Context-Aware Stereotype Detection: Conversational Thread Analysis on BERT-based Models

Pol Pastells[1], Wolfgang S. Schmeisser-Nieto[1,2,3], Simona Frenda[3,4] and Mariona Taulé[1,2]

[1]Centre de Llenguatge i Computació (CLiC), Universitat de Barcelona
[2]Institute of Complex Systems (UBICS), Universitat de Barcelona
[3]Dipartimento di Informatica, Università degli studio di Torino
[4]aequa-tech, Turin, Italy

Abstract

Conversational context plays a pivotal role in disambiguating messages in human communication. In this study, we investigate the impact of contextual information on detecting stereotypes related to immigrants using various BERT-based models. We use two Spanish corpora containing news comments and tweets, together with their conversational threads, annotated with stereotypes related to immigrants in Spain. The results show that the influence of context on stereotype detection varies across different models, corpora and context levels. Although context can enhance performance in specific scenarios, it does not consistently improve stereotype detection across all the levels of contexts. Our comprehensive evaluation underscores the complex relationship between context and stereotype identification when we use BERT-based Language Models. In particular, we found that the number of texts benefiting from contextual analysis may be too limited for the models to effectively learn from.[1]

Warning: *This paper contains derogatory language that may be offensive to some readers.*

Keywords

Stereotype Detection, Context, Conversational Thread, Immigration

1. Introduction

The propagation of misleading information that stigmatizes vulnerable social groups such as immigrants has increased during the last decade [1]. Stereotypes are oversimplified, generalized beliefs or perceptions about particular groups of people, often based on prejudices or misconceptions, and social networks have facilitated and aggravated the spread and reinforcement of these stereotypes about marginalized groups.

The identification of negative stereotypes related to immigrants is not simple and involves knowledge of the situation of the analyzed society and an understanding of the conventional meanings and secondary references used by speakers in that society. These meanings and references can be expressed at the discourse level through, for instance, anaphora and ellipsis. In human communication, context serves as the primary strategy to disambiguate and narrow down the interpretation of a particular message. This is observed in the percentage of data that requires knowledge of the context to identify the presence of negative stereotypes related to immigrants in Spain: each annotator needs to read the context in the 21%–39% of the cases identified as stereotypical[2].

Given the critical role of context and the pervasive impact of stereotypes on marginalized communities, we investigated the impact of context on the detection of stereotypes related to immigrants. For human annotators, detecting stereotypes in textual data is a complex task that requires understanding the underlying context and nuances, especially if the stereotype is implicit, that is when the stereotype is not directly stated in the text and there is an inference process to interpret it. (1) shows an example of a tweet from a Multilingual Stereotypes Corpus (MSC) [2] with an implicit stereotype that requires contextual information to classify. It shows the gold standard annotation for the tweet and its context.

(1) **MSC Tweet:** Y los recortes quien los sufrimos? Los que hemos pagado impuestos toda la vida.[3]
'*And who suffers the cuts? Those of us who have paid taxes all our lives.*'
Annotation: [+stereotype] [+implicit] [+contextual]

Previous tweet: RECETA PARA COCTEL XENOFÓBICO. Toma una medida de "Un ilegal tiene los mismos derechos que tú, pero sin pagar impuestos". Añade una medida de "Que entren todos" Agita bien, y ya tienes un partido anti-inmigración a la europea. Servir bien caliente.
'*RECIPE FOR XENOPHOBIC COCKTAIL. Take a measure of "An*

SEPLN-2024: 40th Conference of the Spanish Society for Natural Language Processing. Valladolid, Spain. 24-27 September 2024.
pol.pastells@ub.edu (P. Pastells); wolfgang.schmeisser@ub.edu (W. S. Schmeisser-Nieto); simona.frenda@unito.it (S. Frenda); mtaule@ub.edu (M. Taulé)
0000-0003-1302-1372 (P. Pastells)

[1]Code available on GitHub: https://github.com/pastells/context-aware-stereotype-detection

[2]This range of percentage is extracted from the annotation of MSC.

[3]All examples have been manually translated.

illegal has the same rights as you, but without paying taxes". Add a measure of "Let them all in". Shake well, and you already have a European-style anti-immigration party. Serve while hot.'
Annotation: [+stereotype] [+implicit] [−contextual]

Fake news: Costear la sanidad de los inmigrantes ilegales cuesta 1.100 millones de euros
'*Paying for the healthcare of illegal immigrants costs 1.1 billion euros.*'

Despite the importance of context in human communication and the evident challenge of stereotype identification, there is a noticeable gap in the literature concerning the influence of context on stereotype detection in Natural Language Processing (NLP). Although there is a growing body of research in related areas such as irony [3] and hate speech detection [4], the role of context in resolving stereotype identification has been largely overlooked.

In this paper, we propose adding context to fine-tuned BERT-based models to **observe whether discursive context plays a role in interpreting and disambiguating a message in NLP**, as it does in natural language. We use the only two existing corpora in Spanish annotated with stereotypes against immigrants that also contain context information: DETESTS [2], consisting on online news comments, and MSC, consisting on tweets. Both corpora feature texts embedded in conversational threads, where the contextual utterances include: 1) preceding sentences, 2) previous comments/tweets, 3) first comment/tweet of the thread, 4) wider discourse, such as the news title or the fake news (or hoax) that generates the conversation.

We propose **adding these different levels of context after the *[SEP]* token of the models**. We evaluate the **quantitative performance of the models and the linguistic characteristics of the texts containing stereotypes**, to understand their impact on the models' performance.

The remainder of this paper is as follows: Section 2 reviews related work in the field of stereotype detection. Section 3 details the methodology, including the dataset, experimental setup, and evaluation metrics. Section 4 presents the experimental results and quantitative analysis, followed by a qualitative analysis in Section 5. Finally, Section 6 concludes the paper and outlines potential directions for future research.

2. Related Work

With the development of virtual communications, such as social media, chats and online news comments, there has been a growth of interactions, accompanied by an increase in abusive language, such as stereotypes.

Stereotypes are cognitive resources that humans use to organize the reality they live in and to categorize social groups that they perceive as different. Social groups undergo a categorization process, in which the features associated with that group are attributed to all of its members [5]. Stereotypes are sets of exaggerated beliefs about a social group [6].

Several studies have been undertaken to mitigate this phenomenon. For instance, every year there are more shared tasks oriented at solving automatic stereotype detection affecting various target groups, such as women and immigrants [7, 8, 9, 10, 11, 12]. Other works have taken into account the different textual expressions in which stereotypes appear, especially focusing on implicit forms of stereotypes that are spread through discourses. [13] propose a conceptual formalism to model pragmatic frames in which people project stereotypes onto others. [14] extract microportraits, i.e., descriptions, of Muslims from texts. [15] present a corpus of stereotypes related to immigrants from mentions at the Spanish Parliament. Nevertheless, to our knowledge, the role of conversational context has not yet been studied within the phenomenon of stereotypes, although there are some studies on context-aware models for the detection of abusive language, with rather inconclusive results.

[16] evaluate toxic language in conversational threads from Wikipedia using two types of GRU, CNN and LSTM models, one trained with single comments and another one considering its context. However, the context-sensitive models did not significantly outperform the single-comment ones. In [17] the authors tried a range of different approaches to add context to LSTM, CNN and BERT-like models for the detection of hate speech, all with negative or neutral results. The authors hypothesized that context-sensitive comments are not frequent enough for the models to learn from them. Therefore, the majority of comments would not need context for the correct classification and those that would require context would not get sufficient attention.

[18] use a dataset of Facebook posts to identify hate speech with a Dutch pre-trained language model, BERTje. On the contrary to the previous works, they obtain positive results when training context-aware models when those contexts are controlled and manually annotated as relevant for the classification of hate speech. On the same line of positive results, [19] explore context-aware models for the detection of hate speech. Their dataset consists of Twitter posts from Argentinian news outlet accounts. For their experiments, they trained BETO, a BERT-based model in Spanish, concluding that some contextual information is beneficial for hate speech detection. In particular, the smallest context, which corresponds to the news title tweet, gave the best results.

In relation to the length of contexts, [20] present their participation in a shared task on context-aware sarcasm detection using BiLSTM, BERT, and SVM classifiers on Twitter and Reddit posts. The models were trained with

five scenarios: zero context, last sentence of the context, two sentences, three sentences, or all the sentences of the context. Likewise, we use different types of contexts, described in Section 3.1. They obtained the best results when only the last sentence was provided.

From this related work, to our knowledge, there are no works so far that inject this type of context into stereotypes detection in Spanish, however, we are aware of the inconclusive results that previous studies show.

3. Methodology

To analyze the models' behavior when provided with different levels of context, we used two existing datasets annotated with the presence of negative stereotypes regarding immigrants. In this section, we describe the used datasets and models.

3.1. Datasets

We used two Spanish corpora annotated with binary values indicating the presence of immigration stereotypes and if the stereotypes are expressed explicitly or implicitly in the text. Table 1 summarizes the two corpora.

DETESTS [12] consists of sentences extracted from comments posted in response to news articles in Spanish newspapers (such as *ABC*, *elDiario.es* and *El Mundo*) and discussion forums (such as *Menéame*). The articles were manually selected based on their immigration-related subject and potential toxicity. Each comment was segmented into sentences. The comment to which every sentence belongs and its position within the comment and thread are indicated in the corpus. Each sentence was annotated by three trained annotators, that had access to the entire comment the sentence belonged to when annotating, along with the news title and the rest of the comment thread. Example (2) shows an implicit stereotype and its contexts:

(2) **DETESTS Sentence:** Y las violaciones.
'*And the rapes.*'
Annotation: [+stereotype] [+implicit]
Previous comment: Y que siga la fiestaaaaa!!!!
'*And let the party continue!!!!*'
News title: Inmigrantes ilegales paralizan el aeropuerto de Palma al huir de un avión marroquí.
'*Illegal immigrants paralyze Palma airport when fleeing a Moroccan plane.*'

MSC [2] is a corpus of Twitter posts (tweets) responding to hoaxes that disseminated fake news against immigrants in newspapers or social media. The tweets were annotated by three trained annotators for the presence of stereotypes and their implicitness. Furthermore, during the annotation process, annotators considered the need to look into the context to decide if there was a stereotype. In those cases, the tweet was annotated as *contextual*. Out of the 1,604 tweets with stereotypes, 590 (37%) were annotated as *contextual*, with 253 (16%) of this subset also categorized as *implicit*. An example of this last case is shown in Example (1). MSC differs from DETESTS in that the corpus does not contain the full Twitter threads, but rather a subset of them (previous tweet, first tweet and the hoax). Therefore, the annotators did not have access to the entire conversational context, as they did in DETESTS.

Another notable distinction between the texts in both corpora is that DETESTS comprises individual sentences, with a median length of 13 words[4], whereas MSC consists of full, unsegmented tweets, with a median of 26 words[5].

The corpora are structured into threads, where the first direct comment or tweet (text from now on) on the article or post is the root of the thread. Each text can then have multiple responses, forming a tree structure. We identified a range of different contexts to which annotators had access, in order to provide them to the models. We structured the contexts into four levels, summarized in Table 2:

1. Previous sentences in the same comment (level 1). This level is only available for DETESTS, as MSC tweets were not split into sentences. Additionally, this level does not apply to the first sentence of each comment, which constitutes 45% of sentences in the DETESTS.
2. Previous text in the thread (level 2). This level is absent for the first comment in each thread,

Table 1
Label distribution for DETESTS and MSC.

	DETESTS	MSC
Total Instances	5,629	5,349
No Stereotype	4,270	3,745
With Stereotype	1,359	1,604
Implicit	1,056	344
Contextual	–	590

Table 2
Context levels for DETESTS and MSC.

	DETESTS	MSC
Text	Sentence	Tweet
Level 1	Previous Sentences	–
Level 2	Parent Comment	Parent Tweet
Level 3	Root Comment	Root Tweet
Level 4	News Title	Hoax

[4]With $Q_1 = 7$ and $Q_2 = 20$.
[5]With $Q_1 = 14$ and $Q_2 = 41$.

accounting for 45% of comments in DETESTS and 8% of tweets in MSC.

3. Root text (level 3). This level does not exist for the first comment of each thread and is identical to the previous comment for the second comment on each thread. It is missing in 45% of comments and 16% of tweets. Note that DETESTS has full threads, so the comments missing level 2 and the ones missing level 3 are the same, while for MSC they are different, although overlapping, sets.
4. News title for DETESTS or fake news text for MSC (level 4). This level is always present and differs from the others in that it does not represent an instance of the dataset, but an external reference.

Even though the contexts for DETESTS are formed by various sentences, they are still smaller (median of 21 words for *previous sentences*, with $Q1 = 13$ and $Q3 = 41$) than the MSC contexts (median of 34 words for *root text*, with $Q1 = 22$ and $Q2 = 49$). This is due to the distribution of the comment threads, most of them having few comments.

3.2. Models

We fine-tuned three pretrained models from the BERT family for the classification task of stereotype detection. The models were trained to output a binary label: 0 for no stereotype, and 1 for stereotype. We are aware of the subjectivity of this task [21], however, considering the evaluative scope of this work, we focused on the gold standard version of the above-mentioned corpora.

We used two different models pretrained in Spanish and also multilingual BERT [22]. The selected models, obtained from the *Huggingface* transformers library (https://huggingface.co/), were:

BETO *dccuchile/bert-base-spanish-wwm-cased* [23], based on the BERT-Base architecture, was trained with the Whole Word Masking technique.

MarIA *PlanTL-GOB-ES/roberta-base-bne* [24], based on the RoBERTa-Base model, pre-trained using 570 GB of Spanish texts, extracted from the Spanish Web Archive crawled by the National Library of Spain.

M-BERT [22] *google-bert/bert-base-multilingual-cased*, based on BERT-Base, pre-trained on the top 104 languages with the largest Wikipedia using the original masked language modeling objective.

For each of the three models and both DETESTS and MSC, we fine-tuned a model without context (as baseline), and a different model incorporating each possible context level. To add the context to the input, we used the sequence *text + [SEP] + context*, where *[SEP]* is the special BERT token that is usually used to split sequences in BERT-based models.

To address the issue of missing contexts during the fine-tuning process, we employed a hierarchical filling strategy. Specifically, if a lower-level context (e.g., level 1) was absent, it was replaced with the next highest level (e.g., level 2). If both level 1 and level 2 were lacking, they were both filled with level 3, and so on. This approach was taken into consideration during the qualitative analysis, ensuring that any observed improvements were attributed to the filled context rather than the missing one.

Both corpora were split in a stratified manner to maintain the same proportion of stereotypes, implicitness and stereotype topics[6][12].

To prevent variability in the results, we decided to use 50 random seeds for training the models and report the average of their results. The data split was the same for all seeds. All models were trained[7] with a 512 token window, using batches of 32 texts and evaluating the results every 50 steps, with early stopping.

4. Quantitative Analysis

We first compared the models with and without context using various metrics. Figures 1 and 2 show the F_1 metric, precision, and recall for both the negative and the positive classes, i.e., the texts with or without stereotypes in the gold standard annotation. The bars represent the median across 50 seeds, with the error bars indicating the first and third quartiles. Furthermore, arrows mark a p-value smaller than 0.05 in a Welch's t-test for each metric, comparing the 50 seeds with and without context. The direction of the arrows denotes an improvement (up) or deterioration (down) in respect to the model without context.

We further examined the texts whose predictions changed upon adding context, in order to focus on the differences between the models. Given the numerous seeds used in our models, we identified texts with consistent classification changes in more than 65% of the seeds. For instance, for true positives (TP), we considered a text classification to have changed if more than 65% of the seeds without context failed to classify it as a stereotype, while more than 65% of the models with a specific context correctly identified it as a stereotype. Moreover, we examined all potential changes, including TP, true negatives (TN), false positives (FP), and false negatives (FN). These cases are shown in Tables 3 and 4 and are the same ones subjected to qualitative analysis in Section 5.

DETESTS predictions. Initially, we looked at the difference in the F_1 metric for the negative and the positive classes. To provide a more comprehensive analysis, we

[6] Although not used for this work, the corpora were also annotated with topics.

[7] We used a single GeForce RTX 4090 GPU, with 24 GB of RAM.

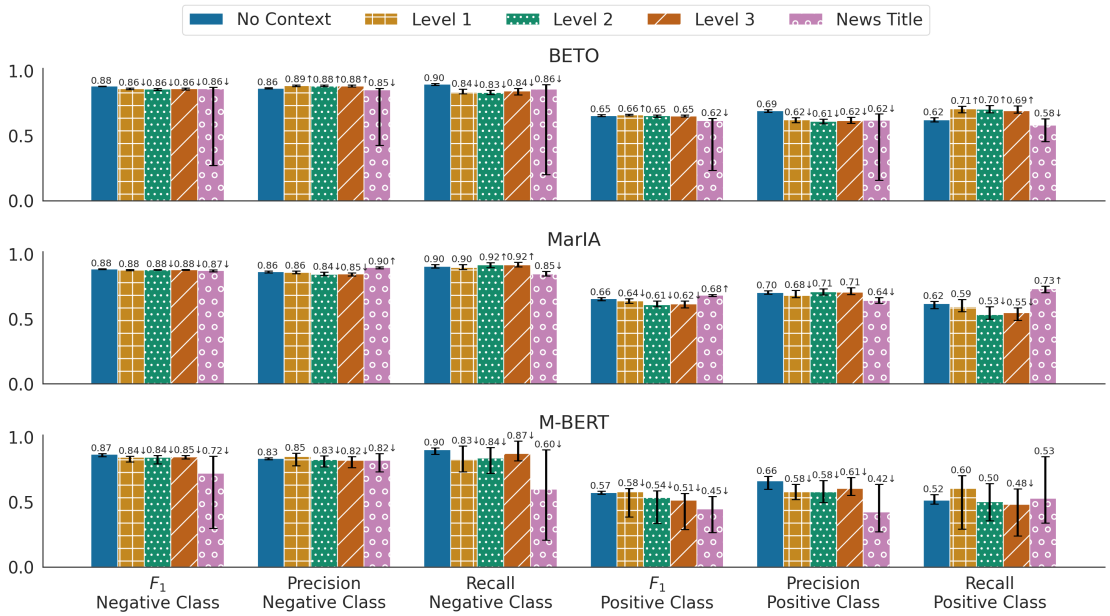

Figure 1: DETESTS models F_1 scores, precision, and recall for the negative and positive classes. The bar values represent the median of the 50 seeds, error bars show the first and third quartiles. Arrows mark statistically significant differences from the model without context.

Table 3
Number of DETESTS sentences with a different model classification for each context compared to the same model without context, grouped by gold label and classification according to the confusion matrix. The percentage of improvement (↑) or worsening (↓) with respect to the models without context, with more than 65% seed-agreement, is shown in parentheses.

Model	Category	No Context > 65% seeds	Level 1 Changes		Level 2 Changes		Level 3 Changes		Level 4 Changes	
BETO	FP	163	**40**	**(25% ↓)**	**50**	**(31% ↓)**	**37**	**(23% ↓)**	4	(2% ↓)
	TN	1173	0		1	(0% ↑)	2	(0% ↑)	3	(0% ↑)
	FN	112	0		1	(1% ↓)	3	(3% ↓)	3	(3% ↓)
	TP	292	16	(5% ↑)	19	(7% ↑)	14	(5% ↑)	2	(1% ↑)
MarIA	FP	169	11	(7% ↓)	7	(4% ↓)	7	(4% ↓)	**41**	**(24% ↓)**
	TN	1187	12	(1% ↑)	14	(1% ↑)	16	(1% ↑)	1	(0% ↑)
	FN	98	8	(8% ↓)	**21**	**(21% ↓)**	**28**	**(29% ↓)**	0	
	TP	277	6	(2% ↑)	3	(1% ↑)	3	(1% ↑)	**35**	**(13% ↑)**
M-BERT	FP	198	6	(3% ↓)	1	(1% ↓)	0		3	(2% ↓)
	TN	1167	0		0		2	(0% ↑)	0	
	FN	86	1	(1% ↓)	1	(1% ↓)	**11**	**(13% ↓)**	0	
	TP	223	3	(1% ↑)	0		0		1	(0% ↑)

also added the precision and recall metrics. This was crucial, as in some instances, a consistent F_1 value obscured variations in precision and recall, either in terms of improvement or decline. These metrics are presented in Figure 1.

For BETO, there was a slight, yet statistically significant, deterioration in performance for the negative class. When evaluating the F_1 metric for the positive class, *level 1* is the only context that improves. The enhancement was driven by an increase in recall, but counterbalanced by a decrease in precision. A comparable trend was observed for *levels 2* and *3*. BETO showed an increase in FP cases and a drop in FN, indicating a tendency to classify more sentences as containing stereotypes when some context is provided.

Models using *news title* as context show a wide vari-

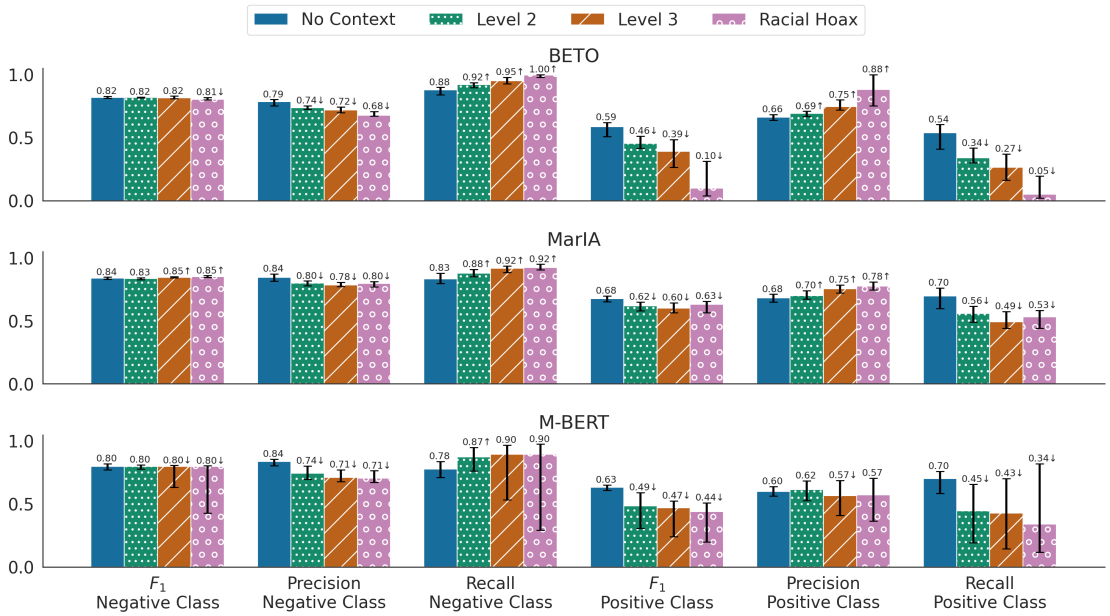

Figure 2: MSC models F_1 scores, precision, and recall for the negative and positive classes. Bar values represent the median over 50 seeds, error bars show the first and third quartiles, and arrows mark statistically significant differences from the model without context.

Table 4
Number of MSC tweets classified differently in more than 65% of the seeds for each context, compared to the same model without context, grouped by gold label and classification according to the confusion matrix. In parentheses, the percentage of improvement (↑) or worsening (↓) is shown, with respect to the models without context with more than 65% seed-agreement. The third comment shows the number of this baseline cases.

Model	Category	No Context > 65% seeds	Level 2 Changes		Level 3 Changes		Level 4 Changes	
BETO	FP	142	12	(8% ↓)	0		0	
	TN	609	28	(5% ↑)	34	(6% ↑)	55	(9% ↑)
	FN	58	**29**	**(50% ↓)**	**51**	**(88% ↓)**	**113**	**(195% ↓)**
	TP	148	2	(1% ↑)	1	(1% ↑)	0	
MarIA	FP	83	1	(1% ↓)	1	(1% ↓)	0	
	TN	566	11	(2% ↑)	18	(3% ↑)	23	(4% ↑)
	FN	68	**22**	**(32% ↓)**	**22**	**(32% ↓)**	**27**	**(40% ↓)**
	TP	226	3	(1% ↑)	0		0	
M-BERT	FP	80	0		0		0	
	TN	504	24	(5% ↑)	34	(7% ↑)	39	(8% ↑)
	FN	112	**30**	**(27% ↓)**	**22**	**(20% ↓)**	**28**	**(25% ↓)**
	TN	227	0		0		0	

ability across the two classes in BETO and M-BERT, as evidenced by the disparity between the first and the third quartiles, with an overall worsening tendency.

In contrast, MarIA behaves differently. It showed a general decline in performance on the F_1 metric for the positive class, primarily due to an increased classification of sentences as not containing stereotypes; except when the model, informed with *news title* context, reports a significant improvement. Lastly, M-BERT's performance, providing the context, shows no significant change in all scenarios.

Table 3 shows the individual texts that change for each

model and context, grouped by category, according to the predictions of the models with context, similarly to a confusion matrix. The arrows denote an improvement (TN and TP) or deterioration (FN and FP). FP and FN changes are misclassified texts with context that were correctly classified without context. Therefore, cases where the context does not help the models. TP and TN changes, instead, are the instances where the context helps the models make the correct prediction. For example, 163 is the number of sentences that did not have stereotypes in their gold label, but were classified as having one (FP) in more than 65% of the seeds for the BETO model without context. The model with *level 1* contexts has 40 more FP (25% increase).

Looking at this table, BETO shows the biggest change in FP, with similar numbers for *level 1*, *level 2* and *level 3* contexts. It also shows a slight improvement in TP for the same contexts. This behavior can be explained by the model just tending to classify more texts as stereotypes, in agreement with the metrics in Figure 1.

MarIA shows a similar behavior, although with the contexts reversed. It tends to classify more sentences as stereotypes when given the *news title* as context, but not so much for the rest of the contexts. Instead, *level 2* and *level 3*, appear to worsen the negative class, with an increase in FN. M-BERT is the model that has less consistent changes, with only a change of more than 10% in the FP with *level 3* context, similarly to MarIA.

MSC predictions. All classification models got biased toward predicting 0, that is, the models tend to predict fewer stereotypes. This can be seen in Figure 2 with the negative class precision and positive class recall worsening, while the negative class recall and positive class precision tending to improve, except for M-BERT. It is also made evident in Table 4, for all three models, adding any context makes the models' FN increase significantly.

Similarly to DETESTS, the metrics for the *level 4* context, the *racial hoax text*, had a big variability for BETO's positive class and M-BERT's negative and positive classes.

5. Qualitative Analysis

In this section, we present a qualitative analysis of the instances that improved or deteriorated their classification on the models trained with different levels of context, as presented in Tables 3 and 4. Our aim is to gain a deeper understanding of the impact of context on the models' predictions from a linguistic perspective. We describe linguistic patterns by comparing three levels of analysis: by models, by datasets, and by levels of contexts.

In the predictions on DETESTS (Table 3), we observed an increase of sensibility towards the positive class, justified by an improvement of the recall (Figure 1). The FP cases had in common that their contexts tended to contain stereotypes.

Example (3) shows a FP for BETO and M-BERT, where the text was annotated with no presence of stereotypes. However, its context does contain a stereotype:

(3) **DETESTS Sentence:** Si aprenden catalán, serán catalanes, quizás catalanistas.
'*If they learn Catalan, they will be Catalan, maybe Catalan nationalists.*'
Annotation: [−stereotype]

Context: Los detuvo, pero quedarán libres y se irán de rositas. Se quedarán en el país para siempre, se llevarán todo tipo de ayudas y traerán a toda la familia.
'*They arrested them, but they will be released and will walk away with ease. They will stay in the country forever, they will take all kinds of aid and they will bring the whole family.*'
Annotation: [+stereotype] [+implicit] [8]

As in the previous example, the classified texts neither focus on immigrants nor evaluate the in-group regarding immigrants. Instead, the topics of these messages predominantly concern evaluations of the in-group, with conclusions that do not necessarily pertain to the target group. Example (4), FP for both BETO and MarIA, shows an evaluation and a consequence derived from previous texts. Although the sentence has no stereotype, both the *previous sentences* and the *previous comment* contexts contain stereotypes.

(4) **DETESTS Sentence:** Dentro de 20 o 30 años, nuestros hijos y nietos nos maldecirán mil veces por el infierno que les hemos dejado.
'*In 20 or 30 years, our children and grandchildren will curse us a thousand times for the hell we have left them.*'
Annotation: [−stereotype]

Previous Sentences: y ya es tarde, el Caballo de Troya lo tenemos dentro.
'*and it's too late, we have the Trojan Horse within us.*'
Annotation: [+stereotype] [+implicit]

Previous Comment: [...] Están moviendo los hilos de esta invasión, que aprovechan para usar a los Ilegales como sicarios, para agredir y amedrentar a los españoles de bien. [...]
'*[...] They are pulling the strings of this invasion, which they take advantage of to use the Illegals as hitmen, to attack and intimidate good Spaniards. [...]*'
Annotation: [+stereotype] [−implicit]

Another case of FP, for BETO, was found in Example (5). Even though the text concerns immigrants with keywords corresponding to the target group, it contains no stereotype according to the annotators. **Its context, however, was annotated with a stereotype**, even though there is no explicit reference to immigrants. This shows that the model attends to enough tokens from the context to determine the presence of a stereotype, **which**

[8] In fact both sentences from the previous comment contain an implicit stereotype.

drives the model to a positive classification.

(5) **DETESTS Sentence:** En Francia, el paro es de 15% en la población general y de 40% en la inmigrada.
'*In France, unemployment is 15% in the general population and 40% in the immigrant population.*'
Annotation: [−stereotype]

Context: Pobres incautos. Salen como locos en vuelo directo a los invoxnaderos a trabajar por 3 € la hora.
'*Poor dupes. They leave like crazy on a direct flight to the invoxnaderos*[9] *to work for €3 an hour.*'
Annotation: [+stereotype] [+implicit]

Nonetheless, out of eleven DETESTS sentences that were classified as FP by BETO with context levels 1 to 3, only two cases have no stereotypes in any of their contexts. For instance, in Example (6), there is no interpretation of stereotypes neither by human annotators nor by the decision of the models without context. However, adding the context, which was previously annotated as containing no stereotype, the prediction of the model yielded a FP.

(6) **DETESTS Sentence:** Que los pececitos coman cachalote franquista.
'*Let the little fish eat Francoist sperm whale.*'
Annotation: [−stereotype]

Context: Pues lanza a tu madre.
'*Then throw your mother.*'
Annotation: [−stereotype]

Furthermore, the opposite phenomenon occurs when MarIA is fine-tuned: it shows a 24% of deterioration on DETESTS's FP when the *news title* is fed as context. It is worth noting that out of the twelve news articles that were used to create DETESTS, six of them contained in their title a word related directly to the target group, such as *immigrant* or *dinghy*, shown in Example (7). The misclassified texts belong to five of these conversation threads with keywords in their title, which might be an indication that the model was affected by the vocabulary used.

(7) **News Title 1:** La otra crisis con la que lidia Ceuta: un tercio de los contagios son de inmigrantes acogidos.
'*The other crisis that Ceuta is dealing with: a third of the infections are from received immigrants.*'
News Title 2: Una "patera aérea", una nueva e insólita manera de entrar en España de forma irregular.
'*A "flying dinghy", a new and unusual way to enter Spain irregularly.*'

Looking at Table 4, we notice an interesting tendency related to FN in all the models informed with context. The model performance worsens if we introduce context, regardless of the level. To understand the behavior of the models, we observed the instances commonly misclassified as not containing stereotypes by the majority of the models (15 instances). We noticed that, in general, the presence of the hoax as context (level 4) affects negatively the decision of the model. Additionally, with further analysis, we consider that most of these instances contain implicit expressions, inducing the need for context to be understood, as seen in Example (8).

(8) **MSC Tweet:** ...fuerzas políticas, ni policiales, ni legales, para empezar a resolver la situación creada. Y yo creo que ni voluntad de hacerlo. Aquello está lejos y a los peninsulares no les preocupa lo más mínimo. Grave error; gravisimo. Una vez controlen las islas vendrán aquí a reclamar..
'*...political forces, neither police nor legal, to begin to resolve the situation created. And I believe that there is no desire to do so. That is far away and the peninsular people are not the least bit worried. Serious mistake; very serious. Once they control the islands they will come here to complain...*'
Annotation: [+stereotype] [+implicit] [+contextual]

Level 2: Canarias ya está "ocupada" por marroquíes y mauritanos. En las islas orientales, Fuerteventura y Lanzarote, el número de moros ya es mayor que el de la población autóctona. Es una estrategia marroquí que empieza a darle resultados: la toma 'pacífica' de territorios ...
'*The Canary Islands are already "occupied" by Moroccans and Mauritanians. On the eastern islands, Fuerteventura and Lanzarote, the number of Moors is already greater than that of the native population. It is a Moroccan strategy that is beginning to give results: the 'peaceful' seizure of territories...*'

Considering this analysis, we plan to investigate further the role played by the context in future work, exploring other models and their common behaviors.

6. Conclusions

Taking into account the importance of context during the identification of stereotypes in online conversational threads, in this work, we analyzed the impact of different levels of context on stereotype detection in news comments and tweets.

In particular, we performed quantitative and qualitative analyses on predictions obtained with fine-tuned language models informed with different context levels. Quantitatively, no general improvement was seen when adding contextual information after the *[SEP]* token to BERT-based models. The results were highly dependent on the dataset used. In DETESTS, only BETO proves to became more sensible to stereotypes when some context is provided, or MarIA when informed with *news title* context. Whereas in MSC, models are biased towards the negative class.

We hypothesize that the number of texts that benefit from looking at the context is too small for the models to learn from, as suggested by the number of contextual-labeled tweets. The models may also be looking into other subtleties other than the presence of stereotypes.

[9]Word play in which the main word *invernadero* 'greenhouse' is embedded with the far-right wing party's name *Vox*, resulting in 'invoxnadero'.

For example, the context in Example (6) has a negative sentiment, even though it does not contain a stereotype.

Future work may require more involved methods of analysis on the quantitative side, using different embeddings for the text and the context or with approaches like mechanistic interpretability.

Limitations Our work was exclusively focused on the Spanish language and employed solely BERT and RoBERTa models. More advanced generative models, such as Llama 2 [25] or Mixtral 8x7B [26], may offer different ways of capturing context.

Among the various levels of context considered, which differed between the two corpora, only *level 4* was consistently present. The other levels had to be filled to prevent the loss of valuable data. Exploring data augmentation techniques, using synthetic data or curating a dataset without missing contexts, could be a promising direction for future research.

Acknowledgments

This work was supported by the international project STERHEOTYPES: STudying European Racial Hoaxes and sterEOTYPES funded by the Compagnia di San Paolo and VolksWagen Stiftung under the Challenges for Europe call (CUP: B99C20000640007); the SGR CLiC project (2021 SGR 00313) funded by the Generalitat de Catalunya, and the FairTransNLP-Language project (PID2021-124361OB-C33) funded by MICIU/AEI/10.13039/501100011033/ and by FEDER, UE.

References

[1] M. Ekman, Anti-immigration and racist discourse in social media, European journal of Communication 34 (2019) 606–618.

[2] T. Bourgeade, A. T. Cignarella, S. Frenda, M. Laurent, W. S. Schmeisser-Nieto, F. Benamara, C. Bosco, V. Moriceau, V. Patti, M. Taulé, A Multilingual Dataset of Racial Stereotypes in Social Media Conversational Threads, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023), 2023.

[3] B. C. Wallace, D. K. Choe, E. Charniak, Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment, in: C. Zong, M. Strube (Eds.), Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Beijing, China, 2015, pp. 1035–1044. URL: https://aclanthology.org/P15-1100. doi:10.3115/v1/P15-1100.

[4] L. Gao, R. Huang, Detecting online hate speech using context aware models, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, INCOMA Ltd., Varna, Bulgaria, 2017, pp. 260–266. URL: https://doi.org/10.26615/978-954-452-049-6_036. doi:10.26615/978-954-452-049-6_036.

[5] G. W. Allport, K. Clark, T. Pettigrew, The nature of prejudice, Addison-wesley Reading, MA, 1954.

[6] D. L. Hamilton, Cognitive processes in stereotyping and intergroup behavior, L. Erlbaum Associates, 1981.

[7] E. Fersini, D. Nozza, P. Rosso, Overview of the evalita 2018 task on automatic misogyny identification (ami), EVALITA Evaluation of NLP and Speech Tools for Italian 12 (2018) 59.

[8] E. Fersini, D. Nozza, P. Rosso, AMI @ EVALITA2020: automatic misogyny identification, in: V. Basile, D. Croce, M. D. Maro, L. C. Passaro (Eds.), Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020, volume 2765 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2765/paper161.pdf.

[9] F. Rodríguez-Sánchez, J. C. de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6443.

[10] P. Chiril, F. Benamara, V. Moriceau, "be nice to your wife! the restaurants are closed": Can gender stereotype detection improve sexism classification?, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2833–2844. URL: https://aclanthology.org/2021.findings-emnlp.242. doi:10.18653/v1/2021.findings-emnlp.242.

[11] M. Sanguinetti, G. Comandini, E. di Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, Haspeede 2 @ EVALITA2020: Overview of the EVALITA 2020 hate speech detection task, in: V. Basile, D. Croce, M. Di Maro, L. Passaro (Eds.), Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), volume 2765, CEUR Workshop Proceedings (CEUR-WS.org), 2020. Conference date: 17-12-2020.

[12] A. Ariza-Casabona, W. S. Schmeisser-Nieto, M. Nofre, M. Taulé, E. Amigó, B. Chulvi, P. Rosso, Overview of DETESTS at IberLEF 2022: DETEction and classification of racial STereotypes in Spanish, Procesamiento del Lenguaje Natural 69 (2022) 217–228.

[13] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, Y. Choi, Social bias frames: Reasoning about social and power implications of language, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5477–5490. URL: https://aclanthology.org/2020.acl-main.486. doi:10.18653/v1/2020.acl-main.486.

[14] A. Fokkens, N. Ruigrok, C. Beukeboom, G. Sarah, W. Van Atteveldt, Studying muslim stereotyping through microportrait extraction, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018, pp. 3734–3741.

[15] J. J. Sánchez-Junquera, B. Chulvi, P. Rosso, S. P. Ponzetto, How do you speak about immigrants? taxonomy and stereoimmigrants dataset for identifying stereotypes about immigrants, Applied Sciences 11 (2021). URL: https://www.mdpi.com/2076-3417/11/8/3610. doi:10.3390/app11083610.

[16] M. Karan, J. Šnajder, Preemptive toxic language detection in wikipedia comments using thread-level context, in: Proceedings of the Third Workshop on Abusive Language Online, 2019, pp. 129–134.

[17] J. Pavlopoulos, J. Sorensen, L. Dixon, N. Thain, I. Androutsopoulos, Toxicity Detection: Does Context Really Matter?, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4296–4305. URL: https://aclanthology.org/2020.acl-main.396. doi:10.18653/v1/2020.acl-main.396.

[18] I. Markov, W. Daelemans, The Role of Context in Detecting the Target of Hate Speech, in: Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022), Association for Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 37–42. URL: https://aclanthology.org/2022.trac-1.5.

[19] J. M. Pérez, F. M. Luque, D. Zayat, M. Kondratzky, A. Moro, P. S. Serrati, J. Zajac, P. Miguel, N. Debandi, A. Gravano, et al., Assessing the impact of contextual information in hate speech detection, IEEE Access 11 (2023) 30575–30590.

[20] A. Baruah, K. Das, F. Barbhuiya, K. Dey, Context-aware sarcasm detection using bert, in: Proceedings of the Second Workshop on Figurative Language Processing, 2020, pp. 83–87.

[21] W. S. Schmeisser-Nieto, P. Pastells, S. Frenda, M. Taule, Human vs. machine perceptions on immigration stereotypes, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 8453–8463. URL: https://aclanthology.org/2024.lrec-main.741.

[22] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.

[23] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[24] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley. doi:10.26342/2022-68-3.

[25] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[26] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, et al., Mixtral of experts, arXiv preprint arXiv:2401.04088 (2024).