Textual Datasets For Portuguese-Brazilian Language Models

Matheus Ferraroni Sanches, Jader M. C. de Sá, Henrique T. S. Foerste, Rafael R. Souza, Julio C. Dos Reis, Leandro A. Villas

[1]Institute of Computing, University of Campinas, Campinas, São Paulo, Brazil

{m212142, j234830, h236651}@dac.unicamp.br

{rroque, jreis, lvillas}@unicamp.br

Abstract. *Advances in Natural Language Processing have generated new models that push forward the state of the art. This reached new heights in complex tasks in handling unstructured texts. Most of the new architectures and models focus on the English language. There is a lack of available datasets that can be used during the training of new models. This investigation presents four new textual datasets for language modeling in Brazilian Portuguese. Our datasets were generated from several specific methodologies that aimed to obtain data of different natures. Two of our sets were originally built from data in online web forums. We also distribute a translated version of MultiWOZ, and a clean version of BrWaC. The original datasets are made available in a structured way to facilitate their use during the training of NLP models, with questions, answers and conversations already identified.*

Resumo. *Avanços em Processamento de Linguagem Natural geraram novos modelos no estado da arte e alcançaram novos patamares em tarefas complexas em tratamento de textos não estruturados. A maioria das novas arquiteturas e modelos foca na língua inglesa. Constatamos uma baixa disponibilidade de conjuntos de dados que podem ser utilizados durante o treinamento de novos modelos. Esta investigação apresenta quatro novos conjunto de dados textuais para modelagem de linguagem no Português-Brasileiro. Nossos conjuntos de dados foram gerados a partir de diversas metodologias específicas que visaram obter dados de diferentes naturezas. Dois de nossos conjuntos foram originalmente construídos a partir dados em forúns Web online. Distribuímos igualmente uma versão traduzida do MultiWOZ, e uma versão limpa do BrWaC. Os conjuntos de dados originais são disponibilizados de maneira estruturada para facilitar sua utilização durante o treinamento de modelos PLN, com perguntas, respostas e conversas já identificadas.*

1. Introduction

The growing computing power allied with an increasing amount of data available allowed state-of-the-art models to achieve higher scores in benchmarks [Wang et al. 2018, Rajpurkar et al. 2018]. The Natural Language Processing (NLP) area is taking advantage of these advances and newly available data. Techniques based on synthetics are becoming obsolete as deep models have explored semantics to understand words and phrases better. The state of the art in NLP is moving fast, with new architectures [Vaswani et al. 2017], models [Devlin et al. 2018, Radford et al. 2019] and datasets

[Budzianowski et al. 2018, Wagner et al. 2018] being created and shared. Advances in this field allow new research topics and can be used as a foundation for novel solutions.

Despite the advances and breakthroughs, train NLP models are still a big challenge. The model architecture and the amount of data used make the training process of NLP models expansive, both in price and in time. A single NLP model can cost up to millions of dollars [Sharir et al. 2020]. Scaling the model size and amount of data used during the training stage can increase performance, whereas reducing over-fitting [Kaplan et al. 2020]. Although the high costs, the benefits of increasing model size and amount of data are still an interesting tradeoff.

In order to get around the high costs related to training NLP models from scratch, researchers can perform only fine-tuning in pre-trained generic models with a smaller dataset. This can be combined with a shorter training with a lower learning rate, preserving most of the patterns already learned by the model. This approach has several advantages but relies on good generic models to be refined, which can be very limited in many languages [Howard and Ruder 2018].

The lack of models in different languages can be directly related to the lack of datasets, as without proper data to train the models, the training process is affected. Furthermore, models that deal with specific problems and contexts, such as models for Question Answering about the medical area, require detailed data to be appropriately trained. Performing fine-tuning must rely on adequate models to use less data during the training stage and achieve good results [Howard and Ruder 2018].

In this investigation, we provide open domain data to pave the way for future researchers to train new generic or contextual specific models. We introduce four new datasets: two entirely new datasets; and two adapted datasets. The approach and tools employed are documented. Our key contributions are two novel datasets with human-to-human conversation with multiple answers to the same message identified; one translated dataset; one cleaned dataset. We demonstrate the defined methodology used to translate and clean the datasets. Our methodology must be helpful for future researchers with similar challenges.

2. Related Work

The increasing demand for NLP models leads to novel research topics by promoting the creation and sharing of new databases to support researchers. We describe datasets available today in Portuguese and two in English related to our study.

The Web corpus for Brazilian Portuguese (brWaC)[Wagner et al. 2018] dataset is one of the most popular datasets in Portuguese due to its size and variety of content. This dataset was constructed adopting an approach called WaCky (Web-As-Corpus Kool Yinitiative), which can obtain data from multiple languages extracting content from the Web. This dataset contains more than 3.5 million pages of 120,000 websites and more than 2.7 billion tokens. Although this is one of the biggest datasets in Portuguese, and the content has been filtered, the quality of many instances is questionable due to the poor writing and offensive/racist content. The model BERTimbau [Souza et al. 2020], based on BERT [Devlin et al. 2018], was trained on this dataset. It is possible for the model to output names of politics-related with offensive words.

Wikipedia creates a standardized manner to share data by enabling to acquire data about articles [Meta 2021]. This data can be used in several ways: NLP models can be trained on high-quality content, with annotations and facts from the articles. The limitation faced in this data is the specific writing style found on Wikipedia. We found that the text is completely correct and written in an informative way. In other words, underlying models trained on this basis understand and generate informative writing style outputs without slang or abbreviations. The model GPorTuguese-2 [Guillou 2020] was trained on this dataset. Tests on the model showed that it is unable to handle slang and abbreviations due to the specific writing style used during training.

The MultiDomain Wizard-of-Oz (MultiWOZ) [Budzianowski et al. 2018] is an English fully labeled dataset. The instances were generated by two humans interacting, simulating a user requesting information or reservations to a system. More than 10,000 dialogues are available. This dataset is a large-scale multi-turn conversation dataset about different domains with annotations. This dataset's major limitation remains on the original language, as every instance is in English.

The Ubuntu Dialogue Corpus [Lowe et al. 2016] is a dataset in English that uses a straightforward methodology to identify answers to other messages and recreate the dialogue between users. The source of this dataset is a technical forum about Ubuntu Operational System. Although the methodology to identify the dialogues is simple, this dataset can identify dialogues with hundreds of turns; the mean messages per dialogue are around seven messages. This work paved the way for new approaches to identify conversations by sharing a unique dataset. Despite the advances, the identification of conversations is very limited, and the generated data is only available in English.

We reveal that the lack of datasets is a problem in many languages. This fact can be checked at *HuggingFace*[1], a website specialized in NLP models and datasets. There are more than 1000 datasets in English[HuggingFace 2022a]; there are around 100 in Portuguese[HuggingFace 2022b]. This problem worsens if the trained model requires a specific context, structure, or a specific type, such as conversations or translated sentences. The training stage usually requires that the data being used are in a specific language and, if possible, inside or close to a specific theme to obtain better results.

The lack of appropriate models for many languages forces the researchers to use alternative approaches, such as translating the original sentence to a specific language to use an already trained model [Poncelas et al. 2020]. This process increases the cost and time for such operations and does not guarantee good results.

3. Datasets

This Section presents the datasets proposed with details about how they were created or adapted. The final result of each dataset is in the Portuguese language. Each dataset maintains specific characteristics, such as different objectives and domains. In the list bellow, we present a short reference to every dataset presented and how they were created.

1. **MultiWOZ-PTBR:** This dataset is a direct translation from the original MultiWOZ [Budzianowski et al. 2018]. The content is a fully annotated multi-domain conversation composed of an exchange between a user and a system with a specific

[1]https://huggingface.co/

goal, such as ask information about places and making reservations (cf. Subsection 3.1).

2. **Cleaned BrWaC:** This dataset is a cleaned version from the original dataset BrWaC [Wagner et al. 2018]. The original dataset collects multiple websites without a specific domain, making this data generic and large. The generic and coverage are a trade for quality, as many politics-related websites or offensive content are indexed (cf. 3.2).
3. **MCCD Generated:** Novel human conversational datasets generated using our proper methodology MCCD [Sanches. et al. 2022] and Miner-XenForo[2] (cf. 3.3).
    (a) **Adrenaline Dataset:** Dataset based on a Web Forum about technology, hardware, and games. We were able to identify high engagement in conversations by users and longer conversations.
    (b) **OuterSpace Dataset:** Dataset based on a Web Forum about games on different platforms and millions of messages generic messages.

We choose not to compare these datasets directly as the final objective, creation method, annotation, and structure may differ and this may lead to wrong conclusions. All the scripts used to generate our datasets, methodologies, software tools, and datasets are available at GitHub [3].

3.1. MultiWOZ-PTBR

The original MultiWOZ is a dataset to create high-quality data to train NLP models to deal with dialogues, where a user and a system talk to each other through turns about different topics. Besides the large size of this dataset, with more than 110.000 turns, the content is fully annotated, allowing researchers to use this dataset for distinct objectives.

This dataset is only available in English because the generated data uses crowdsourcing methods. The generation process was done using *Amazon Mechanical Turk*[4] users following a script to create the initial conversation and later annotate each turn fully.

The utterance contains the actual text obtained through crowd-sourced. It is usually 13 words long, for instance, "*I need a train to stansted airport that leaves after 21:30. I am also going to be looking for local places to eat.*". The slot value only contains specific information that was annotated, for instance, "stansted airport". Meta data and data unrelated to the sentences were not translated as they would not be used by any model directly.

Every slot value is related to a unique utterance and is part of the utterance string. Although this is very useful during the annotation, the matching between the slot value inside the string may be missing after the translation. This behavior is caused by the translation, which may translate an entire utterance differently to just a piece to better match the context. For instance, the utterance "*Yes, I am also looking for a barbeque restaurant in the same area and price range as my hotel.*" has the slot "barbeque".

The translated utterance "*Sim, também estou procurando uma churrascaria na mesma área e faixa de preço do meu hotel." has the slot "churrasco"*. In this case, the

[2]https://github.com/MatheusFerraroni/miner-xenforo

[3]https://github.com/MatheusFerraroni/nlp_ptbr_datasets

[4]https://www.mturk.com/

change occurs because the target language uses the word "*churrascaria*" to "*barbeque restaurant*", causing the loss between these direct matches. To tackle the changes between utterance and slot values after the translation, we identified the mismatches and manually fixed most of these cases. Although there are a few occurrences in the final version, the dataset already is a viable option to train NLP models.

3.2. Cleaned BrWaC

The WaCky - The Web-As-Corpus Kool Yinitiative [Baroni et al. 2009] paved the way for a new set of datasets based on WebSites. This kind of corpus is often very large, as it can collect the content of many different websites. Various websites, including news to personal blogs, collaborate on the corpus generality.

Based on WaCky initiative, the BrWaC dataset was created to be the equivalent of this corpus, but in Portuguese. As the general purpose and the extensive collection of websites, this dataset is one of the biggest available in that language. However, the data does not have any meta-data, and some instances can be labeled as offensive or low-quality. The model BERTimbau[Souza et al. 2020], trained with BrWaC, generates problematic outputs due to examples presented in the training dataset.

Figure 1 presents the output from BERTimbau to replace the token [MASK] in the sentence 'THE PSYCHOPATHY of [MASK] is so big', with the 'is' being written as a slang. The problem occurs as the model suggests the name of a Brazilian politician, Hitler, a USA politician, and God to replace the token.

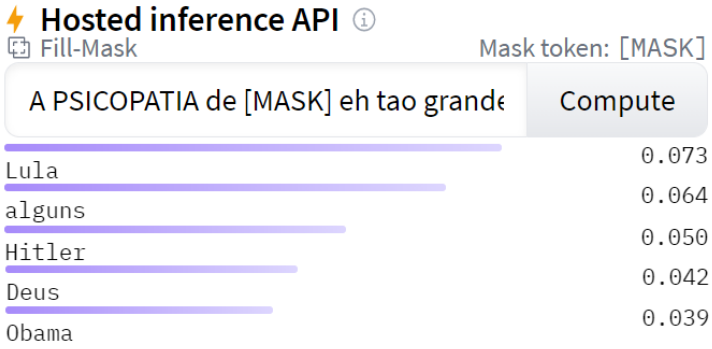

Figure 1. Problematic output from BERTimbau model.

In order to tackle this problem and create a cleaner version of BrWaC corpus, we applied a specific methodology with three steps: 1) Data Exploration; 2) low-quality removal; and 3) offensive content removal.

The BrWaC is separated into two main data types, an URL and the content of that specific WebPage. By exploring the URLs available, we found that many URLs are related to personal blogs and politics. On this basis, we created a list of ignored domains. Based on the findings of the URLs, we removed most of the problematic content, as low-quality content was usually related to personal blogs.

In order to further filter the content, we developed an algorithm capable of inspecting every word in the dataset and searching for offensive terms, which led to discarding

that specific instance from the dataset. A list of offensive terms was created and fed into our algorithm, which searches for variations for these offensive terms. We discarded instances not classified as Portuguese language or that had more than 25% of digits.

The original BrWaC corpus contains 120989 domains and 3530796 pages. After the cleaning process, the cleaned version contained 96814 domains and 3166108 pages, reducing more than 24000 domains and more than 360000 pages.

3.3. MCCD Generated

This Subsection presents the datasets created using our MCCD methodology [Sanches. et al. 2022] and Miner-XenForo [Sanches. et al. 2022]. We created two datasets using the same methodology and software tool but with different data sources. Due to this, the shared datasets have very different content with specific characteristics.

Both datasets use online forums as data sources due to constraints from the Miner-XenForo tool. Although this may limit the domain from the generated datasets to the forum domain, the data generated has some moderation. Such forums rely on rules and administrators to keep order and organization. All data gathered is publicly available and can be found accessing the source directly in the browser.

The datasets are released in two versions, one with raw data acquired, containing only data about messages and categories, and one with the processed files and conversations identified. The conversations have been identified by the *Miner-XenForo*, which implements our methodology MCCD to follow references between messages to identify all the possible conversation flows.

Table 1 compares different metrics of both generated datasets. The raw size of both datasets is similar, but the Processed size is very different, especially if we compare it with the Raw Size. We observe that the Adrenaline gets 2730% larger after processing, whereas OuterSpace increases only 17%. This difference is explained by other metrics, which highly influence the number of conversation flows.

Table 1. Comparison between Adrenaline and OuterSpace Dataset.

Data	Adrenaline	OuterSpace	Adrenaline/OuterSpace
Raw Size (Gb)	2,0	4,6	0,43
Processed Size (Gb)	71	5,4	13,0
Total Topics	356k	570k	0,62
Total Messages	9.5M	24M	0,38
Total Tokens	477M	1T	0,44
Mean Size Per Topic	26	42	0,62
Mean Token Per Post	50	43	1,14
Mean Size Conversation	34	3	11,3
Total Conversation	4.5M	3M	1,47
Unique Tokens	1.7B	2.7B	0,63
Mean Token per Topic	1.3K	1.8k	0,71
Source	https://forum.adrenaline.com.br/	https://forum.outerspace.com.br/	-

Both datasets have similar raw sizes, total topics, total conversations, and total messages. Nevertheless, users from Adrenaline are more engaged in conversations, which the Mean Size Conversation confirms. In the Adrenaline dataset, the discussions are usually longer, with 34 messages; in OuterSpace conversation, the mean size is only 3. This means that the Adrenaline Forum users are more likely to answer other messages

and these conversations are longer than OuterSpace. More persons can join with longer conversations, and more conversational flows were identified.

Figure 2 compares the total of messages per topic in both datasets. We observe that the distribution of messages per topic is almost exponential, with many topics with few messages and few topics with many messages. The OuterSpace dataset has a larger number of topics without any response. Considering the characteristics presented, we note that even with more messages in general and more messages per topic, the OuterSpace fails to engage users in conversations.

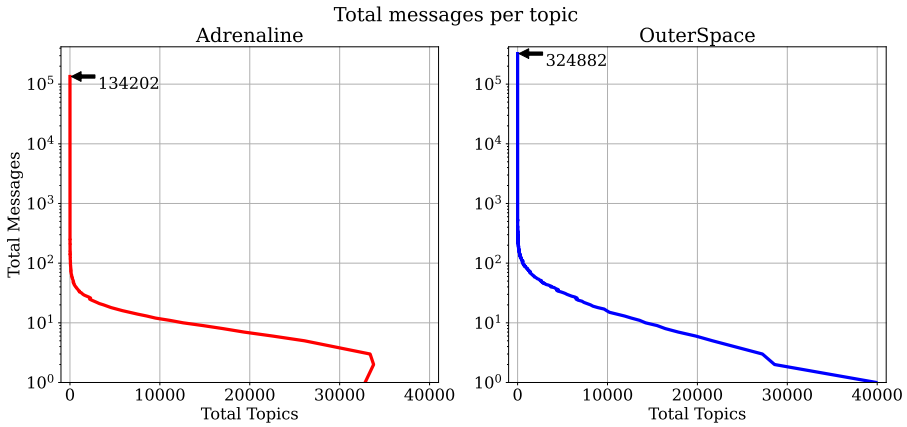

Figure 2. Comparison between total messages per topic between Adrenaline and OuterSpace datasets.

4. Experimental Evaluation

This section evaluates two BERT models trained for two classification tasks. These classification tasks were chosen based on their fitness to highlight the difference in how the models deal with specific contexts. The first model is a literature model, *BERTimbau*; the second is a sample model trained on a portion of the Adrenaline dataset exclusive for this evaluation. This evaluation aims to demonstrate that the proposed datasets can be used to train novel NLP models and which limitations are faced in different situations.

4.1. Procedures

The Figure 3 presents the complete flow during evaluation, including the training of our sample Bert Model to the final classification task performed. The grey box represents the sets to train our sample model and the green box represents the results from the classification tasks to determine the Category and Sub-Category of each instance.

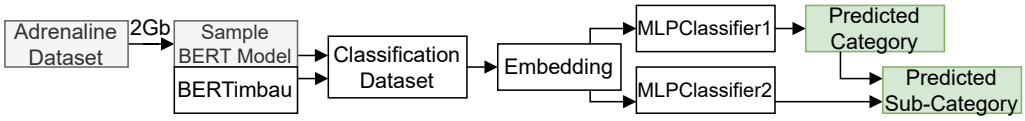

Figure 3. Complete Evaluation flow

The first step in the evaluation was creating a sample BERT model to be compared with an existing model from the literature. We use the raw data from Adrenaline Dataset, presented in Section 3, to train this sample model. Data used from the Adrenaline Dataset represents about 2Gb of textual data.

Our sample BERT model was trained using masked language modeling, with a learning rate of 5^{-05} and AdamW optimizer. As the dataset has a large variety of sentences and this is just a sample model, we train this model for only one epoch with a weight decay of 5^{-07}. As the evaluation task is known, we safely set the max sequence length considered by the model to 512.

Both models, our sample BERT-base model, and *BERTimbau*, were used to generate the embedding for each instance from both classification tasks. This was done by inputting the instance to the model and saving the state of the last layer once the model output the token "[CLS]". The embedded save is formed by and 1D array with 768 floats.

The embedded generated for each instance was then used to train a *Multi-layer Perceptron classifier*, using the embedding as input and the class as label. We used the same architecture for the MLPClassifier in both classification tasks, using 100 layers with 200 nodes, each activated by using a ReLU function. In both classification tasks, 75% of the dataset was used during the training stage, and 25% was used as validation.

The first classification task is to determine which Category and Sub-Category a post belong in a portion of the Adrenaline Dataset, considering only the title and the first message content concatenated. Figure 4 shows how the Categories and Sub-Categories are distributed. There are seven categories with an unbalanced number of Sub-categories. The first Category has 10 Sub-Categories; the second and third Categories have 4 Sub-Categories; the fourth Category has three sub-categories; the fifth and sixth categories have two Sub-Categories; and the last Category has only one Sub-Category. Every Sub-Category has exactly 900 instances. The total instances in the Sub-Categories are balanced, but the instances in the Categories are unbalanced.

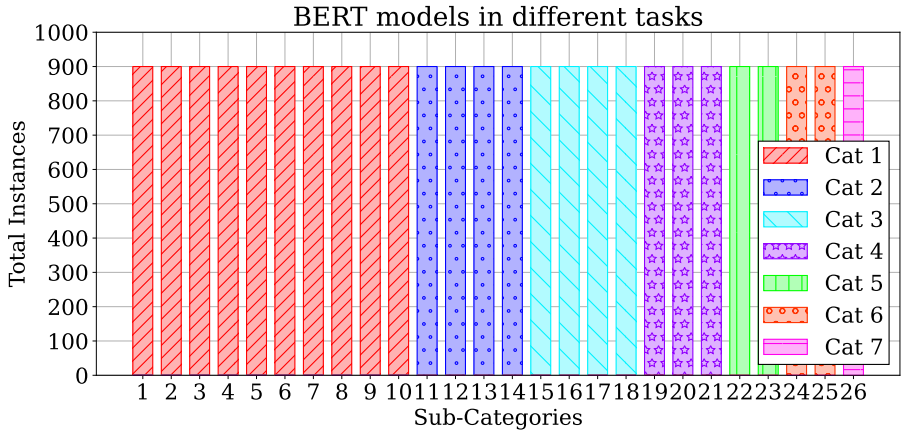

Figure 4. Categories and Sub-Categories distribution.

We used both BERT models to determine the Category using the embedding for

each instance as input. In order to assist the Sub-Category prediction, the predicted Category was input with the embedding to predict the Sub-Category.

The second classification task is to determine if a movie review is positive or negative [Gonçalves 2022] (sentiment analysis). This dataset contains almost 50000 instances and is already balanced between positive and negative reviews. This scenario was chosen to compare the effectiveness of the evaluated BERT models on a scenario with different words and goals as a generic scenario.

4.2. Experimental Results

Figure 5 shows the score obtained for the two classification tasks. The bars at "Category" and "Sub-Category" are the scores obtained at the classification Task in the Adrenaline Dataset; the bars at "Sentiment" are the scores in the sentiment analysis classification dataset. Although we show only the score obtained, the F1 score diverges less than 1pp.

Results obtained from our evaluation indicate that our sample model, which was trained for one epoch, achieves a better result representing the sentences and working as input to an *MLPClassifier*. The sample model scores more than 9% more than *BERTimbau* at classifying the Category and almost 4% more at classifying the Sub-Category. These gains translate to a 4pp at the Category and 3pp at the sub-Category.

Although our sample model achieves better scores at the first classification task, the *BERTimbau* gets better results classifying sentiments in the second classification task. *BERTimbau* was able to score 0.86, while our sample model only scored 0.77, representing about 11% gain. The score difference happens due to the training dataset used by BERTimbau, which is more open-domain, allowing the model to be better suited for generic tasks.

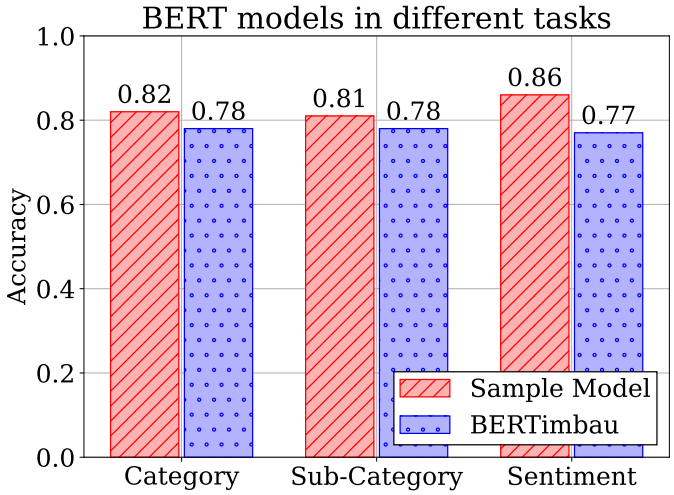

Figure 5. Score comparison between our sample model and BERTimbau.

5. Discussion

The explored datasets play a central role as a factor that influences the results obtained. This happens due to ability of the model to properly understand the semantics of words and phrases to represent a better context at the embedding layer.

The Adrenaline and the OuterSpace datasets were explicitly processed and structured with human conversations between two or more identified users. We employed a specific methodology to translate the MultiWOZ dataset to Portuguese. We described the approach used during the translation and how different portions of the translated sentences were matched. Furthermore, we created a cleaned version of the biggest Web Corpus in Portuguese, BrWaC. We showed how we removed low-quality content and instances classified as offensive by our methodology.

Although the Adrenaline and OuterSpace datasets can be used to train models with distinct objectives, such as Questions & Answers (QA) and text generation, both datasets are limited to the specific themes from the original data source. The Adrenaline dataset was created from an online forum with the main topic related to technology and games. OuterSpace dataset has a larger but limited range of topics, including one sub-category completely open topic. Models being trained on these datasets require attention to this and may require an additional dataset to better represent the sentences into embedding.

To complete the data being used during the training stage, we can use our cleaned version of BrWaC with Adrenaline and OuterSpace datasets. With the composition of datasets being used, it is possible to train an NLP model to understand the sentences better and achieve better results.

The difference in the model's ability to understand the data is directly related to the dataset used and how the training stage was conducted. The dataset utilized by our sample model was directly related to the data being evaluated. In this sense, the model could better represent specific words from the dataset context, which result in better representation at the embedding layer. The *BERTimbau* was trained with BrWaC, a very generic and large dataset, allowing the model to perform well on most tasks but not achieve the best result possible in specific tasks.

The training stage on our sample model was small, as this model was trained only to validate the datasets and show it is possible to obtain an adequate model even with small data. Despite the differences and conclusions already explained, it is possible to refine an already trained version of the model to a specific domain, such as the *BERTimbau*. Using this approach, the model can keep most of the knowledge already learned and achieve even better results. Using context-specific data during the refining stage with a lower learning rate allows the model to learn new patterns related to the new context.

6. Conclusion

The lack of specific datasets to address key tasks in different languages is a challenging research problem. This study presented a set of datasets specifically created or adapted to train or refine NLP models in the Portuguese language. We share two novel human-to-human datasets with multiple answers to the same message in Portuguese, one translated dataset, and one cleaned version of a well-established dataset. We found that NLP models trained with our dataset can outperform literature models in specific tasks. Future studies involve training the first NLP models with BERT and GPT architectures, using the four different datasets presented. These models may be employed to solve various tasks and evaluated against the state-of-the-art models on well-established benchmarks. We will determine the differences in how the state-of-the-art models and our models can represent words and sentences in the embedding layer and how this influences several tasks.

Acknowledgments

This research was supported by CI&T. We are thankful to our colleagues Diego Augusto, Lead Data Scientist at CI&T, and Gabriel Marostegam, Head of Data at CI&T, for their important participation in discussions during the development of this work.

References

[Baroni et al. 2009] Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

[Budzianowski et al. 2018] Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. (2018). Multiwoz – a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling.

[Devlin et al. 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

[Gonçalves 2022] Gonçalves, L. (2022). Imdb pt-br. `https://www.kaggle.com/datasets/luisfredgs/imdb-ptbr`. Accessed: 2022-05-25.

[Guillou 2020] Guillou, P. (2020). Gportuguese-2 (portuguese gpt-2 small): a language model for portuguese text generation (and more nlp tasks...).

[Howard and Ruder 2018] Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification.

[HuggingFace 2022a] HuggingFace (2022a). Hugging face – the ai community building the future. `https://huggingface.co/datasets?languages=languages:en`. Accessed: 2022-05-25.

[HuggingFace 2022b] HuggingFace (2022b). Hugging face – the ai community building the future. `https://huggingface.co/datasets?languages=languages:pt`. Accessed: 2022-05-25.

[Kaplan et al. 2020] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models.

[Lowe et al. 2016] Lowe, R., Pow, N., Serban, I., and Pineau, J. (2016). The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems.

[Meta 2021] Meta (2021). Main page — meta, discussion about wikimedia projects. [Online; accessed 25-May-2022].

[Poncelas et al. 2020] Poncelas, A., Lohar, P., Way, A., and Hadley, J. (2020). The impact of indirect machine translation on sentiment classification.

[Radford et al. 2019] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

[Rajpurkar et al. 2018] Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for squad.

[Sanches. et al. 2022] Sanches., M., C. de Sá., J., M. de Souza., A., Silva., D., R. de Souza., R., Reis., J., and Villas., L. (2022). Mccd: Generating human natural language conversational datasets. In *Proceedings of the 24th International Conference on Enterprise Information Systems - Volume 2: ICEIS,*, pages 247–255. INSTICC, SciTePress.

[Sharir et al. 2020] Sharir, O., Peleg, B., and Shoham, Y. (2020). The cost of training nlp models: A concise overview.

[Souza et al. 2020] Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

[Vaswani et al. 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

[Wagner et al. 2018] Wagner, J., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brwac corpus: A new open resource for brazilian portuguese.

[Wang et al. 2018] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding.