensuring that the decoder can only predict the i-th word by attending to words before it (Vaswani et al. [2017]).

In general, a transformer model can be used in three different ways: *encoder-decoder* (the full transformer architecture), *encoder-only*, and *decoder-only* (Lin et al. [2022]). The full architecture is used generally for machine translation tasks, while the *encoder-only* outputted representation is typically used for text classification tasks (Lin et al. [2022]). Finally, the decoder can be used by itself for generating sequences (Lin et al. [2022]).

2.2.3 BERT

BERT, Bidirection Encoder Representations from Transformers, is a language model based on the transformer architecture which improved the state-of-the-art on eleven different tasks (Devlin et al. [2019]).

BERT is pretrained in two tasks: *Masked Language Modeling* (MLM), which is also known as the *Cloze Task*; and *Next Sentence Prediction* (NSP) (Devlin et al. [2019], Lin et al. [2022]). In MLM, some tokens are randomly substituted for the *[MASK]* token and the model has to learn to predict such tokens, which allows it to learn bidirectional representations (Devlin et al. [2019]). As per NSP, in order for the model to understand cohesion between sentences, it is pretrained on the task of predicting whether two sentences precede each other (Devlin et al. [2019]). Therefore, given sentences A and B, it has to classify whether B is the sentence that follows A (labeled *IsNext*) or not (labeled *NotNext*) (Devlin et al. [2019]).

The pretraining corpora for the original BERT are *BookCorpus*, with 800 million words, and Wikipedia in English, with 2.5 billion words (Devlin et al. [2019]). However, many variations have sprung from the original BERT: some have been trained on specific domain knowledge (Chalkidis et al. [2020]); some have modified its training regimen (Liu et al. [2019]); some have been trained on certain languages, be it monolingual (Souza et al. [2020], Le et al. [2020]) or multilingual (Devlin [2018]). For a more comprehensive evaluation of monolingual BERT models, refer to (Nozza et al. [2020]). An overview of a couple of variations of BERT will be presented below.

RoBERTa

RoBERTa (Robustly optimized BERT approach) is a variation of BERT done attending to more effective training (Liu et al. [2019]). The main adaptations to the original BERT training protocol are: not performing NSP, only MLM; more extensive training (for longer periods of time with more data and bigger batches); using longer sequences; and dynamic masking (Liu et al. [2019]).

BERTimbau

BERTimbau is a BERT model trained for Brazilian Portuguese which achieved state-of-the-art performance in three NLP tasks: *sentence textual similarity* (STS); *recognizing textual entailment* (RTE); and *named entity recognition* (NER) (Souza et al. [2020]). BERTimbau showed to outperform multilingual BERT (Devlin [2018]) on the tasks it was tested, demonstrating that it can be valuable to extensively train specific monolingual models (Souza et al. [2020]). The model was trained on the brWaC corpus (Filho et al. [2018]), which was the largest Portuguese open corpus, with over 2.5 billion tokens from 3.5 million diverse, high-quality documents (Souza et al. [2020], Filho et al. [2018]).

As mentioned earlier, this transformer model was evaluated in three tasks (Souza et al. [2020]). STS consists in using regression to estimate how semantically similar two sentences are, here in a continuous scale from 1 to 5: 1 meaning completely semantically dissimilar sentences; and 5 meaning semantically equivalent sentences (Souza et al. [2020], Nora Raju et al. [2022]). RTE, which is usually synonymous with natural language inference (NLI), is a binary classification task predicting whether a sentence can probably be inferred from another sentence (Souza et al. [2020], Poliak [2020]). For the tasks of sentence textual similarity and recognizing textual

entailment, the ASSIN2 (Real et al. [2020]) dataset was used (Souza et al. [2020]). Finally, named entity recognition consists of identified named entities in text, such as person, location, organization, etc (Souza et al. [2020], Li et al. [2022]). The datasets from the first HAREM contest (Santos et al. [2006]) were used for the task of NER.

BERTimbau performs significantly better than previous methods, and the authors hope their model can help to advance the state-of-the-art in other NLP tasks in Portuguese (Souza et al. [2020]).

2.2.4 GPT-2

GPT-2 (Generative Pre-trained Transformer 2) is a transformer-based language model with 1.5 billion parameters, which achieved state-of-the-art in 7 out of 8 different datasets by "zero-shotting" the varied tasks (Radford et al. [2019]). This model's results indicate that extensive unsupervised training of a LLM on a diverse corpus can achieve surprising performance on various tasks (Radford et al. [2019]).

2.3 CEFR

The Common European Framework of Reference for Languages (CEFR) is a framework of reference for language proficiency (Europarat [2020]). It presents six Common Reference Levels in a proficiency progression, respectively: A1, A2, B1, B2, C1, and C2 (Europarat [2020]). Learners in the levels A1 and A2 are classified as Basic Users, while those in B1 and B2 are classified as Independent Users, and those in C1 and C2 are classified as Proficient Users (Europarat [2020]). Although language proficiency is continuous across various abilities (oral comprehension, reading comprehension, among others), the CEFR levels are a simplification to make learning organized (Europarat [2020]). This scale will be important later when mentioning the task of CEFR classification/prediction by language models.

2.4 Readability Assessment

Readability assessment is important so writers can gauge whether their writings will be understood by a particular target audience (Klare [1974]). For that matter, one can measure readability by testing readers, or one can predict readability by using formulas that take linguistic features as input (Klare [1974]). Several formulas have been proposed over the years for predicting readability in English and in other languages (Klare [1974]). The aspect relevant for this thesis is predicting readability, since that is the goal for the first part, namely in the investigation of RQ1 (1.1.1).

2.5 Automated Essay Scoring

Automated Essay Scoring (AES), also known as Assessment of Proficiency or as Automated Text Scoring (ATS), is the task of assigning a grade to some essay, usually considered a supervised machine learning task (Kusuma et al. [2022], Taghipour and

Ng [2016]). AES has the advantages of saving human effort and time, diminish bias and improving consistency in scores, and its scores are considered more realistic than the ones given by humans (Kusuma et al. [2022]). This task is important for the first research question (1.1.1), since the goal is to classify text difficulty in the CEFR scale (Europarat [2020]) by using a monolingual transformer (Vaswani et al. [2017]) language model. In this sense, the goal would be to automatically assign a score to an arbitrary text discriminating on text difficulty. Further work on this type of AES that is more related to the goal of *RQ1* (1.1.1) is shown in chapter 3.

2.6 Language learning apps

Language learning has evolved in such a way that nowadays one can do it by using an application on their smartphones, such as Duolingo (Duolingo, Inc), Babbel (Babbel, GmbH), or Memrise (Memrise Limited), for example. These aforementioned apps have millions of downloads and reviews on Google Play Store (Google), and are in the top 5 search results for "language learning" in both Google Play Store and Apple App Store (Google, Apple). More specifically, Duolingo has over 100M downloads on Google Play Store and 14M reviews averaging 4.4 out of 5 possible stars (Google); Babbel has more than 50M downloads on the same platform, with 816 thousand reviews averaging 4.5 stars (Google); and Memrise has over 10M downloads with 1M reviews totaling 4.6 stars (Google), and around 65M learners (Memrise Limited). Furthermore, there have been studies testing the efficacy of such applications that conclude that they are indeed helpful for learning languages (Munday [2016], Vesselinov and Grego [2016], Aminatun and Oktaviani [2019]). Additionally, Duolingo has published an article where they explain and justify their methods (Freeman et al. [2023]). In this thesis, the versions of the apps referred are 5.115.4 for Duolingo, 21.32.0 for Babbel, and 2023.07.25.0 for Memrise.

The path of lessons that a learner goes through in a language learning application or method will be referred to as "*skill tree*" or "*language tree*". Since depending on the app, the skill tree can allow the learner to have more or less freedom in the learner's learning path, the terms "*skill tree*" and "*language tree*" will be used more loosely to refer to an organization of target language material into a certain logic. For example, this logic can be in difficulty and/or in topics, which can be semantic or grammatical topics. Language learning apps and their language trees are particularly relevant to the second goal of the thesis (1.1.2), more specifically in how to group relevant language learning material.

2.7 Topic Modeling

In short, topic modeling comprises the techniques to find a topic (or a collection of words) that describes a group of similar documents (Barde and Bainwad [2017]).

2.7.1 BERTopic

BERTopic is a topic model composed of a few steps: getting embedding representations out of textual documents; reducing the dimensionality of such embeddings in order to gain performance; clustering documents; and finally, extracting topic representations from document clusters using a modified TF-IDF method (Grootendorst [2022]). These steps will be further explained below.

Embedding representation

For getting vector representations of text in embedding space, BERTopic by default uses Sentence-BERT (SBERT), which has achieved state-of-the-art performance in sentence embedding tasks (Grootendorst [2022], Reimers and Gurevych [2019], Thakur et al. [2021]). SBERT receives text and returns embedding representations of text which preserve semantic similarity and can be compared by cosine-similarity (Reimers and Gurevych [2019]). Additionally, SBERT significantly beats BERT and RoBERTa in finding the most similar pair of sentences in a collection of 10 thousand sentences (approximately 5 seconds versus 65 hours), which enables a BERT-based model to tap into tasks such as clustering and large-scale semantic similarity comparison, to name a few (Reimers and Gurevych [2019]). Subsequently, BERTopic performs dimensionality reduction on these embeddings (Grootendorst [2022]).

Dimensionality reduction

The dimension of such embeddings can be reduced in order to achieve better performance in the later step of clustering (Grootendorst [2022]). BERTopic makes use of Uniform Manifold Approximation and Projection for Dimension Reduction (McInnes et al. [2020]), in short, UMAP, to reduce the dimensionality of the embeddings generated in the previous step (Grootendorst [2022]). This is done because UMAP can be used for different embedding dimensions, and it can possibly better preserve useful features in its projection than competing dimensionality reduction methods, such as t-SNE and PCA (Grootendorst [2022], McInnes et al. [2020]). For more details about UMAP, refer to its original paper (McInnes et al. [2020]). Thereafter, the lower-dimensional embeddings are clustered by BERTopic (Grootendorst [2022]).

Clustering

The clustering technique used by BERTopic on the reduced embeddings is HDBSCAN - Hierarchical Density-Based Spatial Clustering of Applications with Noise (Grootendorst [2022], McInnes et al. [2017]). It is a hierarchical version of DBSCAN which can model noise as outliers (Grootendorst [2022], McInnes et al. [2017]). Furthermore, HDBSCAN has shown to benefit from UMAP-reduced representations by achieving a faster and also a more accurate performance (Grootendorst [2022], Allaoui et al. [2020]). Topics are then extracted (using a modified version of TF-IDF) from these clusters that were derived from HDBSCAN (Grootendorst [2022]).

Topic representation

After having clusters representing semantically similar topics, it is necessary to know what these topic clusters represent (Grootendorst [2022]). BERTopic uses a modified version of TF-IDF, which is a statistical metric that represents the importance of a word in a certain document from a corpus (Grootendorst [2022], Joachims [1996]).

TF-IDF is the product of *Term Frequency* (TF) and *Inverse Document Frequency* (IDF), which are metrics that respectively measure how many times a word appears in a document, and how much information the word gives considering the context of the corpus with all documents (Grootendorst [2022], Joachims [1996]).

The modified version of this metric used by BERTopic basically makes TF measure the frequency of a word in a cluster of documents, rather than in a single document; and it makes IDF measure the information that a word gives to a cluster of documents, rather than to a single document (Grootendorst [2022]). These clusters with the extra information (what words are important in defining each topic) can then be merged in order to reduce the number of topics to a user-defined amount (Grootendorst [2022]).

2.8 Conclusion

In this chapter some background knowledge has been given in order for the reader to understand RQ1 (section 1.1.1), RQ2 (section 1.1.2), and the subsequent chapters.

Firstly language models were introduced (section 2.1), followed by the transformer architecture (section 2.2). The latter section explained the attention mechanism (section 2.2.1), the encoder-decoder structure (section 2.2.2) and introduced two prominent language models that have this architecture: BERT (section 2.2.3) and GPT-2 (section 2.2.4).

Subsequently, the CEFR scale was introduced (section 2.3), followed by some relevant tasks in NLP, such as Readability Assessment (section 2.4) and Automated Essay Scoring (section 2.5). Then, three prominent language learning mobile applications were introduced (section 2.6), followed by the field of Topic Modeling (section 3.3). Within that field, an overview of what steps BERTopic performs was given (section 2.7.1).

The next chapter (chapter 3) will build on top of the background knowledge exposed in this chapter, presenting related work.

Chapter 3

Related Work

This chapter introduces work related to the work presented in this thesis. Automatic readability assessment is introduced, firstly using feature extraction in a few different languages (section 3.1), and later by using transformers (section 3.2). Finally, the chapter is concluded, paving the way for chapter 4, which relates to the experiments.

3.1 Automatic Readability Assessment with Feature Extraction

There have been papers that performed automatic readability assessment with extracted features, such as sentence length, number of nouns, average word length, among others, particularly in languages other than English (Hancke [2013], Ostling et al. [2013], Vajjala and Lõo [2014], Johan Berggren et al. [2019], Río [2019]). These papers used a few different formulas that take such features as arguments. Such works are particularly relevant to this thesis, since the goal of the first research question entails CEFR level classification of text in Portuguese. The methods that use extracted features are precursor to the methods that use transformers to perform such classification.

3.1.1 German

In the context of a master's thesis, Automatic Readability Assessment has been done for German using various extracted features that capture linguistic knowledge from text to classify its CEFR level (Hancke [2013]). The dataset used is the MERLIN dataset (Boyd et al. [2014], Hancke [2013]). The feature set used is made of many lexical features (diversity, density, variation measures, among others), morphological features (inflectional morphology of verbs and nouns, derivational morphology of nouns, among others), syntactical features (parse tree, dependency features, among others), and language model features (n-grams perplexity scores) (Hancke [2013]). Refer to the original paper for an exhaustive list of the features used, as well as their definitions (Hancke [2013]).

Algorithms

A few algorithms were used to perform CEFR level prediction with the aforementioned features, such as Sequential Minimal Optimization (SMO, an algorithm for training Support Vector Machines - SVMs), Naive Bayes and Decision Tree (J48) (Hancke [2013]). These machine learning algorithms were compared to a majority baseline (as if every text was classified as the CEFR level class with the most examples) (Hancke [2013]). All three classification algorithms achieved significantly higher accuracy and F-measure than the majority baseline, with SMO performing the best (Hancke [2013]). The result was not surprising to the author, since SVMs work well with a large numerical feature space and have previously achieved satisfactory results in similar applications (Hancke [2013], Petersen and Ostendorf [2009], Vajjala and Meurers [2012]).

Feature correlation

In the paper, the author also presents which features are more correlated to the CEFR proficiency level of the text (Hancke [2013]). The highest dependency in syntactical features is the average number of dependent clauses per clause, with a correlation of 0.58, while the most correlated morphological feature is the derived nouns-nouns ratio (0.59) (Hancke [2013]). On the other hand, the most correlated lexical feature is text length (0.81), while language model features all had (absolute) correlations lower than 0.5 (Hancke [2013]). It makes sense that text length is such an important feature in this case, because the MERLIN dataset (Boyd et al. [2014]), used for this experiment, has a divide in text length depending on the proficiency level (Hancke [2013]). For a more comprehensive list of the features and their correlation to proficiency level, refer to the original paper (Hancke [2013]).

Feature group contribution

Lexical features in general seemed to be more correlated to the CEFR level than other kinds of features (Hancke [2013]). Consequently, to investigate which group of features is more relevant, the author trained separate models with each group of features with a holdout scenario and a 10-fold cross-validation scenario (Hancke [2013]). The groups evaluated were syntactic (SYN), lexical (LEX), language model (LM), and morphological (MORPH) features; along with Parse Rule (PR) and Tense features (TEN) that were taken as separate groups, since they differ enough from other syntactic and morphological features respectively (Hancke [2013]). Furthermore, the text length was taken as an extra baseline, since it was the feature with the highest correlation to the CEFR proficiency predictions (Hancke [2013]). Thus, in the end, the groups evaluated were the majority baseline, text length baseline, SYN, LEX, LM, MORPH, PR and TEN (Hancke [2013]). As a result, in the holdout data, lexical features were the best performers, though slightly below the text length baseline for accuracy, but slightly better than the text length baseline for F-measure (Hancke [2013]). On cross-validation on all data, lexical features were the only group of features with better accuracy and F-measure than the text length baseline (Hancke

[2013]). Therefore, lexical features confirm to be the most relevant of the features tested for proficiency prediction (Hancke [2013]).

Feature group combinations

Next, the author investigated the possible combinations of (varying numbers of) feature groups to see if some feature types could complement each other (Hancke [2013]). On cross-validation, the best combination was LEX_LM_MORPH, while on the holdout scenario the best are LEX_LM_MORPH and LEX_MORPH with the same accuracy, but the latter had slightly higher F-measure (Hancke [2013]). Interestingly, combinations with 4 or with all groups performed the worst, worse than the lexical group alone, indicating that sufficient irrelevant features can accumulate, which consequently deteriorates the models' performance (Hancke [2013]).

Feature selection

Since irrelevant features deteriorate models' performance, feature selection is done to keep the most important features that are not too correlated to each other in order to eliminate redundancy (Hancke [2013]). The most performant combination of features consists of 34 features from the 4 different groups (syntactic, lexical, language model, and morphological) (Hancke [2013]). This model is composed by the text length, by measures of vocabulary difficulty and of lexical richness and variation, by features that gauge spelling mistakes, among others (Hancke [2013]).

Conclusion

Among other things, the paper has presented an extensive investigation on feature-based proficiency classification in the CEFR scheme using the MERLIN (Boyd et al. [2014]) dataset for the German language(Hancke [2013]). A few algorithms were compared and SMO brought the best results, hence it was used for the rest of the experiments (Hancke [2013]). A set of over 3000 features of different kinds was devised and they were empirically tested (Hancke [2013]). The combination of features that achieved the best performance was a set of 34 features that encoded most of the linguistic aspects previously tested in a condensed form (Hancke [2013]). All in all, the paper presents insights that can be useful for further investigation. One of the suggested tasks left for future work is to attempt proficiency classification using "more sofisticated machine learning techniques" (Hancke [2013]), which motivates the use of more modern models such as transformer-based ones, relating to the first goal of the thesis (section 1.1.1).

3.1.2 Swedish

Regarding the Swedish language, there is existing work done towards proficiency level prediction of text using the CEFR scale and linguistic features with machine learning algorithms (Pilán [2018]).

Features

The author presents a flexible set of features for the task, which can be used for sentences and for longer texts (Pilán [2018]). This devised set contains 61 features that can be further categorized into "count-based, lexical, morphological, syntactic and semantic" features (Pilán [2018]). A comparison in classifying CEFR proficiency from A1-C1 in the sentence and text level was then made between a majority baseline, a traditional readability measure (*LIX*), and the set of all features (Pilán [2018]). The result was that LIX and the majority baseline had similar metrics (33% accuracy in texts), while the set of all features did significantly better (81.3% accuracy in texts) (Pilán [2018]). The performance achieved by the set of features is comparable to that of human graders (Pilán [2018]). By analyzing the contribution of each feature group, the author found lexical features to get almost the same results as the full feature set at the text level (Pilán [2018]). However, at the sentence level the difference was starker (Pilán [2018]). Further, the paper shows that texts of a specific CEFR level contain many sentences of lower CEFR levels (Pilán [2018]).

Experiments on learner texts

Moreover, this paper investigates the impact of texts written by L2 learners of Swedish versus coursebook texts in training a model to classify proficiency levels of student texts (Pilán [2018]). Three approaches were tried, namely: training a text classifier only on coursebook essays, written by native speakers; training a classifier on both coursebook data and on essays written by L2 learners; and estimating values of lexical complexity features by using coursebook data (Pilán [2018]). Since one of the primary problems with text written by L2 learners is that it might contain mistakes, the author also examined whether correcting, possibly partially, these mistakes would lead to a greater classification performance (Pilán [2018]). What was found was that coursebook data can be successfully used as extra data for training; coursebook texts can be used as the only training data when the mistakes of L2 texts to be classified are corrected; and they can also be used to estimate lexical feature values (Pilán [2018]).

Finding most important features

As eliminating redundant and unimportant features for prediction can be beneficial for performance, the thesis presents experiments to investigate which features of the devised set of features are most important (Pilán [2018]). For that end, a univariate feature selection method was employed, which found 14 features that were considered relevant in all three datasets tested (Pilán [2018]). A comparison was then made between models that use the k-best features and models that use the whole set of features (Pilán [2018]). As expected, the values of accuracy and F_1 using k-best features significantly outperformed models that used the whole set of features (Pilán [2018]).

Conclusion

In conclusion, among other things, the author investigates CEFR proficiency level prediction of texts and sentences for Swedish using machine learning models relying on selected linguistic features (Pilán [2018]). The set of features performs better than the baselines chosen, and has comparable performance to that of humans (Pilán [2018]). A further inspection of which features are the most important for such prediction showed that a smaller feature set can significantly improve performance of the classifier models (Pilán [2018]).

3.1.3 Estonian

Similarly to the above examples of German and Swedish, there is also existing work attempting to predict the CEFR level of text in Estonian using linguistic features (Vajjala and Lõo [2014]). The authors build on top of their own previous work (Vajjala and Loo [2013]) by adding new features to the set of features previously devised (Vajjala and Lõo [2014]). Moreover, the paper compares the results of modeling the problem as a classification and a regression problem (Vajjala and Lõo [2014]). They find that modeling the problem as a classification problem yields better results than as a regression problem (Vajjala and Lõo [2014]).

Corpus

The corpus used is a subset of the Estonian Interlanguage Corpus (EIC ELLE), which is comprised of around 12 thousand texts written by second language learners of Estonian and was made available by Tallinn University (Vajjala and Lõo [2014]). More specifically, the subset used for the experiments is made of 879 documents whose CEFR proficiency level was annotated, ranging from levels A2 to C1 (Vajjala and Lõo [2014]). This is a different version of the same corpus used in the previous work by the same authors (Vajjala and Loo [2013]), with the distinction that in the previous version the grading system of proficiency levels was different (A, B and C vs. A1 to C2) (Vajjala and Lõo [2014]).

Features

The starting set of features was the same as in their previous work (Vajjala and Loo [2013]), and some extra features were added, most of which being lexical richness features (Vajjala and Lõo [2014]). Since surface features such as number of words are highly unbalanced across classes, this can make the algorithm put too much weight on these kinds of features (Vajjala and Lõo [2014]). However, the authors sought to find out the impact of morpho-syntactic features on the prediction task without these surface features (Vajjala and Lõo [2014]). Therefore the authors temporarily removed such features from the feature set for that end (Vajjala and Lõo [2014]). They also used features that measure morphological complexity, and lexical variation (Vajjala and Lõo [2014]). Word and part-of-speech (POS) language models were

discarded because of insufficient results in initial testing (Vajjala and Lõo [2014]). For more details about features refer to the original work (Vajjala and Lõo [2014]).

Experiments

For the task of classification, Sequential Minimal Optimization (SMO) was used with the feature set devised in the authors' previous work (Vajjala and Loo [2013], Vajjala and Lõo [2014]). This approach achieved an accuracy of 73.7% and 72.3% (respectively on an unbalanced and on a balanced dataset), and was taken as a baseline for the experiments (Vajjala and Lõo [2014]). Then the new set of all features was tested and achieved 79% and 76.9% (unbalanced and balanced datasets, respectively), which demonstrates a clear improvement over the baseline (Vajjala and Lõo [2014]). When comparing the models trained on unbalanced data and balanced data, the authors deliberated that the unbalanced version is the superior model for achieving higher performance while not being skewed to favor majority classes too much (Vajjala and Lõo [2014]). Although some performance has been gained compared to the previous work by the same authors (Vajjala and Loo [2013]) by using more fine grained classes (A2 to C1 versus A, B and C), some performance gain can be attributed to the improved feature set (Vajjala and Lõo [2014]).

Additionally to multiclass classification, binary classification has also been experimented (Vajjala and Lõo [2014]). In this case, many classifiers were used to distinguish between every pair of classes, e.g. A2 versus B1, A2 versus B2, and so on (Vajjala and Lõo [2014]). As a result, the accuracies of these models were generally significantly higher than the multiclass approach, which suggests further studies of a multi-level classification can be fruitful (Vajjala and Lõo [2014]). Furthermore, surface features representing text length were added back into the feature set, however the small decrease in accuracy was not significant (Vajjala and Lõo [2014]).

Proficiency level can also be tackled as a regression problem, since proficiency is continuous, even though CEFR levels are discrete (Vajjala and Lõo [2014]). A linear regression model was thus trained for that task, and, with the feature set of the authors' previous work, it achieved a Pearson correlation of 0.77 and an RMSE of 0.58 (Vajjala and Lõo [2014]). Similarly, using the newer feature set, the Pearson correlation was 0.85 and the RMSE was 0.49, both showing an improvement (Vajjala and Lõo [2014]). By comparing classification and regression, the paper finds that classification performs slightly better (Vajjala and Lõo [2014]).

Feature selection

After investigating the above approaches, the next step was to select the most important features for such a prediction (Vajjala and Lõo [2014]). Three methods for feature selection were chosen, namely Information Gain, CfsSubsetEval (Hall M. A [1998]), and ReliefFAttributeEval (Kira and Rendell [1992], Kononenko [1994]). The best method was CfsSubsetEval, achieving an equivalent accuracy to the previous best model, while using much fewer features for the proficiency prediction task (Vajjala and Lõo [2014]). CfsSubsetEval selected 27 features, of which the ten most

important ones represent lexical variability and morphological features (Vajjala and Lõo [2014]). Although CfsSubsetEval accounts for redundancy in features, a correlation measurement between the features selected in this method shows that some feature pairs are highly correlated (Vajjala and Lõo [2014]). The authors thus suggest that it might be possible to discard one of the features in each pair of highly correlated features and still achieve comparable results, but leave it for further analysis (Vajjala and Lõo [2014]).

Comparing features across categories

Finally, the paper investigates which features are the most predictive for certain classes (Vajjala and Lõo [2014]). The reasoning behind this investigation is that perhaps some features might be more important for predicting lower CEFR levels, while some other features might be more important for predicting higher CEFR levels (Vajjala and Lõo [2014]). The hypothesis presented is that morphological features are more important for predicting lower proficiency, but lexical richness features are more important for higher proficiency levels (Vajjala and Lõo [2014]). The findings were that the most important features to predict between the pairs (A2,B1) and (B2,C1) were morphological, while for the pair (B1,B2) lexical richness features did better (Vajjala and Lõo [2014]). Further investigation on the reason why there is a change in the intermediate levels (B1 and B2) is left for future work (Vajjala and Lõo [2014]).

Conclusion

The paper presents feature-based models for the task of proficiency classification of Estonian learner texts in the CEFR scale (Vajjala and Lõo [2014]). The best accuracy achieved was 79%, which is an improvement to previous work in Estonian and to work in other languages, such as German and Swedish (Vajjala and Lõo [2014]). The results achieved thus indicate that further research in the area might be fruitful (Vajjala and Lõo [2014]).

In the paper, a comparison was made treating the problem as a classification and as a regression (Vajjala and Lõo [2014]). The authors found better results when modeling the problem of proficiency prediction as a classification problem (Vajjala and Lõo [2014]). Additionally, different feature selection methods have been tested, and it was found to be possible to achieve accuracy equivalent to the best model with just a small subset of the original features (Vajjala and Lõo [2014]). Moreover, a correlation measurement between the best predicting features was done, and it showed that many features are highly correlated (Vajjala and Lõo [2014]). However, an investigation to eliminate the redundancy of correlated features while maintaining accuracy is left as future work (Vajjala and Lõo [2014]). The authors conclude that researching other learning algorithms for the proficiency assessment task is interesting future work, which motivates the first goal (section 1.1.1) of this thesis (Vajjala and Lõo [2014]).

3.1.4 Norwegian

Similarly to the studies presented above, there is work in Automated Essay Scoring (AES) for the Norwegian language using the CEFR scale (Johan Berggren et al. [2019]). Much like the work for Estonian (Vajjala and Lõo [2014]), this paper investigates the difference between AES as a regression and as a classification problem (Johan Berggren et al. [2019]). Furthermore, the authors investigate the impact of different (linear and neural) models for the task; and also attempt multi-task learning for both AES and for identifying the native language of the text writer (Johan Berggren et al. [2019]). The best results are achieved by a Gated Recurrent Unit (GRU) trained in a single task for Automatic Essay Scoring (Johan Berggren et al. [2019]).

Corpus

The ASK corpus (Tenfjord et al. [2006]) was used in the study, and it is composed of essays for B1 and B2 proficiency tests in Norwegian (Johan Berggren et al. [2019]). The texts in the corpus have annotation indicating the learner's native language, and some later work (Carlsen [2012]) has added CEFR proficiency levels as extra annotation to a subset of the corpus (Johan Berggren et al. [2019]). The corpus has essays with proficiency annotation ranging from A2 to C1, but there are also intermediate gradings between every pair of adjacent levels (such as A2/B1), thus the final number of classes is seven, making the grading scale fine grained (Johan Berggren et al. [2019]). However, the authors also did tests with the class ranges being A2, B1, B2 and C1, which they call "collapsed set of classes" (Johan Berggren et al. [2019]). When analyzing the dataset, the authors find that essays written by learners whose first language is similar to Norwegian (German, English) usually have higher scores than those written by students who speak more dissimilar languages (Johan Berggren et al. [2019]).

Experiments

In the context of investigating the impact of modeling AES as a regression and as a classification problem, three algorithms were tried, namely Logistic Regression, Support Vector Regression and SVM (Johan Berggren et al. [2019]). The authors then combine AES and Native Language Identification (NLI) in the same model in the attempt to improve the performance on AES (Johan Berggren et al. [2019]). On the question of whether AES should be seen as a regression or as a classification problem, the authors identify a problem in using regression: we do not know whether the distance between each pair of adjacent classes is the same (Johan Berggren et al. [2019]). However, they also identify a problem in classification, which is that the order of classes in the CEFR scale is not taken into consideration (Johan Berggren et al. [2019]).

The best performing linear model was Support Vector Regression, and the paper finds that modeling the task as a regression problem gives better results than as a classification problem, at least for the dataset utilized (Johan Berggren et al. [2019]).

Moreover, several neural models are trained, namely convolutional neural networks (CNNs) and gated RNNs, and the authors experiment with both randomly initialized embeddings and fine tuned embeddings (Johan Berggren et al. [2019]). For classification of collapsed labels, the best model was a "CNN ordinal rank regression with POS tags as input" (Johan Berggren et al. [2019]); while for classification of all labels, the best model was a "CNN regression model with mixed POS tags as input" (Johan Berggren et al. [2019]).

Regarding RNNs, the authors modified the architecture of the models used in some previous work (Taghipour and Ng [2016]) for AES as a regression problem (Johan Berggren et al. [2019]). A few experiments were done for RNNs, namely LSTMs versus Gated Recurrent Units (GRUs); bidirectional versus unidirectional; and different attention mechanisms (Johan Berggren et al. [2019]). Similarly to the previous experiment, the authors also experimented with randomly initialized embeddings versus pretrained ones (Johan Berggren et al. [2019]). The experiment resulted in GRUs faring better than LSTMs (Johan Berggren et al. [2019]).

The task of Native Language Identification is modeled as a classification problem (Johan Berggren et al. [2019]). In this respect, CNNs, LSTMs and GRUs were tried for this task (Johan Berggren et al. [2019]). Again, GRU models performed the best out of the attempted ones (Johan Berggren et al. [2019]). However, the best model presented fares worse than previous work on which the paper is based, but the two implementations of the task of NLI are not directly comparable for they do not use equivalent methodologies (Johan Berggren et al. [2019]).

Besides training models separately for each task, this paper experimented with training multitask models, capable of performing both tasks (Johan Berggren et al. [2019]). From this experiment, the paper finds that adding NLI as an auxiliary task to AES at least does not significantly decrease the original task (Johan Berggren et al. [2019]).

All of the experiments mentioned above were tested in a development set (Johan Berggren et al. [2019]). However, a final test on a held out set shows that "a fine-tuned embedding BiGRU model augmented with attention and initiated with FastText word embeddings performs the best" (Johan Berggren et al. [2019]). For more details on the results obtain, refer to the original paper (Johan Berggren et al. [2019]).

Conclusion

To conclude, this paper investigated the task of Automatic Essay Scoring on the ASK dataset using various models (Johan Berggren et al. [2019]). Modeling the task as a regression has shown to be more useful than as a classification problem, at least on the dataset utilized (Johan Berggren et al. [2019]). Support Vector Regression seemed to be the model with the best performance out of the non-linear ones (Johan Berggren et al. [2019]). Neural models have been used for AES and NLI separately, and as a joint task (Johan Berggren et al. [2019]). The paper found that bidirectional GRU models performed the best at AES and at NLI, although with different attention mechanisms (Johan Berggren et al. [2019]). Moreover, the authors

have attempted multitask models, of which the main task would be AES, and NLI would serve as an auxiliary task (Johan Berggren et al. [2019]). However, despite not significantly decreasing the performance of AES, the auxiliary task did not appear to help in AES (Johan Berggren et al. [2019]). All in all, the work presented motivates the use of more recent neural models, such as transformers, for predicting proficiency in the CEFR scale.

3.1.5 Portuguese

A paper experimented with proficiency classification in the CEFR scale for text written by learners of Portuguese (Río [2019]). In this paper, the author tries utilizing different features and different algorithms for such task, besides presenting an improvement to the already existent dataset (Río [2019]).

Corpus

Large learner-written corpora are available in English, but are harder to find in other languages (Río [2019]). With the intent of solving this problem, the paper presents improvements to the NLI-PT dataset (del Río et al. [2018]), which has data for European Portuguese (Río [2019]). As the name indicates, the dataset was originally conceived for the task of Native Language Identification (Río [2019]). Besides raw text data, the NLI-PT dataset also has annotations containing morphological, such as POS, and syntactic features, such as constituency and dependency (Río [2019]). The new version of the dataset presented in the paper, besides being larger, has some improvements in tokenization and in annotation (Río [2019]). Since the NLI-PT dataset is composed of different corpora, the CEFR labeling of each corpus is not the same (Río [2019]). Some corpora in the dataset have texts ranging from A1 to C1, while some other corpora have only A, B or C texts (Río [2019]). In order to standardize the corpora, the author chose to work only with A, B and C labels (Río [2019]). The final dataset contains 3069 texts written by learners who speak 15 different native languages (Río [2019]).

Features

Different linguistic features extracted from the dataset were experimented with (Río [2019]). Some examples include Bag of Words (of which only the best performing word for representation was maintained); POS n-grams of different sizes; dependency triplets n-grams (following the format *head*, *relation*, *dependent*); a set of 39 descriptive and lexical features, which previous studies have proven to be predictive of proficiency, such as number of nouns and of verbs, syllable count, etc (Río [2019]).

Experiments

The task of AES was modeled as a classification problem following the positive results of previous work (Río [2019], Vajjala and Lõo [2014]). The models with which the author experimented were Logistic Regression, Linear Discriminant Analysis,

Support Vector Machines, Random Forests and LogitBoost (Río [2019]). However, the models used with the test set were the three best performing ones in the initial investigation, namely Logistic Regression, Random Forests and LogitBoost (Río [2019]). The models fared better when using the whole dataset as opposed to a balanced version of it (Río [2019]). The best performing model, an ensemble of the whole feature set with Logistic Regression, achieved a 72% accuracy, followed closely (70%) by an LR using Bag of Words (Río [2019]). For more details on the results refer to the original paper (Río [2019]).

Conclusion

The paper presents improvements to the NLI-PT dataset, and, using this improved dataset, the author performs Automatic Essay Scoring in Portuguese on the CEFR scale (Río [2019]). Different features and machine learning algorithms are explored, with the best model achieving 72% accuracy (Río [2019]). The author delves deeper into an investigation on groups of features, different models, balanced versus unbalanced data class-wise, etc (Río [2019]). AES using the CEFR scale for Portuguese relates to the first goal of the thesis (section 1.1.1), and motivates the usage of newer models, such as transformers, for the task, which is done in a paper presented in section 3.2.1 (Santos et al. [2021]).

3.2 Automatic Readability Assessment with Transformers

3.2.1 Automatic CEFR classification in Portuguese with Transformers

One particular paper presented neural network-based approaches in contrast to feature-based traditional methods for the task of proficiency assessment of text in Portuguese using the CEFR scale (Santos et al. [2021]). More specifically, the neural approaches are based on transformers, and the models achieved state-of-the-art performance (Santos et al. [2021]).

Corpus

The corpus is compiled by Camões IP (Camões I.P.), which is responsible for organizing language proficiency assessment tests for Portuguese, targeted at L2 learners (Santos et al. [2021]). It is comprised of 500 texts in Portuguese from various sources such as news, books and articles, and it is considerably larger than previous corpora used in Portuguese (Santos et al. [2021]). The authors experimented with a few different versions of the corpus, such as a balanced version, approximated versions of other previous corpora, among others (Santos et al. [2021]). In total, 5 corpora taken from the original, including itself, were used for the experiments (Santos et al. [2021]).

Methodology

For a feature-based method, they recreate the work from another paper (Curto et al. [2015]), in which 52 features are extracted, including part-of-speech; words; averages and frequencies, to name a few, and a LogitBoost algorithm is trained.

On the other hand, for the transformer-based method, the authors of the paper train two models: GPT-2 and RoBERTa (Santos et al. [2021]). As per the former model, they use a 124 million-parameter GPT-2 model (Guillou [2020], Radford et al. [2019]) from the *Transformers* library (Wolf et al. [2020]), which is fine-tuned from English to Portuguese. This fine-tuning is done on approximately 1GB of data from the Portuguese Wikipedia, in roughly a day, on a NVIDIA V100 32GB GPU. For the latter model, since there was no RoBERTa model in Portuguese, the authors trained a 68 million-parameter model themselves. This model was trained on 10 million English sentences and 10 million Portuguese sentences from the Oscar corpus (Suarez).

In their results, they found the GPT-2 model to have the best performance on their largest balanced dataset. However, the GPT-2 model trained on the largest dataset has seen more examples of data and could represent real world data more accurately. Therefore, it has been made available through an API (noa [a]) where one can input text in Portuguese and have the latter model classify its difficulty in the CEFR scale.

Conclusion

In light of the paper, the authors conclude that neural-based models such as transformers can achieve state-of-the-art performances in text difficulty classification, particularly with under-resourced languages such as Portuguese (Santos et al. [2021]). Additionally, the authors suggest as future work the research of transfer learning from more resourceful languages, and of techniques to artificially increase the dataset used for training (Santos et al. [2021]). This paper, thus, gives motivation for improvements on the task of CEFR level prediction of text in Portuguese using transformer-based language models, by using larger training datasets and models trained more extensively on Portuguese data. This motivation matches the first goal of the thesis (1.1.1), namely investigating *"What impact does a monolingual language model have on classifying the CEFR level of phrases in Portuguese?"*. Therefore this work will be the most relevant in the experiments relating to the first goal of the thesis in chapter 4.

3.3 Topic Modeling

The techniques related to topic modeling can be used in many application domains, such as in the legal field (Silveira et al. [2021]) or in politics (Silva et al. [2021]), however not many closely resemble the second goal of the thesis. Nonetheless, it is possible to find papers that make use of BERTopic to extract topics in languages other than English (Abuzayed and Al-Khalifa [2021]).

3.3.1 Topic Modeling with BERTopic for languages other than English

A paper experiments with BERTopic for performing topic modeling in the Arabic language and compares its results to more popular topic modeling techniques, namely Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) (Abuzayed and Al-Khalifa [2021]).

Setup

The section presenting previous work in the paper finds that LDA is the most used technique for topic modeling in Arabic, more recently being combined with K-means clustering and with word2vec embeddings (Abuzayed and Al-Khalifa [2021]). The authors then use LDA and NMF as baselines to then compare to BERTopic with a varying number of topics and varying word embedding representations for Arabic (Abuzayed and Al-Khalifa [2021]). The dataset used was the Dataset for Arabic Classification (Biniz [2018]), with more than 100 thousand documents from three online newspapers in Modern Standard Arabic (Abuzayed and Al-Khalifa [2021]). For the challenging task of evaluation metrics, although imperfect, the authors used Normalized Pointwise Mutual Information (NPMI), which measures the coherence between the ten highest scoring words in the topic (Abuzayed and Al-Khalifa [2021]).

Results

In their experiments, LDA achieved quite low scores for all total number of topics (Abuzayed and Al-Khalifa [2021]). At the same time, NMF had the best scores for 50 to 150 topics, while the BERTopic topic models are competitive with NMF and outperform it for 200 to 500 topics (Abuzayed and Al-Khalifa [2021]). Most of the word embeddings used with BERTopic achieve similar metrics, with the exception of one, which has a performance drop at around 200 topics, still outperforming NMF but underperforming its BERTopic competitors (Abuzayed and Al-Khalifa [2021]).

Conclusion

The authors conclude that BERTopic has promising results, but that accurately evaluating topic models is not straightforward and is a challenging task that needs more research (Abuzayed and Al-Khalifa [2021]). Through the example of promising results given by the aforementioned paper (Abuzayed and Al-Khalifa [2021]) and the fact that BERTopic (Grootendorst [2022]) provides many analysis and visualization functionalities there is motivation for the use of such techniques for language learning topic modeling.

3.4 Conclusion

By the end of this chapter the reader has got acquainted with work related to this thesis, which will be built upon in the next chapters. The thesis presented some

examples of automatic readability assessment that use feature extraction (section 3.1) or that use transformers (section 3.2); as well as an example of the use of BERTopic for topic modeling in a language other than English (section 3.3).

Of these works, the most relevant one which will be built upon in the next chapters for the first goal of the thesis is the one that makes use of transformers for CEFR classification in Portuguese (Santos et al. [2021]), presented in section 3.2.1. This is because the first goal of the thesis is to investigate the first research question (section 1.1.1). For that matter, the thesis will investigate the impact of a monolingual language model in predicting CEFR levels for Portuguese text, which is related to what is presented in said paper (Santos et al. [2021]).

Chapter 4

Experiments

This chapter presents the experiments done regarding the two tasks in which the goal of the thesis was divided. The first one is to train a monolingual transformer model to classify the CEFR level of text in Portuguese. The second task is to devise an algorithm to order the CEFR-classified sentences. Each section respectively relates to one of these tasks.

4.1 Part I: Training BERTimbau for CEFR level prediction of text in Portuguese

This section describes the setup and results of training BERTimbau for CEFR level prediction of text in Portuguese.

4.1.1 Setup

In a similar goal to a paper described in chapter 3 (section 3.2) (Santos et al. [2021]), one of the aims of this thesis is to investigate the impact of a monolingual transformer model in predicting the CEFR level of sentences in Portuguese. This differs from the work in that paper in the sense that one of their transformer models is not monolingual, while the other is not trained extensively in Portuguese (Santos et al. [2021]), thus they can most likely be improved. To quickly summarize the paper's approach, the authors used a GPT-2 model and a RoBERTa model (Santos et al. [2021]). The RoBERTa model was trained from scratch, with half of the training data being in Portuguese and half being in English (Santos et al. [2021]). The GPT-2 model, which was found to be the best performing one, had its weights initialized from an English version of the model, and was briefly trained on Portuguese data to convert it to a Portuguese model (Santos et al. [2021], Guillou [2020]). This motivates the training of a more extensively trained Portuguese monolingual model. Then it is possible to compare its performance with the GPT-2 model used by the aforementioned paper (Santos et al. [2021]) in order to investigate the impact of a monolingual transformer model in CEFR prediction in Portuguese.

Transformer model

As per the transformer model choice, one can start with a powerful NLP model fully pretrained in Portuguese, namely *BERTimbau* (Souza et al. [2020]), which can be found in the *Transformers* library (Wolf et al. [2020]). The version of *BERTimbau* utilized was *bert-base-portuguese-cased*, which has 12 layers and 110 million parameters. BERTimbau was trained on 17.5GB of data in Portuguese (Souza et al. [2020]), while the GPT-2 model to compare to was trained on only around 1GB of Portuguese data (Guillou [2020]). This BERTimbau model was then fine-tuned on the task of classifying text in Portuguese by CEFR level.

Corpus

Since the datasets used in the aforementioned paper are private, we had to get access to public corpora to train our model. For this reason, COPLE2 (Mendes et al. [2016]) - a learner corpus in Portuguese - was used. This dataset consists of many texts written by L2 learners of Portuguese (Mendes et al. [2016]).

In order to gather the data entries of this corpus, a webscraper was needed to individually download each of the files, since each file could only be downloaded individually in each file's webpage. Furthermore, most entries in the corpus had a few different downloadable versions, such as the student's original version, and additionally an orthographically correct version. Since the goal of the thesis is to later use a classifier to grade orthographically correct sentences in Portuguese, only the orthographically correct versions of the entries were used. Thus the final dataset was comprised of 1020 usable entries. The class distribution can be seen in table 4.1. The class distributions for the training, validation and test sets are equivalent to the one of the total dataset.

Classes	Entries	Approx. Percentage
A1	92	9%
A2	399	39.1%
B1	292	28.6%
B2	199	19.5%
C1	38	3.7%
C2	0	0%
Total	1020	100%

Table 4.1: Class distribution of the utilized entries from COPLE.

Training

The fine-tuning to the task of classifying text in the CEFR scale was performed with a training set, a validation set and a test set. The training set contained 80% of the collected dataset, while the validation and the test sets were each comprised of

Model	Accuracy	Precision	Recall	F1 score
GPT-2	41.2%	49.1%	41.2%	40.1%
BERTimbau	**86.3%**	**87.9%**	**86.3%**	**86.4%**

TABLE 4.2: Accuracy, precision, recall, F1: BERTimbau vs GPT-2

10% of the dataset. This training took less than an hour using a GeForce RTX 3050 Mobile GPU by NVIDIA (noa [b]), which performed 4 epochs (with early stopping) to obtain its minimum loss value on the validation set.

4.1.2 Results

After fine-tuning BERTimbau to the task of CEFR level classification, the model was tested on the aforementioned test set, which makes up 10% of the total dataset (102 out of 1020 instances). These instances have also been fed into the API provided in noa [a], which uses the baseline GPT-2 model (Santos et al. [2021]) to compare to in classification by CEFR level.

The results in terms of accuracy, precision, recall and F1 score metrics can be seen in table 4.2. The precision, recall and F1 score metrics' averages have been set to weighted in order to account for class imbalances. One can see that BERTimbau achieves roughly double of each metric achieved by the baseline.

The confusion matrices of each model, GPT-2 and BERTimbau, respectively, are present in figure 4.1. We can see that, given the true labels A1 and C1, GPT-2 predicts mostly correctly. However, for the true labels A2, B1 and B2, the model seems to give quite a large range of false predictions. On the other hand, BERTimbau's predictions seem to be mostly correct, with only a few misclassifications of adjacent CEFR levels, and even fewer non-adjacent CEFR level misclassifications. The term "*adjacent misclassification*" here is used when the model misclassifies the CEFR level by one level of difference, be it above, or below the true label. One can say an adjacent misclassification is a less serious mistake than its non-adjacent counterpart because CEFR is a scale with progressive difficulty or complexity. The CEFR scale measures difficulty, and its classes have a total ordering of difficulty level, namely: $A1 \preceq A2 \preceq B1 \preceq B2 \preceq C1 \preceq C2$, where "$a \preceq b$" denotes "$a$ is less difficult than b". In this sense, adjacent classes are more similar to each other than to non-adjacent classes, and thus an adjacent misclassification is more acceptable than a non-adjacent one.

To conclude, it seems clear that BERTimbau's predictions are more robust than the ones of GPT-2. Either by comparing their overall metrics or by comparing the type of mistakes made by the models (adjacent versus non-adjacent) BERTimbau performs better. Since both models have similar number of parameters (124M for GPT-2 and 110M for BERTimbau Base), I believe this discrepancy is due to two main causes, which will be quickly explained below but not investigated further since it falls out of the scope of this thesis.

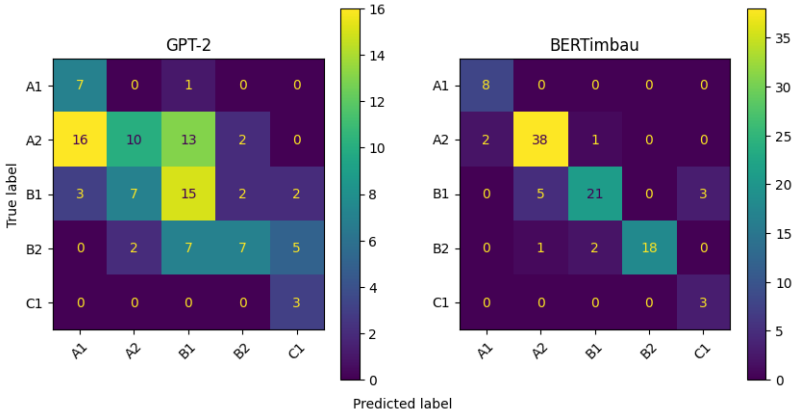

Figure 4.1: Confusion matrices for GPT-2's and BERTimbau's predictions on the test set.

The first cause is a difference in initialization and training. GPT-2 is said to have been fine-tuned from English to Portuguese (Santos et al. [2021], Guillou [2020]). The model's documentation on Hugging Face's Model Hub (Guillou [2020]) states that the model was trained on Portuguese Wikipedia for a little over a day using a GPU NVIDIA V100 32GB and with a little more than 1GB of training data. On the other hand, BERTimbau started with the weights from Multilingual BERT, and has been trained on the brWaC corpus (Filho et al. [2018]), which at the time was the largest corpus in Portuguese, using 17.5GB of high diversity, high quality raw text. BERTimbau was then pretrained for 4 days on a TPU v3-8 instance and performed 8 epochs over the pretraining data. Therefore this difference in the extent of training for each model might be a probable reason for the discrepancy in results of the models.

The second cause is the corpora on which the models were fine-tuned versus the corpora on which they were tested. GPT-2 was fine-tuned for the task of CEFR prediction on a certain private corpus, and then tested on COPLE2. BERTimbau was both fine-tuned and tested for CEFR prediction on the COPLE2 corpus. If one assumes that the corpora are sufficiently different, with different sentence distributions, one could get to the conclusion that this difference is a cause for the discrepancy in results of the models.

4.2 Part II: Developing an algorithm for grouping sentences

This section details the setup and results of the experiment related to the second goal of the thesis (1.1.2), namely a method devised to group sentences by topic. In order to develop a sentence grouping algorithm that leverages text difficulty and topic, we firstly need to classify said sentences by difficulty. Then the goal is to somehow

group sentences in a meaningful way, and compare it to existing language learning skill trees. Section 4.2.1 contains the setup, while section 4.2.2 presents the results.

4.2.1 Setup

This subsection presents the setup for the experiment related to the second goal of the thesis. The setup consists of firstly using the model from the previous experiment (section 4.1) to classify the target sentences by difficulty (section 4.2.1), then of introducing the topic modeling techniques used in BERTopic (section 4.2.1). Next the evaluation is discussed (section 4.2.1), followed by an overview of the topics present in Duolingo (section 4.2.1), Babbel (section 4.2.1), and Memrise (section 4.2.1). Finally, the setup finishes with a section discussing granularity in app topics (4.2.1).

CEFR proficiency level classification

To start with, 273714 sentences in Portuguese with English translation were downloaded from Tatoeba (Tatoeba Association [2023]). Subsequently, the BERTimbau model fine-tuned on CEFR classification was used to predict the difficulty level of each of those phrases. The end product is a dataset with over 273 thousand bilingual Portuguese-English sentences, each with a CEFR proficiency level classification between A1 and C1.

Topic modeling

In order to explore how one might meaningfully group sentences, the task of Topic Modeling was performed using BERTopic (Grootendorst [2022]). BERTopic, as explained before, is a topic model that takes text, makes rich embedding representations for such text, which are immediately reduced in dimension, then clustered, and finally topics are extracted from these resulting clusters (Grootendorst [2022]). The topics that are extracted depend on the parameters passed to the UMAP and HDBSCAN algorithms. Furthermore, BERTopic offers many visualization options to analyze the data, such as graphs with the intertopic distance, similarity matrix between topics, and so on (Grootendorst [2022]).

The dataset used for topic modeling with BERTopic was the over 273 thousand sentences in Portuguese available on Tatoeba (Tatoeba Association [2023]). The default UMAP and HDBSCAN algorithms that BERTopic utilizes were initially used. To generate topics and their probabilities, the BERTopic object calls the *fit_transform* method with the dataset as arguments, which was done on a Lenovo laptop with an AMD Ryzen CPU from the 5000 series in under 10 minutes.

Evaluation

In order to evaluate the quality of the topics extracted by BERTopic, there are a couple of options. One option would be a user study that tracks the users' learning progress using the method devised in this thesis over some period of time, such

as a few months. Then this progress could be evaluated by itself, i.e. "*can users effectively learn Portuguese using this method?*"; or compared to the progress of users in a control group that uses other existing applications, i.e. "*how does the users' progress compare when using the devised method versus using existing apps?*". Although this option would perhaps be a reliable and straightforward manner of testing the quality of the extracted topics and of the method, there are a few problems with this approach. The main problems are that it requires a lot of time and the development of the thesis has time constraints, and that doing user studies can be quite tricky, such as finding enough users, among other things. Alternatively, another approach would be to extract said topics using BERTopic, and compare them to the topics of popular existing language learning apps. Although this approach might be more subjective than the previous one, it does not present the same time and user constraints, hence it will be used for evaluating the second task of this thesis.

In light of the information above, the topics of each unit/module of a few popular language learning apps have been compiled into a table, which can then be used to compare to the topics extracted by BERTopic. The apps chosen were Duolingo, Babbel and Memrise, since they appear to be consistently among the most popular language learning apps in both Apple App Store and Google Play Store, as shown in section 2.6. Next, there will be overviews of the topics in the Portuguese courses provided by these three popular language learning applications.

Duolingo Topic Overview

The Portuguese course on Duolingo consists of 75 units, each of which is composed of a few (around 7) smaller lessons. The lessons range from regular lessons introducing new concepts, to short stories, to personalized practice, and units end with a unit review. Each unit focuses on a certain topic, be it a semantic topic or a grammatical topic. The learner can only move from unit n to unit $n+1$ when he has completed unit n, which delimits a clear path for a less independent student, but can be seen as too constraining for a more independent student who knows what he wants to learn. *Unit 4* can be used as an example of what kind of lessons one can find in a unit, since *Unit 4* has more tangible lesson names than more generic and basic units, such as *Unit 1*. The lessons contained in *Unit 4*, which is described as *"Talk about animals, form the plural"* are: *Describe your food* (a review from the previous unit); *Greet people* (tagged by Duolingo as *"hard"*, also building on top of a previous unit's lesson); *I need money* (a lesson based on a short story); *Personalized practice*; *Form the plural*; *Talk about animals*; and finally *Unit 4 review*. In table 4.3, one can see the titles of the units' topics for the course of Portuguese on Duolingo.

Babbel Topic Overview

Similarly, the Portuguese course offered by Babbel can also be subdivided in topics. However, Babbel does a different subdivision. First of all, Babbel offers the student two possible paths to learn Portuguese: courses by level and courses by theme. The courses by level are divided in three modules, which can be seen on table 4.4. On

Units 1-25	Units 26-50	Units 51-75
1 - Use basic phrases, describe what's around you	26 - Communicate quantities, give orders	51 - Talk about health
2 - Use polite phrases, greet people	27 - Make comparisons	52 - Express a wish, talk about arts
3 - Describe your food	28 - Describe people, talk about abstract things	53 - Say what you are doing
4 - Talk about animals, form the plural	29 - Form adverbs	54 - Talk about abstract things, mention what you would do
5 - Use tu	30 - Talk about the present, form the present continuous	55 - Express regret
6 - Use a gente, describe things	31 - Describe where you are	56 - Talk about science, use modal verbs
7 - Express possession	32 - Say what you were doing, talk about people	57 - Mention possibilities
8 - Use prepositions, use simple contractions	33 - Talk about memories	58 - Talk about nature, discuss communication
9 - Describe clothing	34 - Count up to a million, describe sizes	59 - Describe what was happening
10 - Order food, form basic questions	35 - Say what you need	60 - Say what hasn't happened, discuss business
11 - Describe colors	36 - Express quantity, use participles	61 - Talk about spirituality
12 - Count up to twenty, form the present tense	37 - Say who you are with	62 - Discuss politics
13 - Use verbs with prepositions	38 - Describe quantities	63 - Express quantity, form the present tense, describe countries, describe a trip
14 - Talk about body parts, use more prepositions	39 - Form the present tense, use more adjectives	64 - Talk about people, say what you need, describe quantities, say what had happened
15 - Talk about your family	40 - Talk about the past	65 - Count up to a million, use participles, use more adjectives, talk about school
16 - Describe your home, use common verbs	41 - Say what has been happening	66 - Give directions, describe abstract ideas, say what you will have done
17 - Say what you will do	42 - Describe countries, talk about school	67 - Say what has been happening, talk about the future, use infinitives, express a wish
18 - Name common objects, mention where something is	43 - Tell people what to do	68 - Mention what you would do, mention possibilities, say what hasn't happened
19 - Describe people and things	44 - Say what had happened, describe a trip	69 - Describe where you are, say what you are doing, talk about science, describe sizes
20 - Ask where people are going	45 - Give directions	70 - Talk about memories, discuss sports, talk about relationships, talk about arts
21 - Talk about things around you, talk about your job	46 - Express feelings, talk about the future	71 - Express regret, talk about nature, discuss business, discuss communication
22 - Express opinions	47 - Discuss sports	72 - Say who you are with, talk about the past, tell people what to do, express feelings
23 - Say what you did	48 - Describe abstract ideas, mention options	73 - Say what you were doing, mention options, talk about health, use modal verbs
24 - Combine two sentences, use time prepositions	49 - Use infinitives	74 - Talk about abstract things, describe what was happening, discuss politics
25 - Mention dates	50 - Talk about relationships, say what you will have done	75 - Talk about spirituality, describe where you are, count up to a million

TABLE 4.3: Topics (units) found in the Duolingo Portuguese course

Module	CEFR level	Description
Newcomer	A1	An easy introduction to Portuguese: basic vocab, grammar, and pronunciation
Beginner	A2	Discover more words and expressions for many different life situations (7 courses)
Intermediate	B1	Consolidate what you've learned and start expressing yourself in a more nuanced way (2 courses)

TABLE 4.4: Babbel Portuguese courses by level

Module	Number of Courses	Description
Refresher	1	Ideal for anyone who studied Portuguese a long time ago or wants to test their knowledge
Grammar	5	Grammar practice in easy, understandable steps. Drills and exercises with clear and concrete examples
Countries and Traditions	3	In these courses, you won't just learn the language. You'll also gain useful knowledge about Brazil
Specials	5	Looking for something special? Here you will find a range of courses
Words and Sentences	31	Improve and train your vocabulary with 3000 words and example sentences you can use in everyday life

TABLE 4.5: Babbel Portuguese courses by theme

the other hand, the courses by theme are divided in five modules, which can be seen on table 4.5. Within the courses by theme, the course "*Words and Sentences*" is the one that has the most subcourses, with the most varying topics, which seems to align the most with the second objective of the thesis, namely the part of topic modeling. These subcourses can be seen on table 4.6. Additionally, the "*Countries and Traditions*" course comprises meaningful semantical topics as subcourses, such as "Portuguese for Everyday Life", "Portuguese for Your Vacation", and "Portuguese for Carnival", which can be added to the final set of topics offered by the app. From "*Specials*" the only subcourse that is a semantical topic in itself is "Comunicação no Trabalho" ("Communication at Work" in a loose translation), which will also be used. On the other hand, "*Grammar*" and "*Refresher*", as the names suggest, do not offer semantical topics as subcourses. Babbel differs from Duolingo in the sense that the student has more freedom and independence to choose what course and lesson to learn, while Duolingo requires the student to advance through a certain path.

Course	Number of Lessons	Course	Number of Lessons
First Words and Sentences	10 lessons	Culture	18 lessons
Food and Drinks	21 lessons	Basic Properties	17 lessons
Animals	15 lessons	Academic Fields	13 lessons
Body	17 lessons	Media	16 lessons
Society	17 lessons	Deparments and Services	13 lessons
Sports	19 lessons	Work	13 lessons
Communication	20 lessons	Home	16 lessons
Digital World	14 lessons	Education	13 lessons
Clothes	14 lessons	Landscapes	14 lessons
Vacations	22 lessons	Plants	13 lessons
Feelings and Attitudes	14 lessons	Environment	15 lessons
Relationships	14 lessons	City	13 lessons
Life	12 lessons	Rockstars and Fans	8 lessons
Festivals and Parties	8 lessons	Wine, Food and Gastronomy	9 lessons
Transportation and Travel	13 lessons	Lifestyle	15 lessons
Free Time	17 lessons		

Table 4.6: Babbel Words and Sentences subcourses

Memrise Topic Overview

Finally, the Brazilian Portuguese course for English speakers available on Memrise has a recommended path of topics that the student can choose to follow. However, the student is not bound by it: the student can choose to undertake some paths that are focused on certain topics, such as Education, Health, Food, Relationships, among others. The topics which the student can choose from can be seen in table 4.7. Memrise's course, much like Babbel's and unlike Duolingo's, allows for the student to have more flexibility and independence in which order to learn each topic and lesson.

Topic Granularity in Apps

As can be seen in the tables presented, each of the three language learning apps has a different number of topics. While Duolingo has 75 topics, Babbel has 31 topics and Memrise has 15 of them. However, Duolingo includes grammar topics in these 75 topics, which are not kinds of topics that BERTopic can straightforwardly extract. On the other hand, Babbel's 31 topics are already narrowed down to the course of Words and Sentences, which contains only semantic topics; and Memrise has a topic for "*Miscellaneous*", which can encompass a wide range of subtopics that get lumped together in a single topic. Furthermore, Duolingo possesses some units that mix semantic and grammar topics, such as *Unit 4 (Talk about animals, form the plural)*, which makes it harder to separate grammar and semantic topics. Nevertheless, the units presented by Duolingo whose topic is separated by a comma can be subdivided further into smaller parts, and the obvious grammar points can be identified and ignored for the purpose of comparing semantic topics across apps and with the ones extracted by BERTopic. Some topics in a unit are trickier to pinpoint as a grammatical topic or as a semantic topic, such as *"Use tu"*, from unit 5. However, by the way that BERTopic extracts topics, some topics can be related to a certain

Topics
1 - Activities
2 - Basics
3 - Education
4 - Food
5 - Health
6 - Introductions
7 - Miscellaneous
8 - Opinions
9 - Relationships
10 - Shopping
11 - Social Life
12 - Society
13 - Sports
14 - Travel
15 - Work

TABLE 4.7: Topics found in the Memrise Brazilian Portuguese course for English speakers

word, so these topics will be kept as semantical, since semantics pertains to meaning. Nevertheless, some grammar topics are presented in a way that does not use technical terms, such as *"Say what you will do"* from unit 17, which indicates teaching the future tense, or *"Say what has been happening"* (tense related) and *"Tell people what to do"* (imperative). Although these topics are not clear grammatical topics from their titles, i.e. they are not called "future", "simple present" or "imperative", and so on, they will be removed as they are not clear semantical topics either. Nonetheless, there are still some topics that are not so clear whether they are semantical or grammatical, so those will be taken as semantical topics as a precaution. Repeated topics will be excluded, since for the task of comparing topics only one appearance is needed. As a result, we will have a list of renumbered unique granular topics which are semantical, which can be seen on table 4.8, with 56 semantical topics.

Similarly, some sort of compilation of topics can also be done for the courses offered by Babbel. We take the titles of the subcourses that deal with semantical topics as final semantical topics to later compare to BERTopic-extracted semantical topics. Therefore all subcourses within "Words and Sentences" and some within "Countries and Traditions" and "Specials" are taken as semantical topics. The comprehensive list of semantical topics can be seen in table 4.9, with 35 semantical topics.

Regarding Memrise, the semantical topics are those 15 topics presented in table 4.7. Comparing the amount of semantical topics between apps, one can see that Duolingo has the most (56), followed by Babbel (35) and then by Memrise (15). Related to Duolingo's 56 topics, some of them, while not exactly the same, appear to

Duolingo Semantical Topics 1-19	Duolingo Semantical Topics 20-38	Duolingo Semantical Topics 39-56
1 - Use basic phrases	20 - Describe people and things	39 - Talk about school
2 - Describe what's around you	21 - Ask where people are going	40 - Describe a trip
3 - Use polite phrases	22 - Talk about things around you	41 - Give directions
4 - Greet people	23 - Talk about your job	42 - Express feelings
5 - Describe your food	24 - Express opinions	43 - Discuss sports
6 - Talk about animals	25 - Mention dates	44 - Describe abstract ideas
7 - Use tu	26 - Communicate quantities	45 - Mention options
8 - Use a gente	27 - Describe people	46 - Talk about relationships
9 - Describe things	28 - Talk about abstract things	47 - Talk about health
10 - Express possession	29 - Describe where you are	48 - Talk about arts
11 - Describe clothing	30 - Talk about people	49 - Say what you are doing
12 - Order food	31 - Talk about memories	50 - Express regret
13 - Describe colors	32 - Count up to a million	51 - Talk about science
14 - Count up to twenty	33 - Describe sizes	52 - Talk about nature
15 - Talk about body parts	34 - Say what you need	53 - Discuss communication
16 - Talk about your family	35 - Express quantity	54 - Discuss business
17 - Describe your home	36 - Say who you are with	55 - Talk about spirituality
18 - Name common objects	37 - Describe quantities	56 - Discuss politics
19 - Mention where something is	38 - Describe countries	

Table 4.8: Semantical topics found in the Duolingo Portuguese course (grammar points and repeated topics excluded, renumbered)

be quite adjacent regarding semantics, for example "5 - Describe your food" and "12 - Order food", or "26 - Communicate quantities", "35 - Express quantity" and "37 - Describe quantities". So, while Duolingo has the most courses, some of them appear more correlated with each other than Babbel's and Memrise's topics. It could be beneficial to merge some of these highly correlated semantical topics and reduce the total number of them. However, for that end one would have to carefully go through each lesson in each unit of these Duolingo topics, and verify that they indeed consist of a similar cluster of meaning, which is quite subjective and requires more time than is available in a thesis such as this one.

Finally, in order to evaluate how helpful the topic modeling done by BERTopic is, one should compare its results to the semantical topics present in Duolingo (table 4.8), Babbel (table 4.9) and Memrise (table 4.7).

4.2.2 Results

This section contains the results of the experiment regarding the second goal of the thesis (1.1.2), namely devising a method to group sentences using their semantics and difficulty. After some background has been laid down in the setup section (4.2.1), this section presents the experiments with BERTopic. Firstly the present section describes the results achieved by topic modeling with BERTopic with default parameters (section 4.2.2), followed by two topic reductions (sections 4.2.2). Finally, the section presents topic search and some closing thoughts (section 4.2.2).

Semantical Topic	Semantical Topics
1 - First Words and Sentences	19 - Academic Fields
2 - Food and Drinks	20 - Media
3 - Animals	21 - Deparments and Services
4 - Body	22 - Work
5 - Society	23 - Home
6 - Sports	24 - Education
7 - Communication	25 - Landscapes
8 - Digital World	26 - Plants
9 - Clothes	27 - Environment
10 - Vacations	28 - City
11 - Feelings and Attitudes	29 - Rockstars and Fans
12 - Relationships	30 - Wine, Food and Gastronomy
13 - Life	31 - Lifestyle
14 - Festivals and Parties	32 - Portuguese for Everyday Life
15 - Transportation and Travel	33 - Portuguese for Your Vacation
16 - Free Time	34 - Portuguese for Carnival
17 - Culture	35 - Communication at Work
18 - Basic Properties	

Table 4.9: Babbel semantical topics

BERTopic default parameters

The CPU version of the Topic Model with default UMAP and HDBSCAN parameters in BERTopic (version 0.15.0) extracted 5137 topics from the dataset, with the smallest topic size being 10 sentences (for many topics), and 88004 sentences being categorized as outliers. As a clarification, being classified as an outlier means that these sentences do not bear enough resemblance to be clustered in any of the 5137 topics extracted. That means that from 273714 sentences, over 88 thousand of them (approximately 32%) could not be classified in any clusters with the minimum size of 10 sentences. At the same time that there are probably too many topics (5137 compared to 56, 35 and 15 of Duolingo, Babbel and Memrise, respectively), the minimum size of each topic cluster is large enough to render 32% of the sentences impossible to cluster. In other words, increasing the minimum size of each cluster in order to achieve fewer, more meaningful clusters will result in fewer usable sentences. This is not necessarily negative for two main reasons: firstly, there will still be thousands of usable sentences, which can be enough material for a language learner to be exposed to; and secondly, some possibly meaningless clusters will be removed, which is beneficial for the goal

of providing useful learning material.

BERTopic provides many kinds of analysis and visualization tools for the extracted topics. For example, figure 4.2 represents a sort of two dimensional distance between all of the 5137 topics. For visualizing the 5137 extracted topics, one can use a hierarchical view of them. However, the fact that there are 5137 topics makes the rendering and saving of the image with all topics unfeasible in the machines tested, besides the fact that the image does not adequately fit the pages on this thesis. Hence only the top 100 topics of the topic model will be shown hierarchically in figure 4.3. The top 100 topics are the 100 topics with the most sentences. In the hierarchical view, one can see that more similar topics are grouped together in closer brackets. From figure 4.3, it is clear that many of the topics in close brackets of the same color are close enough in their descriptor's meaning that they could still be merged into a single topic, such as topics 4 and 23 (about Australia and New Zealand); topics 94, 40, 25, 83 and 37 (about beverages); topics 34 and 60 (about love and loving); and so on. This, with the fact that 5137 topics are too many if compared to the amount provided in the aforementioned language learning apps, indicates that it would be ideal to reduce the amount of topics, in order to have very cohese, larger, more meaningful topics through which a learner could usefully browse.

Topic reduction

The module provided by BERTopic allows the user to manually merge whatever topics the user wants, regardless of whether the topics are actually similar. Another possibility is that the user can reduce the number of topics to a specified n amount after having trained the model. This approach is preferred, since BERTopic will merge topics that are semantically similar to each other, and is much faster and requires practically no human effort. In order to see how reducing the number of topics affects the topic model, the topics were reduced from 5137 to 1000, still greatly above the amount of semantical topics in language learning apps. The intertopic distance can be seen in figure 4.4, while the hierarchy of the top 100 topics can be seen in figure 4.5. From the intertopic distance it is clear that the distribution clutter is greatly reduced from the original 5137 topics.

Additionally, using the interactive intertopic distance picture provided by BERTopic on the Jupyter Notebook in which it is run, one can see that some clusters of topics are still quite similar, such as topics with representative words relating to: France, the French language and studying/speaking it (topics 10, 7 and 132), Ireland and Scotland (topic 753), the United Kingdom (topic 309), and so on; multiple topics pertaining to weeks, weekends, week days, today, tomorrow, yesterday, and so forth (topics 57, 533, 58, 136, 558). Intuitively it makes sense that these topics with similar meanings, for example countries/languages or the days of the week, should be grouped in their own separate topics, instead of being formed by multiple smaller topics. At the same time, however, there are some clusters of topics in the interactive image that do not seem to make much sense as a topic. For example, topics 727 (*to repeat, mayuko, how many, waiting/expecting, slow*), 73 (*to return, return, I will return, I returned*), 166 (*to try, new, again, chance, another*) and 306 (*to try, I tried,*

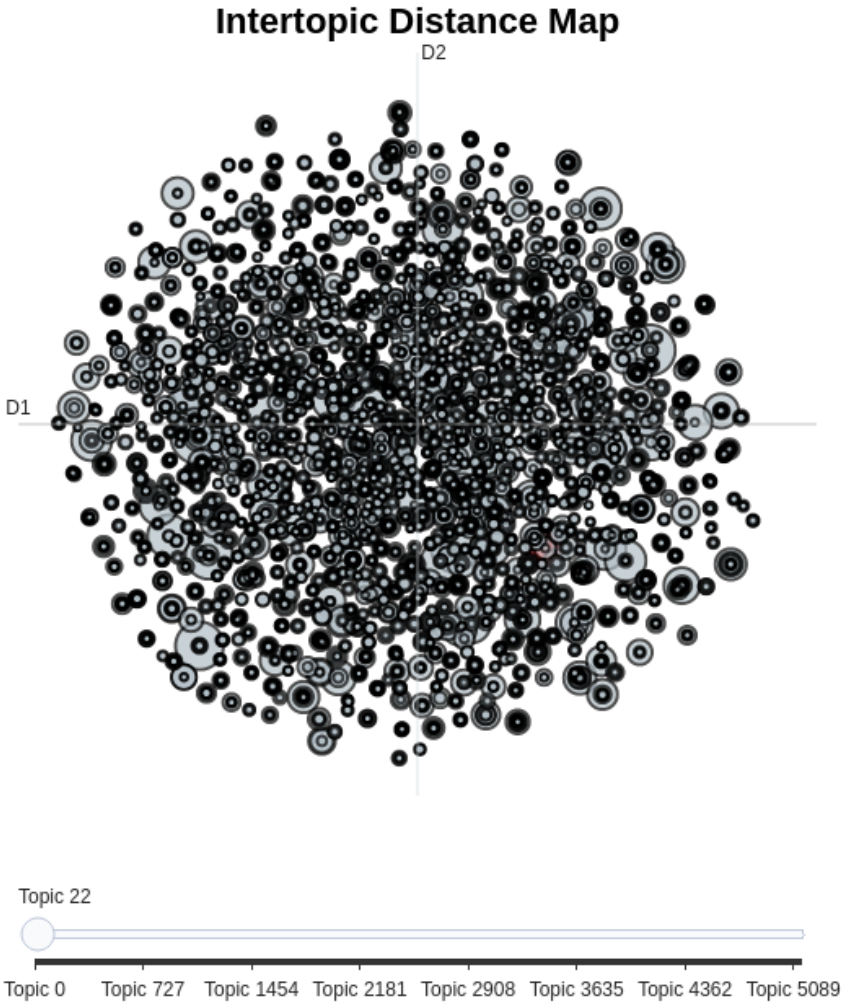

FIGURE 4.2: Distance between 5137 topics extracted by BERTopic with default parameters

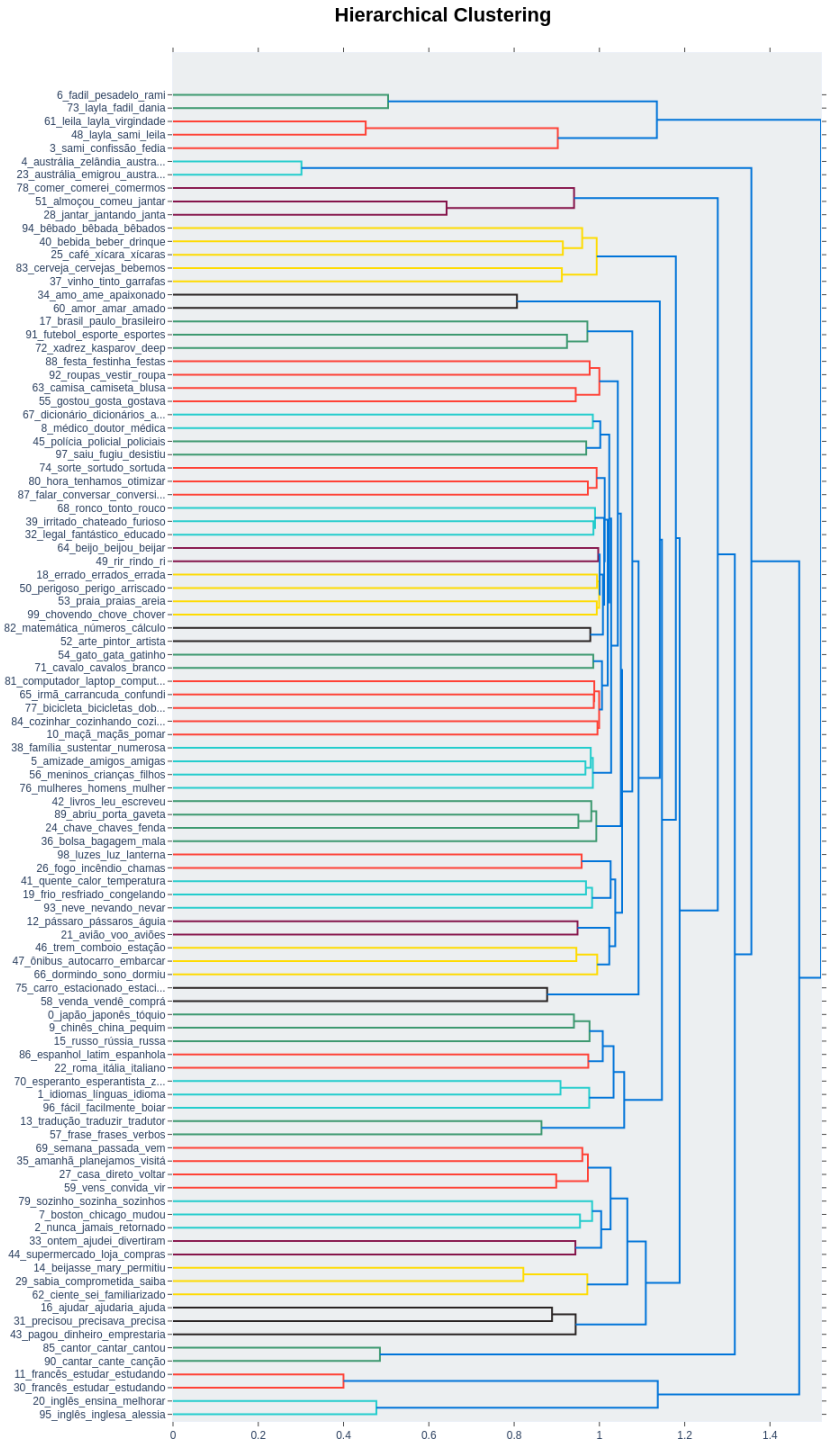

Figure 4.3: Top 100 of 5137 topics extracted by BERTopic with default parameters

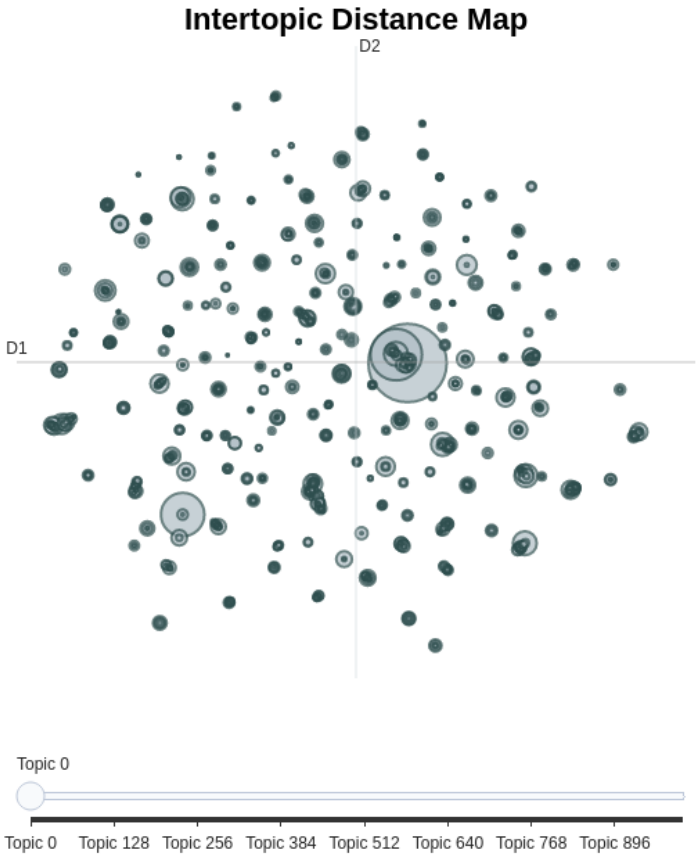

FIGURE 4.4: Intertopic distance of 1000 extracted by BERTopic with default parameters

to experiment, experiment) are plotted very close to each other, but are not clearly meaningful topics for language learning, though it would be important for a learner to eventually learn these words.

With that said, a thousand topics still is vastly greater than the amount usually offered to learners on apps. This, together with the fact that many topics appear to lack merging, indicates that it can be beneficial to further reduce the amount of topics. It is not straightforward to know how many topics extracted by BERTopic is ideal, since there are some topics that are meaningless, and it can form more clusters than there are topics in the apps mentioned. However, because of the high possibility that BERTopic will extract meaningless topics from such a high number of documents, it is probable that the ideal amount of topics to be set is above the amount of topics of an app. In this case, the highest amount of topics from the apps would be the 56 semantical topics of Duolingo. For this reason, more experiments will be made with topic models larger than 56 topics and smaller than 1000 topics, though the choice of n has to be somewhat arbitrary.

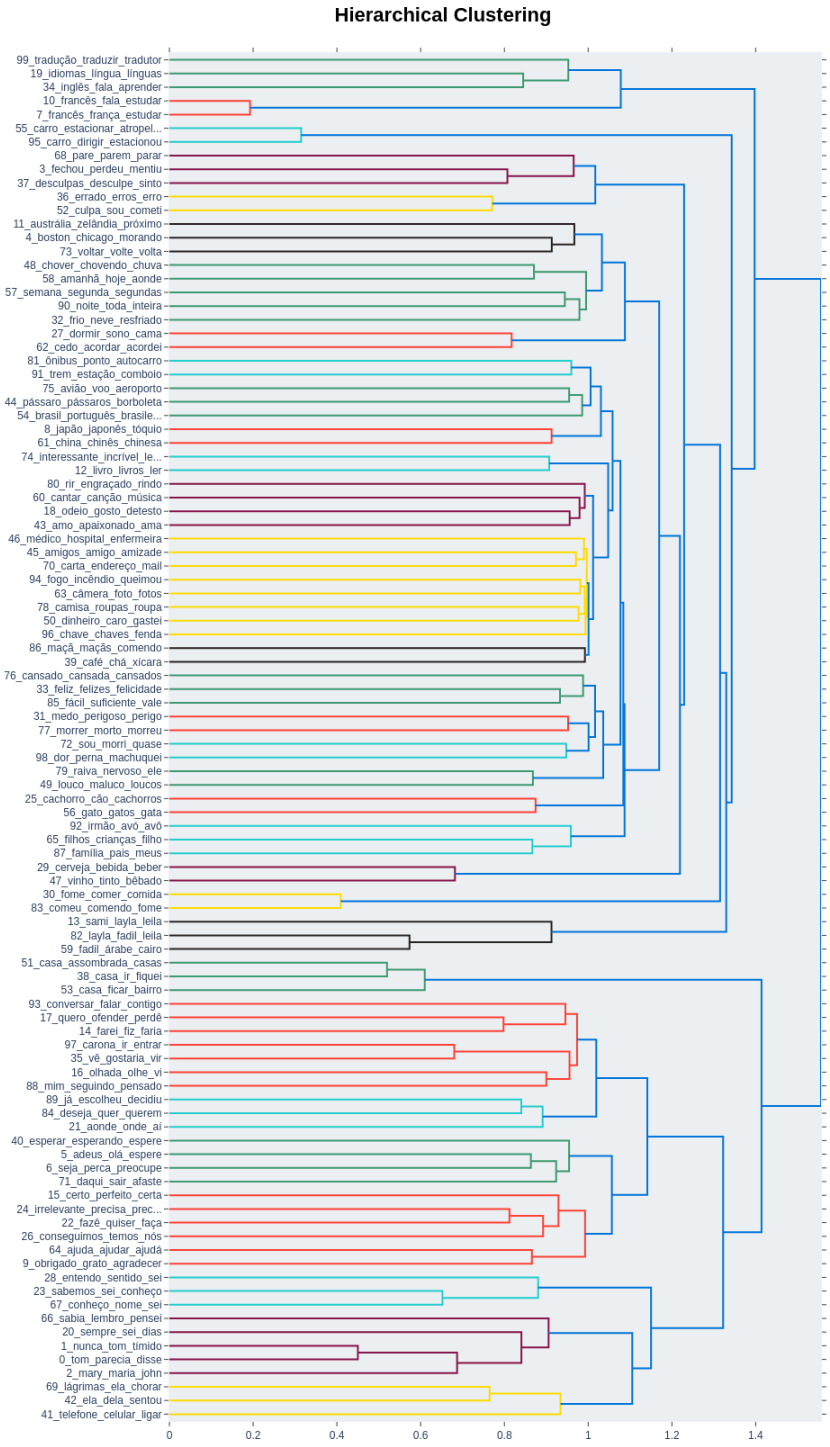

Figure 4.5: Top 100 of 1000 topics extracted by BERTopic with default parameters

Topic search

Regardless, one can use the "Search Topics" section of the BERTopic API to find the most similar topics to a certain search term. The method "*find_topics*" takes the search term and the amount n of most similar topics as input, and returns a tuple with "*similar_topics*" and "*similarity*". The former is a list of the n topics most similar to the search term, while the latter is the cosine similarity distance between the embedding of the search term and of the topic embeddings. Therefore, one can compare how similar a certain search term is to the extracted topics, and can thus use the semantical topics curated from Duolingo, Babbel and Memrise as search terms. In this manner, there is a possibility of numerically comparing the similarity of language learning applications' topics to BERTopic-extracted topics. One can set $n = 1$ to find only the most similar topic, and take note of the similarity score. This allows for finding which topics from apps are covered by BERTopic, and with what similarity percentage.

Tables 4.11, 4.12 and 4.13 show the semantical topics of, respectively, Duolingo, Babbel and Memrise, as well as the most similar topic extracted by BERTopic with 250 topics, its similarity score to the language learning app's topic name, and the amount of sentences belonging to the extracted topic (the frequency). The tables were separated in three to make them fit in the pages of this thesis and to make them more readable, following the semantical topic numbering used earlier with each language learning app.

It appears that the semantical topics that are related to more concrete, less abstract ideas have more similar topics as most similar topic. Some examples of more concrete topics and their most similar counterparts are Duolingo's 5 and 6, and BERTopic's 5 and 29, respectively; Babbel's 6 and BERTopic's 53; and Memrise's 15 and BERTopic's 52. Some examples of more abstract semantic topics that have less similar top BERTopic topic are Duolingo's 7 and BERTopic's 118, Duolingo's 26 and BERTopic's 32; Babbel's 18 and BERTopic's 151; and Memrise's 2 and BERTopic's 184, to name a few. The average similarity of Duolingo's topics to the top topics provided by BERTopic (with $n = 250$ topics) is 56.34%, while for Babbel it is 56.62%, and for Memrise it is 58.20%. The average similarity in total is 56.70%. Although this seems like a small difference across apps, a possible explanation for the difference is that the apps that have fewer semantical topics have more topics that are more concrete in proportion to more abstract topics. Thus they are more similar to BERTopic topics.

When checking the similarity metrics for the topics generated by the original topic model, with $n = 5137$ topics, the average metrics for Duolingo's, Babbel's and Memrise's topics are respectively 68.45%, 65.97% and 66.04%, approximately. The overall average similarity metric is 67.29%. Similarly, for the topic model with $n = 1000$ topics, the average metrics for Duolingo's, Babbel's and Memrise's topics are respectively 63.83%, 61.53%, 61.76%, approximately. The overall average similarity metric is 62.78%. Despite models with a larger amount of topics having more similar topics to the app topics, the sizes of these topics are generally smaller, which introduces a trade off. The values can be seen on table 4.10.

	Avg. Similarity				
Amount of topics	Duolingo	Babbel	Memrise	Total	Avg. Top Topic size
5137	**68.45%**	**65.97%**	**66.04%**	**67.29%**	62.57
1000	63.83%	61.53%	61.76%	62.78%	242.46
250	56.34%	56.62%	58.20%	56.70%	**1079.36**

Table 4.10: Average similarity metrics of most similar BERTopic topic to each app topic

Although the similarity between extracted topics and app topics has varying degrees depending on topic, these results can still be useful. The similarity metric allows one to decide some acceptable threshold to discriminate useful topics. A language course creator or an independent learner could use this topic modeling method in order to streamline the process of mining sentences. By using BERTopic for topic modeling, one has a number of clusters of sentences with common meaning. One can then use the method "*find_topics*" with a search term to find the most semantically similar topics to that search term, and can then go through the sentences that belong to that topic in order to curate a language course for themselves or for others. Moreover, one benefit of using this method is that a student can choose whatever search topic is most interesting for them for which to find sentences. Some attention has to be brought to the fact that the quality of the topics highly depends on the quality of documents being used, in this case, the sentences available on Tatoeba. Additionally, BERTopic is modular and flexible, as it can be used with different embedding models and different techniques with different parameters in order to create the final topic model. Therefore, the quality of its topics is highly dependent on each of its parts and chosen parameters. Nonetheless, BERTopic can be a useful tool to help automate part of the process of creating a language course - be it for oneself or to be distributed. Namely, it can be used to find useful sentences by topic, which can be used either for giving students exposure, or for making exercises, for example by hiding words. Furthermore, BERTopic has many other potentially useful functionalities not explored in this thesis. For example, it is possible to perform topic modeling with topic labels predefined by the user - say, the already existing topic labels from language learning apps -, allowing one to create clusters of documents for predefined topics. This method is not explored in this thesis for potential copyright problems in existing language learning apps' courses. Thus their content is only used as a means of comparison, rather than for creation of new content.

4.3 Conclusion

This chapter presented the experiments done in order to fulfill the goals of the thesis, namely: I) training a model extensively in Portuguese for difficulty classification in the CEFR scale (1.1.1); and II) devising a method to group sentences by difficulty

Duolingo semantical topic	Most similar topic	Similarity	Topic frequency
1 - Use basic phrases	89_dicionário_esperanto_língua_dicionários	47.39%	493
2 - Describe what's around you	2_quem_você_vocês_pergunta	56.6%	5695
3 - Use polite phrases	17_inglês_língua_idiomas_línguas	45.5%	1960
4 - Greet people	6_feliz_obrigado_felizes_muito	56.65%	3055
5 - Describe your food	5_jantar_comer_bolo_pão	63.44%	3384
6 - Talk about animals	29_cachorro_cavalo_animais_cavalos	76.67%	1332
7 - Use tu	118_funcionar_funciona_como_mostrar	37.69%	348
8 - Use a gente	30_nós_temos_juntos_vamos	61.4%	1323
9 - Describe things	138_significa_palavra_pronuncia_pronunciar	59.58%	282
10 - Express possession	28_dinheiro_rico_caro_pagar	34.74%	1342
11 - Describe clothing	58_camisa_casaco_gravata_roupas	66.25%	783
12 - Order food	5_jantar_comer_bolo_pão	59.15%	3384
13 - Describe colors	110_azul_cor_verde_azuis	63.99%	389
14 - Count up to twenty	201_trinta_dólares_30_quarenta	57.96%	102
15 - Talk about body parts	58_camisa_casaco_gravata_roupas	42.34%	783
16 - Talk about your family	14_família_irmã_pai_pais	76.67%	2210
17 - Describe your home	33_casa_ir_em_assombrada	69.96%	1214
18 - Name common objects	89_dicionário_esperanto_língua_dicionários	34.43%	493
19 - Mention where something is	138_significa_palavra_pronuncia_pronunciar	50.18%	282
20 - Describe people and things	138_significa_palavra_pronuncia_pronunciar	43.72%	282
21 - Ask where people are going	2_quem_você_vocês_pergunta	57.47%	5695
22 - Talk about things around you	217_especialidade_depois_livre_passatempo	63.35%	72
23 - Talk about your job	52_trabalho_emprego_trabalhar_escritório	81.2%	852
24 - Express opinions	51_futuro_ideia_plano_pensar	57.98%	857
25 - Mention dates	107_ano_outono_primavera_mês	49.0%	396
26 - Communicate quantities	32_telefone_carta_celular_senha	37.49%	1226
27 - Describe people	138_significa_palavra_pronuncia_pronunciar	43.82%	282
28 - Talk about abstract things	82_matemática_física_células_química	48.99%	522
29 - Describe where you are	2_quem_você_vocês_pergunta	57.3%	5695
30 - Talk about people	2_quem_você_vocês_pergunta	51.02%	5695
31 - Talk about memories	104_lembro_lembra_memória_esquecer	72.04%	409
32 - Count up to a million	168_pessoas_morreram_milhares_havia	56.24%	188
33 - Describe sizes	125_alto_tamanho_alta_altura	54.33%	320
34 - Say what you need	51_futuro_ideia_plano_pensar	57.71%	857
35 - Express quantity	119_explicar_explicação_exemplo_entendo	37.22%	340
36 - Say who you are with	68_amigos_amigo_amiga_amigas	53.89%	685
37 - Describe quantities	119_explicar_explicação_exemplo_entendo	32.07%	340
38 - Describe countries	219_recursos_naturais_países_minerais	69.17%	72
39 - Talk about school	24_escola_professor_aula_professora	78.54%	1467
40 - Describe a trip	112_férias_viajar_viagem_país	69.63%	385
41 - Give directions	181_mapa_triângulo_ângulos_quadrado	56.04%	159
42 - Express feelings	26_amor_amo_coração_beijo	58.11%	1395
43 - Discuss sports	53_tênis_futebol_jogar_beisebol	68.38%	839
44 - Describe abstract ideas	82_matemática_física_células_química	34.65%	522
45 - Mention options	184_formulário_página_site_blog	35.05%	150
46 - Talk about relationships	26_amor_amo_coração_beijo	50.3%	1395
47 - Talk about health	202_saúde_mental_meditando_psicólogo	76.71%	98
48 - Talk about arts	109_arte_parede_pintar_quadro	65.88%	391
49 - Say what you are doing	1_ajudar_te_preciso_ver	66.05%	9355
50 - Express regret	8_errado_esqueci_sinto_ter	66.99%	2431
51 - Talk about science	82_matemática_física_células_química	69.23%	522
52 - Talk about nature	135_vida_universo_alma_mundo	46.36%	293
53 - Discuss communication	32_telefone_carta_celular_senha	44.94%	1226
54 - Discuss business	205_fábrica_engenheiro_empresa_negócios	43.97%	94
55 - Talk about spirituality	95_deus_religião_bíblia_sabedoria	68.25%	451
56 - Discuss politics	98_presidente_votar_eleições_política	71.43%	430

Table 4.11: Most similar topics to Duolingo's semantical topics

Babbel semantical topic	Most similar topic	Similarity	Topic frequency
1 - First Words and Sentences	101_começar_primeira_primeiro_começo	48.33%	422
2 - Food and Drinks	5_jantar_comer_bolo_pão	62.98%	3384
3 - Animals	29_cachorro_cavalo_animais_cavalos	72.97%	1332
4 - Body	128_dançar_exercício_exercícios_dança	41.14%	311
5 - Society	30_nós_temos_juntos_vamos	43.44%	1323
6 - Sports	53_tênis_futebol_jogar_beisebol	68.4%	839
7 - Communication	32_telefone_carta_celular_senha	52.27%	1226
8 - Digital World	139_facebook_internet_google_twitter	53.4%	281
9 - Clothes	58_camisa_casaco_gravata_roupas	75.17%	783
10 - Vacations	112_férias_viajar_viagem_país	72.25%	385
11 - Feelings and Attitudes	26_amor_amo_coração_beijo	45.35%	1395
12 - Relationships	26_amor_amo_coração_beijo	47.03%	1395
13 - Life	135_vida_universo_alma_mundo	67.94%	293
14 - Festivals and Parties	61_festa_natal_aniversário_presente	62.78%	750
15 - Transportation and Travel	43_ônibus_trem_estação_pegar	59.99%	928
16 - Free Time	178_barato_livre_graça_grátis	60.72%	171
17 - Culture	219_recursos_naturais_países_minerais	46.09%	72
18 - Basic Properties	151_pedra_ouro_ferro_anel	31.46%	254
19 - Academic Fields	24_escola_professor_aula_professora	55.69%	1467
20 - Media	93_televisão_rádio_tv_assistir	53.66%	465
21 - Deparments and Services	207_reduzir_preços_despesas_empresa	45.25%	93
22 - Work	52_trabalho_emprego_trabalhar_escritório	67.08%	852
23 - Home	33_casa_ir_em_assombrada	75.52%	1214
24 - Education	24_escola_professor_aula_professora	68.88%	1467
25 - Landscapes	181_mapa_triângulo_ângulos_quadrado	50.27%	159
26 - Plants	120_jardim_batatas_fazenda_quintal	59.46%	339
27 - Environment	45_chuva_chover_guarda_chovendo	38.41%	920
28 - City	163_rua_estrada_atravessar_atravessando	49.17%	209
29 - Rockstars and Fans	225_concerto_show_circo_sucesso	56.41%	53
30 - Wine, Food and Gastronomy	47_café_leite_chá_xícara	47.43%	886
31 - Lifestyle	135_vida_universo_alma_mundo	42.26%	293
32 - Portuguese for Everyday Life	49_brasil_espanhol_português_espanha	69.56%	877
33 - Portuguese for Your Vacation	49_brasil_espanhol_português_espanha	75.15%	877
34 - Portuguese for Carnival	49_brasil_espanhol_português_espanha	57.97%	877
35 - Communication at Work	52_trabalho_emprego_trabalhar_escritório	57.77%	852

Table 4.12: Most similar topics to Babbel's semantical topics

Memrise semantical topic	Most similar topic	Similarity	Topic frequency
1 - Activities	217_especialidade_depois_livre_passatempo	51.89%	72
2 - Basics	184_formulário_página_site_blog	41.8%	150
3 - Education	24_escola_professor_aula_professora	68.88%	1467
4 - Food	5_jantar_comer_bolo_pão	73.31%	3384
5 - Health	202_saúde_mental_meditando_psicólogo	67.86%	98
6 - Introductions	184_formulário_página_site_blog	46.22%	150
7 - Miscellaneous	184_formulário_página_site_blog	44.34%	150
8 - Opinions	51_futuro_ideia_plano_pensar	56.38%	857
9 - Relationships	26_amor_amo_coração_beijo	47.03%	1395
10 - Shopping	127_supermercado_loja_compras_shopping	74.58%	312
11 - Social Life	135_vida_universo_alma_mundo	49.9%	293
12 - Society	30_nós_temos_juntos_vamos	43.44%	1323
13 - Sports	53_tênis_futebol_jogar_beisebol	68.4%	839
14 - Travel	112_férias_viajar_viagem_país	71.99%	385
15 - Work	52_trabalho_emprego_trabalhar_escritório	67.08%	852

Table 4.13: Most similar topics to Memrise's semantical topics

and topic (1.1.2). Section 4.1 deals with the first goal of the thesis, while section 4.2 deals with the second goal.

Section 4.1 starts by explaining the setup of the experiment (4.1.1). The setup starts by briefly recapitulating existing work in text difficulty classification in Portuguese using the CEFR scale and transformer models, of which the authors have made available an API with the best generalized model (Santos et al. [2021]), namely GPT-2. Then, section 4.1.1 contains an explanation of the transformer model extensively trained in Portuguese which can be used for the task of CEFR level classification and compared to the GPT-2 model, namely BERTimbau (Souza et al. [2020]). Next, the section presents the corpus (section 4.1.1) on which to fine-tune the BERTimbau model, and describes the fine-tuning setup (section 4.1.1).

Lastly, section 4.1.2 presents and discusses the results, showing that the BERTimbau model outperforms the GPT-2 model in the task of classifying text difficulty in Portuguese using the CEFR scale. More specifically, the BERTimbau model outperforms the GPT-2 model by a factor of roughly two in all metrics tested, namely accuracy, precision, recall, and F1. Additionally, the misclassifications of BERTimbau are usually adjacent classes to the true label class, while the misclassifications of GPT-2 have a higher range of possibilities, which is not as ideal as the former type of misclassification. The section also presents two hypotheses that can potentially explain the performance difference between models. The first hypothesis is that GPT-2 has not been extensively trained in Portuguese data, differently than BERTimbau, which could cause a significant difference in model performance. The second hypothesis is that the corpus used for the training of GPT-2 could significantly differ to the one used for training BERTimbau and for testing both models, and thus cause a significant difference in model performance.

To conclude, the results of this experiment answer the first research question (1.1.1), namely "What impact does a monolingual language model have on classifying the CEFR level of phrases in Portuguese?". The experiment shows that a "monolingual" Portuguese language model (a model extensively trained in Portuguese) considerably outperforms a model that was trained in English and converted into a Portuguese model.

Much like the section relating to part I, section 4.2 starts by presenting its setup (section 4.2.1), consisting initially of CEFR proficiency classification of a large dataset of sentences in Portuguese (section 4.2.1), followed by topic modeling (4.2.1). The latter section briefly explains how BERTopic works, and is followed by an evaluation section (4.2.1), which exposes two possible options to evaluate the quality of the topics generated by BERTopic. The first option is a user study that would evaluate how much knowledge a language learner can acquire in the learner's target language by studying a course made using the method in the thesis. More specifically, by exposing the student to phrases in Portuguese about a certain topic of interest. The problem with this evaluation option is that it is tricky to do, since it takes time to evaluate and it is not always easy to find study subjects. The second evaluation option is to compare the topics extracted by BERTopic to the existing topics in existing language learning apps. This option does not pose the same problems as the previous option, which is why, despite it potentially being subjective, it was

chosen as the evaluation approach in the thesis. Next, an overview of Duolingo's topics is presented (section 4.2.1), followed by an overview of the topics of Babbel (section 4.2.1) and of Memrise (section 4.2.1), respectively. After that, section 4.2.1 compares the topics of the apps presented, showing their differences in granularity, in semantics versus grammar, to name a few.

Lastly, section 4.2.2 presents and discusses the results of the experiments relating to the second research question (1.1.2). Section 4.2.2 exposes the result of topic modeling with the default parameters of BERTopic, namely clustering the dataset of over 273 thousand sentences in Portuguese into over 5 thousand topics. The largest cluster is made of over 88 thousand phrases considered outliers, which BERTopic cannot cluster with any other topic. For a brief analysis of the topics, the section shows graphs with the intertopic distance map, and the hierarchical clustering of the most frequent topics (topics with the most sentences). A closer look on the topics present in the hierarchical clustering suggests that the topics are too many and too small, since there are topics that are definitely semantically related but are not clustered together. This suggests that a reduction in the number of topics might be beneficial.

Section 4.2.2 presents the feature of topic reduction in BERTopic, which enables merging the most similar topics until the total number of topics reaches a user-specified amount n. An experiment is done with $n = 1000$, which is around 20% of the previous 5137 topics. By analyzing the hierarchical clustering of the top topics again, one can see still the same problem of different topics that are semantically similar. Guided by the number of topics in language learning apps, the number of topics of BERTopic was further reduced to $n = 250$, which is still higher than the number of topics of the analyzed app with the most topics. To this topic model, section 4.2.2 introduces "topic search", which allows one to find the most similar m topics to a certain search term, as well as a numerical measure of their similarity. By using the topic names of app topics as search term, one can find the BERTopic-extracted topics most relevant for a student who wants to learn a certain topic. Comparing the average topic similarity between topic models with different amounts of topics (5137 versus 1000 versus 250 topics), it is clear that the models with more topics get a higher semantic similarity to the topics present in the apps. However, there is a trade off, since these topics which are more semantically similar are composed of fewer sentences.

Although some of these topics that show up in the searches do not seem exactly useful, many with a higher similarity metric do. These topics are clusters of sentences that can be mined by someone creating a language course or by an independent student in order to find authentic material in the target language for language exposure. Furthermore, the sentences in each of these topics are classified by CEFR level, which allows the student or course creator to order sentences within a topic. Therefore, the second research question (1.1.2), namely "How can a language skill tree be automatically built from CEFR-labeled sentences", is answered, with a few caveats. Firstly, the skill tree is more akin to the ones in Babbel and in Memrise, in which the student has more freedom to choose his own topics, as opposed to Duolingo's skill tree, which offers the student a clear path already set. However, the

course creator can also make a skill tree such as Duolingo's if they so desire, but it would take extra human work. Secondly, the automatically extracted topics are not a ready solution, and they would require the course creator or student to go through the sentences that make sense in the same topic. Lastly, the quality of the extracted topics is dependent on the quality of the sentence dataset and on the hyperparameters used in BERTopic. Overall, the full method allows for streamlining the process of creating a language skill tree by classifying sentences regarding difficulty and semantical topic.

Chapter 5

Conclusion

This chapter contains an overall conclusion of the thesis, as well as suggestions for future work.

5.1 Conclusion

Since manually creating language courses is a process that takes resources, such as time, effort and money, the motivation of this thesis is to devise a method to streamline this creative process using Natural Language Processing techniques. Since language learning materials are classified by difficulty and by topic, the thesis goal has to reflect that. Thus, the goal of automating (fully or partially) the creation of language skill trees can be subdivided in two smaller goals. The first goal is, therefore, to automatically classify text difficulty in the target language, and relates to answering the first research question (1.1.1). The second goal relates to the second research question, namely creating a method to automatically build a skill tree from sentences labeled by CEFR level (1.1.2). The latter goal was later focused on the automatic creation of topics, so that at the end of both goals the user has language learning material for exposure (sentences) that is categorized by difficulty and by semantical topic. The target language chosen for this case study was Portuguese.

With the motivation and with the goal in mind, experiments were performed to answer the research questions.

The first experiment consisted of starting from a pre-trained language model extensively trained in Portuguese data (BERTimbau), and fine-tuning it to the task of predicting text difficulty in the CEFR scale using the COPLE2 corpus. After this model was trained, it was tested and compared to a pre-existing model (GPT-2) first trained in English, then converted to Portuguese, and finally fine-tuned for the same task (Santos et al. [2021]). The BERTimbau model achieved a significantly higher performance, while making more acceptable mistakes. Two possible reasons for the significant difference in performance are that BERTimbau was trained in Portuguese data to a much larger extent than the GPT-2 model; and that the test set for evaluating both models was part of the corpus used for fine-tuning BERTimbau but not for GPT-2. Nevertheless, the experiment sheds light into the first research

question (1.1.1), namely the impact that a monolingual model has on predicting CEFR level of phrases in Portuguese. Additionally, the model that resulted from the first experiment has the potential to be useful for language teachers to automatically score essays of students, making the evaluation of a great amount of students more feasible.

The second experiment was to develop a method for grouping sentences by topic and compare these resulting topics to existing topics from language learning apps. This was done using BERTopic (Grootendorst [2022]) as a topic model, which extracts semantically similar topics from documents, and Tatoeba sentences in Portuguese with an English translation as documents for the topic model. The topic names of three main language learning apps were manually collected for evaluating the result of the BERTopic extracted topics by semantical comparison.

Firstly the dataset of sentences had its phrases' difficulty level classified by the model in the first experiment. Next, a topic model was trained with the provided documents and default parameters, which found 5137 clusters (semantical topics). The topic model had its topics reduced to 1000 and to 250. Then, the method "find_topic" was used with the topic names from the three language learning apps as search terms, which resulted in the most similar topic from each of the three topic models (5137, 1000 and 250 topics), along with the similarity score. As a result, it was clear that the more topics a topic model has, the more similar the most similar topic would be when compared to the search terms. However, topic models with more topics have topics with fewer sentences, which introduces a trade off for the user. In the end, a spreadsheet was generated with the over 273 thousand sentences from Tatoeba with their respective CEFR level classification and with their respective topic classification by three topic models of different amounts of topics. This can be used by an independent learner or by a course creator to look for sentences fitting the desired criteria, namely difficulty and semantical topic, though it needs some human input to sift through the spreadsheet and create a skill tree. The second experiment shed light into the second research question (1.1.2), showing a path to building a skill tree from CEFR-labeled sentences. More specifically, the method consists of using topic modeling with BERTopic to extract topics, and having a course creator or an independent student choose topics of interest at the appropriate difficulty.

Furthermore, the method for making language learning material is quite flexible. It is language agnostic and the first part of it requires a language model trained in the target language, as well as a corpus with CEFR level annotation for fine-tuning. For the second part, it requires an embedding model capable in the target language to be used in BERTopic, and a dataset of documents to cluster into topics.

5.2 Future Work

The work presented in the thesis inspires a few possibilities for future work. Firstly, a further investigation on the cause of performance disparity between the GPT-2 and the BERTimbau models would be interesting. Secondly, since the BERTimbau model has been trained on a corpus composed of short essays, the difference in quality of

CEFR prediction of text in Portuguese between essays and single sentences could also be investigated. Additionally, one could examine topic modeling using specified labels by the course creator. And finally, a study could be done to evaluate the language learning progress of students with material collected through the method presented in this thesis.

Bibliography

PORTULAN CLARIN / Workbench / LX-Proficiency, a. URL https://portulanclarin.net/workbench/lx-proficiency/.

World Leader in AI Computing, b. URL https://www.nvidia.com/en-us/.

Abeer Abuzayed and Hend Al-Khalifa. BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique. *Procedia Computer Science*, 189:191–194, January 2021. ISSN 1877-0509. doi: 10.1016/j.procs.2021.05.096. URL https://www.sciencedirect.com/science/article/pii/S1877050921012199. Publisher: Elsevier.

Mebarka Allaoui, Mohammed Lamine Kherfi, and Abdelhakim Cheriet. Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In Abderrahim El Moataz, Driss Mammass, Alamin Mansouri, and Fathallah Nouboud, editors, *Image and Signal Processing*, Lecture Notes in Computer Science, pages 317–325, Cham, 2020. Springer International Publishing. ISBN 978-3-030-51935-3. doi: 10.1007/978-3-030-51935-3_34.

Dyah Aminatun and Lulud Oktaviani. Memrise: Promoting Students' Autonomous Learning Skill through Language Learning Application. *Metathesis: Journal of English Language, Literature, and Teaching*, 3(2):214–223, November 2019. ISSN 2580-2720. doi: 10.31002/metathesis.v3i2.1982. URL https://jurnal.untidar.ac.id/index.php/metathesis/article/view/1982. Number: 2.

Apple. App Store. URL https://www.apple.com/app-store/.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization, July 2016. URL http://arxiv.org/abs/1607.06450. arXiv:1607.06450 [cs, stat].

Babbel, GmbH. Language for Life - Babbel.com. URL https://www.babbel.com/.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate, May 2016. URL http://arxiv.org/abs/1409.0473. arXiv:1409.0473 [cs, stat].

Bhagyashree Vyankatrao Barde and Anant Madhavrao Bainwad. An overview of topic modeling methods and tools. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 745–750, June 2017. doi: 10.1109/ICCONS.2017.8250563.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A Neural Probabilistic Language Model. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. URL https://proceedings.neurips.cc/paper/2000/hash/728f206c2a01bf572b5940d7d9a8fa4c-Abstract.html.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, August 2013. ISSN 1939-3539. doi: 10.1109/TPAMI.2013.50. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

M Biniz. DataSet for Arabic Classification. 2, 2018.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, and Karin Schone. The MERLIN corpus: Learner language and the CEFR. 2014.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Camões I.P. Home - Camões - Instituto da Cooperação e da Língua. URL https://www.instituto-camoes.pt/en/.

Cecilie Carlsen. Proficiency Level—a Fuzzy Variable in Computer Learner Corpora. *Applied Linguistics*, 33(2):161–183, May 2012. ISSN 0142-6001. doi: 10.1093/applin/amr047. URL https://doi.org/10.1093/applin/amr047.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The Muppets straight out of Law School, October 2020. URL http://arxiv.org/abs/2010.02559. arXiv:2010.02559 [cs].

Noam Chomsky. Syntactic Structures. In *Syntactic Structures*. De Gruyter Mouton, September 2009. ISBN 978-3-11-021832-9. doi: 10.1515/9783110218329. URL https://www.degruyter.com/document/doi/10.1515/9783110218329/html.

Pedro Curto, Nuno Mamede, and Jorge Baptista. Automatic Text Difficulty Classifier - Assisting the Selection Of Adequate Reading Materials For European Portuguese Teaching:. In *Proceedings of the 7th International Conference on Computer Supported Education*, pages 36–44, Lisbon, Portugal, 2015.

SCITEPRESS - Science and and Technology Publications. ISBN 978-989-758-108-3. doi: 10.5220/0005428300360044. URL http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0005428300360044.

Iria del Río, Marcos Zampieri, and Shervin Malmasi. A Portuguese Native Language Identification Dataset, April 2018. URL http://arxiv.org/abs/1804.11346. arXiv:1804.11346 [cs].

Anoop Deoras, Tomas Mikolov, and Kenneth Church. A Fast Re-scoring Strategy to Capture Long-Distance Dependencies. 2011.

Jacob Devlin. Multilingual BERT readme document, 2018. URL https://github.com/google-research/bert/blob/a9ba4b8d7704c1ae18d1b28c56c0430d41407eb1/multilingual.md. original-date: 2018-10-25T22:57:34Z.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. A Survey on In-context Learning, February 2023. URL http://arxiv.org/abs/2301.00234. arXiv:2301.00234 [cs].

Duolingo, Inc. Learn a language for free. URL https://www.duolingo.com/.

ELLE. ELLE. URL https://evkk.tlu.ee/.

Europarat, editor. *Common European framework of reference for languages: learning, teaching, assessment ; companion volume*. Council of Europe Publishing, Strasbourg, 2020. ISBN 978-92-871-8621-8.

Jorge A Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. The brWaC Corpus: A New Open Resource for Brazilian Portuguese. 2018.

Cassie Freeman, Audrey Kittredge, Hope Wilson, and Bozena Pajak. The Duolingo Method for App-based Teaching and Learning. 2023.

Google. Apps Android no Google Play. URL https://play.google.com/store/games?hl=pt&gl=BE.

Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure, March 2022. URL http://arxiv.org/abs/2203.05794. arXiv:2203.05794 [cs].

Pierre Guillou. GPorTuguese-2 (Portuguese GPT-2 small): a Language Model. 2020.

Hall M. A. Correlation-based Feature Subset Selection for Machine Learning. *Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato*, 1998. URL https://cir.nii.ac.jp/crid/1570291225445879808.

J. Hancke. Automatic Prediction of CEFR Proficiency Levels Based on Linguistic Features of Learner Language. 2013. URL https://www.semanticscholar.org/paper/Automatic-Prediction-of-CEFR-Proficiency-Levels-on-Hancke/7e822039aaa9ea2ad105efa746e97db6208a0ac6.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. doi: 10.1109/CVPR.2016.90. ISSN: 1063-6919.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models, March 2022. URL http://arxiv.org/abs/2203.15556. arXiv:2203.15556 [cs].

Thorsten Joachims. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. Technical report, March 1996. URL https://apps.dtic.mil/sti/citations/ADA307731. Section: Technical Reports.

Stig Johan Berggren, Taraka Rama, and Lilja Øvrelid. Regression or classification? Automated Essay Scoring for Norwegian. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–102, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4409. URL https://aclanthology.org/W19-4409.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, January 2020. URL http://arxiv.org/abs/2001.08361. arXiv:2001.08361 [cs, stat].

Kenji Kira and Larry A. Rendell. A Practical Approach to Feature Selection. In Derek Sleeman and Peter Edwards, editors, *Machine Learning Proceedings 1992*, pages 249–256. Morgan Kaufmann, San Francisco (CA), January 1992. ISBN 978-1-55860-247-2. doi: 10.1016/B978-1-55860-247-2.50037-1. URL https://www.sciencedirect.com/science/article/pii/B9781558602472500371.

George R. Klare. Assessing Readability. *Reading Research Quarterly*, 10(1):62–102, 1974. ISSN 0034-0553. doi: 10.2307/747086. URL https://www.jstor.org/stable/747086. Publisher: [Wiley, International Reading Association].

Igor Kononenko. Estimating attributes: Analysis and extensions of RELIEF. In J. G. Carbonell, J. Siekmann, G. Goos, J. Hartmanis, Francesco Bergadano, and Luc Raedt, editors, *Machine Learning: ECML-94*, volume 784, pages 171–182. Springer Berlin Heidelberg, Berlin, Heidelberg, 1994. ISBN 978-3-540-57868-0 978-3-540-48365-6. doi: 10.1007/3-540-57868-4_57. URL http://link.springer.com/10.1007/3-540-57868-4_57. Series Title: Lecture Notes in Computer Science.

Jason Sebastian Kusuma, Kevin Halim, Edgard Jonathan Putra Pranoto, Bayu Kanigoro, and Edy Irwansyah. Automated Essay Scoring Using Machine Learning. In *2022 4th International Conference on Cybernetics and Intelligent System (ICORIS)*, pages 1–5, October 2022. doi: 10.1109/ICORIS56080.2022.10031338.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. FlauBERT: Unsupervised Language Model Pre-training for French, March 2020. URL http://arxiv.org/abs/1912.05372. arXiv:1912.05372 [cs].

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, October 2019. URL http://arxiv.org/abs/1910.13461. arXiv:1910.13461 [cs, stat].

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70, January 2022. ISSN 1558-2191. doi: 10.1109/TKDE.2020.2981314. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI Open*, 3:111–132, January 2022. ISSN 2666-6510. doi: 10.1016/j.aiopen.2022.10.001. URL https://www.sciencedirect.com/science/article/pii/S2666651022000146.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019. URL http://arxiv.org/abs/1907.11692. arXiv:1907.11692 [cs].

Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205, March 2017. ISSN 2475-9066. doi: 10.21105/joss.00205. URL http://joss.theoj.org/papers/10.21105/joss.00205.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, September 2020. URL http://arxiv.org/abs/1802.03426. arXiv:1802.03426 [cs, stat].

Memrise Limited. Learn a language. Memrise is authentic, useful & personalised. URL https://www.memrise.com.

Amália Mendes, Sandra Antunes, Maarten Jansseen, and Anabela Gonçalves. The COPLE2 Corpus: a Learner Corpus for Portuguese. *Proceedings of the Tenth Language Resources and Evaluation Conference – LREC'16*, pages 3207–3214, 2016. URL https://repositorio.ul.pt/handle/10451/30692. Accepted: 2018-01-17T16:46:22Z ISBN: 9782951740891 Publisher: European Language Resources Association.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

Pilar Munday. The case for using DUOLINGO as part of the language classroom experience. *RIED: Revista Iberoamericana de Educación a Distancia*, 19(1):83–101, January 2016. ISSN 1138-2783, EISSN: 1390-3306. doi: 10.5944/ried.19.1.14581. URL http://e-spacio.uned.es/fez/view/bibliuned:revistaRied-2016-19-1-7040. Publisher: AIESAD.

Usman Naseem, Imran Razzak, Shah Khalid Khan, and Mukesh Prasad. A Comprehensive Survey on Word Representation Models: From Classical to State-of-the-Art Word Representation Language Models. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 20(5):74:1–74:35, June 2021. ISSN 2375-4699. doi: 10.1145/3434237. URL https://dl.acm.org/doi/10.1145/3434237.

Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, September 2021. ISSN 0925-2312. doi: 10.1016/j.neucom.2021.03.091. URL https://www.sciencedirect.com/science/article/pii/S092523122100477X.

T Nora Raju, P A Rahana, Raichel Moncy, Sreedarsana Ajay, and Sindhya K Nambiar. Sentence Similarity - A State of Art Approaches. In *2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*, pages 1–6, June 2022. doi: 10.1109/IC3SIS54991.2022.9885721.

Debora Nozza, Federico Bianchi, and Dirk Hovy. What the [MASK]? Making Sense of Language-Specific BERT Models, March 2020. URL http://arxiv.org/abs/2003.02912. arXiv:2003.02912 [cs].

Robert Ostling, Andre Smolentzov, Bjorn Tyrefors Hinnerich, and Erik Hoglin. Automated Essay Scoring for Swedish. 2013.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, March 2018. URL http://arxiv.org/abs/1802.05365. arXiv:1802.05365 [cs].

Sarah E. Petersen and Mari Ostendorf. A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1):89–106, January 2009. ISSN 0885-2308. doi: 10.1016/j.csl.2008.04.003. URL https://www.sciencedirect.com/science/article/pii/S0885230808000272.

Ildikó Pilán. *Automatic proficiency level prediction for Intelligent Computer-Assisted Language Learning*. May 2018. ISBN 978-91-87850-68-4. URL https://gupea.ub.gu.se/handle/2077/55895. Accepted: 2018-05-17T11:08:19Z ISSN: 0347-948X.

Adam Poliak. A Survey on Recognizing Textual Entailment as an NLP Evaluation, October 2020. URL http://arxiv.org/abs/2010.03061. arXiv:2010.03061 [cs].

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019.

Livy Real, Erick Fonseca, and Hugo Gonçalo Oliveira. The ASSIN 2 Shared Task: A Quick Overview. In Paulo Quaresma, Renata Vieira, Sandra Aluísio, Helena Moniz, Fernando Batista, and Teresa Gonçalves, editors, *Computational Processing of the Portuguese Language*, Lecture Notes in Computer Science, pages 406–412, Cham, 2020. Springer International Publishing. ISBN 978-3-030-41505-1. doi: 10.1007/978-3-030-41505-1_39.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, August 2019. URL http://arxiv.org/abs/1908.10084. arXiv:1908.10084 [cs].

R. Rosenfeld. Two decades of statistical language modeling: where do we go from here? *Proceedings of the IEEE*, 88(8):1270–1278, August 2000. ISSN 1558-2256. doi: 10.1109/5.880083. Conference Name: Proceedings of the IEEE.

Iria del Río. Automatic proficiency classification in L2 Portuguese. *Procesamiento del Lenguaje Natural*, 63(0):67–74, September 2019. ISSN 1989-7553. URL http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6096. Number: 0.

Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. HAREM: An Advanced NER Evaluation Contest for Portuguese. 2006.

Rodrigo Santos, João Rodrigues, António Branco, and Rui Vaz. Neural Text Categorization with Transformers for Learning Portuguese as a Second Language. In Goreti Marreiros, Francisco S. Melo, Nuno Lau, Henrique Lopes Cardoso, and Luís Paulo Reis, editors, *Progress in Artificial Intelligence*, Lecture Notes in Computer Science, pages 715–726, Cham, 2021. Springer International Publishing. ISBN 978-3-030-86230-5. doi: 10.1007/978-3-030-86230-5_56.

Nádia F. F. da Silva, Marília Costa R. Silva, Fabíola S. F. Pereira, João Pedro M. Tarrega, João Vitor P. Beinotti, Márcio Fonseca, Francisco Edmundo de Andrade, and André C. P. de L. F. de Carvalho. Evaluating Topic Models in Portuguese

Political Comments About Bills from Brazil's Chamber of Deputies. In André Britto and Karina Valdivia Delgado, editors, *Intelligent Systems*, Lecture Notes in Computer Science, pages 104–120, Cham, 2021. Springer International Publishing. ISBN 978-3-030-91699-2. doi: 10.1007/978-3-030-91699-2_8.

Raquel Silveira, Carlos G O Fernandes, João A Monteiro Neto, Vasco Furtado, and Ernesto Pimentel Filho. Topic Modelling of Legal Documents via LEGAL-BERT. 2021.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In Ricardo Cerri and Ronaldo C. Prati, editors, *Intelligent Systems*, Lecture Notes in Computer Science, pages 403–417, Cham, 2020. Springer International Publishing. ISBN 978-3-030-61377-8. doi: 10.1007/978-3-030-61377-8_28.

Pedro Ortiz Suarez. OSCAR. URL https://oscar-project.org/.

Kaveh Taghipour and Hwee Tou Ng. A Neural Approach to Automated Essay Scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1193. URL http://aclweb.org/anthology/D16-1193.

Tatoeba Association. Tatoeba: Collection of sentences and translations, 2023. URL https://tatoeba.org/en/.

Kari Tenfjord, Paul Meurer, and Knut Hofland. The ASK Corpus – a Language Learner Corpus of Norwegian as a Second Language. 2006.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks, April 2021. URL http://arxiv.org/abs/2010.08240. arXiv:2010.08240 [cs].

Sowmya Vajjala and Kaidi Loo. Role of Morpho-Syntactic Features in Estonian Proficiency Classification. 2013.

Sowmya Vajjala and Kaidi Lõo. Automatic CEFR Level Prediction for Estonian Learner Text. 2014.

Sowmya Vajjala and Detmar Meurers. On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. 2012.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, December 2017. URL http://arxiv.org/abs/1706.03762. arXiv:1706.03762 [cs].

Roumen Vesselinov and John Grego. The Babbel Efficacy Study. 2016.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models, October 2022. URL http://arxiv.org/abs/2206.07682. arXiv:2206.07682 [cs].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A Survey of Large Language Models, April 2023. URL http://arxiv.org/abs/2303.18223. arXiv:2303.18223 [cs].