Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 191-205 recibido 05-12-2024 revisado 19-01-2025 aceptado 31-01-2025

Evaluating Galician language models for sentiment analysis on challenging linguistic phenomena

Evaluación de modelos del lenguaje gallegos para el análisis del sentimiento tomando en cuenta fenómenos lingüísticos problemáticos

Anxo Alonso,[1] **Pablo Gamallo**[2]
[1]Universidad Complutense de Madrid (UCM)
[2]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS-USC)
anxoal02@ucm.es, pablo.gamallo@usc.gal

Abstract: Sentiment analysis is still one of the most relevant tasks in NLP. However, low-resource languages lack sufficient datasets and models for this task. In this paper, we present a study on sentiment analysis in Galician, analyzing how linguistic phenomena can influence this task. For this purpose, we developed Senti-Gal, a dataset with 998 sentences including adversative, concessive and conditional sentences, diglossic phenomena, negation and irony. We evaluated Senti-Gal on seven models: a multilingual machine learning model, a multilingual decoder-only (or generative) model, and five encoder-only models (three multilingual and two monolingual), all of them fine-tuned with a training dataset we also developed. The results indicate that the best fine-tuned encoder-only models outperform the decoder-only model, that syntactic and pragmatic phenomena remain a challenge, and that monolingual and multilingual models perform similarly. We release Senti-Gal, the fine-tuned models and the first Galician training corpus for sentiment analysis freely available.
Keywords: sentiment analysis, fine-tuning, galician, evaluation.

Resumen: El análisis de sentimientos sigue siendo una de las tareas más relevantes en PLN. No obstante, las lenguas con escasos recursos carecen de conjuntos de datos y modelos suficientes para esta tarea. En este trabajo, presentamos un estudio sobre el análisis de sentimientos en gallego, analizando cómo los fenómenos lingüísticos pueden influir en esta tarea. Para ello, desarrollamos Senti-Gal, un dataset con 998 oraciones que incluyen oraciones adversativas, concesivas y condicionales, fenómenos diglósicos, negación e ironía. Evaluamos Senti-Gal en siete modelos: un modelo multilingüe de aprendizaje automático, un modelo solo-decodificador (o generativo) multilingüe y cinco modelos solo-codificador (tres multilingües y dos monolingües), todos ellos ajustados con un conjunto de datos de entrenamiento que desarrollamos. Los resultados indican que los modelos solo-codificador ajustados con el conjunto de datos superan a los solo-decodificador, que los fenómenos sintácticos y pragmáticos siguen siendo un desafío y que los modelos monolingües y multilingües tienen rendimientos similares. Liberamos Senti-Gal, los modelos ajustados y el primer corpus gallego de entrenamiento para análisis de sentimientos de libre acceso.
Palabras clave: análisis del sentimiento, ajuste de modelos, gallego, evaluación.

1 Introduction

Sentiment analysis (SA) remains a key NLP task, despite declining hype, evolving in research and industry due to its broad applicability. Raghunathan and Saravanakumar (2023) mention the extraction of public opinion through social networks, product analysis, stock market trend analysis, and enterprise management. SA also serves a purpose in other fields such as the medical and healthcare domain (Wankhade, Rao, and Kulkarni, 2022) or politics (Zhao et al., 2024; Öztürk and Ayvaz, 2018). SA can also be used to tackle social issues such as misogyny (Álvarez-Crespo and Castro, 2024), LGBTQ+ discrimination (Dsouza et al., 2023), racism (Sukanya et al., 2023), and more.

With the advent of Large Language Models (LLMs) to NLP, there are several competing paradigms for classification. Fine-tuning is an approach that involves transfer learning from a pre-trained model to make the model suitable for specific tasks (Zhang, Chai, and Xu, 2023). Fine-tuning a model requires large human-annotated datasets (Prottasha et al., 2022; Geetha

ISSN 1135-5948 DOI 10.26342/2025-74-13

and Renuka, 2021; Tang, Tang, and Yuan, 2020). Nevertheless, creating them for low-resource languages(LRL) remains a challenge. Data augmentation has recently gained popularity. Li et al. (2024) explores how generating synthetic data for text classification could be very beneficial, especially, in low-resourced settings. Alternatively, one of the most recent approaches to SA, based on In Context Learning, involves zero-shot or few-shot settings with generative LLMs (Ghilene et al., 2024; Zhang, Chai, and Xu, 2023). This approach usually makes use of prompt learning, a method which "is to transform the input and output of downstream tasks into an acceptable form of the pre-trained model, so that the model can be used for downstream tasks" (Zhang, Chai, and Xu, 2023). The model dynamically learns patterns from input prompts without fine-tuning. This strategy has the advantage that it does not need a large volume of labeled data, allowing LLMs to adapt dynamically to diverse tasks without retraining.

LLMs remain Anglocentric (their performance is better in English), even if they are multilingual (Yuan et al., 2024). For tasks like machine translation (MT), studies have found that multilingual LLMs perform worse than traditional MT models for Low Resource Languages (LRLs) (Robinson et al., 2023). Making linguistic tools available for LRLs could have meaningful impact on their subsistence (Tsvetkov, 2017, as cited in Magueresse et al. 2020).

In addition to classification strategies and the focus on resource-poor languages, the feasibility of SA depends on the treatment of multiple linguistic phenomena involved in its detection and classification. Sentiment lexicons, linguistic rules, and discourse structures can improve SA and influence sentiment interpretation (Taboada et al., 2011).

The aim of the article is to approach SA taking into account what we have discussed so far: i.e., different methods of classification, languages with few resources, and the study of multiple linguistic phenomena. First, we will compare several classification strategies for SA, mainly focused on the fine-tuning of encoder-only models(EOM), which still seem to outperform decoder-only or generative models on discriminative tasks (Edwards and Camacho-Collados, 2024). Second, we will create resources for SA of a LRL, namely Galician, a language spoken mainly in the autonomous community of Galicia. In this sense, the paper presents a new challenging evaluation dataset for Galician SA, encompassing several linguistic phenomena: adversative, concessive and conditional sentences; diglossic phenomena; negation; and positive and negative irony; it provides a training set for SA in Galician; and it develops new transformed-based models fine-tuned to SA for Galician. All contributions are distributed under open licenses and are available for download.[1] And third, we will explore the difficulties of SA in different linguistic aspects, mainly at the pragmatic and syntactic levels. More precisely, we will focus on the linguistic phenomena more troublesome for the task, on account of the limited number of studies surrounding this area which is subject to improvement. In terms of syntax, we explore four types of sentences: conditional, adversative, concessive, and sentences with negation. On the pragmatic level, we will focus on irony and sarcasm, as well as on diglossic phenomena, due to the diglossic situation of the territories in which Galician is spoken (Torre, 2024). Diglossia is a sociolinguistic phenomenon where two or more languages or language varieties coexist within a speech community, each serving different social functions. Generally, one of them has a high status, in this case Spanish, compared to the other which has a low status, Galician (Jaspers, 2016; Skobel, 2010).

The paper is organized as follows. Section 2 presents related work, Section 3 describes the datasets , and Section 4 details the experiments, including the presentation of results, error analysis, and discussion. Finally, Section 5 provides conclusions, limitations and future work.

2 *Related work*

The main traditional approaches in SA can be divided into two techniques: Machine Learning-based Approach and the Lexicon-based Approach (Raghunathan and Saravanakumar, 2023). In relation to the Lexicon-based Approach, Taboada et al. (2011) present the Semantic Orientation CALculator (SO-CAL), a rule-based SA tool that relies on a lexicon containing the polarity and strength of words. The Machine Learning-based Approach has evolved significantly over the years, transitioning from traditional machine learning(ML) techniques to recent approaches that leverage fine-tuned transformer-based models. In the last decade, transform-based language models with different architectures, encoder-only,

[1] Datasets are available at https://github.com/gamallo/sentiment_analysis_galician_datasets/ and fine-tuned models at https://huggingface.co/collections/anxoanxo/galician-sentiment-analysis-67b266cd0bd4ddcd843e6d33

decoder-only and encoder-decoder, have been the most used strategy for almost every linguistic task. The work of Sun et al. (2020) investigates different strategies for the fine-tuning of BERT models (encoder-only), managing to outperform the previous traditional methods for eight text classification datasets.

Instructed LLMs (mostly decoder-only models) have recently gained importance. In Zhang et al. (2023) a prompt learning-based approach is introduced. The findings of this study show that this method relies heavily on the quality of the hand-crafted prompt, even when a hybrid method, employing a hand-crafted part and an automated part, is used. Edwards and Camacho-Collados (2024) shows that fine-tuned EOMs tend to perform better than decoder-only LLMs for tasks related to text classification, such as SA.

Whereas interest in SA in Galician has been growing, Galician still has limited linguistic tools for this task. To overcome the lack of resources in Galician, Fernández and Campos (2011) leverage the existing resources for Spanish and Portuguese. As far as we know, the only system of SA for the Galician language is LinguaKit (Gamallo and Garcia, 2017), a multilingual suite of linguistic tools for tasks such as SA. The module for SA is based on a hybrid technique consisting of a Bayesian model supported by a polarity dictionary, working for Portuguese, Spanish, English and Galician. Concerning transformer-based models for Galician language, two EOMs based on BERT were trained with relatively small corpora: BERTinho (Vilares, Garcia, and Gómez-Rodríguez, 2021) and BERT-Galician (Garcia, 2021), each available in small (6 layers) and base (12 layers) transformer versions. Recently, a family of generative models for Galician, called Carballo, with 1.3B parameters and trained on the corpus CorpusNós (de Dios-Flores et al., 2024) has been released (Gamallo et al., 2024a). The Galician-Portuguese version is introduced in Gamallo et al. (2024b). In the present work, we will make use of the Galician EOMs with fine-tuning strategies.

Even though most of the publications on SA come from the field of computer science (Wankhade, Rao, and Kulkarni, 2022), the work of linguists on any area of NLP is always important. Many researchers have studied how different linguistic phenomena affect SA. On the topic of negation, the SFU Review$_{SP}$-NEG corpus (Jiménez-Zafra et al., 2017), a corpus annotated at the sentence level with negation cues, their corresponding scopes and events, and the impact of negation on words within the scope, including changes in polarity or shifts in the intensity of their values was developed. It is used in Jiménez-Zafra et al. (2021) to enhance the results of SA. Taboada et al. (2011) deal with negation by changing the polarity and/or strength of the word affected by negation. While these approaches refine sentiment detection with negation, in our work we solely trained models with unannotated sentences in terms of negation, by taking advantage of the LLMs ability to find underlying structures and dependencies.

	Positive	**Neutral**	**Negative**	**Total**
Training dataset	15,610	14,034	16,174	45,818
Senti-GAL	424	224	350	998
Synthetic test dataset	50	50	50	150

Table 1: Number of sentences in the Galician datasets divided by polarity.

Liebrecht et al. (2013) note how sarcasm plays an important role on extracting sentiment out of a text. In the article, they collect a training corpus with tweets that contain the hashtag '#sarcasm' to train a model for sarcasm detection. Similarly, Riloff et al. (2013) analyze sarcasm in tweets attending to the contrast between positive sentiment and a negative situation. Regarding deep learning, Martini et al. (2018) studied the recognition of ironic sentences using attention mechanisms and, although this work is not focused on SA, it was mentioned as one of its main applications. Sarcasm and irony is one of the linguistic phenomena studied in our present work, being one of the factors that most hinders the correct detection of sentiment and polarity.

Several authors have taken into account syntax and compositional rules when working on SA (Diwali et al., 2022; Gómez-Rodríguez, 2020). In Chen et al. (2017), it is described an approach which first classifies sentences into different syntactic types, and then performs SA separately on sentences from each type. Fernández-Gavilanes et al. (2015) remark the importance of the distinction of concessive and adversative sentences and the clauses containing the pragmatic focus for their unsupervised approach for SA. Other work pertaining the analysis of conditional, concessive and adversative sentences include Liang et al. (Liang et al., 2022); Baker and Hashimoto (2024) and Wang et al. (2023). Similarly, in our

work, we address how conditional, concessive and adversative sentences affect the polarity of sentences.

The last phenomenon we study is diglossia. Diglossia is addressed in several articles regarding Arabic NLP (Alomari, ElSherif, and Shaalan, 2017; Jbel et al., 2024). Weidlich (2021) explores how diglossia in Arabic affects sequence labeling, which has been integrated into SA systems. Le et al. (2016) opt for a normalizing approach to overcome diglossia in Indonesian SA. In their approach, they employ a dictionary with words they considered informal and orthographic variants and their corresponding formal versions.

3 Datasets

3.1 Training dataset

To create the training dataset we used two existing datasets, [2] [3] annotated with positive, neutral and negative polarity, and another dataset[4] annotated for five emotions: 'love', 'happy', 'fear', 'anger' and 'sadness'. In order to convert the emotions in polarity classes, we considered the first two classes positive and the other three classes negative. These datasets were chosen because they include a vast range of topics covering, among others, business, technology, market trends, finance and journalism. Furthermore, these datasets cover a wide range of registers. As these datasets were in English, they were translated using the machine translator developed by the Nós Project (Outeirinho et al., 2024). We combined all datasets into one training dataset. In addition, we added 606 sentences that contained the linguistic phenomena studied using GPT-3.5 and GPT-4 via ChatGPT.[5] More precisely, the chat was required to generate polarity sentences with different syntactic types, as well as sentences containing diglossic phenomena and sarcasm. Some of them were generated with negation markers. Although the model was asked to produce positive, neutral or negative sentences, the sentences were revised and manually annotated by one of the authors. The final dataset contains more than 45K sentences (see Table 1) and almost 80K tokens, giving rise to the first Galician training corpus for SA freely available.

[2] https://www.kaggle.com/datasets/ishantjuyal/emotions-in-text

[3] https://www.kaggle.com/datasets/sixlack/finaldf

[4] https://www.kaggle.com/datasets/abhi8923shriv/sentiment-analysis-dataset/data

[5] OpenAI, GPT-3.5 and GPT-4 models accessed through ChatGPT, available at https://chat.openai.com.

3.2 Test datasets

Senti-GAL is a hybrid dataset containing AI-generated sentences and sentences written by humans. It is designed to evaluate sentences with challenging linguistic phenomena.

To build the Senti-GAL test dataset, first, we developed a set of sentences containing the linguistic phenomena we worked on. These sentences were added to the dataset and used as seeds for LLMs to generate similar ones. More precisely, we used GPT-3.5 and GPT-4 via ChatGPT[5](due to the limitations of ChatGPT's free version) to generate simple sentences and sentences containing the different linguistic phenomena we wanted to cover, using as a prompt a direct query with some examples (few shot strategy). The generated sentences were revised by one of the authors. The dataset revision followed guidelines focused on error and hallucination detection while also ensuring sentence diversity. Regarding errors and hallucinations, we discarded any sentences that were ungrammatical or did not align with real-world conditions. In order to maintain diversity, we discarded sentences that repeated the same structure to express the same idea. However, we included sentences that conveyed the same idea using different structures and vice versa.

Concerning the relationship between syntax and polarity, we consider that concessive sentences have positive polarity when the main clause (the thesis) is of positive polarity (see Example 1). In this example, the polarity of the complete sentence is positive because the clause expressing the main thesis (*I'm happy with the result*) is positive, despite the fact that the antithesis expresses a negative sentiment: *it was not what I expected*. By contrast, for adversative sentences, we consider them positive when the antithesis clause (the one following the adversative conjunction) was positive (see Example 2). In this example, the whole sentence inherits the positive polarity of the sentence following the conjunction "pero" / *but*. The Senti-GAL dataset contains occurrences of concessive and adversative sentences with different connectors.

Example 1

Original sentence: Aínda que non é o que esperaba, estou contento co resultado.
Translation: *Even though it was not what I expected, I'm happy with the result.*
Polarity: positive

Example 2

Original sentence: Non é moderno, pero é moi cómodo.
Translation: *It isn't modern, but it is comfortable.*
Polarity: positive

Regarding conditional sentences, we did not follow a specific process for annotating the polarity. While most of the times we focused on the antecedent clause to assign the global polarity (Example 3), in some cases we analyzed the polarity based on the consequent clause (Example 4). For the first type of sentences we consider that the use of the conditional implies the antecedent does not reflect the real state of things, therefore we assigned the polarity based on the real conditions. For instance, in Example 3 not having money, which is expressed in the antecedent, is considered negative, hence the polarity of the whole sentence is also negative. For the second type of conditional sentences, we considered the consequent to be the determining part to assign polarity, given that the antecedent does not provide relevant information about the conditions of reality that are necessary for polarity assignment.

Example 3

Original sentence: Se tivese diñeiro, non pediría prestado.
Translation: *If I had money, I wouldn't borrow.*
Polarity: negative

Example 4

Original sentence: Non viría á festa se soubese que el está alí.
Translation: *I wouldn't come to the party if I knew he was there.*
Polarity: negative

Senti-GAL also contains sentences with irony, to evaluate the model's capacity to handle the pragmatic aspect of language. Although it is difficult to determine whether a sentence contains irony or not, we only include sentences that were not ambiguous in terms of ironicity. This is discussed in the Conclusions (Section 5), as one of the limitations of our approach.

To further explore the topic of irony in all of its complex dimensions, we manage to cover not only irony of negative polarity (Example 5), but also irony with positive polarity (Example 6). We considered a sentence ironic when there was a mismatch between its literal meaning and the usual connotation of the elements in the sentence, following the approach of studies such as Chowdhury and Chaturvedi (2021). To annotate sentiment, we prioritized the connotation of the elements of the sentence over the literal meaning. In Example 5, we considered the burning of the food inherently negative; therefore, even though the speaker literally expresses a positive sentiment by saying "how lucky", we annotated the sentence as negative.

Example 5

Original sentence: Vaia, outra vez a comida queimouse, que sorte.
Translation: *Wow, the food burned again, how lucky.*
Polarity: negative

Example 6

Original sentence: É insoportable ter tanta sorte.
Translation: *It is unbearable to be this lucky.*
Polarity: positive

Lastly, we also consider the diglossic context in which Galician is spoken, so the dataset is provided with sentences containing foreign words from English and Spanish. That way, we can test the models for how they manage foreign words. Example 7 presents a sentence containing a word in Spanish, highlighted in italics.

Example 7

Original sentence: Ese tipo é un paleto
Translation: *That guy is a hick*
Polarity: negative

In addition to Senti-GAL, which is a challenging dataset with complex linguistic phenomena, we build a simpler dataset with synthetic sentences without special difficulties or complex structures that interfere with the SA. The objective is to compare the results of the evaluated models with both datasets and to observe to what extent the linguistic phenomena studied hinder the detection of sentiment/polarity. The synthetic test dataset was also created with the help of GPT-3.5

and GPT-4 via ChatGPT.[5] This dataset is formed by sentences which do not contain as many instances of the linguistic phenomena studied.

Table 1 presents an overview of the polarity of the sentences in the datasets we have introduced. The presence of instances of the three classes, positive, negative and neutral, is balanced in the three datasets.

4 Experiments

4.1 Models

Seven models were evaluated on our test datasets, belonging to three classification paradigms: ML, In-Context Learning, and Fine-Tuning. Five of these models are multilingual and two of them are monolingual.

The first paradigm is represented by the SA module of Linguakit for Galician, which consists of a traditional ML model enhanced with a polarity lexicon, trained on tweets translated into Galician (Gamallo and Garcia, 2017). Since the system does not provide how to train the model, we have used it with the default model and lexicon.

The second paradigm is represented by the open instructed multilingual LLM currently considered state of the art: Llama-3.1-8B-Instruct, a decoder-only model. Being an instructed model, it has been used in a zero-shot configuration, i.e., without examples, and with prompts prepared for sentiment classification. We tested four different prompts (see Appendix B) constructed manually, and, in the experiments, the results of all four were averaged.

The five remaining models are encoder-only LLMs which were fine-tuned with our training dataset. The fine-tuning was made using the Transformers library (Wolf et al., 2020). These five models were configured with a batch size of 16 examples, which required adjustment based on the GPU's memory capacity to avoid out-of-memory errors. The training dataset was split into two partitions, 90% for training and 10% for validation. Two epochs were completed. The learning rate was set to 2e-5, a weight decay of 0.01, and 500 warmup steps. Concerning the tokenizer, the maximum allowable token length for the input sequence was 128, being truncated if any of them exceeds this limit.

Three of the five fine-tuned models use a classical BERT architecture, and the other two use a RoBERTa architecture and a DeBERTa architecture, respectively.

RoBERTa (Robustly optimized BERT approach) architectures were developed inspired by BERT models. Liu et al. (2019) enhanced the BERT architecture by extending the training duration, using larger batch sizes with more data, eliminating the next sentence prediction objective, training on longer sequences, and dynamically adjusting the masking patterns applied to the training data.

Classical BERT architectures represent words using a vector. This vector contains the word embedding and position embedding. DeBERTa, on the other hand, represents this information using two different vectors. This allows DeBERTa to work, not only with absolute positions like classical BERT architectures, but with relative positions too (He et al., 2020).

The following is a more detailed description of the five EOMs.

BERT for Galician (Base)

The base version of BERT for Galician,[6] referred to as BERT-gl in this article, is a 12 layer monolingual model "initialized from the official pretrained mBERT" (Garcia, 2021). The model was trained using the training dataset they specifically developed to identify synonymy and homonymy in context.

BERTinho

The base version of BERTinho[7] is another 12 layer monolingual model for Galician. It uses a BERT architecture and is trained on data from the Galician version of Wikipedia. The authors aimed to improve the performance of mBERT for Galician with the development of this model. BERTinho demonstrated better results for POS-tagging and dependency parsing, while mBERT outperformed it in NER (Vilares, Garcia, and Gómez-Rodríguez, 2021).

BERT multilingual base model (cased)

The cased version of the BERT multilingual base model,[8] referred to as mBERT in this article, is trained on 104 languages, those with the largest amount of data in Wikipedia. The training data comes from Wikipedia dumps. (Devlin and Petrov, 2019; Devlin et al., 2018).

XLM-RoBERTa (base-sized model)

The base version of XLM-RoBERTa,[9] referred to as XLM-RoBERTa in this article, is a multilin-

[6] https://huggingface.co/marcosgg/bert-base-gl-cased

[7] https://huggingface.co/dvilares/bertinho-gl-base-cased

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

[9] https://huggingface.co/FacebookAI/xlm-roberta-base

Type	Model	Accuracy	Precision	Recall	F1-Score
ML Model	Linguakit	0.55	0.59	0.52	0.53
Instructed LLM	Llama3.1-8B-Instruct	0.72 (0.03)	0.77 (0.01)	0.72 (0.03)	0.73 (0.04)
Fine-Tuned Models	BERT-gl	0.77	**0.79**	0.77	0.77
	BERTinho	0.72	0.74	0.72	0.72
	mBERT	0.72	0.73	0.72	0.72
	mDeBERTa	**0.78**	**0.79**	**0.78**	**0.78**
	XLM-RoBERTa	0.76	0.76	0.76	0.76

Table 2: Evaluation metrics for different types of models using the Senti-GAL test dataset. The Llama scores correspond to the mean of four evaluations with different prompts, with standard deviations in parentheses.

Type	Model	Accuracy	Precision	Recall	F1-Score
ML Model	Linguakit	0.75	0.79	0.75	0.75
Instructed LLM	Llama3.1-8B-Instruct	0.96 (0.03)	0.96 (0.02)	0.96 (0.03)	0.96 (0.03)
Fine-Tuned Models	BERT-gl	**0.97**	**0.97**	**0.97**	**0.97**
	BERTinho	0.95	0.95	0.95	0.95
	mBERT	0.93	0.93	0.93	0.93
	mDeBERTa	**0.97**	**0.97**	**0.97**	**0.97**
	XLM-RoBERTa	**0.97**	**0.97**	**0.97**	**0.97**

Table 3: Evaluation metrics for different types of models using the synthetic test dataset. The Llama scores correspond to the mean of four evaluations with different prompts, with standard deviations in parentheses.

gual RoBERTa model that was trained on 100 languages. Unlike mBERT, XLM-RoBERTA is not trained solely on Wikipedia data. It also presents a different tokenization process. While mBERT uses a specific tokenization process for each language, XLM-RoBERTa employs a Sentence Piece model that is applied directly to raw data of any language, optimizing the model (Conneau et al., 2019).

mDeBERTaV3

The multilingual version of DeBERTaV3,[10] referred to as mDeBERTa in this paper (He, Gao, and Chen, 2021; He et al., 2020), is a 12 layer model trained on with CC100 multiliingual data. Galician is included among the languages in this dataset.

Notably that we have selected the *base* (and not *large*) versions of the multilingual models, in order to be able to compare them with the monolingual models in Galician, which only have a base version.

4.2 Results

Table 2 presents evaluation metrics (Accuracy, Precision, Recall, and F1-Score) for the seven models introduced before, applied to the challenge Senti-GAL test dataset. The models are categorized into three groups: ML Model, Instructed LLM, and Fine-Tuned Models.

Linguakit, a traditional ML model provided with a polarity lexicon, achieves modest performance, with an F1-Score of 0.53, indicating limited effectiveness compared to other approaches. It is important to note again that it has been used with the default training provided by the suite tool. Therefore, it has not been trained with the same corpus as the fine-tuned models.

Llama3.1-8B-Instruct performs significantly better, with an average F1-Score of 0.73. Standard deviations (e.g., 0.04 for F1-Score) suggest variability across multiple prompts. In this case it should be noted that four different prompts have been used and the result is an average of each of the metric values. Since this is an instructed model, we have carried out a zero-shot strategy.

Fine-tuned transformer-based models outperform both Linguakit and Llama3.1, achieving the highest overall scores. mDeBERTa leads in all metrics, with an F1-Score of 0.78, followed closely by BERT-gl at 0.77. Other fine-tuned models (e.g., XLM-RoBERTa, mBERT) perform well but slightly trail behind mDeBERTa and BERT-gl. This table highlights the superiority of fine-tuned auto-encoders for this dataset, with mDeBERTa achieving the best balance across all metrics. These results are to be expected since this is

[10] https://huggingface.co/microsoft/mdeberta-v3-base

	BERT-gl	BERTinho	mBERT	mDeBERTa	XLM-RoBERTa
BERT-gl	-				
BERTinho	0.6940	-			
mBERT	0.6703	0.7049	-		
mDeBERTa	0.6992	0.6665	**0.6378**	-	
XLM-RoBERTa	0.7494	0.7042	0.6766	**0.7627**	-

Table 4: Pearson correlation between fined-tuned models on Senti-GAL test dataset.

a classification task, not a generation one, which is the type of task to which encoders-only are best suited, as it also was reported in Edwards and Camacho-Collados (2024).

Table 3 shows the results obtained with the same models on the synthetic dataset, which is less complicated than Senti-GAL, which lacks complex linguistic phenomena related to SA. The results are much better, up to 20 points higher in all models and consistently in all metrics. These results are clear evidence that the linguistic phenomena included in the challenging dataset -i.e., negations, adversatives, concessives, or irony-, make automatic SA a very difficult task. In addition, such high values, achieved by the EOMs on the more basic dataset, indicate that their fine-tuning with our training corpus has yielded high quality classifiers for Galician. In Appendix A, we show a more direct visualization of the F-Score values between all models in both datasets.

Table 4 shows the Pearson correlation coefficients between predictions made by the fine-tuned language models -BERT-gl, BERTinho, mBERT, mDeBERTa, and XLM-RoBERTa- on the Senti-GAL test dataset. Higher correlation values indicate greater similarity in the outputs of two models. Some key observations can be made from this table: First, XLM-RoBERTa demonstrates the highest overall correlation values with other models, including a strong 0.7627 correlation with mDeBERTa. Second, the lowest correlation is 0.6378, observed between mDeBERTa and mBERT, indicating these two models produce the most dissimilar predictions, even if they are both multilingual. These two observations allow us to infer that there is more proximity between RoBERTa and DeBERTa architectures than between these two and BERT-based models. This is surprising, since the pre-training corpus used for mBERT and XLM-RoBERTa is closer than the one used for mDeBERTa, indicating that the type of architecture (including here tokenization) is a very relevant differentiating factor. And third, models fine-tuned on the same architecture (e.g., BERT-gl and BERTinho) do not necessarily have a high correlation, so the pre-training corpus also has a very important influence in qualitatively differentiating these models.

Linguistic Phenomenon	**Predicted Label**		**TOTAL**
	Positive	**Negative**	
Adversative Sentence	16	5	21
Concessive Sentence	7	4	11
Conditional Sentence	6	0	6
Diglossic Phenomena	0	4	4
Positive Irony	4	8	12
Negative Irony	13	58	71
Negation Only	29	-	29
Presence of the studied phenomena	75	79	154
Absence of the studied phenomena			69
TOTAL			223

Table 5: Error analysis of linguistic phenomena for mDeBERTa.

4.3 Error analysis

Table 5 presents a summary of the linguistic phenomena that appeared in the sentences mDeBERTa was unable to classify properly. The number of errors in the assignment of polarity for sentences including the linguistic phenomena studied are more than two times the number of errors for sentences that did not include any of them: 154 *vs* 69. This ratio contrasts with the distribution of sentences with and without those linguistic phenomena, as more than half are sentences without these phenomena. The table also distinguishes errors present in sentences with negation and without negation, including sentences with only negation and no other linguistic phenomenon studied (Negation Only). Analyzing the mistakes

	Neutral	Positive	Negative	Total
Neutral	200	57	51	308
Positive	16	343	67	426
Negative	9	23	232	264
Total	225	423	350	

Table 6: Comparison of predicted (col) vs. real (row) values for mDeBERTa.

Label	Recall	Accuracy
Neutral	0.89	0.65
Positive	0.81	0.81
Negative	0.67	0.88

Table 7: Accuracy and recall values for each label for mDeBERTa.

made by the model deeper, we can see that it handles negation better in sentences with conditionals, diglossic phenomena and irony (both positive and negative irony), i.e. for this type of sentences the model produces more errors when negation is absent than when it is present. By contrast, in concessive and adversative sentences, most mistakes are made in sentences with negation. For instance, in adversative sentences negation is present in more than 3/4 of the errors, namely in 16 out of 21.

The model also shows trouble at handling irony, representing 37% of the errors the model produced. Most of these errors are in sentences containing irony of negative polarity. The presence of negation markers does not represent an important factor of the mismatches for ironic sentences, representing only 1/3 of the errors for positive sentences and less than 1/5 for negative sentences.

Table 6 presents the results in terms of predicted (Y axis) and real (X axis) polarity. Notably, the model errs assigning the negative label to positive sentences in 67 occasions, assigning the positive label to neutral sentences in 57 occasions, and assigning the negative label to neutral sentences in 51 occasions. On the basis of these data, Table 7 shows the accuracy and recall of each label. In general terms, the model has more problems in the recall of negative sentences (67%), meaning it is able to correctly detect only 2/3 of negative sentences, and the accuracy of neutral sentences (65%), i.e. it assigns neutral polarity excessively.

Tables 8 and 9 analyze the low values for the recall of negative sentences and the accuracy of neutral sentences, attending to the linguistic phenomena found in the sentences. Upon initial inspection, we see that ironic sentences represent more than half of the mistakes for sentences of

Linguistic Phenomenon	**Predicted Label**		**TOTAL**
	Positive	**Negative**	
Adversative	9	1	10
Concessive	2	0	2
Conditional	4	2	6
Irony	16	55	**71**
Diglossic Phenomena	2	2	4
Negation Only	13	5	18
Absence of the studied phenomena	5	2	7
TOTAL	51	67	**118**

Table 8: Linguistic phenomena in negative sentences mDeBERTa labeled incorrectly (Recall of negative sentences).

Linguistic Phenomenon	**Predicted Label**		**TOTAL**
	Positive	**Negative**	
Adversative	8	9	17
Concessive	6	2	8
Conditional	0	4	4
Irony	2	16	18
Diglossic Phenomena	0	2	2
Negation Only	4	13	17
Absence of the studied phenomena	37	5	**42**
TOTAL	57	51	**108**

Table 9: Linguistic phenomena in sentences that mDeBERTa labeled neutral incorrectly (Accuracy of neutral sentences).

negative polarity (71 out of 118), while, for errors related to the accuracy of neutral sentences, the absence of linguistic phenomena is the most prominent: 42 out 108. A deeper analysis reveals that 82% of the negative sentences labeled as positive contain irony: 55 out of 67. Another important finding is that the 65% of neutral sentences labeled incorrectly as positive do not contain any of the the linguistic phenomena analyzed here: 37 out of 57. This indicates that the linguistic phenomena studied interfere much more in the detection of positive or negative polarity than in the detection of neutrality.

4.4 Discussion

Our results align with Edwards and Camacho-Collados (2024), showing that fine-tuning smaller

and more efficient language models (encoder-only) can still outperform zero/few-shot approaches of decoder-only LLMs for classification-based tasks.

Another relevant aspect of this experimentation is the fact that it is not clear, in the case of encoder-only fine-tuning, whether the base multilingual models are better than the monolingual ones (e.g., the Galician models). Although the best results are obtained with the multilingual mDeBERTa, the monolingual BERT-gl is almost on a par, and is quite superior to the multilingual one with which it is comparable as it shares the same architecture: mBERT. It will be necessary to build a Galician monolingual mDeBERTa to confirm whether or not Galician monolinguals outperform multilinguals in classification tasks.

When analyzing the mistakes of the best-performing model, we can conclude that it encounters more difficulties in classifying sentiment when complex linguistic phenomena interfere, particularly when pragmatic phenomena, such as irony, are involved. However, specific syntactic constructions also play a role in making classification more difficult, especially in positive and negative sentences, rather than in neutral ones.

5 *Conclusions*

We investigate how linguistic phenomena affect SA in a LRL (Galician) on seven different models: a ML model, an instructed LLM (decoder-only), and five different BERT models (encoder-only) fine-tuned with our training dataset. Two of these models are monolingual, while the rest are multilingual. For this task, we develop 2 datasets. Senti-GAL is a test dataset which is made up from 998 sentences that address several syntactic and pragmatic phenomena. On the other hand, we developed a training dataset with 45818 sentences including these linguistic phenomena. The training dataset was applied during the fine-tuning of the five BERT models we trained for this task. Results show multilingual and monolingual models perform similarly. Furthermore, the most effective fine-tuned EOMs perform better on Senti-GAL than the decoder-only model. Regarding the linguistic phenomena studied, our work exhibits that both syntactic and pragmatic phenomena still constitute a problem for SA. The low results with Senti-GAL show that SA is far from being a solved task and needs further linguistic analysis. Our datasets and fine-tuned models are freely available, being the first free resources and models of SA for the Galician language. Future work will compare monolingual vs. multilingual and encoder-only vs. decoder-only LLMs for classification. Furthermore, we plan to explore more sophisticated prompt selection strategies, such as automatic prompt optimization techniques (Pryzant et al., 2023) or few-shot prompt tuning (Guo et al., 2024) to improve evaluation, specially for LRLs. We also aim to provide a more detailed breakdown of the training and test datasets, attending to the distribution of the linguistic phenomena and the quality control measures applied. Finally, further research should assess whether errors stem from linguistic complexity or model limitations by testing higher-resource languages like English and related languages like Spanish and Portuguese.

Concerning limitations, pragmatic phenomena like irony depend heavily on context, so it is difficult to determine whether a single sentence is ironic or not. Inclusion of a broader context will be a determining factor for polarity assignment. Another limitation is the fact that the sentences of the evaluation dataset, Senti-GAL, have not been annotated with the linguistic phenomena of interest. Only sentences that have been misclassified by the best model in the error analysis have been annotated. We have conducted a partial study on a sample of the dataset and have observed that the ratio between sentences with and without the target linguistic phenomena tends to favor the latter, i.e., there are more sentences without these linguistic phenomena than with them. However, annotation of all sentences has yet to be done, which will allow a more detailed study of the influence of these phenomena on the detection of sentiment/polarity.

Acknowledgments

This publication was produced within the framework of: Nós Project, which is funded by the Spanish Ministry of Economic Affairs and Digital Transformation and by the Recovery, Transformation, and Resilience Plan - Funded by the European Union - NextGenerationEU, with reference 2022/TL22/00215336; LingUMT with ref. PID2021-128811OA-I00, and DeepR with ref, both MEC projects; and grant ED431G-2023/04 by the Galician Ministry of Education, University and Professional Training, and the European Regional Development Fund (ERDF/FEDER program).

References

Alomari, K. M., H. M. ElSherif, and K. Shaalan. 2017. Arabic tweets sentimental analysis using machine learning. In S. Benferhat,

K. Tabia, and M. Ali, editors, *Advances in Artificial Intelligence: From Theory to Practice*, pages 602–610, Cham. Springer International Publishing.

Álvarez-Crespo, L. M. and L. M. Castro. 2024. A galician corpus for misogyny detection online. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 22–31.

Baker, M. J. and B. Hashimoto. 2024. Expression of customer (dis)satisfaction in online restaurant reviews: The relationship between adversative connective constructions and star ratings. *International Journal of Business Communication*, 61(1):148–180.

Chen, T., R. Xu, Y. He, and X. Wang. 2017. Improving sentiment analysis via sentence type classification using bilstm-crf and cnn. *Expert Systems with Applications*, 72:221–230.

Chowdhury, S. B. R. and S. Chaturvedi. 2021. Does commonsense help in detecting sarcasm? *arXiv preprint arXiv:2109.08588*.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

de Dios-Flores, I., S. P. Suárez, C. C. Pérez, D. B. Outeiriño, M. Garcia, and P. Gamallo. 2024. CorpusNÓS: A massive Galician corpus for training large language models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 593–599.

Devlin, J., M. Chang, K. Lee, and K. Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Devlin, J. and S. Petrov. 2019. Github repository: bert/multilingual.md at master · google-research/bert — github.com. `https://github.com/google-research/bert/blob/master/multilingual.md`. [Accessed 28-11-2024].

Diwali, A., K. Dashtipour, K. Saeedi, M. Gogate, E. Cambria, and A. Hussain. 2022. Arabic sentiment analysis using dependency-based rules and deep neural networks. *Applied Soft Computing*, 127:109377.

Dsouza, V. S., P. Rajkhowa, B. R. Mallya, D. Raksha, V. Mrinalini, K. Cauvery, R. Raj, I. Toby, S. Pattanshetty, and H. Brand. 2023. A sentiment and content analysis of tweets on monkeypox stigma among the lgbtq+ community: A cue to risk communication plan. *Dialogues in Health*, 2:100095.

Edwards, A. and J. Camacho-Collados. 2024. Language models for text classification: Is in-context learning enough? In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10058–10072, Torino, Italia, May. ELRA and ICCL.

Fernández, P. M. and J. R. P. Campos. 2011. Generación semiautomática de recursos de opinion mining para el gallego a partir del portugués y el español. In *Workshop on Iberian Cross-Language Natural Language Processing Tasks*.

Fernández-Gavilanes, M., T. Alvarez-López, J. Juncal-Martínez, E. Costa-Montenegro, and F. J. González-Castano. 2015. Gti: An unsupervised approach for sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 533–538.

Gamallo, P. and M. Garcia. 2017. Linguakit: uma ferramenta multilingue para a análise linguística e a extração de informação. *Linguamática*, 9(1):19–28, Jul.

Gamallo, P., P. Rodríguez, I. de Dios-Flores, S. Sotelo, S. Paniagua, D. Bardanca, J. R. Pichel, and M. Garcia. 2024a. Open generative large language models for galician. *arXiv preprint arXiv:2406.13893*.

Gamallo, P., P. Rodríguez, D. Santos, S. Sotelo, N. Miquelina, S. Paniagua, D. Schmidt, I. de Dios-Flores, P. Quaresma, D. Bardanca, J. R. Pichel, V. Nogueira, and S. Barro. 2024b. A galician-portuguese generative model. In M. F. Santos, J. Machado, P. Novais, P. Cortez, and P. M. Moreira, editors, *Progress in Artificial Intelligence*, pages 292–304, Cham. Springer Nature Switzerland.

Garcia, M. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*,

pages 3625–3640. Association for Computational Linguistics.

Geetha, M. and D. K. Renuka. 2021. Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model. *International Journal of Intelligent Networks*, 2:64–69.

Ghilene, R., D. Niaouri, M. Linardi, and J. Longhi. 2024. Analysis of socially unacceptable discourse with zero-shot learning. *arXiv preprint arXiv:2409.13735*.

Gómez-Rodríguez, C. 2020. Syntactically enriched multilingual sentiment analysis. In *CEUR Workshop Proceedings*, volume 2693, pages 5–6. CEUR-WS.

Guo, X., Z. Du, B. Li, and C. Miao. 2024. Generating synthetic datasets for few-shot prompt tuning. In *First Conference on Language Modeling*.

He, P., J. Gao, and W. Chen. 2021. DeBERTav3: Improving deBERTa using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

He, P., X. Liu, J. Gao, and W. Chen. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv preprint arXiv:2006.03654*.

Jaspers, J. 2016. Diglossia and beyond. *Oxford handbook of language and society*, pages 179–196.

Jbel, M., M. Jabrane, I. Hafidi, and A. Metrane. 2024. Sentiment analysis dataset in moroccan dialect: bridging the gap between arabic and latin scripted dialect. *Language Resources and Evaluation*, pages 1–30.

Jiménez-Zafra, S. M., N. P. Cruz-Díaz, M. Taboada, and M. T. Martín-Valdivia. 2021. Negation detection for sentiment analysis: A case study in spanish. *Natural Language Engineering*, 27(2):225–248.

Jiménez-Zafra, S. M., M. Taulé, M. T. Martín-Valdivia, L. A. Ureña-López, and M. A. Martí. 2017. Sfu reviewsp-neg: a spanish corpus annotated with negation for sentiment analysis. a typology of negation patterns. *Language Resources and Evaluation*, 52(2):533–569, May.

Le, T. A., D. Moeljadi, Y. Miura, and T. Ohkuma. 2016. Sentiment analysis for low resource languages: A study on informal Indonesian tweets. In K. Hasida, K.-F. Wong, N. Calzorari, and K.-S. Choi, editors, *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 123–131, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Li, Y., R. Bonatti, S. Abdali, J. Wagle, and K. Koishida. 2024. Data generation using large language models for text classification: An empirical case study. *arXiv preprint arXiv:2407.12813*.

Liang, S., W. Wei, X.-L. Mao, F. Wang, and Z. He. 2022. Bisyn-gat+: Bi-syntax aware graph attention network for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics.

Liebrecht, C., F. Kunneman, and A. van den Bosch. 2013. The perfect solution for detecting sarcasm in tweets #not. In A. Balahur, E. van der Goot, and A. Montoyo, editors, *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 29–37, Atlanta, Georgia, June. Association for Computational Linguistics.

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

Martini, A. T., M. Farrukh, and H. Ge. 2018. Recognition of ironic sentences in twitter using attention-based lstm. *International Journal of Advanced Computer Science and Applications*, 9(8).

Outeirinho, D. B., P. G. Otero, I. de Dios-Flores, and J. R. P. Campos. 2024. Exploring the effects of vocabulary size in neural machine translation: Galician as a target language. In P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G. Oliveira, and R. Amaro, editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 600–604, Santiago de Compostela, Galicia/Spain, March. Association for Computational Lingustics.

Öztürk, N. and S. Ayvaz. 2018. Sentiment analysis on twitter: A text mining approach to the syrian refugee crisis. *Telematics and Informatics*, 35(1):136–147.

Prottasha, N. J., A. A. Sami, M. Kowsher, S. A. Murad, A. K. Bairagi, M. Masud, and M. Baz. 2022. Transfer learning for sentiment analysis using BERT based supervised fine-tuning. *Sensors*, 22(11):4157.

Pryzant, R., D. Iter, J. Li, Y. Lee, C. Zhu, and M. Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7957–7968.

Raghunathan, N. and K. Saravanakumar. 2023. Challenges and issues in sentiment analysis: A comprehensive survey. *IEEE Access*, 11:69626–69642.

Riloff, E., A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 704–714.

Robinson, N. R., P. Ogayo, D. R. Mortensen, and G. Neubig. 2023. ChatGPT MT: Competitive for high-(but not low-) resource languages. *arXiv preprint arXiv:2309.07423*.

Skobel, E. 2010. *Reversing Language Shift in Galicia: A Present-Day Perspective*. Dissertation, Linköping University. Retrieved from Linköping University Electronic Press.

Sukanya, L., J. Aniketh, S. Reddy, H. Kumar, et al. 2023. Racism detection using deep learning techniques. In *E3S Web of Conferences*, volume 391, page 01052. EDP Sciences.

Sun, C., X. Qiu, Y. Xu, and X. Huang. 2020. How to Fine-Tune BERT for Text Classification? *arXiv preprint arXiv:1905.05583*.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, 06.

Tang, T., X. Tang, and T. Yuan. 2020. Fine-tuning BERT for multi-label sentiment analysis in unbalanced code-switching text. *IEEE Access*, 8:193248–193256.

Torre, E., 2024. *Multilayered Diglossia: Identity, Ideology, and Linguistic Entropy in Galicia*, pages 121–142. De Gruyter, Berlin, Boston.

Vilares, D., M. Garcia, and C. Gómez-Rodríguez. 2021. Bertinho: Galician bert representations. *arXiv preprint arXiv:2103.13799*.

Wang, Z., Z. Hu, S.-B. Ho, E. Cambria, and A.-H. Tan. 2023. Mimusa—mimicking human language understanding for fine-grained multiclass sentiment analysis. *Neural Computing and Applications*, 35(21):15907–15921.

Wankhade, M., A. C. S. Rao, and C. Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780, February.

Weidlich, M. 2021. *Sequence Labeling Architectures in Diglossia*. Ph.D. thesis, Humboldt-Universität zu Berlin.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

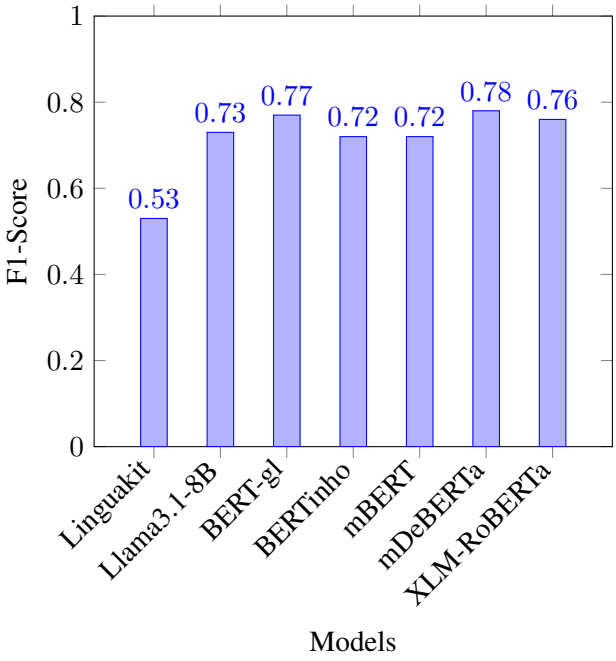
Yuan, F., S. Yuan, Z. Wu, and L. Li. 2024. How vocabulary sharing facilitates multilingualism in llama? *arXiv preprint arXiv:2311.09071*.

Zhang, P., T. Chai, and Y. Xu. 2023. Adaptive prompt learning-based few-shot sentiment analysis. *Neural Processing Letters*, 55(6):7259–7272.

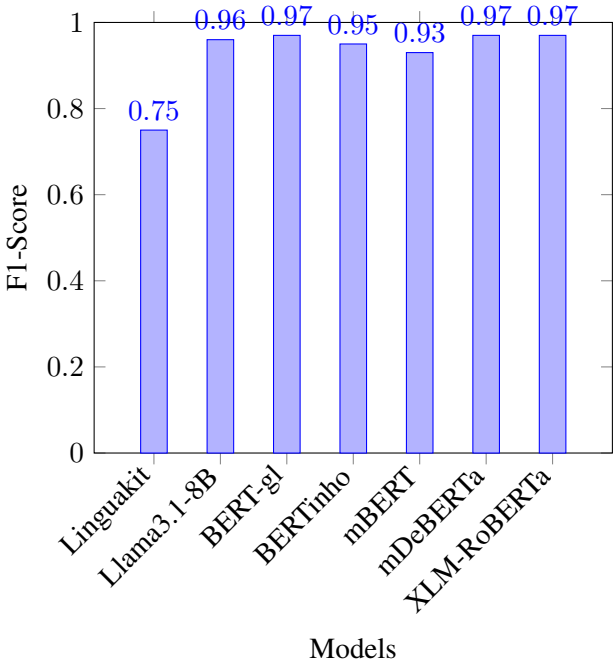
Zhao, B., W. Ren, Y. Zhu, and H. Zhang. 2024. Manufacturing conflict or advocating peace? a study of social bots agenda building in the twitter discussion of the russia-ukraine war. *Journal of Information Technology & Politics*, 21(2):176–194.

A Visualization of F1 Score for all models in the two test datasets

A.1 F1 Score on Senti-GAL



A.2 F1 Score on the synthetic test



B Selected Prompts

Prompt #	Text Prompt	System Role
1	Classify the following sentence as 'neutral', 'negative', or 'positive': '{sentence}'. Answer with one of the following: Positive, Negative, or Neutral.	You are a conversational AI that always responds in Galician and you are helping to classify sentences as positive, negative, or neutral, from those sentences given to you.
2	Determine the sentiment of this sentence as 'Positive', 'Negative', or 'Neutral': '{sentence}'. Respond with only one of these labels: Positive, Negative, or Neutral.	You are an AI assistant that always replies in Galician, helping to analyze the sentiment of the sentences provided to you.
3	Evaluate the sentiment of this sentence: '{sentence}'. Choose one of these categories: Positive, Negative, or Neutral.	You are an AI model that always replies in Galician. Your task is to identify the sentiment of the sentences you are given and respond with Positive, Negative, or Neutral.
4	Classify this sentence '{sentence}' as 'neutral', 'negative', or 'positive'. Reply with one of the following classes: Positive, Negative, or Neutral.	You are a ChatBot who speaks in Galician, and your task is to classify sentences as Positive, Negative, or Neutral.