Language Resources and Evaluation
https://doi.org/10.1007/s10579-023-09661-4

ORIGINAL PAPER

Sentiment analysis in Portuguese tweets: an evaluation of diverse word representation models

Daniela Vianna[1,2] · Fernando Carneiro[3] · Jonnathan Carvalho[4] · Alexandre Plastino[3] · Aline Paes[3]

Accepted: 27 March 2023

Abstract

During the past years, we have seen a steady increase in the number of social networks worldwide. Among them, Twitter has consolidated its position as one of the most influential social platforms, with Brazilian Portuguese speakers holding the fifth position in the number of users. Due to the informal linguistic style of tweets, the discovery of information in such an environment poses a challenge to Natural Language Processing (NLP) tasks such as sentiment analysis. In this work, we state sentiment analysis as a binary (positive and negative) and multiclass (positive, negative, and neutral) classification task at the Portuguese-written tweet level. Following a feature extraction approach, embeddings are initially gathered for a tweet and then given as input to learning a classifier. This study was designed to evaluate the effectiveness of different word representations, from the original pre-trained language model to continued pre-training strategies, to improve the predictive performance of sentiment classification, using three different classifier algorithms and eight Portuguese tweets datasets. Because of the lack of a language model specific to Brazilian Portuguese tweets, we have expanded our evaluation to consider six different embeddings: fastText, GloVe, Word2Vec, BERT-multilingual (mBERT), BERTweet, and BERTimbau. The experiments showed that embeddings trained from scratch solely using the target Portuguese language, BERTimbau, outperform the static representations, fastText, GloVe, and Word2Vec, and the Transformer-based models BERT multilingual and BERTweet. In addition, we show that extracting the contextualized embedding without any adjustment to the pre-trained language model is the best approach for most datasets.

Keywords Sentiment analysis · Word representation · Brazilian Portuguese tweets · Language models

Extended author information available on the last page of the article

Published online: 28 June 2023

1 Introduction

Sentiment analysis or opinion mining is the field of study that aims at identifying opinions or sentiments expressed in the text and the targets of these opinions or sentiments (Liu, 2020). Over the years, social media networks, such as Facebook, Twitter, and Instagram, have emerged and become widely used worldwide. With ease and freedom, users voice their opinions on the most varied subjects, providing researchers with a rich source of digital text data that spearheaded the area of sentiment analysis.

Among all the leading social networks worldwide, by January 2021, Twitter was ranked 16th by the number of active users,[1] which corresponds to more than 10% of the world's overall social media users. Twitter allows users to connect with each other and share their information through short 280-character messages called tweets. Those short messages usually have an informal style, relying heavily on abbreviations, hashtags, and emoticons, among other social media unique features. The discovery of information in such an environment, with incomplete and noisy data, poses a challenge to Natural Language Processing (NLP) tasks, such as sentiment analysis.

In this work, we formulate the sentiment analysis task as a binary classification—and also as a multiclass task—at the tweet level, i.e., given a specific tweet, the goal is to determine whether it reveals a positive or negative opinion (binary mode) or whether it expresses a neutral, positive or negative message (multiclass mode). Furthermore, we focus on tweets written in Brazilian Portuguese, motivated by the fact that Brazil is the fifth country in the absolute number of Twitter users, with 16.2 million active accounts as of January 2021.[2] Moreover, Portuguese is estimated as the ninth most spoken language around the world, with an amount of 258 million Portuguese speakers.[3]

Techniques for classifying user opinions from tweets have been extensively studied, ranging from traditional classification methods to the more recent deep learning-based methods (Kouloumpis et al., 2011; Tang et al., 2014; Severyn & Moschitti, 2015; Machado et al., 2018; Carvalho & Plastino, 2021; Barreto et al., 2021). As a first step, to extract knowledge from the vast amount of text data available using machine learning (ML) techniques, it is essential to select efficient methods to represent these textual content numerically. In the early days, such a numerical representation was commonly extracted from bag-of-words (BoWs) formulations. However, due to their sparsity and lack of semantics, nowadays, BoWs representations are commonly replaced by learning-based vector representations of words, the word embeddings (Mikolov et al., 2013). A common practice when solving tasks relying on embeddings is to leverage pre-trained language models induced from extensive corpora. Initially, those pre-trained word representations were defined statically to individual words or sub-words, i.e., when reused, they would not consider the

[1] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[2] https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/.

[3] https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/.

context in which the word was inserted. Then, state-of-the-art evolved to contextual embedding techniques such as the Transformer-based autoencoder methods, including BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). These language models are pre-trained on large-scale unlabeled text datasets and can be fine-tuned for each application, reducing computational needs and improving classification performance.

Although fine-tuning pre-trained embedding models have shown great advantages over more traditional task-centered approaches, the majority of those models have aimed at the English language, even though English-speaking users amount to only 25.9% of all internet users worldwide as of January 2020.[4]. To overcome the dearth of resources for sentiment analysis in languages other than English, multilingual sentiment analysis techniques have been proposed with noteworthy results (Singhal & Bhattacharyya, 2016; Nankani et al., 2020; Agüero-Torales et al., 2021), despite often being smaller than their monolingual counterparts for English (Martin et al., 2020). Recent efforts have also been made to train monolingual language models for various languages (Virtanen et al., 2019; Nozza et al., 2020; Chan et al., 2020), including Portuguese (Souza et al., 2020; Carmo et al., 2020; Vargas & Moreira, 2020). However, learning those huge models from scratch requires massive data and computational resources that are rarely available for most users. Besides, recent research has discussed the environmental impact of training such huge models (Strubell et al., 2019, 2020; Bender et al., 2021), as a regular Transformer architecture with 213M parameters may emit more than 626,000 pounds of carbon dioxide equivalent (Strubell et al., 2019).

Compelled by those environmental concerns and given the lack of dedicated hardware resources to train a language model from scratch, in this work, we investigate the benefits of adapting pre-trained embeddings for the language of Portuguese tweets. Notably, we follow a feature extraction approach in that we first gather the embeddings for a sentence, and next, we give them as input to learn a classifier. The extracted features may come either from the original language model or from a *continued pre-training strategy* (Gururangan et al., 2020). We adjust the language model using the intermediate self-supervised masked language task in this second case. With those schemas, we would like to point out an effective strategy for relying on pre-trained tweets for sentiment analysis according to the following questions: (i) regarding how the embedding is induced, is a contextualized embedding, in fact, more effective than a static representation? (ii) Regarding the language model, do embeddings trained from scratch for the target Portuguese language work better than a multilingual or a tweets-based language model? (iii) When adjusting the language model with continued pre-training, is it better to rely on more data or focus on the specific data coming from the target dataset?

More precisely, we make the following contributions:

[4] https://www.statista.com/statistics/262946/share-of-the-most-common-languages-on-the-internet/.

(1) A literature review on Twitter sentiment analysis studies for Brazilian Portuguese (Sect. 2).
(2) A survey of annotated datasets for the task of sentiment analysis of Portuguese tweets. This process involved contacting various authors and retrieving tweets using the Twitter API resulting in a set with nine labeled tweet datasets (Sect. 3.1).
(3) A quantitative evaluation of three static embeddings, fastText (Bojanowski et al., 2017), GloVe (Pennington et al., 2014), and Word2Vec (Mikolov et al., 2013), and three context-based embeddings: (1) a multilingual version of BERT (Devlin et al., 2019), (2) the English tweets-based language model BERTweet (Nguyen et al., 2020), and (3) the Portuguese pre-trained model BERTimbau (Souza et al., 2020) (Sect. 4).
(4) A quantitative evaluation of those three contextualized embeddings adapted to Portuguese tweets targeted at sentiment analysis using three different strategies, as follows: (1) using $n - 1$ datasets to the continued pre-training stage and extracting the yielded embedding to learn a classifier to the remaining (target) dataset (*leave-one-out*, LOO), where n is the total number of available datasets; (2) using only the tweets of the target dataset to adjust the language model (*inData*), and (3) adjusting the language model with all the n datasets (*allData*) (Sect. 5).

All the code produced in this manuscript is available at https://github.com/MeLLL-UFF/embeddings-tweets-pt-br-lrev, and the datasets can be downloaded from https://bityli.com/RvhFax.

2 Literature review

This section presents a literature review on Twitter sentiment analysis studies for Brazilian Portuguese (PT-BR). Existing approaches to Twitter sentiment analysis for PT-BR mainly focus on ML techniques (Aguiar et al., 2018; Alves et al., 2014, 2015; Barros et al., 2021; Belisário et al., 2020; Brum & das, 2018; Brum & Nunes, 2018; Carosia et al., 2019, 2020; Carvalho et al., 2020; Correa et al., 2017; Costa et al., 2015; França & Oliveira, 2014; Garcia & Berton, 2021; Gengo & Verri, 2020; Gomes et al., 2018; Grandin & Adan, 2016; Moraes et al., 2016; Nascimento et al., 2015; Neuenschwander et al., 2014; de Oliveira & de Campos Merschmann, 2021; Praciano et al., 2018; Rosa et al., 2013; dos Santos et al., 2018; Silva et al., 2018; Souza et al., 2016a, b; Vargas et al., 2020; Vargas & Moreira, 2020; Vitório et al., 2017, 2019; Yagui & Maia, 2017) or on annotated resources, such as lexicons or dictionaries (Araujo et al., 2016; Araújo et al., 2018, 2020; Belisário et al., 2020; Correa et al., 2017; Cury, 2019; Lauand & Oliveira, 2014; Lima et al., 2016; Malini et al., 2017; Martins et al., 2015; de Melo & Figueiredo, 2021; Moraes et al., 2016; Neuenschwander et al., 2014; Pessanha et al., 2020; Souza et al., 2016b; de Souza et al., 2017; Souza & Vieira, 2012). Fewer studies use different approaches, such as graph- and rule-based methods, and off-the-shelf sentiment analysis systems (Araujo et al., 2016; Araújo et al., 2020; Belisário et al., 2020; Guerra et al., 2014, 2011; Lourenco et al., 2014; Martins et al., 2015; Oliveira et al., 2019; Oliveira & de Souza Bermejo, 2017; Oliveira et al., 2017; Silva et al., 2020, 2011; Souza et al., 2016b; Vilhagra et al., 2020).

Table 1 summarizes Twitter sentiment analysis studies for PT-BR. The *Representations (ML)* column describes the feature representation adopted in ML-based methods, which are the focus of this study. Some of these studies are discussed in Sects. 2.1 and 2.2, focusing on ML- and lexicon-based approaches, respectively.

2.1 Machine learning-based approaches

ML-based methods extract features from texts that are used to train classifiers. So far, the most common feature representation adopted in the sentiment classification of Brazilian Portuguese tweets are still the BoWs and the n-gram models (Aguiar et al., 2018; Alves et al., 2014, 2015; Belisário et al., 2020; Brum & das, 2018; Brum & Nunes, 2018; Carosia et al., 2019, 2020; Carvalho et al., 2020; Correa et al., 2017; Costa et al., 2015; França & Oliveira, 2014; Garcia & Berton, 2021; Gengo & Verri, 2020; Gomes et al., 2018; Grandin & Adan, 2016; Moraes et al., 2016; Nascimento et al., 2015; Neuenschwander et al., 2014; de Oliveira & de Campos Merschmann, 2021; Praciano et al., 2018; Rosa et al., 2013; dos Santos et al., 2018; Silva et al., 2018; Souza et al., 2016a; Vargas et al., 2020; Vitório et al., 2017, 2019; Yagui & Maia, 2017). Nevertheless, it is well known that these types of representations cannot deal with the curse of dimensionality, considering that they make the feature space highly dimensional and often very sparse. For example, in de Oliveira and de Campos Merschmann (2021), Oliveira and Merschmann evaluated various pre-processing tasks for sentiment analysis in the Brazilian Portuguese language, including three datasets of tweets, using the BoW representation. However, they failed at executing some classifiers due to the large number of features that characterized each dataset. Thus, they decided to use as features only the most frequent terms in each dataset.

Regardless of the curse of dimensionality, the BoW model has been used in many applications in different domains, such as politics (Nascimento et al., 2015; Praciano et al., 2018; Souza et al., 2016a; Vitório et al., 2019), entertainment (Brum & das, 2018; Gengo & Verri, 2020; dos Santos et al., 2018), social manifestation (Costa et al., 2015; França & Oliveira, 2014; Yagui & Maia, 2017), financial market (Carosia et al., 2019, 2020; Neuenschwander et al., 2014), technology (Belisário et al., 2020; Moraes et al., 2016), health (Garcia & Berton, 2021; Vargas et al., 2020), consumers' sentiment (Rosa et al., 2013), public security (Carvalho et al., 2020), calamity (Silva et al., 2018), sports (Alves et al., 2014), and environment (Alves et al., 2015). An interesting application in the domain of social manifestation is related to the sentiment analysis of tweets during the Brazilian democratic protests that occurred in 2013, as reported in Costa et al. (2015) and França and Oliveira (2014). For example, França and Oliveira (2014) used the Naive Bayes (NB) classifier with BoW to identify the sentiment expressed in tweets collected from June to August 2013, trying to understand whether the Brazilian citizens were supporting the protests.

Table 1 Summary of Twitter sentiment analysis studies in Brazilian Portuguese in descending order of year

Year	Studies	Domains	ML	Lex.	Other	Representations (ML)
2021	Barros et al. (2021)	Entertainment/Multi-domain	✓			BERTimbau (Souza et al., 2020)
	Garcia and Berton (2021)	Health	✓			N-grams, SBERT (Reimers & Gurevych, 2019) fastText (Bojanowski et al., 2017), mUSE (Yang et al., 2019)
	de Melo and Figueiredo (2021)	Health		✓		–
	de Oliveira and de Campos Merschmann (2021)	Multi-domain	✓			BoW (TF-IDF)
2020	Araújo et al. (2020)	Multi-domain		✓		–
	Belisário et al. (2020)	Technology	✓	✓	✓	BoW, Word2Vec
	Carosia et al. (2020)	Financial market	✓			BoW
	Carvalho et al. (2020)	Public security	✓			BoW (TF-IDF)
	Gengo and Verri (2020)	Entertainment	✓			BoW, hand-crafted features
	Pessanha et al. (2020)	Health		✓		–
	Silva et al. (2020)	Consumers’ sentiment			✓	–
	Vargas et al. (2020)	Health	✓			BoW, hand-crafted features
	Vargas and Moreira (2020)	Multi-domain	✓			BERTPT, ALBERTPT
	Vilhagra et al. (2020)	Multi-domain			✓	–
2019	Carosia et al. (2019)	Financial market	✓			BoW
	Cury (2019)	Politics		✓		–
	Oliveira et al. (2019)	Social manifestation			✓	–
	Vitório et al. (2019)	Politics	✓			BoW

Table 1 (continued)

Year	Studies	Domains	ML	Lex.	Other	Representations (ML)
2018	Aguiar et al. (2018)	Consumers' sentiment	✓			BoW
	Araújo et al. (2018)	Health		✓		–
	Brum and das (2018)	Entertainment	✓			BoW, hand-crafted features
	Brum and Nunes (2018)	Multi-domain	✓			Word2Vec, BoW, hand-crafted features
	Gomes et al. (2018)	Multi-domain	✓			BoW
	Praciano et al. (2018)	Politics	✓			BoW
	dos Santos et al. (2018)	Entertainment	✓			BoW
	Silva et al. (2018)	Calamity	✓			BoW
2017	Correa et al. (2017)	Multi-domain	✓	✓		BoW (TF-IDF), Word2Vec, Doc2Vec hand-crafted features
	Malini et al. (2017)	Politics		✓		–
	Oliveira and de Souza Bermejo (2017)	Social manifestation			✓	–
	Oliveira et al. (2017)	Politics			✓	–
	de Souza et al. (2017)	Politics		✓		–
	Vitório et al. (2017)	Multi-domain	✓			BoW
	Yagui and Maia (2017)	Social manifestation	✓			BoW
2016	Araujo et al. (2016)	Multi-domain		✓		–
	Grandin and Adan (2016)	Multi-domain	✓			N-grams
	Lima et al. (2016)	Financial market		✓		–
	Moraes et al. (2016)	Technology	✓	✓		BoW
	Souza et al. (2016a)	Politics	✓			BoW (TF-IDF)
	Souza et al. (2016b)	Consumers' sentiment	✓	✓	✓	Hand-crafted features

Table 1 (continued)

Year	Studies	Domains	ML	Lex.	Other	Representations (ML)
2015	Alves et al. (2015)	Environment	✓			BoW
	Costa et al. (2015)	Social manifestation	✓			BoW
	Martins et al. (2015)	Financial market/Automobiles		✓	✓	–
	Nascimento et al. (2015)	Politics/Crime/Entertainment	✓			N-grams
2014	Alves et al. (2014)	Sports	✓			BoW
	França and Oliveira (2014)	Social manifestation	✓			BoW
	Guerra et al. (2014)	Sports			✓	–
	Lauand and Oliveira (2014)	Traffic		✓		–
	Lourenco et al. (2014)	Politics/Sports			✓	–
	Neuenschwander et al. (2014)	Financial market	✓	✓		BoW
2013	Rosa et al. (2013)	Consumers' sentiment	✓			N-grams
2012	Souza and Vieira (2012)	Multi-domain		✓		–
2011	Guerra et al. (2011)	Politics/Sports			✓	–
	Silva et al. (2011)	Politics/Sports			✓	–
Total			32	17	12	

Another appealing application is in the health domain to determine the sentiment expressed in tweets about the COVID-19 pandemic in Brazil Vargas et al. (2020). In Vargas et al. (2020), Vargas et al. have recently introduced the OPCovid-BR dataset, which consists of 600 manually labeled tweets about the COVID-19 pandemic posted by Brazilian Twitter users. They also used two other datasets in domains other than health to measure the performance of cross-domain polarity classifiers. They showed that this cross-domain strategy improved the results using an SVM classifier. In addition to BoW, they extracted hand-crafted features that may have helped the classifiers discern between positive and negative tweets. Similarly to Vargas et al. (2020), a few other studies used hand-crafted features, such as emoticons and emojis, sentiment words, Part-Of-Speech (POS) tags, the presence of exclamation and question marks, the presence of words with a sequence of repeated characters, and others (Brum & das, 2018; Brum & Nunes, 2018; Correa et al., 2017; Gengo & Verri, 2020; Souza et al., 2016b).

It is noteworthy that despite the recent and notable advances of deep learning techniques in developing efficient feature representation for NLP tasks, only a few studies in the literature of Twitter sentiment analysis for the Brazilian Portuguese language have explored word embedding-based features (Barros et al., 2021; Belisário et al., 2020; Brum & Nunes, 2018; Correa et al., 2017; Garcia & Berton, 2021; Vargas & Moreira, 2020). Word embeddings are dense feature vectors learned by neural-based techniques to represent words and texts and have been increasingly used to tackle the curse of dimensionality inherited from the BoW-based approaches.

Word embedding-based representations can be roughly divided into static and contextualized models. While static embedding models pre-compute the representation for each word of an input text independently of the context they appear, the contextualized ones are not fixed, leveraging the context and adapting the word representation. Word2Vec (Mikolov et al., 2013) and fastText (Bojanowski et al., 2017) are examples of static embedding models widely adopted in NLP literature. FastText is based on Word2Vec's Skip-Gram model and can deal with tweets containing many uncommon words, considering it learns word representations at sub-words and character levels.

Garcia and Berton (2021) presented a broad evaluation of different classifiers combining N-grams and recent embedding models, including fastText (Bojanowski et al., 2017), mUSE (Yang et al., 2019), and SBERT (Reimers & Gurevych, 2019), in the sentiment classification of tweets related to the COVID-19 pandemic in Brazil and USA. In the experimental evaluation, they used a dataset of 780,000 positive and negative tweets to train the classifiers and showed that combining unigrams, bigrams, and word embedding-based features is beneficial for the sentiment classification of Brazilian Portuguese tweets.

Regarding the use of contextualized embedding models, Barros et al. (2021) proposed a novel methodology based on BERT (Devlin et al., 2019) to classify the sentiment of tweets using not only tokens from texts but also the expressiveness of emojis. BERT is a contextualized language model based on the Transformer architecture (Vaswani et al., 2017) designed to pre-train deep bidirectional representations from unlabeled texts. BERT has achieved state-of-the-art results on several NLP tasks, including sentiment analysis. In Barros et al. (2021), they extracted emojis from the input texts and processed them through the Transformer encoder independently to obtain the maximum information from both the words and emojis sequences. To avoid pre-training the model from scratch, they used BERTimbau (Souza et al., 2020)—a BERT model pre-trained on the brWaC corpus (Filho et al., 2018), which is composed of 2.7B tokens from 120,000 Brazilian Portuguese websites.

Vargas and Moreira (2020) introduced BERTPT and ALBERTPT pre-trained language models for Portuguese based on BERT and ALBERT (Lan et al., 2020), respectively. Their models were trained using 4.8 GB of texts from distinct sources in both Brazilian (BP) and European (EP) Portuguese languages, such as Wikipedia-PT[5] (EP), the Open Subtitles corpus[6] (BP), news articles from Kaggle[7] (BP) and the CHAVE corpus[8] (BP, EP), the EuroParl corpus[9] (EP), and research abstracts from the Brazilian website *Domínio Público*[10] (BP). BERTPT and ALBERTPT training took 33 and 17 h on one cloud TPU v2. They evaluated their pre-trained models on seven NLP tasks using the official BERT-multilingual (Devlin et al., 2019) model as a baseline. For the sentiment analysis task, they compared the classification performances of their models using only one dataset of Portuguese tweets and reported that ALBERTPT achieved the best results, followed by BERTPT. ALBERTPT and BERTPT were not evaluated in this work. Instead, we have chosen BERTimbau since it was pre-trained with a much larger collection of Portuguese texts when compared to ALBERTPT and BERTPT. Besides, both methods, ALBERTPT and BERTPT, were pre-trained using a data collection composed of a mix of Brazilian and European Portuguese texts, while BERTimbau was trained solely using Brazilian Portuguese texts, which is the specific language explored in this work.

2.2 Lexicon-based approaches

Unlike ML-based methods, lexicon-based ones aim at determining the sentiment expressed in texts by relying on annotated resources that contain prior sentiment information of words and phrases. In most cases, positive words are associated with values greater than zero, and negative words are associated with values smaller than zero. In that case, the overall polarity of a text can be determined by summing up the polarity values of its words. Concerning the usage of these methods in the sentiment classification of Brazilian Portuguese tweets, while some studies focus on presenting and evaluating methods particularly designed for Portuguese (Araujo et al., 2016; Araújo et al., 2020; Belisário et al., 2020; Martins et al., 2015; Moraes et al., 2016; Neuenschwander et al., 2014; Souza et al., 2016b; de Souza et al., 2017; Souza & Vieira, 2012), others report their application in different domains, such as health (Araújo et al., 2018; de Melo & Figueiredo, 2021; Pessanha et al., 2020), politics

[5] https://dumps.wikimedia.org/backup-index.html.

[6] http://opus.nlpl.eu/OpenSubtitles-v2016.php.

[7] https://www.kaggle.com/marlesson/news-of-the-site-folhauol.

[8] https://www.linguateca.pt/CHAVE/.

[9] https://www.europarl.europa.eu/.

[10] http://www.dominiopublico.gov.br/.

(Cury, 2019; Malini et al., 2017), financial market (Lima et al., 2016), and traffic (Lauand & Oliveira, 2014). For example, Pessanha et al. (2020) have recently analyzed the sentiment of Brazilian Twitter users towards the COVID-19 pandemic. For this purpose, they used SentiLex (Carvalho & Silva, 2015), which is a sentiment lexicon for Portuguese especially useful for detecting and classifying sentiments targeting human entities.[11] Their study revealed that the negative feeling implied in expressions such as "Brasil pede ajuda" (Brazil asks for help), "morte" (death), and "risco" (risk) was dominant throughout the analyzed period.

In Malini et al. (2017), Malini et al. used a Portuguese version of EmoLex (Mohammad & Turney, 2013) to analyze the sentiment expressed in tweets about the campaign for the impeachment of President Dilma Rousseff between 2015 and 2016. EmoLex is a manually annotated emotion lexicon comprising eight distinct levels of emotions (joy, sadness, anger, fear, trust, disgust, surprise, and anticipation). They show that pro- and anti-Dilma movements are marked by a predominance of anger, fear, and anxiety.

Souza and Vieira (2012) evaluated the impact of different sentiment lexicons for Portuguese and different heuristics to deal with negation. They compared SentiLex (Carvalho & Silva, 2015) and OpLexicon (Souza & Vieira, 2011) lexicons and evaluated two heuristics to handle negation: one based on a pre-fixed 5-word window and another that shifts the polarity of the entire sentence. To perform the experiments, they built a dataset with 1700 tweets collected relying on the positive and negative connotations of the hashtags #win and #fail, respectively, setting the language parameter to Portuguese. The experimental results indicate that OpLexicon outperformed SentiLex, while the treatment of negation had a low impact on the results.

Although most lexicon-based studies in the sentiment classification of Brazilian Portuguese tweets depend on Portuguese resources, a few of them take a different direction and exploit machine translation approaches (Araujo et al., 2016; Araújo et al., 2020; de Melo & Figueiredo, 2021). In Araújo et al. (2020), for example, Araújo et al. evaluated 16 state-of-the-practice sentiment analysis methods specifically designed for the English language in a multilingual study. They showed that simply translating texts in a specific language—including Portuguese tweets—to English, with the use of a machine translation system and then using one of the English methods to perform sentiment analysis, can outperform existing language-specific systems.

Also, in the context of machine translation approaches, de Melo and Figueiredo (2021) compared the sentiment expressed in tweets and news for a better understanding of the COVID-19 pandemic in Brazil. They collected 1,597,934 tweets and 18,413 news articles related to COVID-19 in Brazilian Portuguese from January to May 2020, which were translated to English using Googletrans, a free and unlimited Google Translate API.[12] After translating, they used VADER (Hutto & Gilbert, 2014) lexicon to calculate the degree of positivity and negativity of the texts.

[11] http://b2find.eudat.eu/dataset/b6bd16c2-a8ab-598f-be41-1e7aeecd60d3.

[12] https://pypi.org/project/googletrans/.

In Neuenschwander et al. (2014), Neuenschwander et al. presented a comparison of lexicon- and ML-based methods in Twitter sentiment analysis of Brazilian tweets. Specifically, they compared SO-CAL (Taboada et al., 2011), which is a lexicon-based method that calculates a score for each instance according to the semantic orientation of its words found in a dictionary, and two variants of the NB classifier using the BoW representation. For the SO-CAL method, they adopted the SentiLex (Carvalho & Silva, 2015) lexicon. Additionally, they evaluated different pre-processing techniques in the context of ML-based methods. The experiments were performed using a dataset containing positive and negative Brazilian stock market tweets manually labeled by finance professionals. The results showed that the ML-based method using a Multinomial NB classifier combined with lemmatization and stemming pre-processing techniques outperformed the other methods.

Similarly to Neuenschwander et al. (2014), Moraes et al. (2016) compared lexicon- and ML-based methods but targeted the subjectivity classification task. Subjectivity classification has been widely adopted in sentiment analysis literature to filter out objective sentences that are assumed to imply no opinion or sentiment (Liu, 2020). In Moraes et al. (2016), Moraes et al. tested SentiLex (Carvalho & Silva, 2015) and WordnetAffectBR (Pasqualotti & Vieira, 2008) sentiment lexicons, and SVM and NB algorithms to determine the subjectivity of tweets from a dataset in the domain of technology—the Computer-BR corpus. These tweets were manually labeled according to their polarities, i.e., positive, negative, or neutral. They were considered subjective, those with positive and negative polarities and the remaining ones were objective. Their computational experiments showed that the ML-based method using SVM with the BoW representation achieved the best results.

Belisário et al. (2020) also investigated three methods of different natures for performing subjectivity classification in Brazilian Portuguese texts, including tweets, using the Computer-BR corpus introduced in Moraes et al. (2016). In Belisário et al. (2020), they tested lexicon-, graph-, and ML-based methods. For the lexicon-based method, they used SentiLex (Carvalho & Silva, 2015) and WordnetAffectBR (Pasqualotti & Vieira, 2008) lexicons. Regarding the graph-based method, they adopted a strategy based on word graphs that predicts whether a sentence is objective or subjective based on centrality measures. Lastly, to evaluate the ML-based method, they used two strategies: a traditional BoW representation with NB and SVM classifiers and a multi-layered neural network with word embeddings trained with the Word2Vec's continuous BoWs (CBoWs) model, which achieved the best overall results among all assessed methods.

Although there is a considerable number of studies that use sentiment analysis to detect the sentiment expressed by Brazilian citizens on Twitter, most of them only report its application as a tool in a variety of domains, such as health, politics, sports, social manifestation, consumers' sentiment, financial market, and others. In fact, only the studies presented in Belisário et al. (2020), Garcia and Berton (2021), Moraes et al. (2016), Neuenschwander et al. (2014), and Vargas and Moreira (2020) report a comparison over distinct methods. Nevertheless, most of them focus on determining the best choice between supervised and unsupervised methods for specific domains rather than examining the effect of applying different text representation techniques, especially in ML. Only the studies presented in Garcia and Berton

(2021) and Vargas and Moreira (2020) exploit and evaluate features extracted from different language models, although using only one dataset of tweets. On the other hand, we present a broad evaluation of text representation models, including the static language models fastText (Bojanowski et al., 2017), GloVe (Pennington et al., 2014), and Word2Vec (Mikolov et al., 2013), as well as the most recent Transformer-based models, such as mBERT, BERTweet (Nguyen et al., 2020), and BERTimbau (Souza et al., 2020), for the Brazilian Portuguese language in Twitter sentiment analysis. We conduct the comparative evaluation of such models by using eight datasets of Brazilian Portuguese tweets from various domains and three classification algorithms.

3 Experimental methodology

This section presents the experimental methodology followed in this study. We start by describing the eight Portuguese tweets-based sentiment analysis datasets used to investigate the questions posed in the introduction. Then, we introduce three techniques to continue the pre-training of the contextualized language models and explain how they will be evaluated using three different classifiers: Logistic Regression (LR), Support Vector Machine (SVM), and XGBoost (XGB).

3.1 Datasets

There is a shortage of curated resources for sentiment analysis in Portuguese, even though Portuguese is among the top ten languages used on the Web as of January 2020.[13] Encouraged to build a collection of Portuguese datasets for the task of sentiment analysis, we surveyed the existing literature looking for such data. This process led us to contact authors and retrieve a set of eight datasets varying from COVID-19 domain to automobile-related tweets. From those eight datasets—three binary and five multiclass datasets—two collections were built. In the first one, neutral tweets were removed from the five multiclass datasets to make the task consistent for all the datasets. Statistics for those datasets are described in Table 2. For the second scenario, we isolate the multiclass datasets as described in Table 3. Table 2 shows the total number of tweets (#tweets), the average length of tweets (avg. len), the number of positive tweets (#pos), and finally, the number of negative tweets (#neg) for each dataset. Besides the columns described so far, Table 3 also brings the number of neutral tweets (#neutral) for each of the five multiclass datasets.

The smallest dataset in our collection, called *narr-PT* (Araujo et al., 2016), is part of a set of multilingual datasets composed of human-annotated tweets labeled as neutral, positive, or negative according to their polarity. The *OPCovid-BR* dataset (Vargas et al., 2020) contains 600 tweets collected through the Twitterscraper Python library considering a set of keywords in Portuguese regarding COVID-19.

[13] https://www.statista.com/statistics/262946/share-of-the-most-common-languages-on-the-internet/.

Table 2 Number of examples and the average length of tweets in Portuguese Twitter sentiment datasets with only positive and negative classes (ordered by the number of tweets)

Dataset	#tweets	avg. len	#pos	#neg
narr-PT	510	14.21	297	213
OPCovid-BR	600	27.96	300	300
Computer-BR	601	18.22	197	404
MiningBR	1465	15.61	166	1299
TweetsMG	5746	15.95	3300	2446
TweetSentBR	7769	11.87	4773	2996
UniLex	7912	16.54	3715	4197
FIAT-UFMG	8827	16.68	4437	4390
Total	10,292	–	17,185	16,245

The average length is computed by splitting tweets according to a white space

Table 3 Number of examples and the average length of tweets in Portuguese Twitter sentiment datasets that include the neutral class (ordered by the number of tweets)

Dataset	#tweets	avg. len	#neutral	#pos	#neg
narr-PT	713	15.56	243	297	213
MiningBR	2018	14.9	553	166	1299
Computer-BR	2281	16.8	1677	197	404
TweetsMG	8199	16.15	1453	3300	2446
UniLex	12,655	15.33	4753	3715	4197
Total	25,866	–	8679	7675	8559

The average length is computed by splitting tweets according to a white space

OPCovid-BR was annotated by humans as positive or negative. In Moraes et al. (2016), the authors present a manually labeled (neutral, positive, or negative) corpus of tweets on the area of technology called *Computer-BR*. *MiningBR* (Souza et al., 2016b) contains tweets of the companies with the most number of complaints in the Brazilian Consumer Protection and Defense Program agency (PROCON). At least two researchers manually labeled the collection as neutral, positive, or negative according to their polarity. Although the dataset gathers tweets about problematic companies, it has neutral tweets, for example, "The vice President of the @ BancodoBrasil announces financing agreements with cooperatives and producers to build warehouses @blogplanalto" and also positive tweets, such as "Congratulation to @BancodoBrasil #circuitobancodobrasil for such a high-level event! This bank has history in sports!". The *TweetsMG*,[14] another multiclass dataset, was collected and labeled by the IT staff of Prodemge MG and was used in de Oliveira and de Campos Merschmann (2021). *TweetSentBR* is a corpus of tweets in Brazilian Portuguese introduced in Brum and das (2018). Several annotators labeled each tweet as positive or negative, following an annotation process of eight steps to improve

[14] https://minerandodados.com.br/analise-de-sentimentos-twitter-como-fazer/.

the reliability of the labels. They defined the final sentiment for each tweet based on a major voting of the labels provided by each annotator. While some tweets were only labeled by one annotator, others were annotated by three or seven annotators. In de Souza et al. (2017), the authors introduce a human-annotated set of Portuguese tweets called *UniLex*. Three annotators labeled these tweets as neutral, positive, or negative. *FIAT-UFMG* (Martins et al., 2015) is a dataset composed of tweets related to the "FIAT" brand, manually labeled as positive or negative. Notice that all datasets adopted in this work are composed of human-annotated tweets labeled according to their polarity.

3.2 Evaluation methodology

Task In this work, we are focusing on tweet sentiment classification, i.e., the task of classifying a tweet as either expressing a neutral, positive, or negative sentiment. We follow the general approach of first extracting the tweet's embeddings from a trained model and next those embeddings are used as input features to a classifier. To this end, we run our experiments using three popular and reliable classification algorithms: LR, SVM, and XG, implemented in scikit-learn (Pedregosa et al., 2011)[15]. Formally, we want the classifier to induce a function $f : \boldsymbol{x_i} \rightarrow y$ where $\boldsymbol{x_i} \in \mathbb{R}^d$ is a numerical vector representation (an embedding) of dimension d representing a tweet i, and $y \in \{\textsf{neutral}, \textsf{positive}, \textsf{negative}\}$, or $y \in \{\textsf{positive}, \textsf{negative}\}$, depending on the number of classes in the dataset. The way $\boldsymbol{x_i} \in \mathbb{R}^d$ is assembled depends on the specific embedding model. We follow the default setting for defining the dimensions, making $d = 300$ when the embeddings come from fastText, Word2Vec, and GloVe, and $d = 768$ when the embeddings are extracted from Transformer-based approaches.

Embedding-based text representations Before any classification algorithm can be used, we have to select the best representation for the input words. To this end, we conducted a broad study encompassing six different word embedding methods, namely, fastText, Word2Vec, GloVe, mBERT, BERTweet, and BERTimbau. FastText, Word2Vec and GloVe are static models that provide a fixed representation per word. The remaining models provide contextualized embeddings, which means that a word representation varies depending on its context in the sentence. Following the short description of those models in Sect. 2, we emphasize that the fastText version used in our evaluation was trained on Portuguese Common Crawl and Wikipedia data.[16] For GloVe and Word2Vec we adopted the pre-trained versions available in the repository of word embeddings from NILC.[17] Regarding the Transformer-based language models, we leveraged the Transformers library implemented as part of the Hugging Face framework (Wolf et al., 2020). mBERT is a BERT-based model trained on 104 languages with data from Wikipedia.[18] It may benefit the tweets

[15] https://scikit-learn.org/.

[16] https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.pt.300.bin.gz.

[17] http://www.nilc.icmc.usp.br/embeddings.

[18] https://huggingface.co/bert-base-multilingual-cased.

Table 4 Statistics about the number of tokens and missing vocabulary found by each method in the multiclass datasets

Dataset	Model	min. tokens	max. tokens	avg. tokens	#has OOV	#total OOV	#unique OOV	avg. OOV
narrPT	fastText	2	1109	18,889	28	39	6	0.055
	Word2Vec/ GloVe	2	1080	18,265	512	1417	630	1.760
	mBERT	6	1667	29,899	15	23	7	0.228
	BERTimbau	5	1717	30,112	50	69	17	0.290
	BERTweet	6	1720	30,893	25	26	12	0.194
MiningBR	fastText	3	45	18,889	75	145	6	0.072
	Word2Vec/ GloVe	2	44	16,971	2018	5575	962	2.371
	mBERT	7	56	29,406	0	0	0	0
	BERTimbau	6	83	28,514	156	196	19	0.081
	BERTweet	4	58	31,062	102	103	21	0.051
ComputerBR	fastText	2	62	19,712	130	168	8	0.074
	Word2Vec/ GloVe	1	55	19,131	1847	5853	1151	2.419
	mBERT	6	95	30,952	28	33	7	0.013
	BERTimbau	5	89	33,115	107	134	27	0.050
	BERTweet	5	115	31,166	118	130	19	0.053
TweetsMG	fastText	1	39	18,889	405	607	3	0.074
	Word2Vec/ GloVe	1	38	18,271	8090	19,504	970	2.038
	mBERT	5	58	31,076	101	121	6	0.014
	BERTimbau	6	62	31,201	471	1100	49	0.095
	BERTweet	3	77	33,804	290	295	45	0.035
UniLex	fastText	1	85	20,127	1023	2460	145	0.194
	Word2Vec/ GloVe	1	73	17,143	11,864	48,098	9986	3.512
	mBERT	4	202	33,029	525	1304	108	0.074
	BERTimbau	3	176	32,057	993	1949	206	0.117
	BERTweet	4	204	35,829	1148	3149	459	0.146

sentiment classification due to the variety of its corpus in terms of languages and vocabulary used for training (see Tables 4 and 5 further discussed ahead). BERTweet[19] was chosen for being a language model pre-trained specifically for tweets, even though it was solely trained on English data. BERTweet was trained based on the RoBERTa pre-training procedure, using the same model configuration as BERT. Finally, BERTimbau[20] is a monolingual model trained from scratch to address the

[19] https://huggingface.co/vinai/bertweet-base.

[20] https://huggingface.co/neuralmind/bert-base-portuguese-cased.

Table 5 Statistics about the number of tokens and missing vocabulary found by each method in the binarized datasets

Dataset	Model	min. tokens	max. tokens	avg. tokens	#has OOV	#total OOV	#unique OOV	avg. OOV
narrPT	fastText	2	35	16,896	30	45	5	0.088
	Word2Vec/ GloVe	2	35	16,431	339	642	365	1.186
	mBERT	6	54	26,608	7	11	4	0.021
	BERTimbau	6	59	26,180	46	60	12	0.092
	BERTweet	6	58	28,163	13	13	6	0.025
OPCOvid-BR	fastText	6	84	33,780	23	43	22	0.072
	Word2Vec/ GloVe	6	68	32,842	538	1897	637	2.911
	mBERT	14	146	53,822	46	91	30	0.135
	BERTimbau	13	138	52,268	46	77	36	0.108
	BERTweet	10	141	64,350	81	152	76	0.186
ComputerBR	fastText	3	62	20,900	9	15	4	0.025
	Word2Vec/ GloVe	3	55	20,243	401	831	275	1.202
	mBERT	7	95	30,844	2	2	1	0.003
	BERTimbau	7	89	31,363	51	60	8	0.086
	BERTweet	7	115	33,443	13	21	7	0.021
MiningBR	fastText	3	45	19,619	54	106	4	0.072
	Word2Vec/ GloVe	2	44	17,787	1465	3640	662	2.119
	mBERT	7	55	29,724	0	0	0	0
	BERTimbau	6	54	28,337	124	158	17	0.089
	BERTweet	6	58	32,042	82	83	20	0.056
TweetsMG	fastText	4	36	18,822	142	267	2	0.046
	Word2Vec/ GloVe	4	35	17,974	5723	13,887	446	2.077
	mBERT	8	54	31,153	59	74	5	0.012
	BERTimbau	8	59	31,066	396	1008	39	0.123
	BERTweet	8	57	34,081	212	214	23	0.037
TweetSentBR	fastText	3	48	14,995	319	653	8	0.084
	Word2Vec/ GloVe	2	47	13,601	7769	13,865	1814	1.663
	mBERT	7	97	24,790	8	15	11	0.001
	BERTimbau	6	82	24,257	454	568	43	0.061
	BERTweet	7	138	27,671	173	220	55	0.023
UniLex	fastText	2	71	21,439	704	1624	106	0.205
	Word2Vec/ GloVe	1	69	18,557	7550	28,601	5956	3.317
	mBERT	5	202	34,634	311	743	68	0.070
	BERTimbau	5	176	33,179	675	1146	130	0.115
	BERTweet	5	204	38,176	759	1976	335	0.144

Table 5 (continued)

Dataset	Model	min. tokens	max. tokens	avg. tokens	#has OOV	#total OOV	#unique OOV	avg. OOV
FIAT-UFMG	fastText	2	43	19,720	93	124	15	0.014
	Word2Vec/ GloVe	2	43	19,137	7517	16,089	3074	1.697
	mBERT	7	72	30,384	5	5	3	0.0005
	BERTimbau	7	111	30,911	387	471	72	0.045
	BERTweet	7	73	33,719	298	307	64	0.034

whole-word intermediate task on the BrWaC (Brazilian Web as Corpus), a large Portuguese corpus (Filho et al., 2018).

Figure 1 illustrates our evaluation backbone from the initial preprocessing steps to the final classification. At the bottom of the pipeline, before the tokenization phase, we have a straightforward preprocessing step that removes punctuation, translates emojis into text, converts sentences to lowercase, and replaces user mention and URL links with special tokens (@USER and HTTP, respectively). The tokenization procedure varies with the methods. The Word2Vec and GloVe methods learn static embeddings at the word level. Therefore, we tokenize the tweets into words with the *TweetTokenizer* class of NLTK[21] that takes into account hashtags and emoticons: it separates the symbol '#' from the hashtags and lets the emoticons intact. The other methods tokenize the texts into subwords with their own algorithms. Precisely, fastText tokenizer splits words into subwords based on character N-grams, mBERT and BERTimbau follow the Word Piece tokenization (Schuster and Nakajima, 2012), and BERTweet employs the Byte-Pair encoding (Gage, 1994; Heinzerling & Strube, 2018).

One should notice that splitting words into subwords benefits dealing with tweets, given its informal writing style. Consequently, the number of out-of-vocabulary (OOV) words is much larger in Word2Vec and GloVe than in the other methods. Also, each model has its own vocabulary as they are trained from different corpora. Specifically, the Word2Vec and GloVe models employed here have a vocabulary size of 929,606 words, while fastText is built upon more than 8,000,000 words. The contextualized models also differ as they are trained from distinct corpora and use different tokenization algorithms. While mBERT tokenizer was trained with a vocabulary size of 119,547 words, BERTimbau-base encompasses only 29,794 words, and BERTweet has an even smaller vocabulary of 64,000 words.

Tables 4 and 5 present statistics for each method and dataset regarding the number of tokens after the normalization and tokenization phases (columns min, max, and average tokens) and the number of missing tokens in the method vocabulary

[21] https://www.nltk.org/api/nltk.tokenize.casual.html.

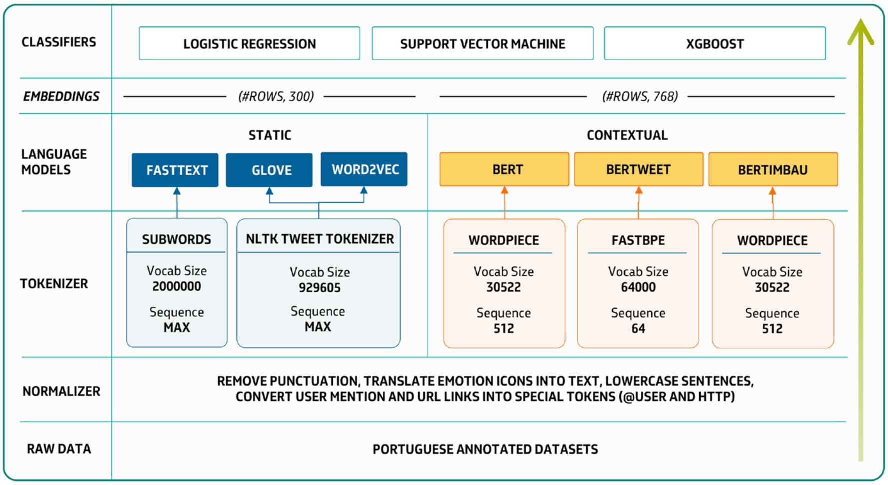

Fig. 1 Evaluation pipeline from data pre-processing to tokenization and classification

(the other columns). In the tables, *#has OOV* is the number of tweets with at least one OOV token, *#total OOV* is the total number of OOV in all the tweets, *avg. OOV* is the average number of OOV per tweet, and *#unique OOV* is the number of OOV without repetition. Notice that the number of tokens of each method is not indicative that they have found a vocabulary closer to the original one. Depending on the pre-trained acquired list of tokens that fastText and the contextualized methods have, they will break a word into more or fewer subwords that may or may not correspond to the original one. Nevertheless, as expected, given its simple tokenization procedure based on white spaces, Word2Vec and GloVe have a more unique and total number of OOV tokens than the other methods consistently. While they have almost six thousand unique OOV in the worst case (UniLex), the others have less than one hundred or a few hundred in this dataset. Also, the average of missing tokens per tweet is always less than one for all the subword-based tokenizers, while Word2Vec and GloVe always have an average of more than one. Although those averages are still small, they may negatively impact Word2Vec and GloVe, as we shall see in the results.

The pre-trained embeddings from fastText, Word2Vec, and GloVe, are the resulting vectors with 300 positions given by the mean of the vector representations of each word of a sentence. On the other hand, the embeddings from the Transformer-based language models are the sum of the last four hidden layers of the embeddings relative to the *[CLS]* token (Devlin et al., 2019). In addition to the models, we explore two strategies for extracting features from the Transformer-based language models: (i) using the pre-trained model to infer the contextualized embedding of a tweet without adjusting any parameter of the pre-trained model; and (ii) first continuing the pre-training of the language model and then extracting the contextualized embeddings for each tweet. This process is repeated for the eight datasets introduced in Sect. 3.1, creating a numerical representation for each one of their tweets.

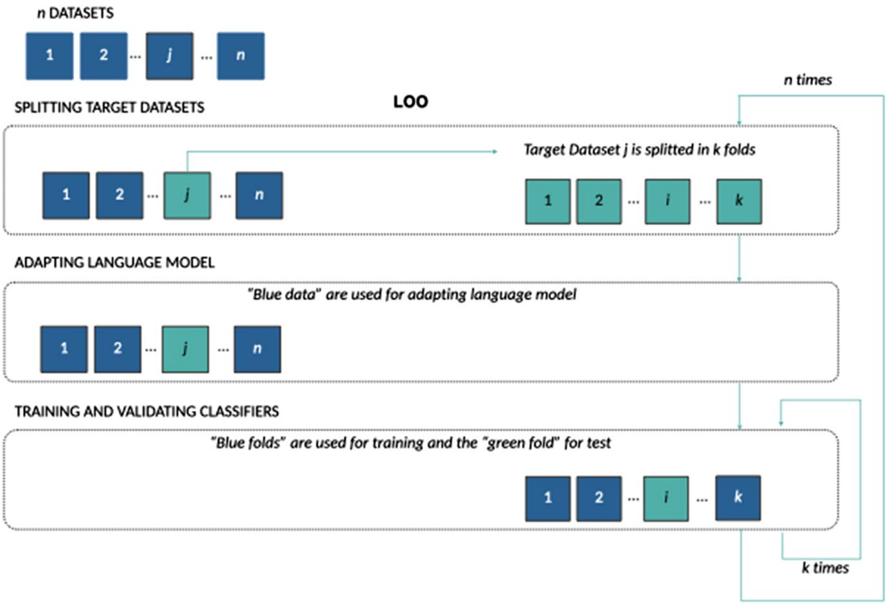
n DATASETS
1
2
j
n
n times
SPLITTING TARGET DATASETS
LOO
Target Dataset j is splitted in k folds
1
2
j
n
1
2
i
k
ADAPTING LANGUAGE MODEL
"Blue data" are used for adapting language model
1
2
j
n
TRAINING AND VALIDATING CLASSIFIERS
"Blue folds" are used for training and the "green fold" for test
1
2
i
k
k times

Fig. 2 Continued pre-training strategy: LOO

Model adjusting approaches For the continued pre-training strategy, embeddings are adjusted using three different approaches: (1) leave-one-out (*LOO*), (2) *inData*, and (3) *allData*. In the LOO approach, all datasets are combined to adjust the language model, except for one of them, the target dataset, that is left out for the classification task, as shown in Fig. 2. For instance, if the goal is to evaluate (classify) the dataset *narr-PT* in the binary scenario, the remaining seven datasets are combined and used to continue the pre-training of the language model. The *LOO* strategy aims to simulate a scenario where an extensive collection of multi-domain datasets is available to adjust the pre-trained language model. After the model adjusting phase, the features related to the target dataset are extracted from the adjusted embeddings and given as input to train the classifier. Following the former example, the embeddings of the dataset *narr-PT* would be collected and a classifier would next be trained from them using a 10-fold cross-validation procedure.

In the *inData* approach, depicted in Fig. 3, the target dataset itself is used to continuing pre-training the language model. In this scenario, each of the eight datasets is used at a time, first to adjust the model and next to create the classifier. This process is done through a 10-fold cross-validation procedure. In this way, in each iteration, tweets from 9-folds of the target dataset are used to adjust the language model and create the classifier, while the remaining fold is used to validate the tuned model. This process repeats for each of the eight datasets on an individual basis.

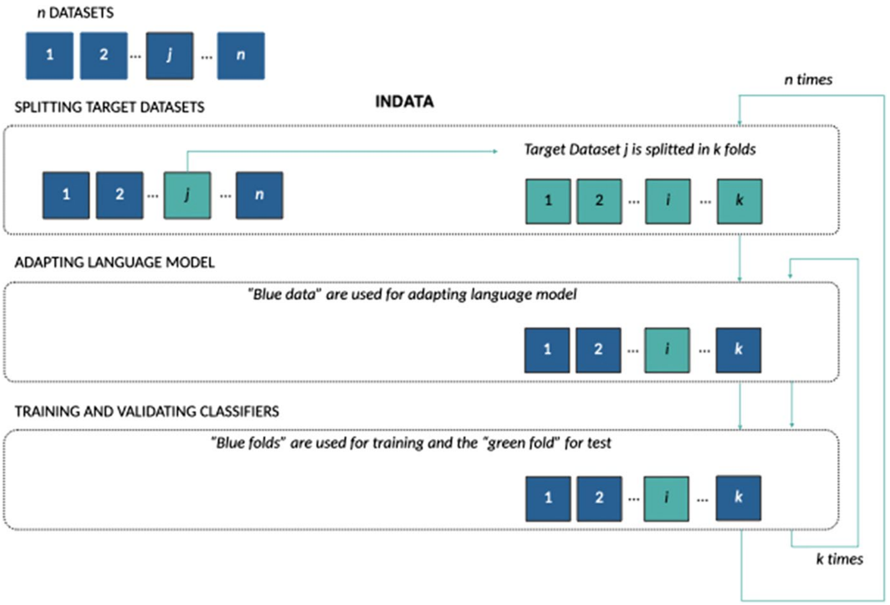
n DATASETS
1
2
j
n
SPLITTING TARGET DATASETS
INDATA
n times
Target Dataset j is splitted in k folds
1
2
j
n
1
2
i
k
ADAPTING LANGUAGE MODEL
"Blue data" are used for adapting language model
1
2
i
k
TRAINING AND VALIDATING CLASSIFIERS
"Blue folds" are used for training and the "green fold" for test
1
2
i
k
k times

Fig. 3 Continued pre-training strategy: inData

Lastly, the *allData* approach (Fig. 4) is a combination of the previous two strategies: as was done in the *inData* strategy, a 10-fold cross-validation procedure is adopted to both adjust the language model and to create the classifier. However, for each dataset and cross-validation iteration, tweets from 9-folds are combined with the remaining seven datasets to continue pre-training the language model, while tweets from the 10th fold are used to validate the resulting model.

In all these cases (LOO, inData, and allData), we set the following hyperparameters: learning rate is 5e−5, the number of epochs is 3, the batch size is 8 and the tokens are randomly masked with a probability of 0.15. All the other hyperparameters are left with the default values of the HuggingFace Transformers Library.

Notice that even though all the datasets contain sentiment labels for each tweet, during the fine-tuning step, none of those labels are taken into consideration since our intention is to leverage the intermediate self-supervised masked language modeling task by adjusting the network weights instead of fine-tuning the model on a downstream task. Note that the same splits for each dataset are used in all the experiments.

In this study, we aim to identify which pair, embeddings type, and data selection strategy, are better suited to the tweets sentiment classification task written in Portuguese. It is worthy to note the lack of a language model specific for *Brazilian Portuguese tweets* in the current literature.

Metrics To measure the sentiment classification predictive performance, we use accuracy (Acc) and F1-score. Classification accuracy is computed as the ratio between the number of tweets correctly classified and the total number of tweets.

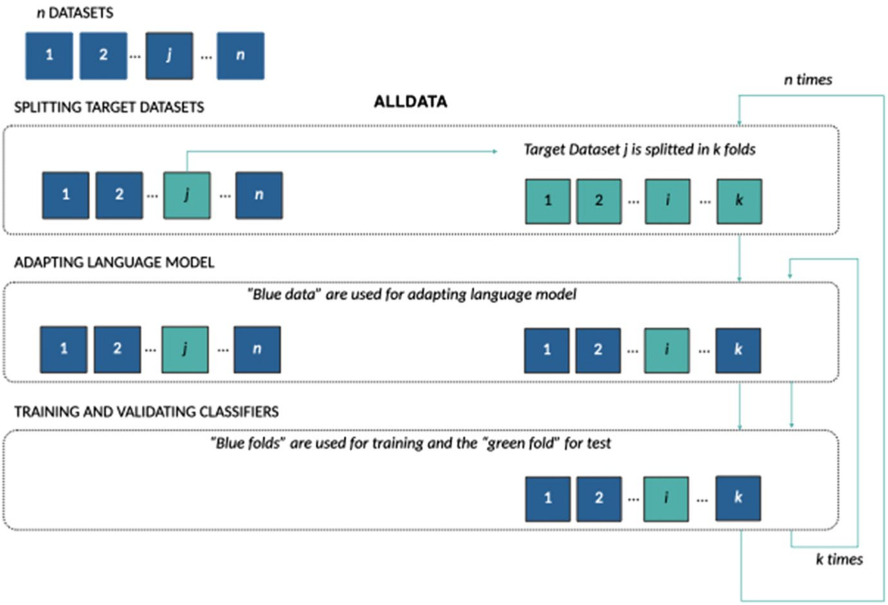
n DATASETS
1
2
j
n
SPLITTING TARGET DATASETS
ALLDATA
n times
Target Dataset j is splitted in k folds
i
k
ADAPTING LANGUAGE MODEL
"Blue data" are used for adapting language model
TRAINING AND VALIDATING CLASSIFIERS
"Blue folds" are used for training and the "green fold" for test
k times

Fig. 4 Continued pre-training strategy: allData

F1-score is the weighted average of the F1-score for the neutral, positive, and negative classes.

A summary of the results for the entire collection of datasets is presented as the number of wins and the sum of ranks. The number of wins (#wins) counts how many times a specific text representation had the best predictive performance for a dataset. The sum of ranks, or #rank sum, is computed considering the position of a text representation in relation to its accuracy (F1-score) value compared to other representations. In our case, the text representation with the highest value for accuracy (F1-score) will be assigned position 1, while the next best value will be in position 2, and so on. If two or more text representations tie for a position in the rank, we assign each of them the same ranking number, which is the mean of what they would have in a regular ordinal ranking.

4 Evaluation of static and Transformer-based embeddings

This section conducts a comparative study between the static embeddings yielded by fastText and contextualized Transformer-based embeddings obtained with mBERT, BERTweet, and BERTimbau. An initial evaluation using three different static embeddings—fastText, Word2Vec, and GloVe—was performed to identify the best static embedding for our scenario. The results can be seen in Appendix with fastText outperforming Word2Vec and GloVe for almost all datasets and classifiers.

Table 6 Accuracy and F1-score for each dataset considering the static and contextualized embeddings for feature extraction and Logistic Regression classifier

Dataset	fastText		mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.843	0.843	0.694	0.693	0.661	0.662	**0.867**	**0.867**
OPCovid-BR	0.770	0.770	0.717	0.716	0.698	0.698	**0.805**	**0.805**
Computer-BR	0.760	0.764	0.690	0.694	0.666	0.671	**0.814**	**0.815**
MiningBR	0.890	0.899	0.835	0.849	0.813	0.834	**0.905**	**0.912**
TweetsMG	0.991	0.991	0.977	0.977	0.980	0.980	**0.992**	**0.992**
TweetSentBR	0.826	0.828	0.735	0.738	0.746	0.749	**0.840**	**0.842**
UniLex	0.778	0.778	0.703	0.703	0.687	0.686	**0.783**	**0.783**
FIAT-UFMG	0.815	0.815	0.775	0.774	0.771	0.771	**0.826**	**0.826**
#wins	0	0	0	0	0	0	**8**	**8**
#rank sums	16.0	16.0	26.0	29.0	30.0	31.0	**8.0**	**8.0**

Neutral tweets were excluded

Table 7 Accuracy and F1-score for each dataset considering the static and contextualized embeddings for feature extraction and the SVM classifier

Dataset	fastText		mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.833	0.833	0.663	0.662	0.643	0.643	**0.853**	**0.852**
OPCovid-BR	0.770	0.769	0.702	0.701	0.662	0.661	**0.783**	**0.783**
ComputerBR	0.725	0.726	0.666	0.663	0.666	0.665	**0.792**	**0.790**
MiningBR	**0.921**	**0.917**	0.873	0.869	0.879	0.871	0.913	0.911
TweetsMG	0.991	0.991	0.981	0.981	0.981	0.981	**0.992**	**0.992**
TweetSentBR	0.826	0.825	0.735	0.733	0.750	0.748	**0.834**	**0.834**
UniLex	0.775	0.775	0.696	0.695	0.689	0.688	**0.778**	**0.777**
FIAT-UFMG	0.814	0.814	0.771	0.771	0.775	0.775	**0.824**	**0.824**
#wins	1	1	0	0	0	0	**7**	**7**
#rank sums	15.0	15.0	28.0	28.5	28.0	27.5	**9.0**	**9.0**

Neutral tweets were excluded

It is important to highlight that, for the contextualized Transformer-based model, the embeddings come from the original pre-trained model, without any adjustments to the intermediate masked language task. By not adjusting the model weights, we save computational resources and evaluate the reliability of pre-computed representations when combined with different classifiers.

Initially, to make the task consistent for all the datasets, we removed tweets annotated as neutral from the multiclass datasets and formulated the task as a binary classification problem. Tables 6, 7, and 8 show Acc and F1-score for LR, SVM, and XGB classifiers, respectively. On both tables, the first eight lines correspond, each one, to a different dataset. For each text representation, we have two columns, one

Table 8 Accuracy and F1-score for each dataset considering the static and contextualized embeddings for feature extraction and the XGB classifier

Dataset	fastText		mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.755	0.751	0.675	0.656	0.624	0.606	**0.808**	**0.805**
OPCovid-BR	0.733	0.732	0.657	0.655	0.657	0.656	**0.748**	**0.748**
ComputerBR	0.719	0.660	0.686	0.626	0.689	0.622	**0.774**	**0.756**
MiningBR	**0.900**	**0.868**	0.886	0.833	0.888	0.837	**0.900**	0.867
TweetsMG	0.963	0.963	0.913	0.912	0.932	0.932	**0.975**	**0.975**
TweetSentBR	0.750	0.746	0.684	0.658	0.697	0.687	**0.778**	**0.775**
UniLex	0.698	0.690	0.648	0.637	0.626	0.608	**0.722**	**0.716**
FIAT-UFMG	0.738	0.736	0.697	0.690	0.707	0.705	**0.778**	**0.777**
#wins	1	0	0	0	0	0	**8**	**7**
#rank sums	15.5	15.0	29.0	29.0	27.0	27.0	**8.5**	**9.0**

Neutral tweets were excluded

for Acc and another one for F1-score (F1). Bold values indicate the best results of Acc and F1-score for a paired dataset and text representation. The last two lines show the overall results for the entire collection of eight datasets, including the number of wins and the sum of ranks, as described in Sect. 3.2. The number of wins (#wins) counts how many times a specific text representation got the best predictive performance for a dataset. For instance, in Table 6, BERTimbau achieved the best Acc and also F1-score for all datasets, so its #wins are equal to 8. The sum of ranks, or #rank sum, is computed considering the position of a text representation in relation to its accuracy (F1-score) value compared to other representations. For instance, in Table 8 for the dataset MiningBR and metric accuracy, fastText, and BERTimbau got the same highest value for accuracy, so in this case, their rank will be $(1 + 2)/2 = 1.5$. In the end, we add up all the ranks for each column to get the final sum of ranks (#rank sum) for each text representation. The smaller the sum of ranks, the better the performance of a strategy.

When the classifier is LR, as shown in Table 6, BERTimbau achieves the best results for Acc and F1-score, followed closely by fastText. Both models were pre-trained with multi-domain Portuguese-written data, which indicates that in the absence of fine-tuning, choosing a model trained solely on the same language as the target dataset is the best approach.

Considering the time and computational resources in our decision-making process, we can safely advocate for the static model fastText, as its results were very close to the heavier Transformer-based model BERTimbau. This conclusion is also confirmed when using the SVM classifier (Table 7), with fastText outperforming BERTimbau for the dataset MiningBR, considering both Acc and F1-score.

The advantage of BERTimbau over fastText increases for the smaller datasets, with less than 1000 tweets, as can be observed for narr-PT, OPCovid-BR, and Computer-BR. This result indicates that BERTimbau combined with LR or SVM is the best candidate approach when working with very small datasets. With larger

Table 9 Accuracy and F1-score for each dataset considering the static and contextualized embeddings for feature extraction and the LR classifier

Dataset	fastText		mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.791	0.791	0.648	0.647	0.633	0.630	**0.815**	**0.815**
MiningBR	0.778	0.776	0.706	0.706	0.698	0.696	**0.782**	**0.784**
Computer-BR	0.772	0.781	0.747	0.752	0.750	0.759	**0.817**	**0.821**
TweetsMG	0.945	0.945	0.927	0.927	0.934	0.934	**0.954**	**0.954**
UniLex	0.662	0.661	0.604	0.602	0.592	0.589	**0.670**	**0.668**
#wins	0	0	0	0	0	0	**5**	**5**
#rank sums	10.0	10.0	17.0	17.0	18.0	18.0	**5.0**	**5.0**

Neutral tweets are also considered

Table 10 Accuracy and F1-score for each dataset considering the static and contextualized embeddings for feature extraction and the SVM classifier

Dataset	fastText		mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
narrPT	0.769	0.769	0.600	0.597	0.610	0.608	**0.793**	**0.792**
MiningBR	**0.777**	0.768	0.698	0.692	0.723	0.709	0.776	**0.773**
ComputerBR	0.792	0.782	0.751	0.745	0.774	0.765	**0.808**	**0.804**
TweetsMG	0.951	0.951	0.931	0.931	0.941	0.941	**0.958**	**0.958**
UniLex	0.661	0.658	0.601	0.596	0.593	0.587	**0.666**	**0.662**
#wins	1	0	0	0	0	0	**4**	**5**
#rank sums	9.0	10.0	19.0	19.0	16.0	16.0	**6.0**	**5.0**

Neutral tweets are also considered

datasets, when the extra cost of adjusting a language model is not a possibility, fastText surfaces as an efficient option with a lower cost than BERTimbau. Regarding the XGB classifier (Table 8), BERTimbau outperforms fastText for seven of the eight datasets. The only exception was for the MiningBR dataset, with BERTimbau and fastText achieving the same accuracy and fastText surpassing BERTimbau for F1-score. The XGB behavior is less predictable than LR and SVM, and we could not find a relationship between dataset size and the performance of the best model, BERTimbau.

When looking into precision and recall, again, fastText and BERTimbau had very competitive results. Considering the number of true positives (tp), true negatives (tn), false positives (fp), and false negatives (fn), the only result that stood out was for the classifier LR with fastText embedding when applied to the imbalanced dataset MiningBR. In this case, the LR classifier struggled to classify positive tweets, being that the minority class. BERTimbau performed slightly better in the same situation.

Table 11 Accuracy and F1-score for each dataset considering the static and contextualized embeddings for feature extraction and the XGB classifier

Dataset	fastText		mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.715	0.716	0.617	0.604	0.606	0.595	**0.718**	**0.718**
MiningBR	0.743	0.695	0.715	0.654	0.721	0.669	**0.747**	**0.703**
Computer-BR	0.741	0.649	0.734	0.637	0.755	0.688	**0.773**	**0.712**
TweetsMG	0.879	0.879	0.805	0.803	0.854	0.854	**0.897**	**0.897**
UniLex	0.574	0.556	0.532	0.505	0.519	0.481	**0.590**	**0.564**
#wins	0	0	0	0	0	0	**5**	**5**
#rank sums	11.0	11.0	18.0	18.0	16.0	16.0	**5.0**	**5.0**

Neutral tweets are also considered

After fastText and BERTimbau, mBERT slightly outperformed BERTweet for the majority of the datasets and classifier LR. In this case, we can see that a multilingual model also trained with multi-domain Brazilian Portuguese data can surpass a model specifically pre-trained to deal with the peculiarities of tweets but in another language, even for datasets full of emoticons e emojis, such as *OPCovid-BR*, *TweetsMG*, and *FIAT-UFMG*. For classifiers SVM, both mBERT and BERTweet presented similar results. However, for classifier XGB, BERTweet slightly outperformed mBERT.

Overall, the classifiers LR and SVM performed very stably across all datasets, with BERTimbau posing as the best embedding, followed closely by fastText.

Considering that five of the eight datasets also contain neutral tweets, besides the positive and negative tweets addressed so far, we proceed to evaluate the effect of different embeddings and classifiers in the multiclass scenario. Tables 9, 10, and 11, bring the results of Acc and F1-score for five multiclass datasets, four different embeddings—fastText, mBERT, BERTweet, and BERTimbau—and the classifiers, LR, SVM, and XGB, respectively. As was observed in the binary evaluation, BERTimbau outperforms all the other embeddings for all three classifiers, LR, SVM, and XGB. In general, fastText keeps its position as second best embedding, followed by BERTweet and mBERT. The exception is the LR classifier, with BERT slightly outperforming BERTweet for rank sums. Overall, LR and SVM present the best results for all datasets when compared against XGB.

When looking into the numbers of true/false positives and true/false negatives for the three classes, neutral, positive, and negative, we see that, in general, BERTimbau achieves better results than fastText. For instance, the classifier LR combined with embedding fastText had difficulty classifying the negative and positive tweets from the imbalanced dataset Computer-BR. Note that 73% of this entire dataset is composed of neutral tweets. For the same dataset and classifier, BERTimbau dealt well with the negative class and performed slightly better than fastText for the positive class. Those results, combined with what we have seen for Acc and F1-score, indicate that the combination of classifiers LR and SVM with embedding BERTimbau is more robust than the other combinations of classifiers and embeddings when dealing with imbalanced datasets.

Comparing the three classifiers, LR, SVM, and XGB, with the best embedding model, BERTimbau, we noticed that LR is clearly the best classifier for the binary and multiclass sets of datasets.

5 Continued pre-training of Transformer-based models

This section evaluates the effectiveness of adjusting a language model by continuing its pre-training with a corpus focused on the task at hand. The evaluation includes relying on only the tweets of the specific target dataset or combining multi-domain unlabeled datasets. With that, we can also observe the impact of the number of examples when adjusting the weights of the language model with a continued pre-training strategy. Notably, three different approaches (Sect. 3.2) will be explored: leave-one-out (*LOO*), *inData*, and *allData*. In short, our goal is to study whether and how different continued pre-training strategies help improve the predictive performance of a variety of language models, from the more general mBERT to a more specific BERTweet trained for English tweets concerning the sentiment classification task of Portuguese tweets. Following the same methodology adopted in Sect. 4, for the first set of results, we formulate the sentiment analysis task as a binary classification problem by removing tweets annotated as neutral (Table 2). This will guarantee that the task is consistent for all datasets. Then, we proceed by analyzing the five multiclass datasets (Table 3) respecting their original class distribution, which includes the neutral class (neutral tweets).

5.1 Continued pre-training of the Transformer-based models: the leave-one-out strategy

Given the time and resources required to continue pre-training a language model, our first step is to analyze the performance of extracting the features to learn a classifier from a specific dataset, from a language model previously adjusted with other datasets related to the same task irrespective of their domain. As mentioned in Sect. 3.2, the leave-one-out (*LOO*) approach combines $n-1$ datasets from a set of n datasets to be used for continuing the pre-training, while the remaining dataset will be used to validate the model during the classification step. In our case, considering the set of datasets presented in Sect. 3.1 (Table 2), we have a total of eight datasets, so for each target dataset, the remaining seven datasets will be used for adjusting the weights of the language model. The predictive Acc and F1-score (F1) of a pair of target dataset and language model are presented in Tables 12, 13, and 14, for the classifiers LR, SVM, and XGB, respectively.

We observe from Tables 12, 13, and 14 that for all classifiers, LR, SVM, and XGB, the language model BERTimbau outperformed mBERT and BERTweet. From all datasets, the smallest ones, namely *narr-PT*, *OPCovid-BR*, and *ComputerBR*, are the ones that took more advantage of the combination of datasets

Table 12 Accuracy and F1-score for each dataset considering the combination of continued pre-training the language models and the Logistic Regression classifier: leave-one-out

Dataset	mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.684	0.683	0.696	0.697	**0.861**	**0.861**
OPCovid-BR	0.712	0.711	0.688	0.687	**0.798**	**0.798**
Computer-BR	0.642	0.648	0.692	0.694	**0.814**	**0.814**
MiningBR	0.835	0.848	0.833	0.850	**0.923**	**0.927**
TweetsMG	0.977	0.977	0.984	0.984	**0.991**	**0.991**
TweetSentBR	0.735	0.738	0.751	0.753	**0.832**	**0.833**
UniLex	0.742	0.742	0.688	0.688	**0.777**	**0.776**
FIAT-UFMG	0.789	0.789	0.781	0.781	**0.824**	**0.824**
#wins	0	0	0	0	**8**	**8**
#rank sums	20.0	21.0	20.0	19.0	**8.0**	**8.0**

Neutral tweets were excluded

Table 13 Accuracy and F1-score for each dataset considering the combination of continued pre-training the language models and the Support Vector Machine classifier: leave-one-out

Dataset	mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.663	0.662	0.643	0.643	**0.853**	**0.852**
OPCovid-BR	0.702	0.701	0.662	0.661	**0.783**	**0.783**
Computer-BR	0.666	0.663	0.666	0.665	**0.792**	**0.790**
MiningBR	0.873	0.869	0.879	0.871	**0.913**	**0.911**
TweetsMG	0.981	0.981	0.981	0.981	**0.992**	**0.992**
TweetSentBR	0.735	0.733	0.750	0.748	**0.834**	**0.834**
UniLex	0.696	0.695	0.689	0.688	**0.778**	**0.777**
FIAT-UFMG	0.771	0.771	0.775	0.775	**0.824**	**0.824**
#wins	0	0	0	0	8	8
#rank sums	20.0	20.5	20.0	19.5	8.0	8.0

Neutral tweets were excluded

Table 14 Accuracy and F1-score for each dataset considering the combination of continued pre-training the language models and the XGBoost classifier: leave-one-out

Dataset	mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.622	0.590	0.657	0.640	**0.804**	**0.800**
OPCovid-BR	0.662	0.661	0.675	0.674	**0.732**	**0.731**
ComputerBR	0.676	0.591	0.716	0.661	**0.794**	**0.777**
MiningBR	0.887	0.834	0.885	0.838	**0.908**	**0.882**
TweetsMG	0.923	0.922	0.941	0.941	**0.967**	**0.967**
TweetSentBR	0.662	0.605	0.695	0.686	**0.768**	**0.762**
UniLex	0.679	0.669	0.630	0.612	**0.713**	**0.706**
FIAT-UFMG	0.740	0.739	0.707	0.706	**0.773**	**0.772**
#wins	0	0	0	0	**8**	**8**
#rank sums	21.0	22.0	19.0	18.0	**8.0**	**8.0**

Neutral tweets were excluded

Table 15 Accuracy and F1-score for each dataset considering the combination of continued pre-training the language models and the LR classifier: leave-one-out

Dataset	mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.644	0.642	0.658	0.657	**0.801**	**0.800**
MiningBR	0.708	0.703	0.711	0.709	**0.794**	**0.794**
Computer-BR	0.730	0.738	0.759	0.768	**0.814**	**0.819**
TweetsMG	0.915	0.915	0.938	0.938	**0.952**	**0.952**
UniLex	0.640	0.637	0.597	0.594	**0.659**	**0.657**
#wins	0	0	0	0	**5**	**5**
#rank sums	14.0	14.0	11.0	11.0	**5.0**	**5.0**

Neutral tweets are also considered

Table 16 Accuracy and F1-score for each dataset considering the combination of continued pre-training the language models and the SVM classifier: leave-one-out

Dataset	mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1
narrPT	0.641	0.639	0.644	0.644	**0.774**	**0.773**
MiningBR	0.697	0.687	0.726	0.713	**0.773**	**0.769**
ComputerBR	0.743	0.735	0.768	0.758	**0.818**	**0.813**
TweetsMG	0.929	0.928	0.946	0.946	**0.956**	**0.955**
UniLex	0.642	0.638	0.595	0.590	**0.655**	**0.650**
#wins	0	0	0	0	**5**	**5**
#rank sums	14.0	14.0	11.0	11.0	**5.0**	**5.0**

Neutral tweets are also considered

during the *LOO* fine-tuning process. It indicates that in scenarios where the target dataset is too small, one can leverage a language model adjusted to examples of the same target task, by combining other available datasets, even when they are from different domains. The exceptions were for the dataset *ComputerBR* and classifier XGB. In this case, compared against mBERT and BERTweet,

Table 17 Accuracy and F1-score for each dataset considering the combination of continued pre-training the language models and the XGB classifier: leave-one-out

Dataset	mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.577	0.562	0.614	0.605	**0.738**	**0.737**
MiningBR	0.707	0.633	0.723	0.666	**0.741**	**0.689**
Computer-BR	0.746	0.649	0.746	0.672	**0.775**	**0.717**
TweetsMG	0.811	0.808	0.844	0.843	**0.891**	**0.891**
UniLex	0.561	0.538	0.516	0.469	**0.587**	**0.559**
#wins	0	0	0	0	**5**	**5**
#rank sums	13.5	14.0	11.5	11.0	**5.0**	**5.0**

Neutral tweets are also considered

Table 18 The F1-score for each dataset considering the model BERTimbau, strategies contextualized and LOO, and classifiers LR, SVM, and XGB

Dataset	Contextualized		LOO	
	F1	Classifier	F1	Classifier
narr-PT	**0.867**	LR	0.861	LR
OPCovid-BR	**0.805**	LR	0.798	LR
Computer-BR	**0.815**	LR	0.814	LR
MiningBR	0.912	LR	**0.927**	LR
TweetsMG	**0.992**	LR/SVM	**0.992**	SVM
TweetSentBR	**0.842**	LR	0.834	SVM
UniLex	**0.783**	LR	0.777	SVM
FIAT-UFMG	**0.826**	LR	0.824	LR/SVM

Neutral tweets were excluded

Table 19 The F1-score for each dataset considering the model BERTimbau, strategies contextualized and LOO, and classifiers LR, SVM, and XGB

Dataset	Contextualized		LOO	
	F1	Classifier	F1	Classifier
narrPT	**0.815**	LR	0.800	LR
MiningBR	0.784	LR	**0.794**	LR
ComputerBR	**0.821**	LR	0.819	LR
TweetsMG	**0.958**	SVM	0.955	SVM
UniLex	**0.668**	LR	0.657	LR

Neutral tweets are also considered

BERTimbau gain was less prominent. Between those three smallest datasets, *ComputerBR* is the only imbalanced dataset.

Tables 15, 16, and 17 bring the results of Acc and F1-score for the five multiclass datasets, three different embeddings, mBERT, BERTweet, and BERTimbau, and three classifiers, LR, SVM, and XGB, respectively. Similarly to the LOO strategy when applied to the binary datasets, BERTimbau considerably outperformed mBERT and BERTweet for all datasets and classifiers. *UniLex*, a slight imbalance dataset that is also the largest in this collection presented the lowest value for precision and recall across all embeddings and classifier LR, around 0.64 for both precision and recall. We also noticed that those values degrade from BERTimbau to mBERT and then BERTweet. Comparing precision against recall, we did not observe any huge gaps between them, indicating that both precision and recall contribute equally to the final F1 score.

Table 18 compares the *LOO* strategy against the purely contextualized approach from Sect. 4, when all datasets are considered as binary datasets. We observed that for seven of the eight datasets, the contextualized approach outmatched or tied with the *LOO* approach. The exception was for the dataset *MiningBR*, a heavily imbalanced dataset with 89% of negative tweets. In this case, *LOO* outperformed the contextualized approach, obtaining a smaller number of false positives. A similar trend can be observed in Table 19 that shows the best results for BERTimbau, classifiers

Table 20 Accuracy and F1-score for each dataset considering the combination of continued pre-training the language models and the LR classifier: inData

Dataset	mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.694	0.695	0.677	0.676	**0.867**	**0.867**
OPCovid-BR	0.735	0.735	0.672	0.671	**0.785**	**0.784**
Computer-BR	0.681	0.685	0.709	0.712	**0.782**	**0.783**
MiningBR	0.813	0.834	0.821	0.841	**0.915**	**0.922**
TweetsMG	0.974	0.974	0.980	0.980	**0.991**	**0.991**
TweetSentBR	0.762	0.764	0.751	0.754	**0.836**	**0.837**
UniLex	0.725	0.724	0.685	0.684	**0.773**	**0.772**
FIAT-UFMG	0.780	0.780	0.774	0.774	**0.830**	**0.830**
#wins	0	0	0	0	**8**	**8**
#rank sums	19.0	19.0	21.0	21.0	**8.0**	**8.0**

Neutral tweets were excluded

Table 21 Accuracy and F1-score for each dataset considering the combination of continued pre-training the language models and the SVM classifier: inData

Dataset	mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.680	0.681	0.655	0.652	**0.835**	**0.835**
OPCovid-BR	0.713	0.713	0.655	0.655	**0.750**	**0.749**
Computer-BR	0.684	0.682	0.712	0.712	**0.795**	**0.793**
MiningBR	0.869	0.860	0.877	0.871	**0.918**	**0.919**
TweetsMG	0.977	0.977	0.985	0.985	**0.992**	**0.992**
TweetSentBR	0.763	0.762	0.756	0.755	**0.833**	**0.833**
UniLex	0.724	0.723	0.687	0.685	**0.773**	**0.772**
FIAT-UFMG	0.779	0.779	0.781	0.781	**0.831**	**0.831**
#wins	0	0	0	0	**8**	**8**
#rank sums	20.0	20.0	20.0	20.0	**8.0**	**8.0**

Neutral tweets were excluded

LR, SVM, and XGB, and the multiclass set of datasets. Again, the contextualized approach surpasses the best results from *LOO*, except for the *MiningBR* dataset, the heavily imbalanced dataset with twice more negative tweets than neutrals, and almost ten times more negative tweets than positives. With that, we assume that overall, a *LOO* strategy that combines various datasets from different domains to continue the pre-training of a language model does not incur any gain to the final classification task, unless the dataset is heavily imbalanced, then the *LOO* strategy is potentially a better solution.

Table 22 Accuracy and F1-score for each dataset considering the combination of continued pre-training the language models and the XGBoost classifier: inData

Dataset	BERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.661	0.636	0.667	0.648	**0.786**	**0.783**
OPCovid-BR	0.700	0.698	0.665	0.664	**0.733**	**0.733**
Computer-BR	0.694	0.628	0.681	0.620	**0.787**	**0.760**
MiningBR	0.889	0.842	0.891	0.843	**0.898**	**0.866**
TweetsMG	0.922	0.922	0.946	0.945	**0.978**	**0.978**
TweetSentBR	0.700	0.679	0.693	0.683	**0.780**	**0.777**
UniLex	0.669	0.657	0.631	0.612	**0.717**	**0.708**
FIAT-UFMG	0.708	0.705	0.716	0.713	**0.767**	**0.767**
#wins	0	0	0	0	**8**	**8**
#rank sums	20.0	21.0	20.0	19.0	**8.0**	**8.0**

Neutral tweets were excluded

Table 23 Accuracy and F1-score for each dataset considering the combination of continued pre-training the language models and the LR classifier: inData

Dataset	mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.667	0.668	0.641	0.641	**0.806**	**0.806**
MiningBR	0.716	0.714	0.710	0.710	**0.790**	**0.791**
Computer-BR	0.743	0.753	0.751	0.760	**0.814**	**0.821**
TweetsMG	0.926	0.926	0.938	0.938	**0.949**	**0.949**
UniLex	0.611	0.608	0.599	0.597	**0.668**	**0.665**
#wins	0	0	0	0	**5**	**5**
#rank sums	12.0	12.0	13.0	13.0	**5.0**	**5.0**

Neutral tweets are also considered

Table 24 Accuracy and F1-score for each dataset considering the combination of continued pre-training the language models and the SVM classifier: inData

Dataset	mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.651	0.648	0.609	0.609	**0.767**	**0.768**
MiningBR	0.714	0.700	0.721	0.710	**0.780**	**0.776**
Computer-BR	0.744	0.736	0.762	0.749	**0.819**	**0.814**
TweetsMG	0.935	0.935	0.944	0.944	**0.956**	**0.956**
UniLex	0.609	0.602	0.600	0.595	**0.661**	**0.657**
#wins	0	0	0	0	**5**	**5**
#rank sums	13.0	13.0	12.0	12.0	**5.0**	**5.0**

Neutral tweets are also considered

Table 25 Accuracy and F1-score for each dataset considering the combination of continued pre-training the language models and the XGB classifier: inData

Dataset	mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.631	0.626	0.609	0.598	**0.747**	**0.746**
MiningBR	0.712	0.653	0.727	0.671	**0.748**	**0.702**
Computer-BR	0.744	0.661	0.747	0.669	**0.777**	**0.721**
TweetsMG	0.834	0.833	0.854	0.853	**0.894**	**0.893**
UniLex	0.542	0.523	0.531	0.487	**0.584**	**0.560**
#wins	0	0	0	0	**5**	**5**
#rank sums	13.0	13.0	12.0	12.0	**5.0**	**5.0**

Neutral tweets are also considered

5.2 Continued pre-training of the Transformer-based models: the inData strategy

Continuing our analysis, we move our focus towards a more traditional approach of adjusting the weights of a model with the target dataset itself, which we call in this work the *inData* approach. The main idea is to use a specific sentiment dataset to continue the pre-training of a language model and also to extract their features to build the resulting classifier. As explained in Sect. 3.2, a 10-fold cross-validation procedure is adopted. In each iteration, data from 9-folds are used to adjust the language model and collect the embeddings to train the classifier, while the remaining fold is used to validate the model.

Tables 20, 21, and 22 present the Acc and F1-score for each binary dataset and language model, all using the *inData* approach, and classifiers LR, SVM, and XGB, respectively. As we have observed for the *LOO* approach, the adjusted BERTimbau embeddings outmatched mBERT and BERTweet for all eight datasets. The same was observed with the collection of multiclass datasets as presented in Tables 23, 24, and 25.

For both collections, binary and multiclass, the smallest datasets, *narr-PT* and *Computer-BR* are the ones that benefit the most from BERTimbau when used together with the strategy inData, while large datasets, such as *UniLex* and *FIAT-UFMG*, present very small gains. However, *OPCovid-BR*, the second smallest dataset in the binary collection, did not benefit as much when compared to the datasets *narr-PT* and *Computer-BR*. This result can be explained by the fact that *OPCovid-BR* is a perfectly balanced dataset and, in this case, the text representation adopted had a lesser impact in the final classification.

Table 26 presents the best F1-score for each binary dataset and each of the three strategies, contextualized (Sect. 4), *LOO* (Sect. 5.1), and inData. The *LOO* approach outperformed or matched the results obtained with the inData approach for almost all binary datasets, except for the smallest and largest dataset, *narr-PT* and *FIAT-UFMG* respectively. However, when comparing the three approaches, contextualized, *LOO*, and inData, we observed that the contextualized approach usually outperforms or matches the other two approaches, except for the largest binary dataset, *FIAT-UFMG*. The same analysis is done in Table 27 for the multiclass collection of

Table 26 The F1-score for each dataset considering the model BERTimbau, strategies Contextualized, LOO, and inData, and classifiers LR, SVM, and XGB

Dataset	Contextualized		LOO		inData	
	F1	Classifier	F1	Classifier	F1	Classifier
narr-PT	**0.867**	LR	0.861	LR	**0.867**	LR
OPCovid-BR	**0.805**	LR	0.798	LR	0.784	LR
Computer-BR	**0.815**	LR	0.814	LR	0.793	SVM
MiningBR	0.912	LR	**0.927**	LR	0.922	LR
TweetsMG	**0.992**	LR/SVM	**0.992**	SVM	**0.992**	SVM
TweetSentBR	**0.842**	LR	0.834	SVM	0.837	LR
UniLex	**0.783**	LR	0.777	SVM	0.772	LR/SVM
FIAT-UFMG	0.826	LR	0.824	LR/SVM	**0.831**	SVM

Neutral tweets were excluded

Table 27 The F1-score for each dataset considering the model BERTimbau, strategies contextualized, LOO, and inData, and classifiers LR, SVM, and XGB

Dataset	Contextualized		LOO		inData	
	F1	Classifier	F1	Classifier	F1	Classifier
narr-PT	**0.815**	LR	0.800	LR	0.806	LR
MiningBR	0.784	LR	**0.794**	LR	0.791	LR
Computer-BR	**0.821**	LR	0.819	LR	**0.821**	LR
TweetsMG	**0.958**	SVM	0.955	SVM	0.956	SVM
UniLex	**0.668**	LR	0.657	LR	0.665	LR

Neutral tweets are also considered

datasets. Again, the contextualized approach was superior to the other approaches for most of the datasets, with the exception of the very imbalanced dataset *MiningBR*, which performed better with *LOO*. The conclusion so far is that a purely contextualized approach, when the language model is trained exclusively using the target Portuguese language, performs well enough that we can disregard any adjustments to the intermediated language task. By not adjusting the model weights, we save computational resources.

5.3 Continued pre-training of the Transformer-based models: the allData strategy

This section evaluates the continued pre-training approach *allData*. As we have described in Sect. 3, the *allData* strategy is a combination of the two previous approaches: *LOO* and *inData*, evaluated in Sects. 5.1 and 5.2, respectively. The goal is to investigate whether a more robust dataset (as with *LOO*) together with a more

Table 28 Accuracy and F1-score for each dataset considering the combination of continued pre-training the language models and the Logistic Regression classifier: allData

Dataset	mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.706	0.701	0.729	0.730	**0.845**	**0.844**
OPCovid-BR	0.675	0.673	0.645	0.643	**0.768**	**0.766**
Computer-BR	0.714	0.695	0.720	0.705	**0.827**	**0.820**
MiningBR	0.777	0.809	0.828	0.848	**0.900**	**0.904**
TweetsMG	0.948	0.948	0.956	0.956	**0.973**	**0.973**
TweetSentBR	0.750	0.748	0.743	0.742	**0.812**	**0.812**
UniLex	0.675	0.668	0.665	0.658	**0.748**	**0.745**
FIAT-UFMG	0.744	0.744	0.755	0.754	**0.792**	**0.792**
#wins	0	0	0	0	**8**	**8**
#rank sums	21.0	21.0	19.0	19.0	**8.0**	**8.0**

Neutral tweets were excluded

Table 29 Accuracy and F1-score for each dataset considering the combination of continued pre-training the language models and the SVM classifier: allData

Dataset	mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.702	0.696	0.726	0.725	**0.843**	**0.842**
OPCovid-BR	0.675	0.674	0.648	0.647	**0.752**	**0.749**
Computer-BR	0.704	0.689	0.712	0.699	**0.820**	**0.815**
MiningBR	0.766	0.801	0.832	0.851	**0.900**	**0.904**
TweetsMG	0.953	0.953	0.961	0.961	**0.975**	**0.975**
TweetSentBR	0.748	0.746	0.747	0.746	**0.812**	**0.812**
UniLex	0.673	0.667	0.669	0.663	**0.748**	**0.745**
FIAT-UFMG	0.745	0.745	0.761	0.760	**0.792**	**0.792**
#wins	0	0	0	0	**8**	**8**
#rank sums	21.0	21.5	19.0	18.5	**8.0**	**8.0**

Neutral tweets were excluded

Table 30 Accuracy and F1-score for each dataset considering the combination of continued pre-training the language models and the XGB classifier: allData

Dataset	mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.633	0.634	0.649	0.649	**0.765**	**0.765**
OPCovid-BR	0.625	0.624	0.597	0.582	**0.653**	**0.651**
Computer-BR	0.596	0.593	0.636	0.614	**0.717**	**0.684**
MiningBR	0.637	0.701	0.655	0.716	**0.874**	**0.876**
TweetsMG	0.897	0.897	0.898	0.897	**0.928**	**0.928**
TweetSentBR	0.660	0.641	0.671	0.667	**0.723**	**0.713**
UniLex	0.617	0.601	0.611	0.598	**0.677**	**0.657**
FIAT-UFMG	0.666	0.665	0.682	0.679	**0.698**	**0.696**
#wins	0	0	0	0	**8**	**8**
#rank sums	22.0	21.5	18.0	18.5	**8.0**	**8.0**

Neutral tweets were excluded

Table 31 Accuracy and F1-score for each dataset considering the combination of continued pre-training the language models and the LR classifier: allData

Dataset	mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.599	0.587	0.621	0.609	**0.727**	**0.724**
MiningBR	0.628	0.643	0.672	0.680	**0.765**	**0.763**
Computer-BR	0.748	0.732	0.740	0.729	**0.800**	**0.793**
TweetsMG	0.874	0.873	0.853	0.852	**0.910**	**0.909**
UniLex	0.578	0.573	0.571	0.568	**0.637**	**0.633**
#wins	0	0	0	0	**5**	**5**
#rank sums	12.0	12.0	13.0	13.0	**5.0**	**5.0**

Neutral tweets are also considered

Table 32 Accuracy and F1-score for each dataset considering the combination of continued pre-training the language models and the SVM classifier: allData

Dataset	mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1
narrPT	0.590	0.576	0.610	0.598	0.714	0.710
MiningBR	0.631	0.643	0.690	0.694	0.764	0.762
ComputerBR	0.743	0.729	0.746	0.733	0.811	0.805
TweetsMG	0.884	0.883	0.864	0.862	0.918	0.917
UniLex	0.584	0.576	0.577	0.570	0.635	0.630
#wins	0	0	0	0	5	5
#rank sums	13.0	13.0	12.0	12.0	5.0	5.0

Neutral tweets are also considered

Table 33 Accuracy and F1-score for each dataset considering the combination of continued pre-training the language models and the XGB classifier: allData

Dataset	mBERT		BERTweet		BERTimbau	
	Acc	F1	Acc	F1	Acc	F1
narr-PT	0.404	0.308	0.447	0.354	**0.455**	**0.375**
MiningBR	0.513	0.510	0.625	0.620	**0.686**	**0.665**
Computer-BR	0.703	0.657	0.707	**0.689**	**0.744**	0.682
TweetsMG	0.784	0.782	0.759	0.754	**0.833**	**0.831**
UniLex	0.499	0.426	0.495	0.445	**0.526**	**0.459**
#wins	0	0	0	1	**5**	**4**
#rank sums	13.0	14.0	12.0	10.0	**5.0**	**6.0**

Neutral tweets are also considered

domain-specific set of tweets (as with *inData*) can improve the sentiment classification predictive performance.

As we have observed so far, the language model BERTimbau again outperformed the mBERT and BERTweet for almost all datasets, classifiers, and evaluation metrics. The only exception was for dataset *Computer-BR*, classifier XGB, and metric

Table 34 The F1-score for each dataset considering the model BERTimbau, strategies contextualized, LOO, inData, and allData, and classifiers LR, SVM, and XGB

Dataset	Contextualized		LOO		inData		allData	
	F1	Classifier	F1	Classifier	F1	Classifier	F1	Classifier
narr-PT	**0.867**	LR	0.861	LR	**0.867**	LR	0.844	LR
OPCovid-BR	**0.805**	LR	0.798	LR	0.784	LR	0.766	LR
Computer-BR	0.815	LR	0.814	LR	0.793	SVM	**0.820**	LR
MiningBR	0.912	LR	**0.927**	LR	0.922	LR	0.904	LR/SVM
TweetsMG	**0.992**	LR/SVM	**0.992**	SVM	**0.992**	SVM	0.975	SVM
tweetSentBR	**0.842**	LR	0.834	SVM	0.837	LR	0.812	LR/SVM
UniLex	**0.783**	LR	0.777	SVM	0.772	LR/SVM	0.745	LR/SVM
FIAT-UFMG	**0.826**	LR	0.824	LR/SVM	0.831	SVM	0.792	LR/SVM

Neutral tweets were not considered

Table 35 The F1-score for each dataset considering the model BERTimbau, strategies contextualized, LOO, inData, and allData, and classifiers LR, SVM, and XGB

Dataset	Contextualized		LOO		inData		allData	
	F1	Classifier	F1	Classifier	F1	Classifier	F1	Classifier
narrPT	**0.815**	LR	0.800	LR	0.806	LR	0.724	LR
MiningBR	0.784	LR	**0.794**	LR	0.791	LR	0.763	LR
ComputerBR	**0.821**	LR	0.819	LR	**0.821**	LR	0.805	SVM
TweetsMG	**0.958**	SVM	0.955	SVM	0.956	SVM	0.917	SVM
UniLex	**0.668**	LR	0.657	LR	0.665	LR	0.633	LR

Neutral tweets are also considered

F1-score. BERTimbau gains are more expressive for the smallest datasets, *narr-PT* and *MiningBR*. Tables 28, 29, and 30, present Acc and F1-score for the *allData* strategy for LR, SVM, and XGB, respectively. The same behavior was observed with the multiclass datasets, as can be seen in Tables 31, 32, and 33, that show the F1-score for the *allData* strategy and classifiers LR, SVM, and XGB, respectively.

When comparing mBERT and BERTweet, we observed that for the classifier LR, mBERT slightly outperforms BERTweet for #rank sums; while, for classifiers SVM and XGB, BERTweet is the one with a higher #rank sums. Both text representations performed the same way with the inData strategy.

Comparing *allData* with all the previous approaches—*contextualized*, *LOO*, and *inData*, the only case in which allData surpassed all the other approaches was for the dataset *Computer-BR* when considering only two classes, positive and negative (Table 34). Overall, considering the best values for F1-score, which was achieved using BERTimbau and classifiers LR and/or SVM (Tables 34 and 35), we observed that the contextualized strategy is still the best choice for both sets of datasets, binary and multiclass. That is, combining all datasets to adjust the weights of any

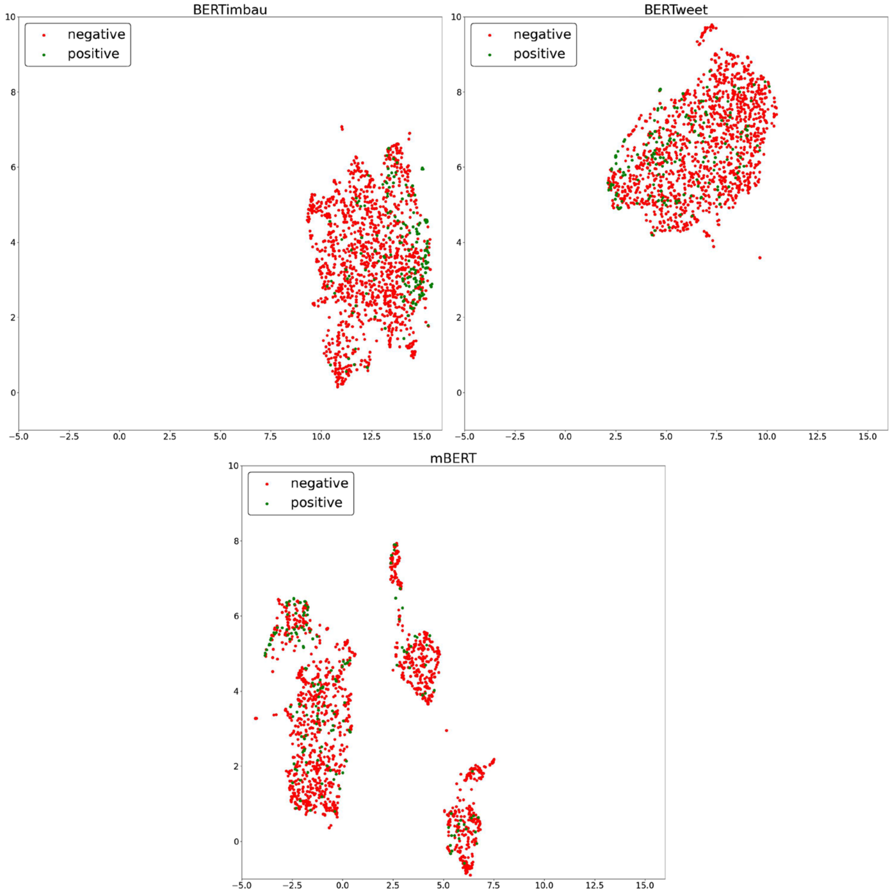

Fig. 5 BERTimbau, BERTweet, and mBERT embeddings for the binary dataset *MiningBR* using the contextualized strategy. UMAP (Uniform Manifold Approximation and Projection) (Ghojogh et al., 2023) was used for dimension reduction

Table 36 Best results (F1-score) for each dataset considering the combination of classifier and feature extraction method (Strategy)

Dataset	F1-score	Classifier	Strategy
narr-PT	0.867	LR	Contextualized
		LR	inData
OPCovid-BR	0.805	LR	Contextualized
Computer-BR	0.820	LR	allData
MiningBR	0.927	LR	LOO
TweetsMG	0.992	LR/SVM	Contextualized
		LR/SVM	LOO
		LR/SVM	inData
TweetSentBR	0.842	LR	Contextualized
UniLex	0.783	LR	Contextualized
FIAT-UFMG	0.826	LR	Contextualized

For all datasets, BERTimbau achieved the best results

Table 37 Best results (F1-score) for each dataset considering the combination of classifier and feature extraction method (Strategy)

Dataset	F1-score	Classifier	Strategy
narrPT	0.815	LR	Contextualized
MiningBR	0.794	LR	LOO
ComputerBR	0.821	LR	Contextualized
	0.821	LR	inData
TweetsMG	0.958	SVM	Contextualized
UniLex	0.668	LR	Contextualized

For all datasets, BERTimbau achieved the best results

of the three language models, in general, did not improve the quality of the embeddings to the point of enhancing the classifier Acc and F1 score. Considering that the largest dataset is the *UniLex* multiclass, with 12,655 tweets, it raises the question of how large should be a dataset to efficiently adjust the weights of a language model, incurring in high-quality classification models.

5.4 Overall results

To summarize the results obtained for each dataset, Tables 36 and 37 present the best results achieved regarding F1-score, highlighting the best combination of classifier (LR and SVM) and feature extraction strategy (*Contextualized*, *LOO*, *inData*, and *allData*).

Concerning the language model, we observed that embeddings trained from scratch solely using the target Portuguese language, BERTimbau, outperformed both the multilingual BERT (mBERT) and the tweet-based model BERTweet. To illustrate, in Fig. 5, we show the embeddings extracted for the binary dataset MiningBR using the contextualized strategy and language models BERTimbau, BERTweet, and mBERT. We can notice that, in the case of BERTimbau, the positive samples are easier to identify than when using BERTweet or mBERT. In general, mBERT was the language model with the least number of unique OOV tokens (Tables 4 and 5, *#unique OOV* column); however, when comparing the average number of OOVs (Tables 4 and 5, *avg. OOV* column) between mBERT, BERTimbau, and BERTweet we realize that the difference between them is not significative. For the static embeddings Word2Vec and GloVe, they consistently have a more unique and average number of OOV tokens than any other method across all datasets since they are based on classical word tokenization. It can, to some extent, explain their low performance for all classifiers and datasets (Appendix). On the other hand, a similar conclusion in the opposite direction cannot be reached for the contextualized embeddings since mBERT has not achieved good results when compared to other methods, even if it has the least number of OOV tokens. As BERT-based models adopt subwords tokenization, it could be that mBERT only breaks some words into more pieces. Concerning how the embedding is induced, we concluded that a contextualized embedding is more effective than a static representation for all scenarios tackled in this study. Even though identifying the best classifier was not a goal of this study, we noticed that, overall, LR achieved the best results for F1-score. For simplicity, the language model BERTimbau was omitted from Tables 36 and 37.

Taking into account the contextualized and the *LOO*, *inData*, and *allData* strategies for all combinations of classifier, language model, and strategies, we observed that extracting the contextualized embedding without any adjustment to the pre-trained model is the best approach for most of the datasets. The exceptions were for binary datasets *Computer-BR* and *MiningBR*, as can be seen in Table 36. *Computer-BR* achieved its best F1-score with the *allData* strategy, while *MiningBR* achieved its best score with the *LOO* strategy. When treated as a binary dataset, *Computer-BR* is small and imbalanced, with twice as many negative tweets than positive tweets. This result may indicate that the *LOO* strategy may improve the final classification for a small and imbalanced dataset, which keeps the target dataset away from the language model adjustment. *MiningBR*, also treated as a binary dataset, is heavily imbalanced, with almost ten times more negative than positive tweets. However, *MiningBR* is more than twice the size of *Computer-BR*, which may explain its good results with the *allData* strategy since *allData* also considers a portion of the target dataset when fine-tuning the language model.

Concerning the largest binary datasets, *TweetSentBR*, *FIAT-UFMG*, and *UniLex*, Table 26 points out that *TweetSentBR* and *FIAT-UFMG* achieved similar results. However, *UniLex* achieved a smaller F1-score, which can be explained by the fact that *UniLex* is a more imbalanced dataset than the others. We also verified that TweetSentBR and UniLex achieved precision and recall of around 0.82, while UniLex had a slightly lower precision and recall of around 0.77. For all those datasets, there was no considerable gap between both metrics, precision and recall. Moreover, we can observe in Table 5 that *UniLex* has significantly more missing tokens than the other datasets, according to all models. In addition, it also has more extensive examples than *TweetSentBR* (Table 2). All the tokenizers found more tokens to it (Tables 5 and 4), either because of that length or because they had more difficulty matching tokens and breaking the words between subwords. Those points potentially make *UniLex* a more challenging dataset.

When comparing mBERT with the tweet-based model BERTweet, we observed that the former model outperformed the latter for the largest datasets from the binary set when combined with strategies *inData* and *LOO*. For the smallest datasets, strategies *allData* and *inData* worked best, usually combined with BERTweet. For the multiclass scenario, both models, mBERT and BERTweet, alternated without any specific pattern concerning dataset size and class distribution. mBERT and BERTweet performed better after fine-tuning the model using any of the three strategies, *LOO*, *inData*, or *allData*. On the other hand, BERTimbau achieved the best results without any tuning (contextualized approach).

6 Conclusions

In this work, we investigated an effective strategy for the problem of sentiment analysis in Portuguese tweets. Precisely, we state sentiment classification as a binary classification and also as a multiclass task, in which embeddings are initially gathered for a tweet and then used as input to a classifier. First, we reviewed the literature concerning sentiment analysis for Portuguese-written tweets. We have collected eight datasets in various sizes and domains from the review, all manually annotated for Portuguese sentiment analysis. Then, for each dataset, a combination of language model and feature extraction schema was evaluated to pinpoint a practical approach for tweet sentiment analysis in the absence of a language model specific to Brazilian Portuguese tweets.

Our experiments demonstrated that regarding the language model, BERTimbau—a model trained exclusively using the target Portuguese language—surpassed three static models, GloVe, Word2Vec, and fastText, and the other Transformer-based models BERT multilingual and BERTweet. Depending on dataset size and continued pre-training strategy, mBERT, and BERTweet alternate as the second-best approaches.

Given the results obtained during this study, as the next step, we intend to investigate if a language model trained from scratch with Portuguese tweets benefits the task of sentiment analysis in Portuguese tweets, considering that the specificities of the Portuguese language and the noisy and informality of tweets would equip this new language model. In addition, we will look closer at the OOV tokens and the intersection of training vocabulary and the dataset vocabulary to draw correlations among those features and the quantitative results.

Transformer-based pre-trained models can have their weights adjusted according to a specific task. This is achieved by adding an extra classification layer at the top of the model. The question of whether this fine-tuning of the downstream task would be a better approach than the ones explored in this study remains open, and we plan to address it in future work.

Ethics statement

Datasets All the datasets considered in this manuscript were gathered from previous work that made them publicly available. Although we have not directly collected any tweets, we are aware that using data collected from the Twitter platform should raise ethical reflections. Even though Twitter users assume their posts are not private, they are usually not explicitly informed that what they write can be used for scientific—our case—or commercial—not our case—purposes. Besides, they might usually assume that their tweets are ephemeral whilst they, in fact, can be collected and stored by anyone anywhere. We tried our best not to include sensitive content in our examples and not disclose the identity of their authors.

Language model Given that this work strongly relies on large-scale language models and datasets composed of social media texts, despite the best intentions, we anticipate possible ethical and social risks by perpetuating social biases and providing false or misleading information. In the case of language models, these risks usually spring from the chosen training *corpora* used to pre-train such large models.

With those aspects in mind, the models resulting from this work are released for research purposes only.

Appendix: Evaluation of static embeddings

An initial evaluation using three different static embeddings, fastText, Word2Vec, and GloVe, was performed to identify the best static embedding for our scenario. Tables 38, 39, and 40 present Acc and F1-score (F1) for all three static embeddings and classifiers LR, SVM, and XGB, respectively. The superiority of fastText is clear for all datasets when combined with classifiers LR and SVM, and almost all datasets when used with the XGB classifier. The same can be seen with the multiclass set of datasets, as shown in Tables 41, 42, and 43. The superiority of fastText can be explained by its ability to deal with subwords, an important feature when working with tweets since those usually follow an informal style with heavy use of abbreviations and misspelling words.

Table 38 Accuracy and F1-score for each dataset considering the static embeddings for feature extraction and Logistic Regression classifier (binary datasets)

Dataset	fastText		GloVe		Word2Vec	
	Acc	F1	Acc	F1	Acc	F1
narr-PT	**0.843**	**0.843**	0.784	0.785	0.741	0.741
OPCovid-BR	**0.770**	**0.770**	0.748	0.747	0.757	0.756
Computer-BR	**0.760**	**0.764**	0.719	0.723	0.729	0.731
MiningBR	**0.890**	**0.899**	0.868	0.881	0.859	0.874
TweetsMG	**0.991**	**0.991**	0.989	0.989	0.989	0.989
TweetSentBR	**0.826**	**0.828**	0.783	0.785	0.772	0.774
UniLex	**0.778**	**0.778**	0.755	0.755	0.750	0.750
FIAT-UFMG	**0.815**	**0.815**	0.802	0.801	0.796	0.796
#wins	**8**	**8**	0	0	0	0
#rank sums	**8.0**	**8.0**	18.5	18.5	21.5	21.5

Table 39 Accuracy and F1-score for each dataset considering the static embeddings for feature extraction and SVM classifier (binary datasets)

Dataset	fastText		GloVe		Word2Vec	
	Acc	F1	Acc	F1	Acc	F1
narr-PT	**0.810**	**0.810**	0.745	0.746	0.743	0.743
OPCovid-BR	**0.777**	**0.776**	0.750	0.749	0.773	0.773
Computer-BR	**0.764**	**0.765**	0.734	0.735	0.739	0.739
MiningBR	**0.918**	**0.918**	0.894	0.899	0.866	0.879
TweetsMG	**0.992**	**0.992**	0.991	0.991	0.991	0.991
TweetSentBR	**0.832**	**0.833**	0.788	0.790	0.776	0.778
UniLex	**0.789**	**0.788**	0.762	0.762	0.761	0.761
FIAT-UFMG	**0.836**	**0.836**	0.824	0.824	0.810	0.810
#wins	**8**	**8**	0	0	0	0
#rank sums	**8.0**	**8.0**	18.5	18.5	21.5	21.5

Table 40 Accuracy and F1-score for each dataset considering the static embeddings for feature extraction and XGB classifier (binary datasets)

Dataset	fastText		GloVe		Word2Vec	
	Acc	F1	Acc	F1	Acc	F1
narr-PT	**0.755**	**0.751**	0.712	0.709	0.678	0.661
OPCovid-BR	**0.733**	**0.732**	0.692	0.690	0.672	0.671
ComputerBR	**0.719**	0.660	**0.719**	**0.671**	0.716	0.655
MiningBR	**0.900**	**0.868**	0.885	0.835	0.892	0.846
TweetsMG	0.963	0.963	**0.967**	**0.967**	0.962	0.962
TweetSentBR	**0.750**	**0.746**	0.706	0.695	0.689	0.657
UniLex	**0.698**	**0.690**	0.663	0.650	0.676	0.666
FIAT-UFMG	0.738	0.736	**0.743**	**0.742**	0.732	0.730
#wins	**6**	**5**	3	3	0	0
#rank sums	**10.5**	**11.0**	15.5	15.0	22.0	22.0

Table 41 Accuracy and F1-score for each dataset considering the static embeddings for feature extraction and Logistic Regression classifier (multiclass datasets)

Dataset	fastText		GloVe		Word2Vec	
	Acc	F1-score	Acc	F1-score	Acc	F1-score
narr-PT	**0.791**	**0.791**	0.697	0.697	0.697	0.697
Computer-BR	**0.772**	**0.781**	0.751	0.758	0.749	0.758
MiningBR	**0.778**	**0.776**	0.749	0.744	0.734	0.727
TweetsMG	**0.945**	**0.945**	0.933	0.933	0.938	0.939
UniLex	**0.662**	**0.661**	0.624	0.622	0.619	0.616
#wins	**5**	**5**	0	0	0	0
#rank sums	**5.0**	**5.0**	11.5	12.0	13.5	13.0

Table 42 Accuracy and F1-score for each dataset considering the static embeddings for feature extraction and SVM classifier (multiclass datasets)

Dataset	fastText		GloVe		Word2Vec	
	Acc	F1-score	Acc	F1-score	Acc	F1-score
narr-PT	**0.760**	**0.760**	0.689	0.689	0.693	0.693
Computer-BR	**0.747**	**0.769**	0.702	0.729	0.696	0.725
MiningBR	**0.761**	**0.766**	0.739	0.744	0.711	0.720
TweetsMG	**0.961**	**0.962**	0.955	0.955	0.959	0.959
UniLex	**0.676**	**0.675**	0.648	0.648	0.643	0.642
#wins	**5**	**5**	0	0	0	0
#rank sums	**5.0**	**5.0**	12.0	12.0	13.0	13.0

Table 43 Accuracy and F1-score for each dataset considering the static embeddings for feature extraction and XGB classifier (multiclass datasets)

Dataset	fastText		GloVe		Word2Vec	
	Acc	F1-score	Acc	F1-score	Acc	F1-score
narr-PT	**0.715**	**0.716**	0.658	0.655	0.648	0.641
Computer-BR	0.741	**0.649**	**0.744**	0.648	0.740	0.634
MiningBR	**0.743**	**0.695**	0.715	0.651	0.700	0.626
TweetsMG	0.879	0.879	**0.894**	**0.894**	0.892	0.893
UniLex	**0.574**	**0.556**	0.550	0.533	0.546	0.516
#wins	**3**	**4**	2	1	0	0
#rank sums	**8.0**	**7.0**	**8.0**	9.0	14.0	14.0

Acknowledgements This research was partially financed by CNPq (National Council for Scientific and Technological Development) under Grants 311275/2020-6 and 315750/2021-9 and FAPERJ—*Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro*, Process SEI-260003/000614/2023, E-26/202.914/2019 and E-26/201.139/2022.

References

Agüero-Torales, M. M., Salas, J. I. A., & López-Herrera, A. G. (2021). Deep learning and multilingual sentiment analysis on social media data: An overview. *Applied Soft Computing, 107*, 107373.

Alves, A. L., Baptista, C. D. S., Andrade, L. H. D., & Paes, R. (2015). Uso de técnicas de análise de sentimentos em tweets relacionados ao meio-ambiente. In *Anais do Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais (WCAMA)*, 2015 (pp. 37–46). Sociedade Brasileira de Computacao.

Alves, A. L., Baptista, C. D. S., Firmino, A. A., Oliveira, M. G. D., & Paiva, A. C. D. (2014). A comparison of SVM versus Naive-Bayes techniques for sentiment analysis in tweets: A case study with the 2013 FIFA confederations cup. In *WebMedia 2014—Proceedings of the 20th Brazilian symposium on multimedia and the web*, 2014 (pp. 123–130). Association for Computing Machinery, Inc.

Araújo, G., Teixeira, F., Mancini, F., Guimarães, M., & Pisa, I. (2018). Sentiment analysis of Twitter's health messages in Brazilian Portuguese. *Journal of Health Informatics, 10*, 17–24.

Araújo, M., Pereira, A., & Benevenuto, F. (2020). A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences, 512*, 1078–1102.

Araujo, M., Pereira, A., Reis, J., & Benevenuto, F. (2016). An evaluation of machine translation for multilingual sentence-level sentiment analysis. In *Proceedings of the ACM symposium on applied computing*, 2016, August 4 (pp. 1140–1145). Association for Computing Machinery.

Barreto, S., Moura, R., Carvalho, J., Paes, A., & Plastino, A. (2021). Sentiment analysis in tweets: an assessment study from classical to modern text representation models. *Data Min Knowl Disc, 37*, 318–380 (2023). https://doi.org/10.1007/s10618-022-00853-0

Belisário, L. B., Ferreira, L. G., & Pardo, T. A. S. (2020). Evaluating methods of different paradigms for subjectivity classification in Portuguese. In *Proceedings of the 14th international conference on the computational processing of Portuguese*, LNAI, 2020 (Vol. 12037, pp. 261–269). Springer.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In: M. C. Elish, W. Isaac & R. S. Zemel (Eds.), *FAccT '21: 2021 ACM conference on fairness, accountability, and transparency, virtual event*, Toronto, Canada, March 3–10, 2021 (pp. 610–623). ACM. https://doi.org/10.1145/3442188.3445922.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146.

Brum, H. B., & das Graças Volpe Nunes, M. (2018). Building a sentiment corpus of tweets in Brazilian Portuguese. In N. C. C. (chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (Eds.), *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018. European Language Resources Association (ELRA).

Brum, H. B., & Nunes, M. D. G. V. (2018). Semi-supervised sentiment annotation of large corpora. In *Proceedings of the 13th international conference on the computational processing of Portuguese*, 2018 (pp. 385–395). Springer.

Carmo, D., Piau, M., Campiotti, I., Nogueira, R., & Lotufo, R. (2020). PTT5: Pretraining and validating the t5 model on Brazilian Portuguese data. arXiv preprint. arXiv:2008.09144

Carosia, A., Coelho, G. P., & da Silva, A. E. A. (2019). The influence of tweets and news on the Brazilian Stock Market through sentiment analysis. In *Proceedings of the 25th Brazilian symposium on multimedia and the web*, 2019. ACM.

Carosia, A., Coelho, G. P., & Silva, A. E. A. (2020). Analyzing the Brazilian financial market through Portuguese sentiment analysis in social media. *Applied Artificial Intelligence, 34*, 1–19.

Carvalho, J., & Plastino, A. (2021). On the evaluation and combination of state-of-the-art features in Twitter sentiment analysis. *Artificial Intelligence Review, 54*(3), 1887–1936.

Carvalho, P., & Silva, M. J. (2015). Sentilex-pt: Principais características e potencialidades. *Oslo Studies in Language*. https://doi.org/10.5617/osla.1444

Chan, B., Schweter, S., & Möller, T. (2020). German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, 2020 (pp. 6788–6796). International Committee on Computational Linguistics (Online).

Correa, E. A., Marinho, V. Q., Santos, L. B. D., Bertaglia, T. F. C., Treviso, M. V., & Brum, H. B. (2017). PELESent: Cross-domain polarity classification using distant supervision. In *Proceedings—2017 Brazilian conference on intelligent systems, BRACIS 2017*, 2017, January 2018 (pp. 49–54). Institute of Electrical and Electronics Engineers, Inc.

Costa, J. M. R., Rotabi, R., Murnane, E. L., & Choudhury, T. (2015). It is not only about grievances: Emotional dynamics in social media during the Brazilian protests. In *Proceedings of the international AAAI conference on web and social media*, 2015 (Vol. 9).

Cury, R. M. (2019). Oscillation of tweet sentiments in the election of João Doria Jr. for Mayor. *Journal of Big Data, 6*, 1–15.

da Silva, A. M., Bastos, R. D. M., & de Azevedo da Rocha, R. L. (2018). Sentiment analysis in Brazilian Portuguese tweets in the domain of calamity: Application of the summarization method and semantic similarity in polarized terms. In *IJCCI 2018—Proceedings of the 10th international joint conference on computational intelligence*, 2018 (pp. 225–231). SciTe Press.

De Aguiar, E. J., Faiçal, B. S., Ueyama, J., Silva, G. C., & Menolli, A. (2018). Análise de sentimento em redes sociais para a língua portuguesa utilizando algoritmos de classificação. In *Anais do XXXVI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*. SBC.

De Barros, T. M., Pedrini, H., & Dias, Z. (2021). Leveraging emoji to improve sentiment classification of tweets. In *Proceedings of the 36th annual ACM symposium on applied computing*, 2021 (pp. 845–852). ACM.

de Carvalho, V. D. H., Nepomuceno, T. C. C., & Costa, A. P. C. S. (2020). An automated corpus annotation experiment in Brazilian Portuguese for sentiment analysis in public security. In *Lecture notes in business information processing, LNBIP* (Vol. 384 pp. 99–111). Springer.

de Melo, T., & Figueiredo, C. M. (2021). Comparing news articles and tweets about COVID-19 in Brazil: Sentiment analysis and topic modeling approach. *JMIR Public Health and Surveillance, 7*, e24585.

de Oliveira, D. N., & de Campos Merschmann, L. H. (2021). Joint evaluation of preprocessing tasks with classifiers for sentiment analysis in Brazilian Portuguese language. *Multimedia Tools and Applications, 80*, 15391–15412.

de Souza, K. F., Pereira, M. H. R., & Dalip, D. H. (2017). Unilex: Método léxico para análise de sentimentos textuais sobre conteúdo de tweets em português brasileiro. *Abakós, 5*(2), 79–96.

de Vargas Feijó, D., & Moreira, V. P. (2020). Mono vs. multilingual transformer-based models: A comparison across several language tasks. CoRR **abs/2007.09757**. arxiv:2007.09757

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019: Long and Short Papers*, Minneapolis, MN, USA, June 2–7, 2019. (Vol.1, pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/n19-1423.

dos Santos, A., Júnior, J. D. B., & de Arruda Camargo, H. (2018). Annotation of a corpus of tweets for sentiment analysis. In *Lecture notes in computer science (including subseries lecture notes* *in artificial intelligence and lecture notes in bioinformatics), LNAI* (Vol. 11122, pp. 294–302). Springer.

Filho, J. A. W., Wilkens, R., Idiart, M., & Villavicencio, A. (2018). The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of the eleventh international conference on language resources and evaluation, LREC 2018*, Miyazaki, Japan, May 7–12, 2018. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2018/summaries/599.html

França, T., & Oliveira, J. (2014). Análise de sentimento de tweets relacionados aos protestos que ocorreram no brasil entre junho e agosto de 2013. In *Anais do III Brazilian workshop on social network analysis and mining*, 2014 (pp. 128–139). SBC.

Gage, P. (1994). A new algorithm for data compression. *The C Users Journal Archive, 12*, 23–38.

Garcia, K., & Berton, L. (2021). Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Applied Soft Computing, 101*, 107057.

Gengo, P., & Verri, F. A. (2020). Semi-supervised sentiment analysis of Portuguese tweets with random walk in feature sample networks. In *Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and Lecture notes in bioinformatics), LNAI* (Vol. 12319, pp. 595–605). Springer.

Ghojogh, B., Crowley, M., Karray, F., & Ghodsi, A. (2023). *Uniform manifold approximation and projection (UMAP)* (pp. 479–497). Springer. https://doi.org/10.1007/978-3-031-10602-6.

Gomes, F. B., Adán-Coello, J. M., & Kintschner, F. E. (2018). Studying the effects of text preprocessing and ensemble methods on sentiment analysis of Brazilian Portuguese tweets. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), LNAI* (Vol. 11171, pp. 167–177). Springer.

Grandin, P., & Adan, J. M. (2016). Piegas: A systems for sentiment analysis of tweets in Portuguese. *IEEE Latin America Transactions, 14*, 3467–3473.

Guerra, P. H. C., Meira, W., & Cardie, C. (2014). Sentiment analysis on evolving social streams: How self-report imbalances can help. In *WSDM 2014—Proceedings of the 7th ACM international conference on web search and data mining* (pp. 443–452). Association for Computing Machinery.

Guerra, P. H. C., Veloso, A., Meira, W., & Almeida, V. (2011). From bias to opinion: A transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining—KDD '11*, 2011. ACM Press.

Gururangan, S., Marasovic, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics, ACL 2020*, Online, July 5–10, 2020 (pp. 8342–8360). Association for Computational Linguistics.

Heinzerling, B., & Strube, M. (2018). BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (Eds.), *Proceedings of the eleventh international conference on language resources and evaluation, LREC 2018*, Miyazaki, Japan, May 7–12, 2018. European Language Resources Association (ELRA).

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, 2014. The AAAI Press.

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the international AAAI conference on web and social media*, 2011 (Vol. 5).

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International conference on learning representations, ICLR 2020*, Addis Ababa, Ethiopia, April 26–30, 2020. OpenReview.net. https://openreview.net/forum?id=H1eA7AEtvS

Lauand, B. P., & Oliveira, J. (2014). Inferindo as condições de trânsito através da análise de sentimentos no Twitter. *iSys - Revista Brasileira de Sistemas de Informação, 7*(3), 56–74.

Lima, M. L., Nascimento, T. P., Labidi, S., Timbó, N. S., Batista, M. V. L., Neto, G. N., Costa, E. A. M., & Sousa, S. R. S. (2016). Using sentiment analysis for stock exchange prediction. *International Journal of Artificial Intelligence and Applications*. https://doi.org/10.5121/ijaia.2016.7106

Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emoticons*. Cambridge University Press.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692**. arxiv:1907.11692

Lourenco Jr., R., Veloso, A., Pereira, A., Meira Jr., W., Ferreira, R., & Parthasarathy, S. (2014). Economically-efficient sentiment stream analysis. In *Proceedings of the 37th international ACM SIGIR conference on research and development in information retrieval, SIGIR '14*, 2014 (pp. 637–646). Association for Computing Machinery.

Machado, M. T., Pardo, T. A. S., & Ruiz, E. E. S. (2018). Creating a Portuguese context sensitive lexicon for sentiment analysis. In *Computational processing of the Portuguese language—13th International conference, PROPOR 2018*, Canela, Brazil, September 24–26, 2018, Proceedings, lecture notes in computer science (vol. 11122, pp. 335–344). Springer.

Malini, F., Ciarelli, P., & Medeiros, J. (2017). O sentimento político em redes sociais: big data, algoritmos e as emoções nos tweets sobre o impeachment de dilma rousseff. *Liinc em Revista, 13*, 323–342.

Martin, L., Müller, B., Suárez, P.J.O., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., & Sagot, B. (2020). CamemBERT: A tasty French language model. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics, ACL 2020*, Online, July 5–10, 2020 (pp. 7203–7219). Association for Computational Linguistics.

Martins, R., Pereira, A., & Benevenuto, F. (2015). An approach to sentiment analysis of web applications in Portuguese. In *Proceedings of the 21st Brazilian symposium on multimedia and the web, WebMedia '15*, 2015 (pp. 105–112). Association for Computing Machinery.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013) Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th international conference on neural information processing systems, NIPS'13*, 2013 (Vol. 2, pp. 3111–3119).

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence, 29*(3), 436–465.

Moraes, S. M., Santos, A. L., Redecker, M., Machado, R. M., & Meneguzzi, F. R. (2016). Comparing approaches to subjectivity classification: A study on Portuguese tweets. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 9727, pp. 86–94). Springer.

Nankani, H., Dutta, H., Shrivastava, H., Krishna, P. R., Mahata, D., & Shah, R. R. (2020). Multilingual sentiment analysis. In *Deep learning-based approaches for sentiment analysis* (pp. 193–236). Springer.

Nascimento, P., Osiek, B., & Xexéo, G. (2015). Análise de sentimento de tweets com foco em notícias. *Revista Eletrônica de Sistemas de Informação, 14*, 2.

Neuenschwander, B., Pereira, A., Meira, W., & Barbosa, D. (2014). Sentiment analysis for streams of web data: A case study of Brazilian financial markets. In *WebMedia 2014—Proceedings of the 20th Brazilian symposium on multimedia and the web*, 2014 (pp. 167–170). Association for Computing Machinery, Inc.

Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2020). BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations*, 2020 (pp. 9–14).

Nozza, D., Bianchi, F., & Hovy, D. (2020). What the [mask]? Making sense of language-specific BERT models. CoRR **abs/2003.02912**. arxiv:2003.02912

Oliveira, D. J. S., & de Souza Bermejo, P. H. (2017). Mídias sociais e administração pública: análise do sentimento social perante a atuação do governo federal brasileiro. *Organizações & Sociedade, 24*, 491–508.

Oliveira, D. J. S., de Souza Bermejo, P. H., & dos Santos, P. A. (2017). Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls. *Journal of Information Technology and Politics, 14*, 34–45.

Oliveira, D. J. S., Souza Bermejo, P. H., Pereira, J. R., & Barbosa, D. A. (2019). The application of the sentiment analysis technique in social media as a tool for social management practices at the governmental level. *Revista de Administracao Publica, 53*, 235–251.

Pasqualotti, P. R., & Vieira, R. (2008). Wordnetaffectbr: uma base lexical de palavras de emoções para a língua portuguesa. *RENOTE-Revista Novas Tecnologias na Educação, 6*(1).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pennington, J., Socher, R., & Manning, C. (2014) GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014 (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162.

Pessanha, G. R. G., Fidelis, T. O., Freire, C. D., & Soares, E. A. (2020). Fiqueemcasa: Análise de sentimento dos usuários do twitter em relação ao covid19. *HOLOS, 5*, 2020.

Praciano, B. J. G., da Costa, J. P. C. L., Maranhao, J. P. A., de Mendonça, F. L. L., de Sousa Junior, R. T., & Prettz, J. B. (2018). Spatio-temporal trend analysis of the Brazilian elections based on Twitter data. In *IEEE international conference on data mining workshops*, November 2018 (pp. 1355–1360). IEEE Computer Society.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, EMNLP-IJCNLP 2019*, Hong Kong, China, November 3–7, 2019 (pp. 3980–3990). Association for Computational Linguistics.

Rosa, R. L., Rodriguez, D. Z., & Bressan, G. (2013). SentiMeter-Br: A social web analysis tool to discover consumers' sentiment. In *Proceedings—IEEE international conference on mobile data management*, 2013 (Vol. 2, pp. 122–124).

Schuster, M., & Nakajima, K. (2012). Japanese and Korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing, ICASSP 2012*, Kyoto, Japan, March 25–30, 2012 (pp. 5149–5152). IEEE.

Severyn, A., & Moschitti, A. (2015). Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 2015 (pp. 959–962).

Silva, A. N. D., Souza, O. D., & Souza, J. N. D. (2020). Sentiment parser based on x-bar theory to Brazilian Portuguese. In *Proceedings of the 2020 international conference on computing, electronics and communications engineering*, 2020 (pp. 166–171). Institute of Electrical and Electronics Engineers, Inc.

Silva, I. S., Gomide, J., Veloso, A., Meira, W., & Ferreira, R. (2011). Effective sentiment stream analysis with self-augmenting training and demand-driven projection. In *SIGIR'11—Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval*, 2011 (pp. 475–484). Association for Computing Machinery.

Singhal, P., & Bhattacharyya, P. (2016). Borrow a little from your rich cousin: Using embeddings and polarities of English words for multilingual sentiment classification. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, 2016 (pp. 3053–3062).

Souza, B. A., Almeida, T. G., Menezes, A. A., Nakamura, F. G., Figueiredo, C. M., & Nakamura, E. F. (2016). For or against? Polarity analysis in tweets about impeachment process of Brazil President. In *Proceedings of the 22nd Brazilian symposium on multimedia and the web*, 2016 (pp. 335–338). ACM.

Souza, E., Alves, T., Teles, I., Oliveira, A. L., & Gusmão, C. (2016). TOPIE: An open-source opinion mining pipeline to analyze consumers' sentiment in Brazilian Portuguese. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 9727, pp. 95–105). Springer.

Souza, F., Nogueira, R., & Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Brazilian conference on intelligent systems*, 2020 (pp. 403–417). Springer.

Souza, M., & Vieira, R. (2011). Construction of a Portuguese opinion lexicon from multiple resources. In *Anais do Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, 2011, Brasil.

Souza, M., & Vieira, R. (2012). Sentiment analysis on Twitter data for Portuguese language. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), LNAI* (Vol. 7243, pp. 241–247). Springer.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In A. Korhonen, D. R. Traum & L. Màrquez (Eds.), *Proceedings of the 57th conference of the Association for Computational Linguistics, ACL 2019: Long papers*, Florence, Italy, July 28–August 2, 2019 (Vol. 1, pp. 3645–3650). Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-1355.

Strubell, E., Ganesh, A., & McCallum, A. (2020). Energy and policy considerations for modern deep learning research. In *The thirty-fourth AAAI conference on artificial intelligence, AAAI 2020, the* *thirty-second innovative applications of artificial intelligence conference, IAAI 2020, the tenth AAAI symposium on educational advances in artificial intelligence, EAAI 2020*, New York, NY, USA, February 7–12, 2020 (pp. 13693–13696). AAAI Press. https://aaai.org/ojs/index.php/AAAI/article/view/7123

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*(2), 267–307.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics: Long papers*, 2014 (Vol. 1, pp. 1555–1565).

Vargas, F. A., Santos, R. S. S. D., & Rocha, P. R. (2020). Identifying fine-grained opinion and classifying polarity on coronavirus pandemic. In *Proceedings of the 9th Brazilian conference on intelligent systems*, LNAI, 2020 (Vol. 12319, pp. 511–520). Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need, 5998–6008.

Vilhagra, L. A., Fernandes, E. R., & Nogueira, B. M. (2020). TextCSN: A semi-supervised approach for text clustering using pairwise constraints and convolutional Siamese network. In *Proceedings of the ACM symposium on applied computing*, 2020 (pp. 1135–1142). Association for Computing Machinery.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. CoRR **abs/1912.07076** . arxiv:1912.07076

Vitório, D., Souza, E., & Oliveira, A. L. (2019). Evaluating active learning sampling strategies for opinion mining in Brazilian politics corpora. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*, LNAI (Vol. 11805, pp. 695–707). Springer.

Vitório, D., Souza, E. P. R., Pereira, I., & Oliveira, A. (2017). Investigating opinion mining through language varieties: A case study of Brazilian and European Portuguese tweets. In *Proceedings of the 11th Brazilian symposium in information and human language technology*, 2017 (pp. 43–52). Sociedade Brasileira de Computação, Uberlândia.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations*, 2020 (pp. 38–45). Association for Computational Linguistics (Online). https://www.aclweb.org/anthology/2020.emnlp-demos.6

Yagui, M., & Maia, L. (2017). Data mining of social manifestations in Twitter: An ETL approach focused on sentiment analysis. In *XIII Brazilian symposium on information systems*, 2017 (pp. 1–8). Sociedade Brasileira de Computacao.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G. H., Yuan, S., Tar, C., Sung, Y. H., Strope & B., Kurzweil, R. (2020). Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Daniela Vianna[1,2] · Fernando Carneiro[3] · Jonnathan Carvalho[4] · Alexandre Plastino[3] · Aline Paes[3]

✉ Daniela Vianna
dvianna@icomp.ufam.edu.br

✉ Aline Paes
alinepaes@ic.uff.br

Fernando Carneiro
fernandocarneiro@id.uff.br

Jonnathan Carvalho
joncarv@iff.edu.br

Alexandre Plastino
plastino@ic.uff.br

[1] Institute of Computing, Universidade Federal do Amazonas (UFAM), Manaus, AM, Brazil

[2] Jusbrasil, Salvador, Brazil

[3] Institute of Computing, Universidade Federal Fluminense (UFF), Niterói, RJ, Brazil

[4] Instituto Federal Fluminense (IFF), Itaperuna, RJ, Brazil