IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 13 July 2023, accepted 3 August 2023, date of publication 14 August 2023, date of current version 17 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3304633

RESEARCH ARTICLE

The Importance of Context for Sentiment Analysis in Dialogues

ISABEL CARVALHO, HUGO GONÇALO OLIVEIRA, AND CATARINA SILVA, (Senior Member, IEEE)

Centre for Informatics and Systems of the University of Coimbra, Department of Informatics Engineering, University of Coimbra, 3030-290 Coimbra, Portugal

Corresponding author: Isabel Carvalho (isabelc@dei.uc.pt)

This work was supported in part by the Project FLOWANCE, Co-Financed by the European Regional Development Fund (FEDER), through Portugal 2020 (PT2020) under Grant POCI-01-0247-FEDER-047022; in part by the Competitiveness and Internationalization Operational Program (COMPETE 2020), Project POWER, Co-Financed by the European Regional Development Fund (FEDER), through Portugal 2020 (PT2020) under Grant POCI-01-0247-FEDER-070365; in part by the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI; in part by the Competitiveness and Internationalization Operational Program (COMPETE 2020); in part by the National funds through Foundation for Science and Technology (FCT), within the Scope of the Project Centre for Informatics and Systems of the University of Coimbra (CISUC) under Grant UID/CEC/00326/2020; and in part by the European Social Fund, through the Regional Operational Program Centro 2020.

ABSTRACT Sentiment Analysis (SA) can be applied to dialogues to determine the emotional tone throughout the conversation. This is beneficial for dialogue systems because it may improve human-computer interaction. For instance, in case of negative sentiment, the system may switch to a human operator who can handle the situation more effectively. However, given that dialogues are a series of utterances, the context, including the previous text, plays a crucial role in analyzing the current sentiment. Our aim is to investigate the importance of context when monitoring the sentiment of every utterance during a conversation. To accomplish this goal, we assess sentiment analysis in dialogues with varying levels of context, specifically differing in the number and author of preceding utterances. We conduct experiments on Portuguese customer-support conversations, with each utterance manually labeled as having negative or non-negative sentiment. We test a wide range of text classification approaches, from traditional, as simplicity should not be overlooked, to more recent methods, as they are more likely to achieve better performances. Results indicate that the relevance of context varies. However, context assumes particular value in human-computer dialogues, when considering both speakers, and in shorter human-human conversations, when focusing on the client. Moreover, the best classifier for both scenarios, based on BERT, achieves the highest scores when considering the context.

INDEX TERMS Sentiment analysis, dialogue analysis, context awareness, natural language processing, deep learning, machine learning.

I. INTRODUCTION

The use of dialogue systems has become increasingly common, as businesses are interested in optimizing their workflow and human resources while keeping their clients satisfied. Sentiment Analysis (SA) classifies the sentiment conveyed in natural language. Arava et al. [1] presents an overview of SA, including its units of analysis (commonly document and sentence) and applications. The domains of application for SA differ, from more critical uses, e.g., stock market and electoral predictions, to more trivial ones, e.g., recommendation systems and box-office revenue prediction.

In this work, the unit of analysis is the dialogue. When applied to dialogue systems, there are several scenarios where the inclusion of sentimental context information can be relevant, i.e. it can be important whenever a dialogue or thread is involved. Hence, some scenarios include SA in product reviews, since current reviews can be influenced by previous ones; SA in social media, since this type of platform is essentially a large chat; or SA in healthcare helplines, since context can be critical to predict the sentiment of someone in a crisis. In the scenario of this work, SA may help businesses determine when a client is or is not pleased with their service.

The associate editor coordinating the review of this manuscript and approving it for publication was Renato Ferrero.

In the latter case, personalised attention may be required. SA contributes to the identification of this type of situation and, hopefully, to decrease the number of unsatisfied clients, as it allows for a quick reaction upon the determination of negative sentiment. In this work, we collaborate with a telecommunications (TeleCom) company that will allow us to assess SA in an industry scenario.

Machine Learning (ML) is becoming increasingly effective in solving the aforementioned problems. Natural Language Processing (NLP) has also made significant strides in developing a range of techniques to address SA. These approaches vary from traditional methods such as logistic regression to more recent approaches such as transformers, providing a plethora of options. More specifically, Language Models (LM) like Bidirectional Encoder Representations from Transformers (BERT) [2] are being increasingly utilized in a wide range of scenarios, from natural language understanding and generation to recommendation systems and machine translation [3], [4], [5].

Sentiment Analysis (SA) has potential applications in analyzing the importance of context in dialogues. Since dialogues typically involve multiple utterances with inherent contextual dependencies, it is crucial to determine the number of relevant utterances for effective SA. Additionally, as each speaker in the conversation (i.e., the client and service representative) plays a distinct role, it is beneficial to investigate the relevance of each speaker for the classification of the sentiment.

In this work, we focus on SA applied to customer support in Portuguese, using the background of a TeleCom helpdesk, where dialogue systems are largely used and a lack of understanding by the system could cause the loss of the client, harming the businesses' revenue and reputation.

Hence, this work addresses the following research questions:

- RQ1 When does the inclusion of context improve the model's performance compared to excluding previous utterances?
- RQ2 How does speaker selection assist in classifying sentiment in dialogues?
- RQ3 How many prior utterances are necessary to enhance the model's performance?
- RQ4 Which classifier and representation technique is better suited for SA in dialogues?

To address these questions, we design different types of context and compare the performances of several ML models in each situation, from Logistic Regression and Conditional Random Fields (CRF) [6] to fine-tuned BERT [2] models, BERT-CRF, and Few-Shot Learning [7] using GPT-3 [8] and OPT [9]. We use datasets created and labeled by a team of annotators, in the Portuguese language, as there is a lack of annotated data in this language and domain. The main contributions of this work are summarized as follows:

- We propose seven levels of context to evaluate their impact on sentiment analysis in customer-service dialogues. We have determined the scenarios where context is beneficial and how much of it is needed;
- We extend and make available a Portuguese dialogue dataset annotated for sentiment analysis, doubling the number of samples from the previous version. This facilitates further research in NLP and SA for the Portuguese language and dialogue/context analysis;
- We develop and compare classifiers, including traditional, BERT-based, and Few-Shot Learning (FSL) approaches, to determine the best model and representation suited for SA in dialogues, evaluating their performance in the seven levels of context and with different representation techniques.

The rest of this paper is organized as follows. We first review related works in the areas of SA, dialogue, and context in Section II. In Section III, we describe the proposed approach to tackle the research questions from the data curation to the model evaluation stages. Then, in Section IV, we present and discuss the performance of different models. Finally, in Section V, we summarize the main findings and discuss potential future work in this domain.

II. RELATED WORK

This work is based on customer-support dialogues in the Portuguese language, for which there are major restrictions regarding the data used. Hence, in this section, we present related work that faces some of our challenges: the use of Portuguese, the use of Twitter or manually-labelled data, and the consideration of context. Table 1 presents a summary of the surveyed works.

A survey [10] that addresses the work developed in the domain of SA in Portuguese states that some advances are still required to make better use of the language. In fact, due to the difference of maturity of available tools, they claim that the translation of data into English and the use of tools for that language may lead to better results, yet, this is hard to affirm.

Pak and Paroubek [11] and Duarte et al. [12] proposed approaches that could possibly reduce the impact of using other languages by exploiting emojis and emoticons. The former extracted a corpus of tweets and annotated their sentiment based on emoticons. They applied classifiers such as SVM, CRF, and Naive Bayes, the latter achieving the highest performance, with F-score close to 70%. The latter followed the same approach to annotate emotion using emojis. They applied Naive Bayes and SVM for classifying and predicting emojis, obtaining their highest F1-score of 70.8% for emotion classification and 23.7% for emoji prediction, using the former classifier. This type of approach can be limiting as not all tweets contain emoticons or emojis and these icons may not be able to express the correct sentiment in the whole tweet.

A more recent study [13] used a public corpus of tweets for SA and traditional classifiers (SVM, Random Forest, Decision Trees, and Logistic Regression). They conclude that Decision Trees outperformed the remaining algorithms, achieving an F1-Score of 86%. This work employed Term

TABLE 1. Summary of the related works.

Authors	Year	Source of Data	Dialogues	Task	Language	Representation Technique	Classifier(s)	Evaluation Metric(s)
Pak and Paroubek	2010	Twitter	No	SA	English	N-grams	SVM, CRF, Naive Bayes	F-score, Accuracy
Duarte *et al.*	2019	Twitter	No	Emotion Recognition	Portuguese	TF-IDF	SVM and Naive Bayes	F1-score, Accuracy
Saad and Yang	2019	Twitter	No	SA	English	TF-IDF	SVR, Random Forest, Decision Trees, Logistic Regression	F1-score, Accuracy
Souza and Filho	2022	User Reviews	No	SA	Portuguese	BERTimbau, m-BERT, TF-IDF	Logistic Regression	AUC
Roy *et al.*	2023	Twitter	No	SA and Hate Detection	English	TF-IDF, N-grams and Word Cloud	LSTM, SVM, Naive Bayes, Random Forest,...	F1-score, Accuracy, Precision, Recall
Fernandez *et al.*	2022	Trading Applications	No	SA and Slang Recognition	Indonesian	IndoBERT	Fine-tuned IndoBERT	Accuracy
Souza *et al.*	2019	Several texts	No	NER	Portuguese	BERTimbau, m-BERT	Fine-tuned BERT-CRF, m-BERT	F1-score
Tan *et al.*	2022	Social Media	No	SA	English	RoBERTa	RoBERTa-LSTM	F1-score, Accuracy
Ling *et al.*	2022	Dialogue Systems, Crowd-sourcing	Yes	Response Generation	English	Hierarchical Encoding (BiGRU)	Variants of CARG	BLEU
Wang *et al.*	2020	Online Customer Service	Yes	Topic-aware SA	Chinese	BERT-LSTM, GloVe	BERT, LDA-LSTM, LDA-BERT, TML, ...	F1-score
Song *et al.*	2022	News and Chit-Chat Dialogue Datasets	Yes	Aspect-based SA	Chinese	RoBERTa WWM Ext Self-attention-based	RoBERTa WWM Ext Self-attention-based	F1-score, Accuracy
Hosseini-Asl *et al.*	2022	Movie Reviews, Customer Reviews	No	Aspect-based SA	English	Text	GPT-2	Accuracy

Frequency-Inverse Document Frequency (TF-IDF) to represent each tweet, a traditional approach that weights the importance of each word in a document, given a collection of documents. Another study [14] compared this type of representation to the one produced by a BERT [2] model, a popular approach for NLP problems. They used a pre-trained BERT model for the Portuguese language, BERTimbau [15], and concluded that, while TF-IDF presents a good balance of computational cost and performance, BERT representations achieve the highest scores in most cases.

Several current studies rely on transformers such as BERT. Roy et al. [16] compared several traditional approaches with a BERT-based model for SA, in particular hate detection. They determined that approaches like Naive Bayes, Logistic Regression, k-NN, and Long Short-Term Memory Networks (LSTM) can perform equally or better than BERT, which stresses the importance of including solutions of different complexities in any machine learning study. The work by Fernandez et al. [17] used manually-labeled data and a fine-tuned Indonesian BERT model [18] for SA on Indonesian stock messages that included slang. This outperformed previous studies and achieved 60.5% accuracy in the prediction of sentiment and recognition of slang. Transformers have also been combined with other models such as CRF, LSTM, or simple fully connected layers [19], [20].

CRF has been commonly used for sequence tagging tasks, such as Named Entity Recognition (NER) or Part-of-Speech tagging, due to their architecture that allows the exploitation of context [21]. Souza et al. [22] used a BERT-CRF architecture for NER in Portuguese, achieving better performances than previous studies with their fine-tuning approach. They suggest the future experimentation of another transformer, RoBERTa [23], which they claim could be more efficient. In fact, a more recent work by Tan et al. [24] combined this model with an LSTM classifier for the task of SA in English tweets, and achieved F1-scores of 93%, 91%, and 90% over three different datasets. However, none of the previous works has considered context for SA.

Poria et al. [25] wrote on the challenges of a similar task to sentiment analysis, emotion recognition in conversation and claim that it is dependent on three factors: (i) the current utterance and its context, i.e., previous utterances, intent, and topic; (ii) the speaker's state and personality; and (iii) the emotions expressed in the previous utterances. They also suggest that speaker-specific emotion recognition could be an interesting approach.

Ling et al. [26] used a concatenation strategy to include context in short dialogues, which tackles the first of the three factors mentioned, however, their work was focused on response generation. Wang et al. [27] applied SA to customer service dialogues, and propose a topic-aware approach, which also seems to tackle the first factor. They experimented with several classifiers, including BERT, LDA-LSTM, and LDA-BERT, and several multi-task scenarios involving topic information, which allowed them to outperform several baselines. Song et al. [28] developed several baseline models based on RoBERTa or self-attention to perform aspect SA in dialogues. They determined that RoBERTa has better performance, further validating the idea that this model is a good option for this type of task. Moreover, the authors consider speaker selection but only as a binary feature signaling if a previous token in the dialogue history is from the same speaker as the current utterance, in order to add more contextual information. However, they do not analyse the effect of each speaker on the model's performance. Regarding context, it is present in the form of detecting mentions of the current sentiment expressions in the dialogue history, not considering the full context. Nevertheless, they found that looking for mentions in previous utterances improved the performance when compared to only considering the current utterance, suggesting that previous utterances may improve the models' performance in SA.

Another interesting approach for considering context is to use Few-Shot Learning (FSL) [7], which allows for the meta-training of classifiers with just a few labeled samples. Usually, large models, e.g., GPT-4 [29], GPT-3 [8], GPT-Neo [30], GPT-2 [31], and Meta-OPT [9], are used for this

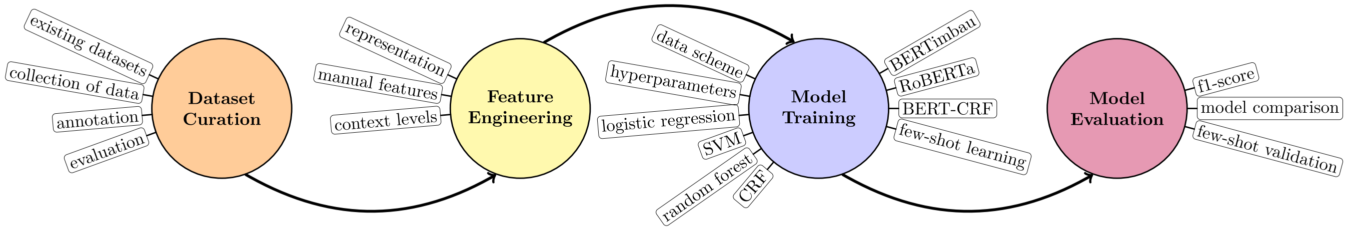

FIGURE 1. Representation of the general approach followed.

type of approach and their main task is text generation. These large models are trained on extensive data, mostly based on The Pile [32], a combination of 22 diverse datasets that culminates in 825GiB of text. Due to their broad and intensive training, these models are more likely to be able to generalise with little extra information. Hosseini-Asl et al. [33] employed GPT-2 in a few-shot learning strategy to perform SA, but in an aspect-based approach. Nevertheless, GPT-2 outperformed BERT-based approaches while using less than 20% of the training data.

In the summary presented in Table 1, it is visible that there is much work on SA, using shallow classification approaches and more recent ones, based on fine-tuning transformers. Both types of approaches are included in our study. On the other hand, only a minority of works consider SA in dialogue. Out of them, none is in Portuguese and none analyses speaker-specific context, which are the focus of this study. One could argue that the work by Song et al. [28] presents some sort of speaker analysis, but there is no specific identification of the speaker and no analysis of its effect on performance, which this work provides.

III. PROPOSED APPROACH

In this section, we present the approach adopted for answering the defined research questions. Figure 1 showcases the general process, typical of ML problems. We go in-depth into each of the four steps: dataset curation, feature engineering, model training, and model evaluation.

A. DATASET CURATION

There are several options of data sources for SA, depending on the desired application. A common option is to crawl through social networks or use existing datasets. Other possibilities include online portals, blogs, news websites, and reviews [1].

Given the nature of this study and the scenario of the TeleCom company, there were some requirements regarding our data, namely:

1) Written dialogues;
2) In Portuguese;
3) In domains where customer-support is especially relevant (e.g., Telecommunications, TV, e-Commerce);

A search for available datasets showed that none of the existing sources was a match for all requirements, as seen in Table 2, which includes whether they were annotated or not. Further details on DailyDialog [34], Mastodon [35],

TABLE 2. Analysis of dataset requirements.

Dataset	Dialogues	Portuguese	Annotation	Domain(s)
DailyDialog	Yes	No	Yes	No
Mastodon	Yes	No	Yes	No
Friends' Emotion Detection	Yes	No	Yes	No
CORAA	Yes	Yes	No	No
Emotion in News	No	Yes	Yes	No
ReLi	No	Yes	Yes	No
Sentituites-PT	No	Yes	Yes	No
Wizard of Wikipedia	Yes	No	No	No
Multi-WoZ	Yes	No	Yes	No
CamRest	Yes	No	Yes	No
Ubuntu Dialog	Yes	No	No	Yes

Friends' Emotion Detection [36], CORAA [37], Emotion in News [38], ReLi [39], Sentituites-PT [40], Wizard of Wikipedia [41], Multi-Woz [42], CamRest [43], and Ubuntu Dialog [44] can be found in Appendix B.

Hence, there was a need to create a dataset that would meet our needs. This resulted in the curation of two datasets, one with data provided by the company, and that cannot be disclosed, and another extracted from Twitter.

Starting with the former, hereafter TelecomSA, it contains conversations from the company's clients with their customer-support dialogue system and is focused on the domain of telecommunications. Sensitive information was filtered by the company and the need for preprocessing was minimal.

Regarding the other dataset, hereafter TwitterDialogue-SAPT, domains include not only Telecom-related data, but also conversations about Television (TV), Healthcare, eCommerce, and Finance & FinTech. A table presenting the accounts from which we extracted data and their most representative domain can be found in Appendix A. Some topics are not as common, making them less represented. However, eCommerce could be included in some of the TV or TeleCom domains. The creation process included four steps:

1) The selection of Portuguese customer-support accounts from the relevant domains, as mentioned;
2) The extraction of tweets that these accounts had replied to, using the Twitter API[1], and ensuring a two-way conversation, as services do not reply to every tweet identifying them;

[1] https://developer.twitter.com/en/docs/twitter-api

TABLE 3. Size of each dataset.

Name	Version	Date	#Dialogues	#Utterances	Avg. Turns per Dialogue
TeleComSA	1.0	Oct. 2022	1,000	5,312	5.32
TwitterDialogueSAPT	1.0	Oct. 2022	318	954	2.52
TwitterDialogueSAPT	2.0	Feb. 2023	916	2,285	2.49

TABLE 4. Class frequency and agreement scores for each dataset.

Name	Version	Date	#Utterances	#Negative Utterances	#Other Utterances	Fleiss' Kappa	Krippendorff's Alpha
TeleComSA	1.0	Oct. 2022	5,312	972	4,340	0.62	0.67
TwitterDialogueSAPT	1.0	Oct. 2022	1,055	448	607	0.67	0.67
TwitterDialogueSAPT	2.0	Feb. 2023	2,285	915	1,370	0.67	0.67

3) The collection of the full conversations containing the identified tweets;
4) The selection of dialogues involving only two accounts, the service and the user, as Twitter dialogues usually involve third parties.

Steps 3 and 4 were achieved through the Twitter API's parameters *Conversation ID*, which matches the ID of the tweet that started the conversation and is present in all subsequent tweets, *In Reply to User ID*, which contains the ID of the author of the tweet it replies to, and *Author ID*, which identifies the author of the current tweet. The two latter parameters allowed us to collect the conversation between the user and the service, without third parties.

To remove sensitive information, reduce bias, and make the data more general, the preprocessing for this dataset includes the replacement of the user handles and of the URLs by specific placeholders.

Table 3 presents the dimension of each dataset. TeleComSA and the first version of TwitterDialogueSAPT (v. 1.0) were presented in a previous work [45], but, in the scope of this work, the latter was extended (v 2.0) and was made available on GitHub[2]. Whereas this new version nearly matches the number of dialogues in the largest dataset (TeleComSA), the number of utterances is 43% lower due to the shorter conversations (see the average number of turns per dialogue). TwitterDialogueSAPT (v. 1.0) presents tweets collected during April–May 2022 whereas v 2.0 includes data extracted in November–December 2022.

With the data collected and the datasets defined, in order to adopt a supervised approach, we still require utterances to have their sentiment labeled. Hence, each dataset was annotated by three people for a multiclass scenario where sentiment ranged from -2 (very negative) to 1 (positive). Each person involved received a guideline containing sentence examples of each level of sentiment, a brief explanation of the purpose of this work and what to look for, and were made comfortable to express any doubts to the authors. They were specifically asked to consider the context and not to label the sentiment of each utterance without regard for the previous. The annotator's backgrounds, gender, and age are diverse, but people involved in computer science and/or belonging to the 18-30 age group are better represented.

Despite the availability of the multiclass annotations, we use a binary version that defines sentiment as negative (0) and non-negative (1). This makes the classification process less complex and still answers our needs, because the main goal of our task is to discriminate negative sentiment from the rest. Table 4 presents the class frequency in each dataset and the evaluation of the agreement between the annotators, using Fleiss' Kappa [46] and Krippendorff's Alpha [47], common metrics for this purpose [48], which consider more than two annotators per sample. From the observation of the table, we verify that whereas the dimension of the Twitter dataset has more than doubled, the level of agreement has not been negatively affected by this expansion. In fact, for all datasets, the agreement between annotators is considered substantial (Fleiss' Kappa in between 0.61–0.80) [49] and acceptable for tentative conclusions (Kripperndorff's Alpha over 0.667) [50]. It is also visible that the negative class is less represented in our datasets, especially in the TeleComSA, where it is only in 18% of the total utterances. This unbalance is much lower in the TwitterDialogueSAPT data, which contains 40% of negative samples. The other utterances in both datasets mainly showcase a neutral sentiment: TeleComSA presents less than 1% of the data labeled as positive, whereas TwitterDialogueSAPT (v. 2.0) contains 11% of positive utterances, doubling the presence of this class in comparison with the previous version (v. 1.0), where it was only present in 5.6% of the dataset.

For a better perception of the contents of our datasets, some examples of dialogues can be seen in Tables 7 and 6, respectively regarding TeleComSA and TwitterDialogueSAPT. For those that do not speak Portuguese, we include a rough translation of each utterance. The Portuguese utterances were not omitted and are identified with the colour blue. Table 7 presents an example of the importance of context in Turn 3 of Dialog ID 1, which was verified by comparing the predictions of the RoBERTa model (to be introduced in Section III-C) with and without context analysis. When considering context, it is noticeable that the user is simply replying to the service's question. However, when analysing only that utterance, the classifier would signal it as containing negative sentiment,

[2]https://github.com/NLP-CISUC/TwitterDialogueSAPT

TABLE 5. Context scopes levels.

Utterance(s)	Speaker(s)
Current	Customer
Current and previous	Service
Current and two previous	Customer and service

possibly because the user is seemingly having difficulties with their cellphone.

At the end of this step, we have two complete datasets, but still require their data to be represented in such a way that it is usable by the classifiers. This is discussed next.

B. FEATURE ENGINEERING

The primary objective of this study is to assess the importance of context in SA for dialogues and determine the optimal amount of context for this task. To achieve this goal, we begin by defining the scope of context in our study, which we establish as different levels derived from the combinations of values specified in Table 5. Regarding the levels using the service as the only speaker, we only consider the current utterance, as for our task, the focus is on the sentiment of the customer. Hence, we will consider seven levels of context: the three utterance-levels for the customer (3), the three utterance-levels for both speakers (3), and the current utterance for the service (1). There are four main reasons for considering these seven levels of context due to:

1) Token limitations in some of the models, no more than two previous utterances could be considered. Hence, at the utterance level, there are only three possible options: the current utterance, i.e., no context, the current and previous utterances, and the current and two previous utterances;
2) The nature of this work, the dialogues were limited to two speakers. Hence, at the speaker level, there are also three options: the customer utterances, the service utterances, and the utterances from both speakers;
3) The focus on the customer, so that customer-support services can intervene when needed, we remove the context analysis from the service utterances, as mentioned;
4) The existing studies [25], [26] which reinforce the importance of including context in SA in dialogues, and of considering the speaker's state.

Since the dialogues in our datasets are relatively short, especially in the TwitterDialogueSAPT dataset, we consider that the use of up to two previous utterances is enough contextual information. Furthermore, some models have token limitations that do not allow for more than two or three utterances to be considered.

The classifiers are able to use the data based on their features or representations.

We explored two different representations for the shallow learning classifiers: Term Frequency - Inverse Document Frequency (TF-IDF) and a Portuguese BERT (BERTimbau) fine-tuned for Semantic Textual Similarity (STS)[3]. The former is considered a traditional approach that weights the importance of each word in a document based on its occurrences in the training data. The latter is a sentence transformer, fine-tuned in pairs of Portuguese sentences and their semantic similarity, which can be used for encoding sequences of text. For the BERT-based classifiers, the samples were represented using the corresponding tokenizers, which break some words in subwords, and includes not only a token embedding for each but also a position embedding and a segment embedding identifying each sentence.

It should be noted that some of the models used inherently consider context (CRF and BERT-CRF), but for the remainder of the classifiers, context is simulated by concatenating the number of utterances to consider, when they exist in the dialogues. Hence, for the CRF-based approaches, we also include manual features, as to represent the previous sentences and to include their associated sentiment (whether labeled, during training, or predicted, during testing), which, as seen in Section II, is an important part of defining context in a dialogue.

Furthermore, these are the handcrafted features defined for the CRF model:

- **Turn/Longest Dialog**: Current turn number in comparison with the largest number of turns in the training data;
- **Number of Words**: Number of tokens in the sentence, obtained with the spaCy[4] toolkit;
- **Has Question**: Whether the sentence contains a question mark;
- **Has Exclamation**: Whether the sentence contains an exclamation mark;
- **Beginning of Speech (BOS)**: Whether the sentence is the first in the dialogue;
- **Previous labels**: The sentiment of the previous utterances, if applicable. In training, the label is used, otherwise, the predicted sentiment is used;
- **Encodings**: The TF-IDF or STS embeddings of the previous utterances, if applicable, and the current.

Regarding the BERT-CRF model, the handcrafted features considered when feeding the CRF part of the model are as follows:

- **Class Probabilities**: The probability of the current utterance belonging to each of the possible classes;
- **BERT's Prediction**: The classification label of the previous utterances, if applicable, and the current, using the BERT model;
- **BOS**: Whether the previous utterances, if applicable, and the current are the first in the dialogue;
- **Encodings**: The BERT embeddings of the previous utterances, if applicable, and the current, obtained from the model's last hidden layer.

At the end of this step, the data can now be represented in a way that is computable by the classifiers, but they still need to learn how to perform the SA task. The training stage is discussed next.

[3]https://huggingface.co/rufimelo/bert-large-portuguese-cased-sts

[4]https://spacy.io/models/pt#pt_core_news_sm

TABLE 6. Twitter's dialogue examples, in Portuguese (blue) and English (black), labeled for sentiment (S: 0 for negative, 1 for non-negative).

Dialog ID	Speaker	Turn	Utterance	S
1	USER	1	@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 @SERVICE My internet upload speed is very low (1MB). I tried to call the client-support line and they told me there are anomalies affecting all services. Is my situation related to this anomaly? Thank you	0
	SERVICE	2	Olá @USER, respondemos à sua mensagem privada. Obrigado e até breve! Hello @USER, we have replied to your private message. Thank you and see you soon!	1
2	SERVICE	1	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 Direção-Geral da Saúde. Os resultados são fruto do trabalho das equipas e da adesão da população à vacinação. #DGS The vaccination coverage of children up to one year old reached 99% in 2021 and surpassed, up to seven years old, the target of 95% of the National Vaccination Programme (PNV) of the Direção-Geral da Saúde. These results are the product of the teams' work and the population's participation in vaccination. #DGS	1
	USER	2	@SERVICE Sensacional !! Um bom trabalho gera bons resultados ! Triste ver em outros lugares, a vacinação perdendo a adesão da população, devido a falsas informações e ações de grupos antivacinas!! @SERVICE Outstanding !! Good work produces good results ! Sad to see that in some places, vaccination is losing the participation of the population, due to false information and the actions of anti-vaccine groups!!	1

TABLE 7. TeleComSA's dialogue examples, in Portuguese (blue) and English (black), labeled for sentiment (S: 0 for negative, 1 for non-negative).

Dialog ID	Speaker	Turn	Utterance	S
1	USER	1	vai buscar o livro o caderno o telemóvel o telemóvel bloqueou go get the book, the notebook, the cellphone blocked	0
	SERVICE	2	Precisa do pin e puk ou quer bloquear o seu cartão por perda ou furto? Se pretende tratar de outro assunto diga voltar atrás Do you need the pin and puk or do you want to block your card due to loss or theft? If you want to talk about another topic say go back	1
	USER	3	sim é o telemóvel bloqueou tenho aqui o puk e o e o número e o pin yes, the cellphone blocked, I have here the puk, the number, and the pin	1
	SERVICE	4	Lamento não ter percebido corretamente o que me disse. Vou transferir a sua chamada para o menu de atendimento automático I am sorry but I could not understand what you said. I will transfer your call to the automatic service menu	0
2	USER	1	oi eu quero saber porque é que este mês pago uma fatura tão grande 25 e 76 eu nunca paguei isto hi, I want to know why I'm paying such a big bill this month, 25 and 76, I never paid this	0
	SERVICE	2	Lamentamos, mas não entendemos o que disse, pode repetir por outras palavras We are sorry but we did not understand what you said, could you repeat in other words	0
	USER	3	outras palavras é isso é sobre a faturação deste mês como apareceu aquilo essa fatura 2576 e eu nunca paguei este valor e diz que não sei o que é que foi in other words it's about this month's bill, how did that bill appear 2576 and I never paid this value and it says I don't know what it was	0
	SERVICE	4	Pretende saber os dados pagamento do numero de onde está a ligar? Se pretende de outro numero diga outro numero, ou se não pretende saber dados de pagamento diga voltar Atrás Do you wish to know the payment data of the number you're calling from? If you wish another number's say other number, or if you don't wish to know payment data say go back	1
	USER	5	queres saber o nosso pagamento sim porque é que eu estou a pagar caraças pá You want to know our payment, yes, why am I paying this, man, damn, man	0
	SERVICE	6	Tem a pagar .. por débito bancário até ao dia... Posso ajudar em mais alguma questão? You owe ... by bank debit until the day.... Is there anything else I can help you with?	1
	USER	7	olha essa questão saiu que tenho aqui na fatura eu quero saber o porquê o porquê o porquê dessa fatura estar tão elevada que nunca Look, one question, it says here in the bill, I want to know why, the reason why, why this bill is high as never	0
	SERVICE	8	Obrigado por utilizar os nossos serviços. Thank you for using our services.	1

C. MODEL TRAINING

Having defined two datasets, seven levels of context, and how to represent our data, we focus on the used classifiers and their training.

We experiment with multiple solutions, from shallow learning classifiers, such as Logistic Regression, Radial Basis Function (RBF) SVM, Random Forest, and CRF, to deep learning classifiers, mostly based on BERT (fine-tuned BERTimbau, fine-tuned RoBERTa, BERT-CRF), but we also explored GPT-3 and OPT in a Few-Shot Learning (FSL) approach. These options were mostly inspired by the related work, presented in Section II.

The training-testing split (75%-25%) is the same for all experiments, meaning that each trained model learned from and is evaluated in the same data, allowing for a fair comparison between each approach.

In the remainder of this subsection, we briefly present each classifier and its hyperparameters. We do not go in-depth about the meaning and choice of each hyperparameter as they are only presented here for replicability. A table presenting the packages used and respective versions is available in Appendix C. Furthermore, Appendix D presents the specifications of the machines used in the development of the mentioned approaches.

- Logistic Regression:
  This traditional type of classifier uses a linear combination of the features to reach a conclusion on which class a sample belongs to. Regarding hyperparameters, we used the L-BFGS solver, an L2 penalty, a maximum value of iterations of 100, and a C value of 1.
- Support Vector Machines (SVM):
  This traditional type of classifier uses a kernel function to transform the data and find a hyperplane that will allow it to split the data by its class. Regarding hyperparameters, we used the RBF kernel, a scaled gamma, and a C value of 1.
- Random Forest:
  This traditional type of classifier is an ensemble. It combines the output of several decision trees to reach a final decision on the class of a given sample. Regarding hyperparameters, we used the Gini criterion, did not define a maximum depth, used the square root for defining the maximum number of features, considered 100 estimators, enabled bootstrapping, and set a minimum number of samples to be a leaf of 1 and a minimum number of samples required to split an internal node of 2.
- Conditional Random Fields (CRF):
  This traditional type of classifier is especially interesting for this work because it is suited for sequence labelling tasks, where contextual information is important. It produces a global probability for the whole utterance considering the probability of the label for each word and the transition probability between labels, returning the most likely class. Regarding hyperparameters, we used a 0 coefficient for L1 and L2 regularization and did not limit the number of maximum iterations.
- BERT:
  This type of classifier is based on transformers, which are encoder-decoder models that contain self-attention layers, allowing them to use information from large contexts. In this study, we use BERTimbau [15], a BERT model that was pretrained in Portuguese datasets. Regarding hyperparameters, we fine-tuned the model for 2 epochs, using a batch size of 16, a learning rate of 0.0001, and an epsilon of 0.00000001.
- RoBERTa:
  This type of classifier is an optimized version of BERT, trained on five English corpora. In this study, we use Twitter-XLM-Roberta-base [51], a RoBERTa model pre-trained on nearly 200 million tweets in multiple languages, including Portuguese, making it suitable for our short and informal data. Regarding hyperparameters, this model was fine-tuned in the same manner as the BERTimbau model.
- BERT-CRF:
  This type of classifier is a combination of a BERT model (in this study, BERTimbau) and a CRF model [20], [22]. CRF is best suited when contextual information affects the current state, as its graph-like nature inherently considers context. BERT provides contextual embeddings and is a powerful model with higher performance than CRF. BERT-CRF combines CRF's sequence modeling characteristics with the high power of BERT. The latter produces a representation of each sample and their predicted label and feeds this information to the former, which exploits these characteristics to provide a decision based on its contextual capabilities. Regarding hyperparameters, this model was fine-tuned in the same manner as the previous BERT-based models and retains the settings used for training the CRF model.
- GPT-3:
  This transformer-based language model is used, in this work, to explore a FSL approach. We use the Curie GPT-3 model [8], *text-curie-001*, which was fine-tuned on human-written demonstrations, and was, at the time, the largest model after DaVinci models, containing 6.7 billion parameters. The choice for this model was due to budget limitations. To generate text, we employ the sampling algorithm, with a temperature value of 0.1.
- OPT:
  Similarly, this transformer-based language model is used to explore a FSL approach. Meta-OPT [9] matches the size of GPT-3 and was trained predominantly in the English language, with an emphasis on human-generated text. We use the 2.7 billion parameters version, as preliminary experimentation showed worse results with other versions. To generate text, we employ the sampling algorithm, with a temperature value of 0.1.

In a FSL approach, the models, in this case, GPT-3 and OPT, will generate text based on a few examples received as input, e.g. *my television stopped working. sentiment negative*, *thank you for you help. sentiment non-negative*, *my phone is broken. sentiment*, and hopefully assign the correct label, which would be negative for this example. Usually, very large models are chosen for this, due to their high generalisation capability, which is expected to allow them to infer the output from a few structured examples. For our study, we model the examples in the following pattern: `|bos| current utterance`

`|eos| |sentiment| label, |bos| previous utterance |pad| current utterance |bos| |sentiment| label`, for as many dialogues as defined, and then follow with |*bos*| *new utterance* |*eos*| |*sentiment*|, at which point we expect the model to be able to complete the prompt with one of the following labels: 0, for negative sentiment, or 1, for non-negative sentiment. The `|bos|` tag represents the beginning of a dialogue, whereas the `|eos|` tag represents its end. The `|pad|` tag represents the beginning of a new turn, used when we consider context, and the `|sentiment|` tag indicates that the sentiment label (given or predicted) is represented ahead. It should be noted that this type of approach requires some processing of the text completion to assess its validity as it may generate something that is not valid, i.e., not one of the defined classes.

Whereas in the other approaches the models learn from the training dataset, in the FSL approach, we defined that the models only receive three examples of dialogues.

These examples were randomly selected from the training dialogues where all annotators agreed on the labels of all utterances. To ensure a broader range of sentiments in the examples, we use the multiclass labels and select one sample out of the very negative, negative, and neutral classes. We dismiss the positive class as this is a very under-represented label and not common in customer-service.

At the end of this step, the classifiers have been trained to perform SA, but we have not yet assessed how they perform on unseen data. The evaluation stage is discussed next.

D. MODEL EVALUATION

Once the classifiers are trained, we need a way to evaluate and compare their performances. We selected the F1 Score metric since we want to avoid both the False Negative and the False Positive scenarios. In our scenario, a False Negative would mean that a client expressing negative sentiment was wrongly classified and may not get the required assistance, possibly resulting in a loss for the business; A False Positive would mean that a client expressing non-negative sentiment was wrongly classified and the company could be spending human resources in a client that does not require the personalised support. It is common to report the Accuracy metric in SA, however, the datasets used in this work are not balanced. TeleComSA in particular only contains 18% of the samples labeled with negative sentiment. In situations like this, the accuracy metric can be misleading, e.g., if a model predicts that all utterances present a non-negative sentiment, it would result in an accuracy of 82%, which could make the reader mistakenly believe this was a good model. Hence, for simplicity and because the F1-score considers the impact of both classes by combining the recall and precision metrics, it was deemed the best option to represent performance in our datasets.

As mentioned earlier (Subsection III-C), the FSL approach may produce too much or undesired text that does not fit our needs. To validate the classification assigned, we simply determine if there is a class label present in the whole generated text, and consider that to be the determined sentiment. If no label is found, that sample is dismissed. If the percentage of valid samples is too low, this can critically reduce the applications of this approach. For a better perception of the FSL approach to assign a valid class to an utterance, in section IV we present the percentage of valid samples for each model, dataset, and context level.

At the end of this step, we have completed the general approach presented and produced comparable performance results for each model, which will allow us to take conclusions on the importance of context for SA in dialogues. The next section presents the F1 Score for each experiment and discusses the obtained results, including the percentage of valid samples in the FSL approach.

IV. RESULTS AND DISCUSSION

In order to answer our four research questions, presented in Section I, we must first compare and analyse the performances of each setting of experiments. As such, we split our analysis into four subsections, each related to one or more of the questions.

To represent each context level, the figures will use the following legend:

- **1-both**: Considers only the current utterance (1) and both speakers (full dialogue). The sentiment of both speakers' utterances is predicted;
- **2-both**: Considers the current and previous utterances (2) and both speakers (full dialogue). The sentiment of both speakers' utterances is predicted;
- **3-both**: Considers the current and the two previous utterances (3) and both speakers (full dialogue). The sentiment of both speakers' utterances is predicted;
- **1-user**: Considers only the current utterance (1) and the only speaker considered is the user (user's dialogue). Only the sentiment of the user's utterances is predicted;
- **2-user**: Considers the current and previous utterances (2) and the only speaker considered is the user (user's dialogue). Only the sentiment of the user's utterances is predicted;
- **3-user**: Considers the current and the two previous utterances (3) and the only speaker considered is the user (user's dialogue). Only the sentiment of the user's utterances is predicted;
- **1-service**: Considers only the current utterance (1) and the only speaker considered is the service (service's dialogue). Only the sentiment of the service's utterances is predicted;

There are three other aspects to take into consideration when analysing the figures:

1) The CRF-based models were not evaluated in any of the context levels considering only the current utterance, as the goal of using these models was to make use of their inherent contextual capabilities;
2) Due to token limitations, the RoBERTa model could not correctly process three utterances at once as it truncated

TABLE 8. Percentage of valid predictions for each model and context levels, for both datasets.

	TeleComSA		TwitterDialogueSA	
	GPT-3	OPT	GPT-3	OPT
1-both	99.25%	95.64%	89.63%	98.76%
2-both	98.42%	82.41%	84.23%	95.85%
3-both	99.40%	80.68%	76.76%	92.53%
1-user	100.00%	93.98%	91.04%	98.51%
2-user	98.35%	84.96%	87.31%	98.51%
3-user	99.70%	93.08%	76.87%	98.51%
1-service	98.80%	95.34%	89.72%	95.33%

the input, which is why very low results were obtained in the TwitterDialogueSAPT dataset, and no results at all in the TeleComSA dataset, as it was not able to classify these dialogues;

3) For ease of visualisation, model names are presented in short forms, such as RF for Random Forest and Log Reg for Logistic Regression.

Before delving into our SA results, and as mentioned earlier, it is important to consider how many of our samples were validly predicted by the FSL approaches. FSL uses decoding algorithms, e.g., beam search, greedy search, top-p and top-k sampling, that analyse and select the most likely tokens to generate. Table 8 presents the percentage of valid predictions for each model and context level considered. Whereas these results show that most of the time the models will generate a valid prediction ("1", non-negative sentiment, or "0", negative sentiment), other times they will generate unexpected tokens that will not provide a label to the input, e.g., symbols like "." or "|". The analysis of valid predictions is important to assess how fair the comparison with other models is and to evaluate how adequate the prompts were.

Looking at the table, it seems that GPT-3 is able to produce more valid predictions for the TeleComSA dataset and OPT is more suited to the TwitterDialogueSAPT data. There is mostly a percentage of over 98% of valid predictions when using GPT-3 in the TeleComSA dataset and OPT in the TwitterDialogueSAPT dataset. Even when using the other model for each of the datasets, the percentages are mostly over 80%, with two exceptions at around 76%, meaning that overall, in terms of the ability to generate outputs in the desired format, these models and prompts seem adequate. However, this does not mean they are assigning the right sentiment to each input, only that they are classifying the utterances with a negative or non-negative label.

A. CONTEXT EFFECT

In this analysis, we present the full results, considering all models, representations, and context levels. The context effect provides an overview of all the approaches, but it is harder to interpret due to the large number of experiments and overlaps in the results. Hence, we will focus on specific parts of the context levels in the remaining subsections, which will make the charts easier to read and to take conclusions on the impact of each effect.

Figure 2 presents the performance comparison between all approaches using the TeleComSA dataset, whereas Figure 3

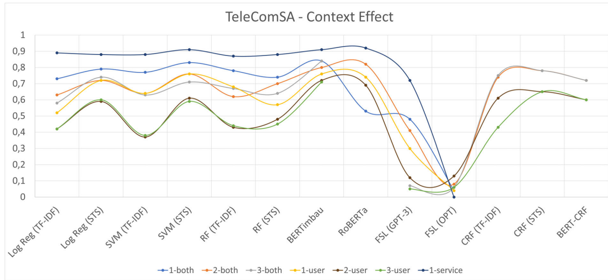

FIGURE 2. F1-scores for the experiments, using the TeleComSA dataset.

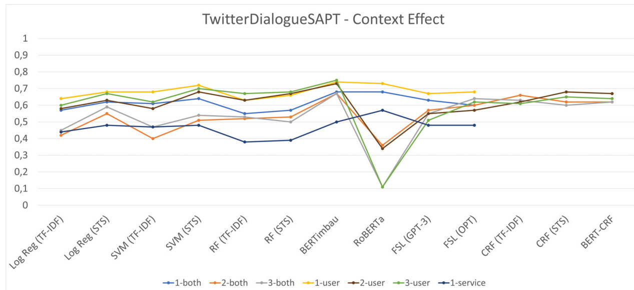

FIGURE 3. F1-scores for the experiments, using the TwitterDialogueSAPT dataset.

presents the same comparison using the TwitterDialogue-SAPT dataset.

Regarding the TeleComSA dataset, it is noticeable that every model, except GPT-3 and OPT, is very good at classifying the service lines. This is likely due to their very repetitive nature, as they are automatic and specific responses. None of the FSL approaches presented results better than random guessing, as they are all below 50%, so these do not seem like adequate choices for our task. This also happened in five out of nine models, excluding GPT-3 and OPT, when considering the 2 user and 3 user context levels, meaning the classifiers are not good at determining the sentiment when considering only the user's lines, especially if including context. Overall, for this type of data, no use of context seems to be the best scenario for most classifiers. However, this is not the case for the BERTimbau and RoBERTa models, and for the CRF-based models, where a larger context seems to improve results. In fact, for the RoBERTa approach, there is an astounding difference of 29 percentage points between not considering context and considering the previous utterance (52% for 1-both vs. 81% for 2-both). There are several possible reasons for this performance difference. RoBERTa is not trained with contextual information, it was trained on tweets, which are self-contained, not on dialogues. In the TeleComSA dataset, the data's nature is inherently different, as it represents real dialogues. This could be the reason for the difference in performance, as just one utterance may not provide enough information for the model to classify such a different structure of data as dialogue. In the FSL approach using OPT the use of context also improved the results, but seeing as these are very low, we will not consider it. A possible explanation for the low performance of OPT in this dataset is that the model itself "tends to be repetitive and can easily get stuck in a loop" [9]. Adding to this that TeleComSA is also a repetitive dataset, it is possible that this made the model get stuck more often, as it would be more often faced with similar dialogues that could produce wrong outputs. Another limitation of OPT is that it may not work well with declarative instructions, and while a task description was not included in our FSL scheme, there is a similarity between the prompts that could justify the worse performance.

Regarding the TwitterDialogueSAPT dataset, and contrary to what was seen with the TeleComSA dataset, it seems like most models are worse than random guessing when considering the service lines, reaching their best performance at 57%. This does, however, make sense, as the nature of the service utterances is much different from the repetitive and specific responses from the Telecom company's dialogue system. On Twitter, services sometimes also use automatic responses, but even then, they offer more variability, which could make the classification of these samples harder. This also applies to the remaining context levels, as overall, the models' performance worsens in this more diverse and multiple domains dataset. Also differently from what was previously seen, the FSL approaches presented acceptable results, close to 70%, and could be an option for this type of scenario. Similarly to the TeleComSA results, overall, no use of context seems to be the best option for most classifiers. However, this is not true when using Random Forest, BERTimbau, or FSL using the OPT model. Again, differing from TeleComSA, the CRF-based models do not seem to benefit from an increased context level.

Overall, and dismissing the service context level, as it is not the focus of this work, the approach with the best overall performance for both datasets is the BERTimbau model considering full context, with F1 scores of 84% and 75%, respectively for the TeleComSA and TwitterDialogueSAPT datasets, followed by RoBERTa, with scores of 82% considering context and 73% not considering it. They are followed by SVM (STS) not considering context, with scores of 83% and 72%.

B. SPEAKER EFFECT

In this analysis, we focus on the results considering the different speaker levels. Figure 4 presents the performance comparison between all approaches in the TeleComSA dataset without context, whereas Figure 5 presents the same comparison in the TwitterDialogueSAPT dataset. In these figures, we removed the CRF-based models as they did not consider scenarios with no context.

When not considering context, it is noticeable that the results for each dataset are opposing: in the TeleComSA data, the service lines are better classified than the user lines, whereas in the TwitterDialogueSAPT, the opposite happens. There is only one model that breaks the trend of the user lines achieving lower performances in the former dataset, which is the RoBERTa classifier, by the largest overall difference in percentage points.

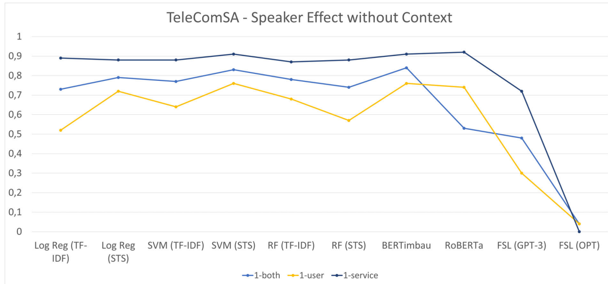

FIGURE 4. F1-scores for the classifiers, considering each speaker level without context, using the TeleComSA dataset.

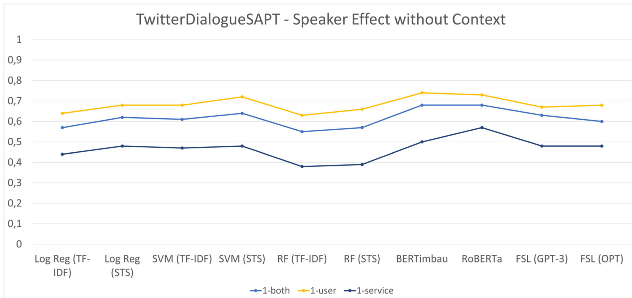

FIGURE 5. F1-scores for the classifiers, considering each speaker level without context, using the TwitterDialogueSAPT dataset.

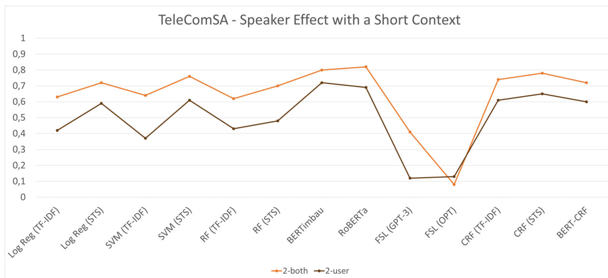

FIGURE 6. F1-scores for the classifiers, considering each speaker level with short context, using the TeleComSA dataset.

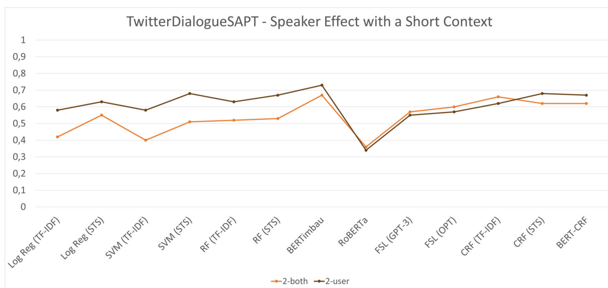

FIGURE 7. F1-scores for the classifiers, considering each speaker level with short context, using the TwitterDialogueSAPT dataset.

Figure 6 presents the performance comparison between all approaches using the TeleComSA dataset with a short amount of context, whereas Figure 7 presents the same comparison using the TwitterDialogueSAPT dataset.

When considering a short amount of context, meaning the current and previous utterances, we verify that the speaker

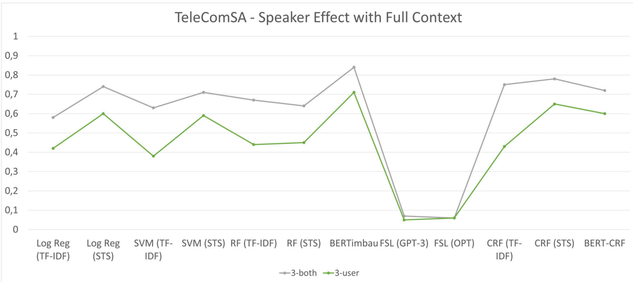

FIGURE 8. F1-scores for the classifiers, considering each speaker level with full context, using the TeleComSA dataset.

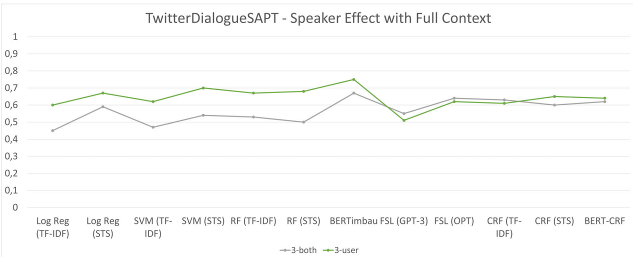

FIGURE 9. F1-scores for the classifiers, considering each speaker level with full context, using the TwitterDialogueSAPT dataset.

effect trends continue in each dataset, with exceptions in the the TeleComSA dataset when using OPT, but this could be discarded due to the low scores, and in the TwitterDialogueSAPT dataset when using the CRF (TF-IDF), RoBERTa, and the FSL approaches.

Figure 8 presents the performance comparison between all approaches using the TeleComSA dataset with a full level of context, whereas Figure 9 presents the same comparison using the TwitterDialogueSAPT dataset. In these figures, we removed the RoBERTa scores as this model did not provide results for the TeleComSA dataset and very low results for the TwitterDialogueSAPT dataset.

When considering a full amount of context, meaning the current and the two previous utterances, the speaker effect trends remain overall. Three exceptions, common with the previously mentioned ones, are the CRF (TF-IDF) and the FSL approaches when using the TwitterDialogueSAPT dataset.

C. UTTERANCES EFFECT

In this analysis, we focus on the results considering the different amounts of context, represented by the number of utterances considered. Figure 10 presents the performance comparison between all approaches using the TeleComSA dataset, whereas Figure 11 presents the same comparison using the TwitterDialogueSAPT dataset. We do not include the no-context levels as we want to focus on a comparison between the two levels of context, and also remove the RoBERTa approach for the same reasons mentioned earlier.

Looking at both figures, we see that the advantage of using more or less contextual information varies. When using the traditional approaches, most classifiers benefit from a higher

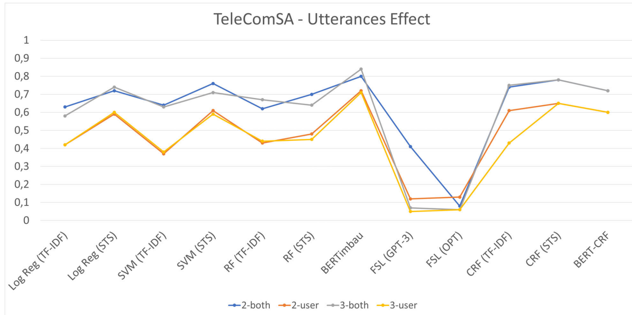

FIGURE 10. F1-scores for the classifiers, considering the two levels of context, using the TeleComSA dataset.

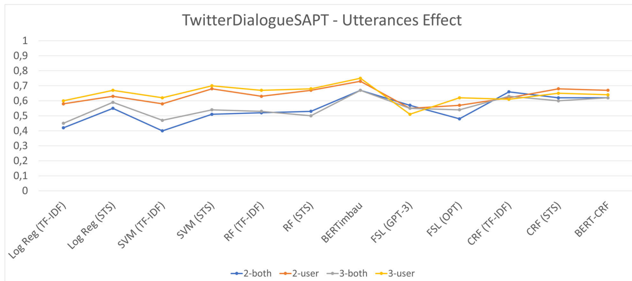

FIGURE 11. F1-scores for the classifiers, considering the two levels of context, using the TwitterDialogueSAPT dataset.

level of context when considering the user lines, with the exception of the SVM (STS) and the Random Forest (STS) in the TeleComSA dataset. When using CRF-based approaches, in the TwitterDialogueSAPT dataset, most seem to benefit from a lower level of context, although the difference in performance is not large. In the TeleComSA dataset, the difference is mostly irrelevant, except in the CRF (TF-IDF) approach, which performs particularly badly when considering the user's full context. When using BERTimbau, the performance slightly increases when considering the context in both datasets, with a slight decrease when considering the user dialogues and the TeleComSA dataset. Finally, when using the FSL approaches, less context is usually better, except when using the OPT model in the TwitterDialogueSAPT dataset.

D. APPROACH EFFECT

In this analysis, we focus on the results considering the approaches in general, meaning the combination of representation technique and classifier. We do not include the BERT-based and FSL approaches, as for the same model we did not experiment with different representations. Figure 12 presents the performance comparison between these approaches using the TeleComSA dataset, whereas Figure 13 presents the same comparison using the TwitterDialogueSAPT dataset. We focus on the impact of the representation technique, hence the figures present disconnected lines between different classifiers, for ease of visualisation.

Looking at both figures, we can verify that overall, the STS representation allows for better performance, with the

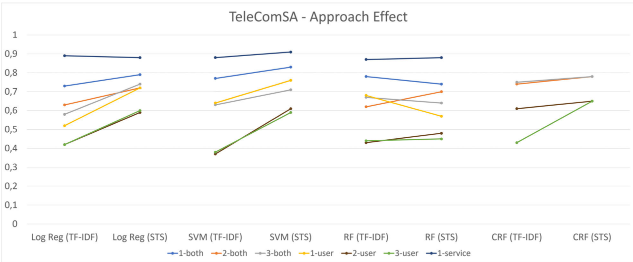

FIGURE 12. F1-scores for the traditional classifiers, considering all context levels, using the TeleComSA dataset.

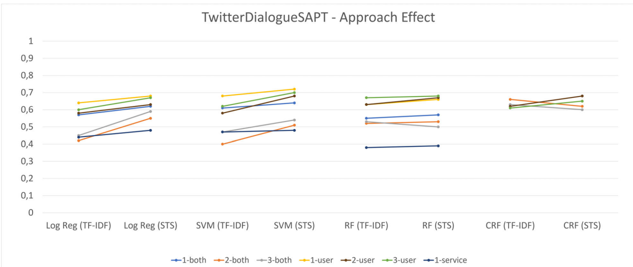

FIGURE 13. F1-scores for the traditional classifiers, considering all context levels, using the TwitterDialogueSAPT dataset.

exception of the Random Forest model, where it mostly remains the same in the TwitterDialogueSAPT dataset, but where performances decrease in the TeleComSA dataset.

It is important to note that there are no high-tech requirements for the development of these solutions, as can be seen in Appendix D. Furthermore, the speaker selection process is computationally inexpensive and quick, and it can provide a boost to the performance of the models. Hence, this is an option that should be particularly considered, as results may improve with low effort.

In the next section, and based on the analyses performed, we answer our research questions and discuss the importance of context for SA in dialogues.

V. CONCLUSION AND FUTURE WORK

At the start of this study, we defined four research questions to answer through our experiments. In this section, we propose an answer to each question and present possible options for future work. The research questions defined were the following:

RQ1 **When does the inclusion of context improve the model's performance compared to excluding previous utterances?**
For any of the datasets, the BERT-based classifier (BERTimbau), pre-trained using differents types of text in Portuguese, performs better or equally when considering the context. Furthermore, in the TeleComSA dataset, the RoBERTa classifier, pre-trained in multilanguage tweets, achieved the highest performance when changing from a no-context level to a two-sentence context level, improving its score by 29 percentage points. If the dataset used contains more variability, our results suggest that a Random Forest model can also benefit from the use of more contextual information.

RQ2 **How does speaker selection assist in classifying sentiment in dialogues?**
It seems that this is highly dependent on the type of reply provided by the service. If the customer-support dialogue system provides mostly repetitive and specific utterances (human-machine dialogues), our results suggest that the consideration of both speakers benefits the classification task. However, if the service utterances regard several domains, and may not be fully automatic (human-human dialogues), meaning they are not as repetitive, our results suggest that considering only the user utterances will benefit the performance of the classifiers.

RQ3 **How many prior utterances are necessary to enhance the model's performance?**
Our results suggest that this depends on the type of classifier. However, most approaches, including the traditional classifiers and the BERT-based models seem to benefit from more contextual information (two previous utterances), while the CRF-based models seem to benefit from less contextual information (one previous utterance), although with a small difference in their performances. The same happens with most of our FSL approaches. Overall, when considering context, it is likely that the classification task will improve if we provide the models with more information.

RQ4 **Which classifier and representation technique is better suited for SA in dialogues?**
Considering what we called the "Approach Effect", our results suggest that a more recent representation, i.e., sentence embeddings obtained from a Sentence Transformer, is beneficial to our task. In fact, the SVM (STS) approach, which achieved the better performance among the traditional classifiers, is also present in the top three overall approaches, the other two being the BERT-based models, BERTimbau and RoBERTa, which also use a similar type of representation, as all three are computed through BERT-based tokenizers.

TABLE 9. Selected Twitter accounts and their most representative domain.

Name	Handle	Domain
Altice Portugal	@altice_portugal	Telecom
Deco Proteste	@decoproteste	Finance & Fintech
Direção-Geral de Saúde	@DGSaude	Healthcare
MEO Portugal	@MEOpt	Telecom
Netflix Portugal	@NetflixPT	TV
NOWO Portugal	@nowoportugal	Telecom
RTP	@rtppt	TV
RTP Notícias	@RTPNoticias	TV
RTP Play	@playrtp	TV
Vodafone Portugal	@VodafonePT	Telecom
VOST Portugal	@VOSTPT	Healthcare
Worten Portugal	@WortenPT	eCommerce

TABLE 10. Analysis of the available datasets.

Dataset	Size	Annotation	Language	Data Source(s)	Domain(s)
DailyDialog	13.118 dialogues	Dialogue act and emotion	English	Websites for the practice of English dialogues	Daily life topics (e.g., buying goods from a shop, summer vacation)
Mastodon	505 dialogues	Dialogue act and sentiment	English	Social Media (Mastodon)	General dialogues in the octodon.social Mastodon instance
Friends' Emotion Detection	12.606 utterances	Emotion	English	TV Show Friends	Daily life topics common in comedy TV shows
CORAA	400.000+ audio transcriptions	-	Brazilian Portuguese	Five audio corpora	Spontaneous speech (e.g., interviews, informal conversations)
Emotion in News	1.750 news	Emotion	Brazilian Portuguese	News extracted from Globo news website	News (international, national, politics, economics, law enforcement)
ReLi	1.600 reviews	Opinion and sentiment	Brazilian Portuguese	Book reviews posted on the Internet	Book reviews
Sentituites-PT	30.470 tweets	Sentiment	Portuguese	Tweets posted during the 2011 Portuguese elections	Politics
Wizard of Wikipedia	22.311 dialogues	-	English	WoZ framework	Wikipedia
Multi-WoZ	8.438 dialogues	Dialogue act and slots	English	WoZ framework	Restaurants, Attractions, Hotels
CamRest	680 dialogues	Dialogue act and slots	English	WoZ framework	Restaurant search
Ubuntu Dialog	Nearly 1 million dialogues	-	English	Ubuntu chat logs (2004-2015)	Technical support for Ubuntu-related problems

This work may offer a guideline for researchers wondering if and how they should consider the context in their studies and industry developers that employ or want to use dialogue systems in their company. Moreover, the expansion of TwitterDialogueSAPT contributes to the computational processing of the Portuguese language, which is currently a field lacking resources. To the best of our knowledge, this is the first work applying SA to dialogues in the Portuguese language and considering speaker-selection and context levels. And while, overall, the SA task in our scenarios does not benefit from the use of contextual information, there are some cases in which they do. In fact, our highest performances come from considering not only context but our highest level of it, with the BERT-based approach.

In the future, it would be interesting to apply summarization techniques to grasp the most important aspects of what was previously mentioned. This approach could hopefully allow us to include more than the two previous sentences, as this number was constrained by the models' token limitations, and a summary could contain more information in the same number of tokens. However, it should be noted that errors in the summarization will have a snowball effect and cause increased mistakes in our main task. Furthermore, the FSL approach could benefit from the use of even larger models such as GPT-4 [29], which could also be employed to explore Zero-Shot Learning [52]. A new BERT-based model has recently been released, Albertina PT-* [53], trained in the Portuguese language, and particularly in European Portuguese that could be an interesting option for future work. Despite these new advancements, the use of traditional models should never be overlooked as simplicity can provide high value. In fact, the SVM and RF models obtained adequate results in both datasets, with F1 scores of around 80% in TeleComSA and around 70% in TwitterDialogueSAPT. Furthermore, the latter model has the advantage of transparency, which could promote a better understanding of SA in our scenarios. As such, we will soon begin exploring the application of explainability techniques to our models, which would hopefully allow us to better perceive what kind of information is being focused on in the different context levels, and verify if the models are making the best use of it or if our approach could be improved in that way.

TABLE 11. List of packages used in the development of this work.

Package	Description	Version
TwitterAPI	Allows access to Twitter	2.7.5
requests	Allows HTTP requests to be sent	2.27.1
json	Allows writing and reading .json files	2.0.9
os	Allows access to operating system functionalities (e.g., setting environment variables)	3.8.13
time	Allows the measure of execution time	3.8.13
pandas	Allows access to more data structures (e.g., dataframes)	1.4.2
numpy	Supports multi-dimensional data structures (e.g., tensors)	1.22.3
torch	Allows tensor computation and an auto-gradient system for deep neural networks	1.11.0
nltk	Allows access to Natural Language Processing techniques (e.g., tokenization)	3.7
sklearn	Allows access to machine learning models and techniques (e.g., SVM)	1.1.1
matplotlib	Allows the visualization of data	3.5.1
seaborn	Allows statistical data visualization	0.11.2
pickle	Allows easy storage of many data types	4.0
transformers	Allows access to pretrained transformer models	4.19.2
tensorflow	Allows high performance numerical computation	2.8.0
spacy	Allows access to Natural Language Processing techniques (e.g., lemmatization)	3.3.0
krippendorff	Allows fast computation of Krippendorff's alpha agreement measure	0.5.1
statsmodels	Allows fast computation of Fleiss' kappa agreement measure	0.13.2
random	Allows access to random-number generators	3.8.13
sentence_transformers	Allows access to state-of-the-art sentence embeddings	2.2.0
sklearn-crfsuite	Allows access to a sklearn-compatible CRF classifier	0.3.6
tqdm	Allows access to progress bars and functions	4.64.0
math	Allows access to mathematical functions (e.g., ceil)	3.8.13
openai	Allows access to GPT-3	0.18.0

TABLE 12. Specifications of the machines used in the development of this work.

Machine Name	Operative System	CPU	GPU	RAM
Personal_PC	Windows 10	i7-12700H	-	16GB
Castor	Debian GNU/Linux 10	i7-4930K	-	24GB
Flower	Ubuntu 20.04.1	i9-11900F	NVIDIA GeForce RTX 3090	128GB

APPENDIX A
ANALYSIS OF THE SELECTED TWITTER ACCOUNTS

Table 9 presents the Twitter accounts chosen to gather data for the TwitterDialogueSAPT dataset, including their most representative domain.

APPENDIX B
ANALYSIS OF AVAILABLE DATASETS

Table 10 presents further information on the available datasets considered for this study, including size, for which task they are annotated, if labelled, in which language, from where they were extracted, and which domains are represented.

APPENDIX C
INFORMATION ABOUT THE PACKAGES USED

Table 11 presents information on the packages used in this study, including a short description and the versions installed.

APPENDIX D
INFORMATION ABOUT SPECIFICATIONS OF THE MACHINES USED

Table 12 presents some specifications of the machines used in this study.

ACKNOWLEDGMENT

The authors would like to thank everyone involved in the annotation process that allowed the creation and expansion of the TwitterDialogueSAPT dataset.

REFERENCES

[1] K. Arava, R. S. K. Chaitanya, S. Sikindar, S. P. Praveen, and D. Swapna, "Sentiment analysis using deep learning for use in recommendation systems of various public media applications," in *Proc. 3rd Int. Conf. Electron. Sustain. Commun. Syst. (ICESC)*, Aug. 2022, pp. 739–744.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[3] Z. Zhang, Y. Wu, H. Zhao, Z. Li, S. Zhang, X. Zhou, and X. Zhou, "Semantics-aware BERT for language understanding," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, 2020, pp. 9628–9635.

[4] G. Penha and C. Hauff, "What does BERT know about books, movies and music? Probing BERT for conversational recommendation," in *Proc. 14th ACM Conf. Recommender Syst.*, Sep. 2020, pp. 388–397.

[5] J. Zhu, Y. Xia, L. Wu, D. He, T. Qin, W. Zhou, H. Li, and T. Liu, "Incorporating BERT into neural machine translation," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–18.

[6] C. Sutton, K. Rohanimanesh, and A. McCallum, "Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data," in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*. San Francisco, CA, USA: Morgan Kaufmann, 2004, pp. 282–289.

[7] M. G. Miller, N. E. Matsakis, and P. A. Viola, "Learning from one example through shared densities on transforms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2000, pp. 464–471.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, and S. Agarwal, "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Sys.*, vol. 33, 2020, pp. 1877–1901.

[9] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. Victoria Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer, "OPT: Open pre-trained transformer language models," 2022, *arXiv:2205.01068*.

[10] D. A. Pereira, "A survey of sentiment analysis in the Portuguese language," *Artif. Intell. Rev.*, vol. 54, no. 2, pp. 1087–1115, Feb. 2021.

[11] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proc. 7th Int. Conf. Lang. Resour. Eval. (LREC)*, 2010, pp. 1320–1326.

[12] L. Duarte, L. Macedo, and H. Gonçalo Oliveira, "Exploring emojis for emotion recognition in Portuguese text," in *Proc. 19th Conf. Artif. Intell. (EPIA)*, vol. 11805. Vila Real, Portugal: Springer, Sep. 2019, pp. 719–730.

[13] S. E. Saad and J. Yang, "Twitter sentiment analysis based on ordinal regression," *IEEE Access*, vol. 7, pp. 163677–163685, 2019.

[14] F. D. Souza and J. B. D. O. E. S. Filho, "BERT for sentiment analysis: Pre-trained and fine-tuned alternatives," in *Proc. Int. Conf. Comput. Process. Portuguese Lang.*, vol. 13208, Jan. 2022, pp. 209–218.

[15] F. Souza, R. Nogueira, and R. Lotufo, "BERTimbau: Pretrained BERT models for Brazilian Portuguese," in *Proc. Brazilian Conf. Intell. Syst. (BRACIS)*, vol. 12319. Cham, Switzerland: Springer, 2020, pp. 403–417.

[16] S. S. Roy, A. Roy, P. Samui, M. Gandomi, and A. H. Gandomi, "Hateful sentiment detection in real-time tweets: An LSTM-based comparative approach," *IEEE Trans. Computat. Social Syst.*, early access, May 10, 2023, doi: 10.1109/TCSS.2023.3260217.

[17] E. Fernandez, Anderies, M. G. Winata, F. H. Fasya, and A. A. S. Gunawan, "Improving IndoBERT for sentiment analysis on Indonesian stock trader slang language," in *Proc. IEEE Int. Conf. Internet Things Intell. Syst. (IoTaIS)*, Nov. 2022, pp. 240–244.

[18] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 757–770. [Online]. Available: https://aclanthology.org/2020.coling-main.66

[19] M. Li, L. Chen, J. Zhao, and Q. Li, "Sentiment analysis of Chinese stock reviews based on BERT model," *Int. J. Speech Technol.*, vol. 51, no. 7, pp. 5016–5024, Jul. 2021.

[20] R. Bensoltane and T. Zaki, "Towards Arabic aspect-based sentiment analysis: A transfer learning-based approach," *Social Netw. Anal. Mining*, vol. 12, no. 1, pp. 1–16, Dec. 2022.

[21] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015, *arXiv:1508.01991*.

[22] F. Souza, R. Nogueira, and R. Lotufo, "Portuguese named entity recognition using BERT-CRF," 2019, *arXiv:1909.10649*.

[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[24] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A hybrid model for sentiment analysis with transformer and recurrent neural network," *IEEE Access*, vol. 10, pp. 21517–21525, 2022.

[25] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100943–100953, 2019.

[26] Y. Ling, Z. Liang, T. Wang, F. Cai, and H. Chen, "Sequential or jumping: Context-adaptive response generation for open-domain dialogue systems," *Appl. Intell.*, vol. 53, pp. 11251–11266, Sep. 2022.

[27] J. Wang, J. Wang, C. Sun, S. Li, X. Liu, L. Si, M. Zhang, and G. Zhou, "Sentiment classification in customer service dialogue with topic-aware multi-task learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, 2020, pp. 9177–9184.

[28] L. Song, C. Xin, S. Lai, A. Wang, J. Su, and K. Xu, "CASA: Conversational aspect sentiment analysis for dialogue understanding," *J. Artif. Intell. Res.*, vol. 73, pp. 511–533, Feb. 2022.

[29] OpenAI, "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[30] S. Black, L. Gao, P. Wang, C. Leahy, and S. Biderman, "GPT-Neo: Large scale autoregressive language modeling with Mesh-Tensorflow," Zenodo, 2021, doi: 10.5281/zenodo.5297715.

[31] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[32] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy, "The pile: An 800GB dataset of diverse text for language modeling," 2021, *arXiv:2101.00027*.

[33] E. Hosseini-Asl, W. Liu, and C. Xiong, "A generative language model for few-shot aspect-based sentiment analysis," 2022, *arXiv:2204.05356*.

[34] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," 2017, *arXiv:1710.03957*.

[35] C. Cerisara, S. Jafaritazehjani, A. Oluokun, and H. Le, "Multi-task dialog act and sentiment recognition on mastodon," 2018, *arXiv:1807.05013*.

[36] S. Zahiri and J. D. Choi, "Emotion detection on TV show transcripts with sequence-based convolutional Neural Networks," in *Proc. AAAI Workshop Affect. Content Anal.*, 2018, pp. 44–51.

[37] A. C. Junior, E. Casanova, A. Soares, F. S. D. Oliveira, L. Oliveira, R. C. F. Junior, D. P. P. D. Silva, F. G. Fayet, B. B. Carlotto, L. R. S. Gris, and S. M. Aluísio, "CORAA: A large corpus of spontaneous and prepared speech manually validated for speech recognition in Brazilian Portuguese," 2021, *arXiv:2110.15731*.

[38] M. M. Dosciatti, L. Ferreira, and E. Paraiso, "Identificando emoçoes em textos em Português do Brasil usando máquina de vetores de suporte em soluçao multiclasse," in *Proc. Encontro Nacional de Inteligência Artif. e Computacional (ENIAC)*, Fortaleza, Brasil, 2013, pp. 1–12.

[39] C. Freitas, E. Motta, R. Milidiú, and J. César, "Sparkling vampire…lol! Annotating opinions a book review corpus," in *New Language Technologies and Linguistic Research: A Two-Way Road*. Newcastle Upon Tyne, U.K.: Cambridge Scholars, 2014, pp. 128–146.

[40] P. Carvalho, L. Sarmento, J. Teixeira, and M. J. Silva, "Liars and saviors in a sentiment annotated corpus of comments to political debates," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Human Lang. Technol.*, 2011, pp. 564–568.

[41] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Wizard of Wikipedia: Knowledge-powered conversational agents," 2018, *arXiv:1811.01241*.

[42] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić, "MultiWOZ—A large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 5016–5026.

[43] T.-H. Wen, M. Gasic, N. Mrkšić, L. M. R. Barahona, P.-H. Su, S. Ultes, D. Vandyke, and S. Young, "Conditional generation and snapshot learning in neural dialogue systems," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 2153–2162.

[44] R. Lowe, N. Pow, I. Serban, and J. Pineau, "The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems," in *Proc. 16th Annu. Meeting Special Interest Group Discourse Dialogue*, 2015, pp. 285–294.

[45] I. Carvalho, H. G. Oliveira, and C. Silva, "Sentiment analysis in Portuguese dialogues," in *Proc. IberSPEECH*, Nov. 2022, pp. 176–180.

[46] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, Nov. 1971.

[47] K. Krippendorff, "Estimating the reliability, systematic error and random error of interval data," *Educ. Psychol. Meas.*, vol. 30, no. 1, pp. 61–70, Apr. 1970.

[48] A. Zapf, S. Castell, L. Morawietz, and A. Karch, "Measuring inter-rater reliability for nominal data— Which coefficients and confidence intervals are appropriate?" *BMC Med. Res. Methodol.*, vol. 16, no. 1, p. 93, Dec. 2016.

[49] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[50] K. Krippendorff, *Content Analysis: An Introduction to its Methodology*. Thousand Oaks, CA, USA: Sage publications, 2018.

[51] F. Barbieri, L. E. Anke, and J. Camacho-Collados, "XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond," in *Proc. Lang. Resour. Eval. Conf.* Marseille, France: European Language Resources Association, 2022, pp. 258–266. [Online]. Available: https://aclanthology.org/2022.lrec-1.27

[52] M.-W. Chang, L.-A. Ratinov, D. Roth, and V. Srikumar, "Importance of semantic representation: Dataless classification," in *Proc. AAAI*, vol. 2, 2008, pp. 830–835.

[53] J. Rodrigues, L. Gomes, J. Silva, A. Branco, R. Santos, H. L. Cardoso, and T. Osório, "Advancing neural encoding of Portuguese with transformer Albertina PT," 2023, *arXiv:2305.06721*.

ISABEL CARVALHO received the B.Sc. and M.Sc. degrees in informatics engineering from the University of Coimbra (UC), Portugal, where she is currently pursuing the Ph.D. degree in informatics engineering, under the supervision of Prof. Catarina Silva and Prof. Hugo Gonçalo Oliveira. She has been a Researcher with the Centre for Informatics and Systems, University of Coimbra (CISUC), since 2021, and has worked on topics, such as semantic databases and sentiment analysis. She has been invited to start lecturing on NLP with UC, in the first semester of 2023 and 2024. Her research interests include machine learning, natural language processing, and explainable artificial intelligence. She was recognized with the Best Paper Award from the Portuguese Conference on Pattern Recognition, the RECPAD, in 2022, and the Merit Award from UC.

HUGO GONÇALO OLIVEIRA is currently an Associate Professor with the Department of Informatics Engineering, University of Coimbra (UC), and a Researcher with CISUC. He has participated in European and nationally-funded projects in computational creativity (CC), natural language processing (NLP), and data science, and lead the development of computational resources, methods and systems, mostly for the Portuguese language. He is the author of circa 150 papers in peer-reviewed journals and conferences, regularly part of the program committee of the main NLP and CC venues. He won the Best Doctoral Thesis in the area of computational processing of Portuguese, from 2011 to 2014, and a special recognition in Education by the World Cultural Council, in 2022. He is an Executive Editor of the journal *Linguamática*.

CATARINA SILVA (Senior Member, IEEE) received the Ph.D. degree in computer engineering. She is currently an Assistant Professor with the Department of Informatics Engineering, University of Coimbra. With 20 years experience in teaching computer engineering for the B.Sc. and M.Sc., while supervising the M.Sc. and Ph.D. students. She is a Senior Researcher with the Adaptive Computation Group, CISUC, with machine learning and pattern recognition as main areas of research. Skilled at managing different sized projects and scientific entrepreneurship, involving people with different backgrounds, namely faculty, students, alumni and companies. She is the author and coauthor of four books, circa 20 journal articles, and 80 conference papers. She is a scientific committee member and a paper reviewer of several conferences and journals. She is a member of the Board of the Portuguese Association of Pattern Recognition and a IEEE Senior Member of the Computational Intelligence Society. She is the IEEE Past Chair of the Portugal Section.