Sentiment Analysis in Portuguese Dialogues

Isabel Carvalho¹, Hugo Gonçalo Oliveira¹, Catarina Silva¹

¹University of Coimbra, CISUC - Department of Informatics Engineering, Coimbra, Portugal

isabelc@student.dei.uc.pt, hroliv@dei.uc.pt, catarina@dei.uc.pt

Abstract

Sentiment analysis in dialogue aims at detecting the sentiment expressed in the utterances of a conversation, which may improve human-computer interaction in natural language. In this paper, we explore different approaches for sentiment analysis in written Portuguese dialogues, mainly related to customer support in Telecommunications. If integrated into a conversational agent, this will enable the automatic identification and a quick reaction upon clients manifesting negative sentiments, possibly with human intervention, hopefully minimizing the damage. Experiments were performed in two manually annotated real datasets: one with dialogues from the call-center of a Telecommunications company (TeleComSA); another of Twitter conversations primarily involving accounts of Telecommunications companies. We compare the performance of different machine learning approaches, from traditional to more recent, with and without considering previous utterances. The Fine-tuned BERT achieved the highest F1 Scores in both datasets, 0.87 in the Twitter dataset, without context, and 0.93 in the TeleComSA, considering context. These are interesting results and suggest that automated customer-support may benefit from sentiment detection. Another interesting finding was that most models did not benefit from using previous utterances, suggesting that, in this scenario, context does not contribute much, and classifying the current utterance can be enough.

Index Terms: Natural Language Processing, Sentiment Analysis, Text Classification, Dialogue

1. Introduction

Sentiment Analysis (SA) is a classification technique to determine the sentiment expressed in natural language, which, in our scenario, means dialogues. Given an utterance, dialogue systems should reply in natural language. However, the replacement of human interactions by chatbots may lead to a lack of understanding of the user’s wishes. In our work, we focus mostly on the Telecommunications (TeleCom) domain, where communication failures may lead to the loss of clients. In some cases, a customer-system conversation is going so bad that it can only be saved if a human replaces the system. SA is a contribution to suggesting the replacement, ideally, without the customer realizing it. This assessment would hopefully contribute to decreasing the number of unsatisfied clients. Hence, our objective is to explore and compare different Machine Learning (ML) models to perform SA using datasets created and labeled by us, in the Portuguese language and with a main focus on TeleCom services. We start by experimenting with shallow learning classifiers, and move on to deep learning models, focusing on the use of Transformers like BERT [1]. The work developed can be valuable, given that there is a lack of annotated datasets in this language and domains, and the application of SA to dialogues is not standard. With this in mind, one of the datasets, extracted from Twitter, has been made publicly available.

In the remainder of this article, we summarize related works, present our approach regarding the creation and annotation of the datasets and the experimentation of different models, evaluate their performance, and stress the main conclusions taken from the work and analysis developed, as well as possible directions for future work.

2. Related Work

One of the major restrictions of this work is its application to Portuguese. Hence, we focused on related work for this language, as well as work using Twitter as a source of data.

There is a survey of SA for the Portuguese language [2] that categorizes and describes works related to the task of sentiment classification. This work claims that often translating texts into English and using tools developed for that language may be more effective than efforts in Portuguese, however, this is highly debatable. Cunha et al. [3] proposed a deep neural network model to classify the sentiment polarity in YouTube video comments. Hammers and de Freitas [4] fine-tuned the pre-trained BERT model for the Portuguese language (BERTimbau) [5] to classify emotions in utterances. The model was trained on an automatic translation of the GoEmotions dataset, labeled in 28 types of emotion, and achieved a macro-averaged F1 Score of 48%. Carvalho and Silva [6] used a lexicon-based approach to classify sentiment polarity on a collection of tweets and a Portuguese book. Dosciani et al. [7] classified emotions using a Support Vector Machine (SVM) model applied to a collection of newspapers’ headlines. Souza and Filho [8] explored different strategies using BERT for binary sentiment classification of user reviews. They compared BERT’s representation with Term Frequency - Inverse Document Frequency (TF-IDF), and while the former allowed for better performance in most cases, achieving a 99% A.U.C. score, they mention that the latter represents a good trade-off between performance and computational cost. Souza et al. [9] employed a BERT-CRF architecture, combining the transfer capabilities of BERT and the predictions of Conditional Random Field (CRF). They applied this combined model to the task of Named Entity Recognition in the Portuguese language, and not SA, but it is still an interesting approach to explore. Duarte et al. [10] classified emotions present in tweets in the Portuguese language, exploiting the use of emojis, which are often connected with emotions.

Regarding the use of Twitter as a data source, one of the most studied Natural Language Processing (NLP) tasks on this social network’s data is sentiment recognition [7]. But the social network can also be used for extracting dialogues. Both of these considerations are important for our work, and along with the type of language and size, they make Twitter a good option as a data source for our work.

Pak and Paroubek [11] collected tweets, some containing emoticons, for SA and used three shallow learning classifiers, Naïve Bayes, SVM, and CRF. The latter has the particularity of considering context, which could be interesting for this work.
since a dialogue is a sequence of utterances that evolve in accordance with what was previously said. More recently, Madanian et al. [12] used Twitter to explore reactions and perceptions of the population towards Covid-19 health messaging. The three of the mentioned works explored different types of classifiers. Duarte et al. [10] and Pak and Paroubek [11] used shallow learning classifiers, but the latter considered context in one of the models, Madanian et al. [12] used a deep learning Long Short-Term Memory (LSTM) model.

De Velasco Vázquez et al.[13] used another Iberian language in their work. Spanish. They detect emotion using transcriptions from speech, as is the case with one of our datasets, but they also consider acoustic features. Their best result was using a classifier that considers context, a Recurrent Neural Network (RNN). They also created and labeled their own dataset, making their approach somewhat similar to ours.

Considering these works, we see there are several approaches to SA, and we will experiment with some of the mentioned models. Since part of the works considered context and this could be relevant for our work, we will explore two classifiers that consider previous utterances, CRF and BERT-CRF.

3. Datasets

Because we are working with a Portuguese company, with Portuguese clients, we aim at exploring SA approaches in Portuguese dialogues, ideally covering domains like Telecom, eCommerce, or Healthcare, as requested by the company.

This work had the following data requirements: the Portuguese language, the presence of dialogues and domains such as the ones mentioned above, and the association of each utterance with a sentiment. We explored several existing datasets but found no option that matched all four requirements, which justifies the choice of creating our own datasets.

Regarding annotation, we will consider a binary scenario, where 0 stands for an utterance transmitting a negative sentiment, and 1 for non-negative, since the focus is on determining when the sentiment is negative. When manually annotating the datasets, we considered four classes, ranging from very negative to positive, which were then converted to the mentioned binary labels.

We created two datasets, which will be referred to as TeleComSA and Twitter. The annotation effort involved 14 adults with varying backgrounds, and so we used the Fleiss and Krippendorff metrics to assess the agreement between the annotators. Table 1 presents an analysis of the datasets created and their agreement scores. It should be mentioned that groups of three annotators went independently through blocks of data, following a common framework containing examples of sentences to be annotated, their agreement scores. It should be mentioned that groups of three annotators went independently through blocks of data, following a common framework containing examples of sentences to be annotated, their agreement scores.

As seen in Table 1, this is the largest dataset, however, it is very unbalanced (only 18% of the samples have negative sentiment), and due to the nature of the dialogues and the automatic responses, the data tends to be repetitive.

For classification purposes, we only considered the utterance and not the speaker.

The dataset contains real conversations, so it is not available for distribution. Table 2 presents some examples of dialogues.

The Twitter dataset was created from Twitter accounts in the TeleCommunications, TV, and Health Care domains. The extraction process was done using the Twitter API[1] and during the months of April and May 2022. From each account, only tweets to which the account replied were collected, in order to ensure there was a dialogue involving the service account. Associated tweets were then collected, using the Conversation ID. A filter was applied to reduce the number of conversations involving more than two speakers.

As seen in Table 1, the samples in this dataset are much more balanced than for the TeleComSA dataset (42% are labeled as negative) and the conversations are more diverse, but the number of dialogues is smaller.

The preprocessing consisted of replacing data with placeholders, such as usernames and URLs. There were two reasons for this: they were not expected to contain much informational value, and it was a way to generalize these entities.

The distribution of this dataset is restricted by Twitter’s policy,[2] which allows only for the Twitter IDs to be shared. However, some examples of dialogues are present in table 3.

### Table 1: Analysis of the two datasets used

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Dialogues</th>
<th>#Turns</th>
<th>Avg. Turns Per Dialog</th>
<th>#Negative Samples</th>
<th>Avg. Fleiss</th>
<th>Binary Avg. Fleiss</th>
<th>Binary Avg. Krippendorff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>418</td>
<td>1,055</td>
<td>2.52</td>
<td>448</td>
<td>0.67 pm</td>
<td>0.67 pm 0.16</td>
<td>0.67 pm 0.16</td>
</tr>
<tr>
<td>TeleComSA</td>
<td>1,000</td>
<td>5,312</td>
<td>3.32</td>
<td>972</td>
<td>0.62 pm</td>
<td>0.67 pm</td>
<td>0.67 pm 0.16</td>
</tr>
</tbody>
</table>

### Table 2: TeleComSA’s dialogue examples, labeled for sentiment (S)

```
<table>
<thead>
<tr>
<th>Speaker</th>
<th>Turns</th>
<th>Utterances</th>
</tr>
</thead>
<tbody>
<tr>
<td>USER</td>
<td>1</td>
<td>Eu não gosto de pagar telefonar e o serviço é caro.</td>
</tr>
<tr>
<td>SERVICE</td>
<td>2</td>
<td>Posso ajudar em mais alguma questão?</td>
</tr>
<tr>
<td>USER</td>
<td>3</td>
<td>Obrigado por utilizar os nossos serviços.</td>
</tr>
<tr>
<td>USER</td>
<td>4</td>
<td>Lamento, mas não entendi o que quis dizer.</td>
</tr>
<tr>
<td>USER</td>
<td>5</td>
<td>Não posso ajudar com essa consulta.</td>
</tr>
<tr>
<td>USER</td>
<td>6</td>
<td>Lamento, não posso ajudar com essa consulta.</td>
</tr>
<tr>
<td>SERVICE</td>
<td>7</td>
<td>Você quer saber o saldo de sua fatura?</td>
</tr>
<tr>
<td>USER</td>
<td>8</td>
<td>Obrigado por utilizar os nossos serviços.</td>
</tr>
</tbody>
</table>
```


Table 3: Twitter’s dialogue examples, labeled for sentiment (S)

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Turn</th>
<th>Utterance</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>USER 1</td>
<td>1</td>
<td>@SERVICE Estou com velocidade de upload de internet baixíssima (1MB). Tentei ligar para a linha de apoio ao cliente e falam que há anomalias que afetam todos serviços. Esta minha situação está relacionada com essa anomalia? Obrigado</td>
<td>0</td>
</tr>
<tr>
<td>SERVICE</td>
<td>2</td>
<td>A cobertura vacinal das crianças até um ano atingiu 99% em 2021 e ultrapassou, até aos sete anos, a meta de 95% do Programa Nacional de Vacinação (PNV) da Direcção-Geral da Saúde. Os resultados são fruto do trabalho das equipes e da adesão da população à vacinação. #DGS</td>
<td>0</td>
</tr>
<tr>
<td>USER 2</td>
<td>-1</td>
<td>Tente ve em outros lugares, a vacinação perdendo a adesão da população, devido a falhas informações e ações de grupos anti-vacinas!!</td>
<td>1</td>
</tr>
</tbody>
</table>

5. Results

It is important to define a metric to evaluate the performance of the classifiers. Hence, and considering our scenario, we chose the F1 Score (macro-averaged), since it reflects the balance on how many False Negative and False Positive classifications occur. In our context, the former means that a client with negative sentiment was labeled as negative (which could mean they do not get the needed assistance and may cause them to leave the company), and the latter means that a client with non-negative sentiment was labeled as negative (which could mean they get personalized treatment, but if this happens a lot, the company could be spending resources trying to support clients that do not require much effort). Since neither scenario is beneficial, the F1 Score seems like a good metric to evaluate the performance of the classifiers.

For ease of visualization, we present graphics in which we compare the F1 Score of each classifier using context and not using context. The CRF and BERT-CRF models only present results considering context since they are classifiers that originally consider the previous utterances, unlike the remaining models. F1 Scores are presented for both datasets, TeleComSA and Twitter, and can be viewed in Figures 1 and 2, respectively.

![Figure 1: Comparison of the F1 Scores for each classifier, with and without context, using the TeleComSA dataset](image)

An analysis of the performance of each classifier reveals that the Fine-tuned BERTimbau, BERT-CRF and CRF classifiers achieve better performances than any other model when considering context, in both datasets. In the standard application (not considering context), the Fine-tuned BERTimbau model still takes the lead, but despite being a shallow classifier, the Random Forest model also presents quite high scores, in both datasets.

Regarding the Few-SL, and keeping in mind that it does not always generate the desired output, only 47% and 73% of the classifications were valid for the Twitter and TeleComSA datasets, respectively. The larger percentage of valid predictions in the TeleComSA datasets suggests that this data is better for text generation, probably because there is less variance in the samples. We did not present the results for this model without context, as the percentage of valid classifications was close to zero. Nevertheless, the score for the Twitter dataset is...
considerably worse than for the TeleComSA dataset. Yan et al. [16] alerted to the fact that in short texts, such as tweets, abbreviations are commonly used, and the sentence is normally informal, with poor grammar and misspellings, which could make it harder for a Few-SL model to generalize and gather sufficient semantic information, and possibly justify the significantly lower results.

Notably, most classifiers perform considerably worse when considering context. The only improvement was in the Fine-tuned BERTimbau model, in the Twitter dataset, by 0.01.

6. Conclusion

SA has many interested stakeholders, but it still faces many challenges, mainly because there are many different languages, dialects, and words that are continuously being invented and dropped, or their meanings keep evolving, making it a very dynamic field. Besides this, the perception of sentiments and emotions can be very subjective, which can make the classification process even more complex.

In this work, we achieved our goals of creating datasets that fit our data requirements and exploring and evaluating different classifiers for the task of SA in Portuguese dialogues. The data creation (and sharing of the Twitter dataset) is a big contribution of this work, as well as the experimentation of different models in types of text that are not often considered, the Portuguese language and the context of dialogues. The Twitter IDs of each tweet and the corresponding annotations (binary and multi-class) have been made publicly available on GitHub.4

An analysis regarding the use of context revealed that, for these datasets, the inclusion of the previous utterances actually harmed the performance of most models, contrary to what could have been expected. We note that the BERT-CRF combination performed better than the CRF model alone, but worse than the Fine-tuned BERTimbau, which may have been due to the consideration of context by the CRF part of the model. Only one model performed slightly better when considering context, the Fine-tuned BERTimbau model, but this only happened in the Twitter dataset. In the TeleComSA dataset, we can argue that a possibility for context not being helpful in the classification of sentiment in dialogues, is that the conversations are with a chatbot, making them less emotionally rich, especially when the user knows they are not speaking with another person, as is the case. A possibility for these results in the Twitter dataset, is that the dialogues are mostly small (an average of 2.52 turns per dialogue), which could make it so that the conversation does not develop enough for context to be beneficial to the classification task.

Despite these experiments not being presented in this article, we note that the models using TF-IDF encoding did not benefit from the use of multi-grams over unigrams, except when context (the concatenation of the current and previous sentences) was involved, possibly due to the use of similar expressions or words over the dialogue. The use of dense representations (in this case, BERTimbau NLI)4 instead of the sparse representations given by TF-IDF improved the performance of the classifiers. The Fine-tuned BERTimbau still outperformed these models, but the CRF model with this dense representation performed similarly to it, enforcing the importance of semantically meaningful embeddings for NLP.

Overall, and despite the difference in the size of both datasets, the models trained on the Twitter dataset generally perform better than those trained on the TeleComSA dataset. This may be due to the quality of the data samples, since the latter dataset suffers from some limitations regarding the underlying speech-to-text conversion, making some utterances hard to understand even for people (confirmed during the annotation process). In all datasets and scenarios, the Fine-tuned BERT achieved the highest F1 scores, 0.87 and 0.93: the first in the TeleComSA dataset, without considering context, and the second in the Twitter dataset, while considering it.

Future work could involve a review of the annotations with a lower agreement, so as to improve the quality of the datasets. We also intend to enlarge our datasets, which could be achieved manually, similarly to what we have done, or through active learning, as a way to accelerate the annotation process. A GPT model (e.g., GPT-2 or GPT-3) could also be used to generate new data based on the TeleComSA dataset, making it available for the public. Regarding the usually promising Few-SL and Zero-SL, further experimentations could be performed, using different base models, decoding algorithms (in the case of the former), and more semantically rich hypotheses (in the case of the latter).

7. Acknowledgements

This work was supported by: project FLOWANCE (POCI-01-0247-FEDER-047022), co-financed by the European Regional Development Fund (FEDER), through Portugal 2020 (PT2020), and by the Competitiveness and Internationalization Operational Programme (COMPETE 2020); project POWER (POCI-01-0247-FEDER-070365), co-financed by the European Regional Development Fund (FEDER), through Portugal 2020 (PT2020), and by the Competitiveness and Internationalization Operational Programme (COMPETE 2020); and by national funds through FCT, within the scope of the project CISUC (UID/CEC/00326/2020) and by European Social Fund, through the Regional Operational Program Centro 2020.

4https://huggingface.co/ricardo-filho/bert-portuguese-cased-nli-assin-assin-2

---

Figure 2: Comparison of the F1 Scores for each classifier, with and without context, using the Twitter dataset

---

3https://github.com/NLP-CISUC/twitter_sentiment_analysis/
8. References


