BPA: A Multilingual Sentiment Analysis Approach based on BiLSTM

Iago C. Chaves, Antônio Diogo F. Martins, Francisco D. B. S. Praciano, Felipe T. Brito,
Jose Maria Monteiro and Javam C. Machado
Computer Science Department, Universidade Federal do Ceará, Fortaleza, Brazil

Keywords: Natural Language Processing, Sentiment Analysis, LSTM, Pooling, Attention Mechanism.

Abstract: Sentiment analysis (SA) is the automatic process of understanding people's feelings or beliefs expressed in texts such as emotions, opinions, attitudes, appraisals and others. The main task is to identify the polarity level (positive, neutral or negative) of a given text. This task has been the subject of several research competitions in many languages, for instance, English, Spanish and Arabic. However, developing a multilingual sentiment analysis method remains a challenge. In this paper, we propose a new approach, called BPA, based on BiLSTM neural networks, pooling operations and attention mechanism, which is able to automatically classify the polarity level of a text. We evaluated the BPA approach using five different data sets in three distinct languages: English, Spanish and Portuguese. Experimental results evidence the suitability of the proposed approach to multilingual and domain-independent polarity classification. BPA's best results achieved an accuracy of 0.901, 0.865 and 0.923 for English, Spanish and Portuguese, respectively.

1 INTRODUCTION

Sentiment analysis (SA), also known as opinion mining, is the automatic process of understanding people's feelings in written text. Its primary task is to determine the polarity level (positive, neutral or negative) of a text. Recently, sentiment analysis expanded this task in order to identify the emotional status of a sentence, e.g., sadness, excitement, joy, anger, and whether a text is humorous or not.

Every day, many positive and negative comments are shared on the Internet. These comments are present in product reviews, social media posts and survey responses. The impact of these comments on the economy is real. In this context, organizations and enterprises can take advantage and extract knowledge from their customer's opinions about their products and services, improving their marketing strategies and decision policies. The goal of text sentiment classification is to automatically determine whether a comment's sentiment polarity is negative, positive or neutral. This problem, also called polarity prediction task, has been subject of several international research challenges, such as SemEval, TASS and IberEval, and it is present in many languages, e.g., English, Spanish and Arabic. Nevertheless, developing a multilingual sentiment analysis method remains a challenge.

In this paper, we propose a new approach, called BPA, based on BiLSTM neural networks, pooling operations and attention mechanism in order to classify the polarity level of a text in an automatic manner. We evaluated the BPA approach using five data sets in three different languages: English (SST-2 and SemEval 2017 Subtask A data sets), Spanish (TASS 2017 Task 1 General Corpus and CCMD-ES data set) and Portuguese (CCMD-PT data set).

Experimental results evince the suitability of the BPA approach to multilingual and domain-independent polarity classification. BPA's best results achieved an accuracy of 0.901 (SST-2 data set), 0.865 (CCMD-ES data set) and 0.923 (CCMD-PT) for English, Spanish and Portuguese, respectively. It is important to highlight that BPA outperformed the SemEval 2017 Subtask A competition winner model (DataStories System) in terms of accuracy, F1-Score (Macro) and Recall. Using this data set, BPA obtained an accuracy of 0.659, an F1-Score of 0.682 and a Recall of 0.682, while DataStories System obtained an accuracy of 0.6515, an F1-Score of 0.6772 and a Recall of 0.6811. For TASS 2017 Task 1 General Corpus data set, BPA approach had a similar performance to the competition winner model (INGEOTEC-Cevodag_003 System). BPA obtained an accuracy of 0.791, an F1-Score (Macro) of 0.576. The winner of the competition achieved an accuracy of 0.645 and F1-Score (Macro) of 0.577.

The remainder of this paper is organized as follows: Section 2 presents the related work. Section 3

Chaves, I., Martins, A., Praciano, F., Brito, F., Monteiro, J. and Machado, J.
BPA: A Multilingual Sentiment Analysis Approach based on BiLSTM.
DOI: 10.5220/0011071400003179
In *Proceedings of the 24th International Conference on Enterprise Information Systems (ICEIS 2022) - Volume 1*, pages 553-560
ISBN: 978-989-758-569-2; ISSN: 2184-4992

presents the proposed approach for multilingual and domain-independent polarity classification. Section 4 discusses the results of performed case study. Finally, Section 5 concludes this paper.

2 RELATED WORK

In (Bonadiman et al., 2017), the authors have studied the use of neural networks for the Sentiment Analysis of Twitter text associated with a real application scenario. They modified the network architecture by applying a recurrent pooling layer enabling the learning of longer dependencies between words in tweets. The recurrent pooling layer makes the network more robust to unbalanced data distribution.The results showed that the proposed approach worked well for both English and Italian languages.

In (Jianqiang et al., 2018), the authors used a depth convolution neural network for sentiment classification on tweets. First, the proposed method creates global vectors for word representation by unsupervised learning on large Twitter corpora. Next, these word embeddings are combined with n-grams features and word sentiment polarity score features to form a sentiment feature set of tweets. Finally, the feature set is integrated into a deep convolution neural network. They reported their experimental results in five data sets.

Laerte et al. (Letarte et al., 2018) introduced the Self-Attention Network (SANet), a flexible and interpretable architecture for text classification. The authors showed that gains obtained by self-attention are task-dependent. Experiments on sentiment analysis tasks showed an improvement of around 2% when using self-attention compared to a baseline without attention. Experiments on topic classification showed no gain.

In (Graff et al., 2018), the authors proposed EvoMSA, a multilingual and domain-independent sentiment analysis system. EvoMSA is a classifier, based on Genetic Programming that works by combining the output of different text classifiers to produce the final prediction. Furthermore, it is worth to mention that EvoMSA was developed using Python and is available as open-source software.

In (Sarlis and Maglogiannis, 2020), the authors evaluated several algorithms on various sentiment-labeled data sets, creating two vector space models. The goal was to obtain the model with the highest accuracy and the best generalization. To measure how well these models generalize in other domains, several data sets were used.

Cai et al. (Cai et al., 2020) proposed a series of progressively enhanced multi-task models for sentiment analysis, where each model is an enhanced version of the former and the last model is the best. By combining a pooling layer and a bidirectional RNN, the model could comprehensively extract semantic text information. In addition, the attention mechanism linking the task-specific layer and the shared layer empowers the model to intelligently select effective features from the shared layer. These new features allowed the authors to better conduct sentiment analysis tasks, particularly on small data sets.

In (Araújo et al., 2020), the authors evaluated 16 methods for sentence-level sentiment analysis proposed for English, and compared them with 3 language-specific methods. Based on 14 data sets, they provide an extensive quantitative analysis of existing multilingual approaches. The results suggested that simply translating the input text in a specific language to English and then using one of the existing best methods developed for English can be better than the existing language-specific approach evaluated.

Chen et al. (Chen et al., 2018) proposed a deep neural network model combining convolutional neural network and regional long short-term memory (CNN-RLSTM) for the task of sentiment analysis. The proposed approach reduces the training time of neural network model through a regional LSTM and uses a sentence-level CNN to extract sentiment features of the whole sentence. Experiments showed that the proposed approach presented better performance than SVM and several other neural network models.

In (Guo et al., 2018), the authors introduced PERSEUS – a personalized framework for sentiment categorization on user-related data. The proposed framework provides a deeper understanding of user behavior in determining the sentiment orientation. The proposed framework explores a recurrent neural network with long short-term memory to leverage the assumptions. Experiments showed the effectiveness of the components used in PERSEUS.

3 THE PROPOSED APPROACH TO POLARITY CLASSIFICATION

In this section we will describe a new approach, called BPA, based on BiLSTM neural networks, pooling operations and attention mechanism in order to classify the polarity level of a text in an automatic manner. First, we will present the BPA architecture. Next, we will show a framework to apply the BPA approach.

3.1 BPA Architecture

The BPA approach uses different kinds of neural networks and some advanced techniques, such as BiLSTMs, BERT Embeddings and Pooling layers. The main idea is to gather different approaches to address several challenges and to classify short instant messages into three different classes (positive, negative and neutral). The BPA architecture has several layers, which are depicted in Figure 1. The first layer is the BERT representation model. The BERT (Devlin et al., 2018) layer works like an embedding component. The last hidden state of the BERT layer acts as a word vector, representing the input sentence. This vector is the input of a bidirectional LSTM. In this way, we model the correlation between the words in the input sentence, in both directions, forward and backward. We stacked two BiLSTMs so that the complexity of our model is greater and, consequently, allowing the sentences to be represented with more complex patterns, making the model have a greater variety of these patterns to be able to distinguish the sentiment presented in the sentence. Next, we perform the max pooling and the average pooling operations. The pooling results are concatenated together with the last hidden state of the BiLSTM. The resulting value is the input of a linear layer. The rationale under the use of these pooling operations is that the most (maximum value) and the less (average value) important BiLSTM's outputs will be provided to the model together with the output of the last layer of the BiLSTM, allowing the model to be able to distinguish the sentiment of the sentences based on more specific outputs. In other words, the model will receive the value corresponding to the word that has the most significant impact on the sentence sentiment (maximum value) and two values representing the word context (average value and the result of the last hidden layer).

3.2 BPA Framework

The BPA approach comprises an architecture and an usage framework. Figure 2 shows how the BPA framework is structured. The BPA framework has two main phases: Data Processing and Sentiment Analysis. We will describe each one of these phases next.

3.2.1 Data Processing

This phase handles all the data processing steps since the corpus labeling until the application of data augmentation techniques (if it is necessary). Next we will describe each one of these steps.

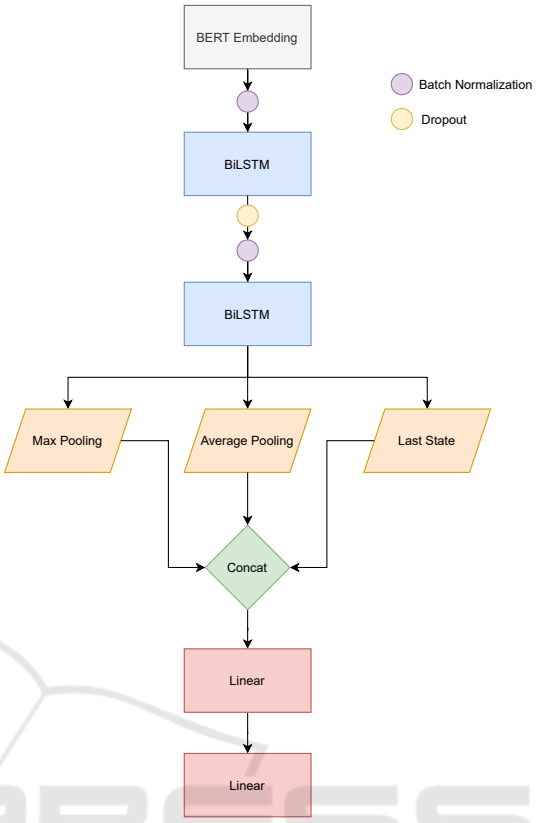

Figure 1: The BPA architecture.

Data Labeling. When the data is not labeled, this step is mandatory. Some online services are available for data annotation, including sentiment analysis labeling. However, the data annotator must follow a well-written and strict labeling guideline to avoid predictive models with poor performance. This guideline must provide accurate, simple and straightforward information to guide the data annotator through the labeling process. Following a guideline will make the data labeling process reliable and less error-prone.

Labeling Revision. Even with a well-written and strict guideline, labeling disagreements will emerge if more than one person labels the corpus. There are different ways to deal with the disagreements, depending on how the labeling process was performed. An interesting approach is to ask each annotator to review all messages labeled by other annotators. In case of labeling disagreements, ask the annotators to discuss the suitable label for the message. Sometimes it is necessary to update the labeling guideline to assess recurrent issues. Solving the labeling disagreement problem by revising the corpus helps the classifier avoid missing predictions due to labeling errors.

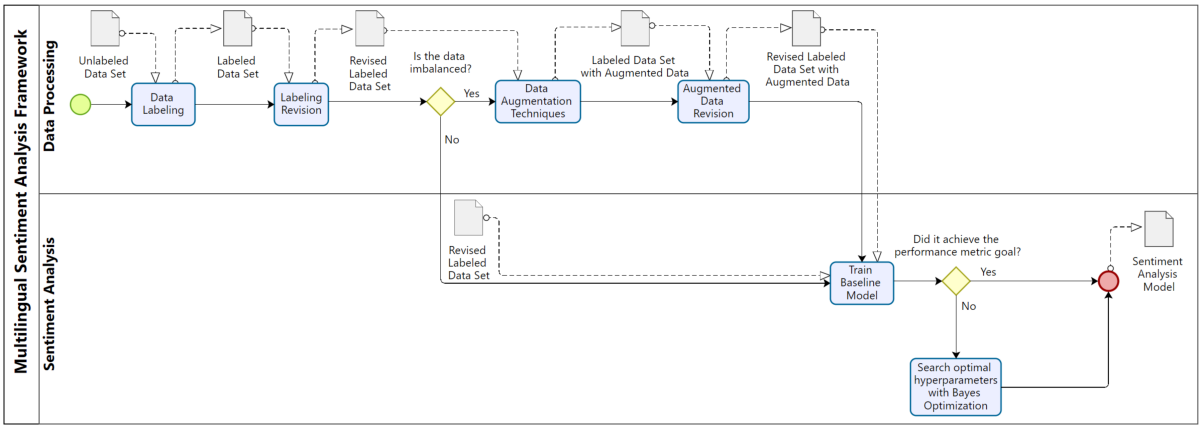

Figure 2: An overview of the proposed BPA framework.

Data Augmentation. Another common issue in real-world applications is the class imbalance. There are several techniques to assess this problem, but we need to be cautious in the sentiment analysis context because some techniques' results may completely change a message's context. When working on the sentiment analysis task and the data set is imbalanced or do not have enough instances, some of these data augmentation techniques can be applied:

- Deleting random words or swapping words randomly (Wei and Zou, 2019). This technique is called "Easy Data Augmentation" (EDA).
- Masking a random word in the message and using BERT predictions for contextual augmentation(Wu et al., 2018).
- Replacing words by their synonyms.

If the classes are balanced, there is no need to perform these augmentation techniques. Then, we can go to the next phase of the BPA framework, which is called sentiment analysis. It is essential to highlight that sometimes it is interesting to build a predictive model and evaluate it with both data sets: the original imbalanced data set and the balanced data set obtained with the data augmentation techniques. This strategy can be used to check if the model is robust enough to deal with the data imbalance problem.

Data Augmentation Revision. Depending on the text data augmentation technique, we need to revise its outputs since the new data may have a completely distinct context. For this reason, we need to revise the outputs of the data augmentation step. However, there is no need to revise the EDA outputs since this technique conserve the label of the original sentence (Wei and Zou, 2019).

3.2.2 Sentiment Analysis

This phase aims to build a predictive model to classify a text's polarity level automatically. More precisely, the model will classify short instant messages into three different classes (positive, negative and neutral).

Train Baseline Model. A baseline is a machine learning model that is simple to set up and has a reasonable chance of providing acceptable results. So, building a baseline model is usually quick and low cost. When starting on a project, the first priority is learning about what potentially unforeseen challenges will stand in the way. So, the baseline will probably not be the best-evaluated model in the project. However, it allows us to obtain initial results very quickly while wasting minimal time. In this context, we will use the original (and probably unbalanced) data set to create the baseline model. Moreover, we will initialize the baseline model with the default hyperparameters. After evaluating the baseline model, we can explore more complex models to perform better.

Search Optimal Hyperparameters with Bayes Optimization. One of the most costly steps in developing a classifier is finding the optimal values for hyperparameters. There are several methods to optimize the values of the hyperparameters, such as Grid Search, Random Search and Bayes Optimization Search. If the objective function is cheap to evaluate, we can use Grid Search or Random Search, methods that explore a large number of candidate values from the search space. However, if the objective function is expensive to evaluate, we can use Bayesian Optimization since it attempts to find the global optimum in a minimum number of steps. In the BPA architecture, we need to set many hyperparameters (batch size, class weights, dropout, learning rate etc). For this reason, we used Bayesian Optimization.

4 CASE STUDY

We evaluated the BPA approach in three different languages: English, Spanish and Portuguese. For each language, we searched for public sentiment analysis data sets and for polarity classification methods available in competitions or related works. After this initial search, we selected the Stanford Sentiment Treebank (SST-2) (Socher et al., 2013) data set available on GLUE Benchmark (Wang et al., 2019) and the SemEval 2017 Subtask A data set (Rosenthal et al., 2017) to evaluate BPA approach in the English language. For the Spanish language, we selected the TASS 2017 Task 1 General Corpus (Martínez-Cámara et al., 2017) data set and our own Spanish data set, called CCMD-ES. Unfortunately, as far as we know and searched, there is no suitable data set in Portuguese. So, we only used our own data set, called CCMD-PT, to evaluate BPA approach in Portuguese. Table 1 shows a summary of the data sets we used in the case studies.

The SST-2 data set consists of sentences from movie reviews from *RottenTomatoes* and their respective sentiment that can be positive or negative. This data set is provided with *train/dev/test* splits, since the test golden labels are not available we used the *dev* split as test in our experiments. There are 29780 negative sentences and 37659 positive sentences in the *train* split while in the *dev* split there are 428 negative and 444 positive sentences.

The SemEval 2017 Subtask A data set consists on tweets about the current trending topics at the time they were collected. The authors used CrowdFlower to perform the tweets sentiment annotation that can be positive, neutral or negative. This data set is also provided with *train/dev/test* splits. We merged the *train* and *dev* splits to build our train set and used the *test* split as is. There are 19799 positive, 22524 neutral and 7809 negative tweets in the train set while in the test there are 2352 positive, 5743 neutral and 3811 negative tweets.

The TASS 2017 Task 1 General Corpus data set has tweets about politics, economy, communication, media and culture personalities and celebrities collected between November 2011 and March 2012. The tweets sentiment annotation can be positive, neutral or negative. It is available with only the train and test splits. The train split has 2714 positive, 313 neutral and 1986 negative tweets while the test split has 22233 positive, 1305 neutral and 15844 negatives tweets.

We built Spanish (CCMD-ES) and Portuguese (CCMD-PT) data sets. These data sets consist of customer's messages from a multinational company customer service chat bot application. We manually annotated the data sets following a strict guideline provided by the company. We labeled the messages sentiments as positive, neutral or negative. Three annotators conducted the process labeling different splits of the data set. After they finished, each annotator revised the data from the other to check for disagreements. We solved disagreements by performing a round of review with all annotators.

The CCMD-ES has 21452 messages in the train split being 624 positive, 19587 neutral and 1241 negative while the test split has 518 messages being 62 positives, 255 neutral and 201 negative. Meanwhile, CCMD-PT data set has 14213 messages in the train split being 1124 positive, 11832 neutral and 1257 negative while in the test split it has 1161 messages being 112 positive, 970 neutral and 79 negative. The number of messages in the train splits, for both CCMD-ES and CCMD-PT, is already considering the data augmentation process.

Specifically for our Spanish model we also used a public data set[1] for pre-training purposes. This data set consists of positive and negative customer's reviews from Decathlon, Ebay, Tripadvisor and Filmaffinity. Since we only used it to train, we did not split into train and test sets. This data set has 38254 positive and 38254 negative reviews, but we sampled only 19127 reviews from each class due to performance limitations. In addition, we used 18996 neutral messages from CCMD-ES that we had access later in the experiments.

4.1 Results for English Language

The SemEval 2017 Subtask A competition ordered the competing models for the sentiment analysis task using Recall (Macro) as the main performance metric. The data set is imbalanced, so Recall (Macro) is a suitable evaluation metric. We also report F1-Score (Macro), accuracy and present the confusion matrix. Figure 3 shows the confusion matrix and the accuracy value for BPA approach using the SemEval 2017 Subtask A data set. It is important to highlight that BPA outperformed the SemEval 2017 Subtask A competition winner model (DataStories System) in terms of accuracy, F1-Score (Macro) and Recall. Using this data set, BPA obtained an accuracy of 0.659, an F1-Score of 0.682 and a Recall of 0.682, while DataStories System obtained an accuracy of 0.6515, an F1-Score of 0.6772 and a Recall of 0.6811. Analyzing the confusion matrix, we can observe that we have an interesting result for both positive and negative classes,

[1] https://github.com/sentiment-analysis-spanish/sentiment-analysis-model-neural-network

Table 1: Data sets summary.

Data Set	Language	Number of sentences (Train / Test)	Positive (Train / Test)	Neutral (Train / Test)	Negative (Train / Test)
CCMD-PT	Portuguese	15374 (14213 / 1161)	1236 (1124 / 112)	12802 (11832 / 970)	1336 (1257 / 79)
CCMD-ES	Spanish	21970 (21452 / 518)	686 (624 / 62)	19842 (19587 / 255)	1442 (1241 / 201)
TASS 2017	Spanish	44395 (5013 / 39382)	24947 (2714 / 22233)	1618 (313 / 1305)	17830 (1986 / 15844)
SST-2	English	68311 (67439 / 872)	38103 (37659 / 444)	- (- / -)	30208 (29780 / 428)
SemEval 2017	English	62038 (50132 / 11906)	22151 (19799 / 2352)	28267 (22524 / 5743)	11620 (7809 / 3811)

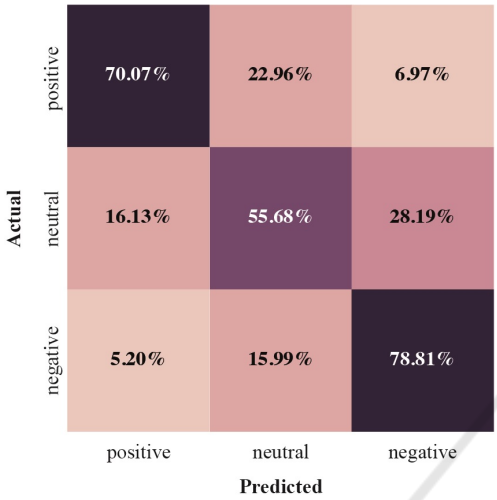

Figure 3: SemEval 2017 Subtask A confusion matrix.

but for the neutral one it performs poorly. It is related to the class weights we had to add to the classifier to tackle the class imbalance problem. Since BPA outperformed the competition winner model using the imbalanced data set we did not perform the data augmentation process. We also used the *bert-large-uncased* as the BERT model.

The GLUE Benchmark reports the evaluation of the models using the SST-2 for the sentiment analysis task using accuracy as the main performance metric. Since this data set is balanced, we can use accuracy without bias concerns. We also report the F1-Score (Macro) and present the confusion matrix. Figure 4 shows the confusion matrix and the accuracy value for BPA approach using the SST-2 data set. BPA obtained an accuracy of 0.901 and an F1-Score (Macro) of 0.901. Besides, analyzing the confusion matrix we can see that the predictions are well balanced. The best model present in the GLUE Benchmark is ERNIE (Sun et al., 2019) with an accuracy of 0.978. The BPA did not outperform ERNIE, but BPA performed well with the SST-2 data set and the english language. As we stated before, this data set is balanced, so there is no need to apply data augmentation techniques. We used the *bert-large-uncased* as the BERT model.

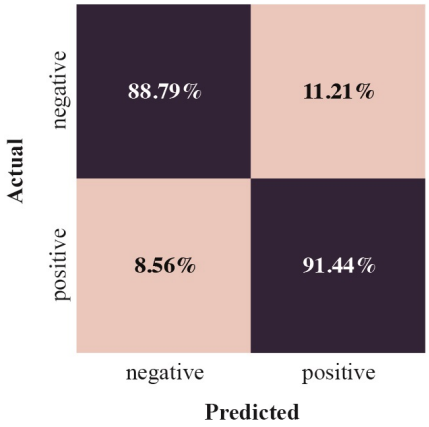

Figure 4: SST-2 confusion matrix.

4.2 Results for Spanish Language

The TASS 2017 ordered the competitors results for the sentiment analysis task using F1-Score (Macro) and accuracy. The data set is heavily imbalanced, so F1-Score (Macro) is the main performance metric for this experiment. Alongside the F1-Score and accuracy results, we also present the confusion matrix. Figure 5 shows the confusion matrix and the accuracy for the BPA approach using TASS 2017 data set. For TASS 2017 Task 1 General Corpus data set, BPA approach had a similar performance to the competition winner model (INGEOTECevodag_003 System). BPA obtained an accuracy of 0.791, an F1-Score (Macro) of 0.576. The winner of the competition achieved an accuracy of 0.645 and F1-Score (Macro) of 0.577. Even though we know that accuracy is not the best metric for this data set, they achieved an accuracy score of 0.645 while we obtained 0.791. Analyzing the confusion matrix, we can see that the minority class is not being classified properly, even with the class weights we added to tackle the class imbalance problem. Moreover, the difference between BPA approach and INGEOTECevodag_003 system is of 0.01 in terms of F1-Score (Macro). Since BPA had a similar performance to the competition winner using the original (imbalanced) data set, we decided not to perform data augmentation. We used the *BETO* (Cañete et al., 2020) as the BERT model, since it is trained with Spanish corpora.

CCMD-ES is heavily imbalanced, so F1-Score

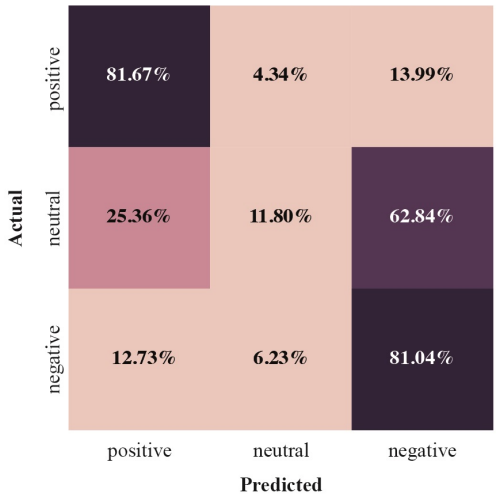

Figure 5: TASS 2017 Task 1 General Corpus data set confusion matrix.

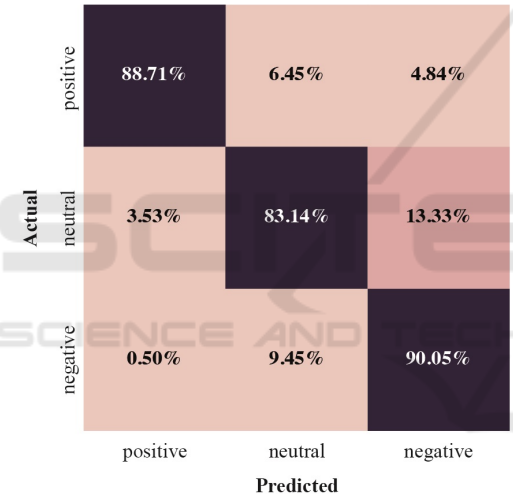

Figure 6: CCMD-ES confusion matrix.

(Macro) is a suitable evaluation metric. We also report Recall (Macro) and accuracy. Figure 6 shows the confusion matrix and the obtained accuracy for the CCMD-ES data set. BPA achieved an accuracy score of 0.865 which is an excellent result for an imbalanced data set. By the confusion matrix, we can see that the predictions are well balanced. We achieved an F1-Score (Macro) of 0.865 and a Recall (Macro) of 0.873. Thus, we can argue that BPA is a suitable model for the sentiment analysis in the Spanish language. We also used the *BETO* as the BERT model. To achieve these results, we performed all steps of the BPA framework, including data augmentation and hyperparameters optimization.

4.3 Results for Portuguese Language

CCMD-PT is also heavily imbalanced, so F1-Score (Macro) is a suitable evaluation metric. We also report Recall (Macro) and accuracy. Besides, we present the confusion matrix. Figure 7 shows the confusion matrix and the obtained accuracy for the CCMD-PT data set. BPA achieved an accuracy score of 0.923 for the CCMD-PT, which is an excellent result for a heavily imbalanced data set. In terms of F1-Macro and Recall-Macro, BPA achieved 0.91 and 0.84, respectively. So, we can argue that BPA is suitable for the sentiment analysis of chat bots messages in the Portuguese language. Since BPA achieved an excellent performance, we decided not to perform the hyperparameter optimization. We used the *BERTimbau* (Souza et al., 2020) as the BERT model, since it is trained with Portuguese corpora.

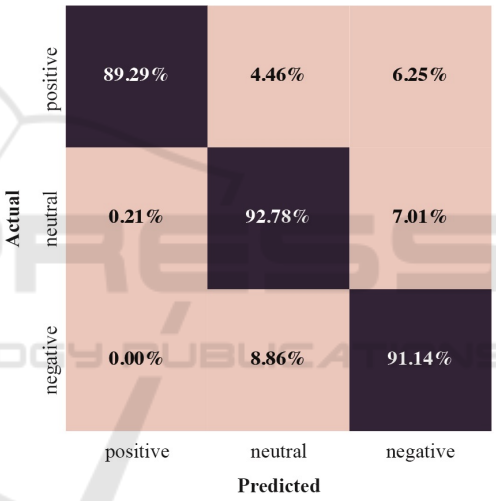

Figure 7: CCMD-PT confusion matrix.

5 CONCLUSION

In this work, we propose a new approach, called BPA, to classify the polarity level of a text in an automatic manner. We evaluated BPA using five data sets in three different languages: English, Spanish and Portuguese. Experimental results evince the suitability of BPA to multilingual and domain-independent polarity classification. Our best results achieved an accuracy of 0.901, 0.865 and 0.923 for English, Spanish and Portuguese, respectively. As future work, we intend to apply BPA in the audio messages domain and in other tasks in the customer service domain, such as predicting the level of criticality of a message to prioritize the service order in a call center.

ACKNOWLEDGMENTS

This work was partially funded by Lenovo, as part of its R&D investment under Brazil's Informatics Law, CAPES/Brazil (under grant numbers 88887.609129/2021 and 88881.189723/2018-01) and LSBD/UFC.

REFERENCES

Araújo, M., Pereira, A., and Benevenuto, F. (2020). A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences*, 512:1078–1102. https://repositorio.ufmg.br/bitstream/1843/ESBF-AQ2PSM/1/matheuslimadiniz.pdf.

Bonadiman, D., Castellucci, G., Favalli, A., Romagnoli, R., and Moschitti, A. (2017). Neural sentiment analysis for a real-world application.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Cai, Y., Huang, Q., Lin, Z., Xu, J., Chen, Z., and Li, Q. (2020). Recurrent neural network with pooling operation and attention mechanism for sentiment analysis: A multi-task learning approach. *Knowledge-Based Systems*, 203:1–12.

Chen, S., Peng, C., Cai, L., and Guo, L. (2018). A deep neural network model for target-based sentiment analysis. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Graff, M., Miranda-Jiménez, S., Tellez, E. S., and Moctezuma, D. (2018). Evomsa: A multilingual evolutionary approach for sentiment analysis. *CoRR*, abs/1812.02307.

Guo, S., Höhn, S., Xu, F., and Schommer, C. (2018). Perseus: A personalization framework for sentiment categorization with recurrent neural network. In *ICAART*.

Jianqiang, Z., Xiaolin, G., and Xuejun, Z. (2018). Deep convolution neural networks for twitter sentiment analysis. *IEEE Access*, 6:23253–23260.

Letarte, G., Paradis, F., Giguère, P., and Laviolette, F. (2018). Importance of self-attention for sentiment analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 267–275, Brussels, Belgium. Association for Computational Linguistics.

Martínez-Cámara, E., Díaz-Galiano, M., García-Cumbreras, M., García-Vega, M., and Villena-Román, J. (2017). Overview of tass 2017.

Rosenthal, S., Farra, N., and Nakov, P. (2017). SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Sarlis, S. and Maglogiannis, I. (2020). On the reusability of sentiment analysis datasets in applications with dissimilar contexts. In Maglogiannis, I., Iliadis, L., and Pimenidis, E., editors, *Artificial Intelligence Applications and Innovations*, pages 409–418, Cham. Springer International Publishing.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., and Wang, H. (2019). Ernie 2.0: A continual pre-training framework for language understanding. *arXiv preprint arXiv:1907.12412*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

Wei, J. W. and Zou, K. (2019). EDA: easy data augmentation techniques for boosting performance on text classification tasks. *CoRR*, abs/1901.11196.

Wu, X., Lv, S., Zang, L., Han, J., and Hu, S. (2018). Conditional BERT contextual augmentation. *CoRR*, abs/1812.06705.