Procesamiento del Lenguaje Natural, Revista nº 67, septiembre de 2021, pp. 83-94 recibido 06-05-2021 revisado 08-06-2021 aceptado 10-06-2021

Masking and BERT-based Models for Stereotype Identification

Modelos Basados en Enmascaramiento y en BERT para la Identificación de Estereotipos

Javier Sánchez-Junquera[1], Paolo Rosso[1], Manuel Montes-y-Gómez[2], Berta Chulvi[1]

[1]PRHLT Research Center, Universitat Politècnica de València, València, Spain;
[2]Laboratorio de Tecnologías del Lenguaje, Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico
{juasanj3@doctor.; prosso@dsic.; berta.chulvi@}upv.es; mmontesg@inaoep.mx

Abstract: Stereotypes about immigrants are a type of social bias increasingly present in the human interaction in social networks and political speeches. This challenging task is being studied by computational linguistics because of the rise of hate messages, offensive language, and discrimination that many people receive. In this work, we propose to identify stereotypes about immigrants using two different explainable approaches: a deep learning model based on Transformers; and a text masking technique that has been recognized by its capabilities to deliver good and human-understandable results. Finally, we show the suitability of the two models for the task and offer some examples of their advantages in terms of explainability.
Keywords: social bias, immigrant stereotypes, BETO, masking technique.

Resumen: Los estereotipos sobre inmigrantes son un tipo de sesgo social cada vez más presente en la interacción humana en redes sociales y en los discursos políticos. Esta desafiante tarea está siendo estudiada por la lingüística computacional debido al aumento de los mensajes de odio, el lenguaje ofensivo, y la discriminación que reciben muchas personas. En este trabajo, nos proponemos identificar estereotipos sobre inmigrantes utilizando dos enfoques diametralmente opuestos prestando atención a la explicabilidad de los mismos: un modelo de aprendizaje profundo basado en Transformers; y una técnica de enmascaramiento de texto que ha sido reconocida por su capacidad para ofrecer buenos resultados a la vez que comprensibles para los humanos. Finalmente, mostramos la idoneidad de los dos modelos para la tarea, y ofrecemos algunos ejemplos de sus ventajas en términos de explicabilidad.
Palabras clave: sesgo social, estereotipos hacia inmigrantes, BETO, técnica de enmascaramiento.

1 Introduction

Nowadays, social media, political speeches, newspapers, among others, have a strong impact on how people perceive reality. Very often, the information consumers are not aware of how biased is what they are exposed to. To mitigate this situation, many computational linguistics efforts have been made to detect social bias such as gender and racial biases (Bolukbasi et al., 2016; Garg et al., 2018; Liang et al., 2020; Dev et al., 2020). The immigrant stereotype is another type of social bias that is present when a message about immigrants disregards the great diversity of this group of people and highlights a small set of their characteristics. This process of homogenization of a whole group of people is at the very heart of the stereotype concept (Tajfel, Sheikh, and Gardner, 1964). As (Lipmann, 1922) said in his seminal work

ISSN 1135-5948. DOI 10.26342/2021-67-7

about stereotypes, stereotyping, as a cognitive process, occurs because "we do not first see and then define, we define first and then see". In short, we can say that a stereotype is being used in language when a whole group of people, itself very diverse, is represented by appealing to a few characteristics.

Unfortunately, the use of stereotypes promotes undesirable behaviors among people from different nationalities; an example is the violence against Asian Americans that have taken place recently (Tessler, Choi, and Kao, 2020). Moreover, political analysts have associated the success of anti-immigration parties with the even more negative attitudes to the immigration phenomenon (Dennison and Geddes, 2019). These stereotypes have received little attention to be automatically identified, despite the harmful consequences that prejudices and attitudes, in many cases negative, may have.

There have been some works related to the problem of immigrant stereotypes identification (Sanguinetti et al., 2018), but they are mainly focused on the expressions of hate speech; or social bias in general that involves racism (Nadeem, Bethke, and Reddy, 2020; Fokkens et al., 2018). However, it is necessary to have a whole view of immigrant stereotypes, taking into account both positive and negative beliefs, and also the variability in which stereotypes are reflected in texts. This more refined analysis of stereotypes would make it possible to detect them not only in clearly dogmatic or violent messages, but also in other more formal and subtle texts such as news, institutional statements, or political representatives' speeches in parliamentary debates.

Similar to applications of healthcare, security, and social analysis, in this task is not enough to achieve high results, but it is also mandatory that results could be understood or interpreted by human experts on the domain of study (e.g., social psychologists). Taking into account these two aspects, performance and explainability, the objective of this work is to compare two approaches diametrically opposite to each other in the text classification state of the art. On the one hand, a transformer-based model, which has shown outstanding performance, but high complexity and poor explainability; and, on the other hand, a masking-based model, which requires fewer computational-resources and showed a good performance in related tasks like profiling.

We aim to find explainable predictions of immigrant stereotypes with BETO by using its attention mechanism. In this way, we derive the explanation by investigating the importance scores of different features used to output the final prediction. With the other approach that we use, the masking technique (Stamatatos, 2017; Sánchez-Junquera et al., 2020), it is possible to know what are the most important words that the model preferred to highlight. We compare these approaches using a dataset of texts in Spanish, which contains annotated fragments of political speeches from Spanish Congress of Deputies.

The research questions aim to answer in this work are:

RQ1: Is the transformer more effective than the masking technique at identifying stereotypes about immigrants?

RQ2: Is it possible to obtain local explanations on the predictions of the models, to allow human interpretability about the immigrant stereotypes?

The rest of the paper is organized as follow: Section 2 presents related work concerning the immigration stereotype detection and the models that we propose. Section 3 describes the two models, and Section 4 the dataset used in the experiments. Sections 5 and 6 contain the experimental settings and the discussion about the results. Finally, we conclude the work in Section 7 where we mention also future directions.

2 Related Work

2.1 Immigrant Stereotype Detection

There have been attempts to study stereotypes from a computational point of view, such as gender, racial, religion, and ethnic bias detection do (Garg et al., 2018; Bolukbasi et al., 2016; Liang et al., 2020). Those works predefine two opposite categories (e.g., men vs. women) and use word embeddings to detect the words that tend to be more associated with one of the categories than with the other. In (Nadeem, Bethke, and Reddy, 2020), the authors propose two different level tests for measuring bias. First,

the intra-sentence test, with a sentence describing the target group and a set of three attributes that correspond to a stereotype, an anti-stereotype, and a neutral option. Second, the inter-sentence test, with a sentence containing the target group; a sentence containing a stereotypical attribute of the target group; another sentence with an anti-stereotypical attribute; and lastly, a neutral sentence. These tests are similar to the idea of (Dev et al., 2020) that consist in using natural language inference to measure entailment, contradiction, or neutral inferences to quantify the bias. To evaluate their proposal, in (Nadeem, Bethke, and Reddy, 2020), the authors collected a dataset (StereoSet) for measuring bias related to gender, profession, race, and religion domains.

On the other hand, stereotypes are not always the (explicit) association of *words* (seen as attributes or characteristics) from two opposite social groups, like women vs. men in the context of gender bias. Such is the case of immigrant stereotypes, in which sentences like *¿Por qué ha muerto una persona joven?* (Why did a young person die?) do not contain an attribute of the immigrant group although from its context[1] it is possible to conclude that here immigrants are placed as victims of suffering. Also, it is not clear the representative word of the social group, since *persona joven* (young person) is neutral to immigrants and non-immigrants.

Other works have built annotated data to foster the development of supervised approaches. In (Sanguinetti et al., 2018), was presented an Italian corpus focused on hate speech against immigrants, which includes annotations about whether a tweet is a stereotype or not. This corpus was used in the HaSpeeDe shared task at EVALITA 2020 (Sanguinetti et al., 2020). Most participant teams only adapted their hate speech models to the stereotype identification task, thus, representing (and reducing) stereotypes to characteristics of hate speech. One of the conclusions was that the immigration stereotype appeared as a more subtle phenomenon, which also needs to be approached as non-hurtful text. Additionally, in (Nadeem, Bethke, and Reddy, 2020), it was proposed a dataset that includes the domain of racism (additionally to gender, religion, and profession). Although this dataset does not focus on the study of stereotypes about immigrants, its authors reported the word "immigrate" as one of the most relevant keywords that characterized the racism domain.

[1]Fragment of a political speech from a Popular Parliamentary Group politician in 2006. The speaker is mentioning some of the conditions of immigrants in Spain in that period.

2.2 On the explainability of AI models

Since eXplainable Artificial Intelligence (XAI) systems have become an integral part of many real-world applications, there is an increasing number of XAI approaches (Islam et al., 2021) including white and black boxes. The first group, which includes decision trees, hidden Markov models, logistic regressions, and other machine learning algorithms, are inherently explainable; whereas, the second group, which includes deep learning models, are less explainable (Danilevsky et al., 2020). XAI has been characterized according to different aspects, for example, (i) by the the level of the explainability, for each single prediction (*local explanation*) or the model's prediction process as a whole (*global explanation*); (ii) and if the explanation requires post-processing (*post-hoc*) or not (*self-explaining*).

XAI has also been characterized in accordance to the source of the explanations, for example: (i) *surrogate models*, in which the model predictions are explained by learning a second model as a proxy, such is the case of LIME (Ribeiro, Singh, and Guestrin, 2016); (ii) *example-driven*, in which the prediction of an input instance is explained by identifying other (labeled) instances that are semantically similar (Croce, Rossini, and Basili, 2019); (iii) *attention layers*, which appeal to human intuition and help to indicate where the neural network model is "focusing"; and (iv) *feature importance*, in which the relevance scores of different features are used to output the final prediction (Danilevsky et al., 2020).

Taking into account this characterization, we frame our approach in the *self-explaining* scope, and consider two different models to obtain *local explanations* of the predicted texts. In this sense, we use the *attention layers* which have been commonly applied by local self-explaining models (Mullenbach et al., 2018; Bodria et al., 2020). For example, in

(Mathew et al., 2020) the attention weights were used to compare the posts' segments on which the labeling decision was based, highlighting the tokens that the models found the most relevant. Similarly, in (Clark et al., 2019) the authors used datasets for tasks like dependency parsing, to evaluate attention heads of BERT, and found relevant linguistic knowledge in the hidden states and attention maps, such as direct objects of verbs, determiners of nouns, and objects of prepositions. Finally, in (Jarquín-Vásquez, Montes-y-Gómez, and Villaseñor-Pineda, 2020) attention was used to prove that some swear words are inherently offensive, whilst others are not, since their interpretation depends on their context.

The other self-explaining model that we use to obtain the local explanations, is a masking technique which can be described as a white box. In this case, the explainable strategy is based on the *feature importance* idea, by measuring and observing the relevant words used in its masking process. The masking technique used in this work incorporates an additional way to explain decisions (Stamatatos, 2017; Granados et al., 2011; Sánchez-Junquera et al., 2020). It allows highlighting content and style information from texts, by masking a predefined and task-oriented set of irrelevant words.

3 Models

In this section, we briefly describe the two models that we use in our experiments.

BETO: it is based on BERT, but it was pre-trained exclusively on a big Spanish dataset (Cañete et al., 2020). The framework of BETO consists of two steps: pre-training and fine-tuning, similar to BERT (Devlin et al., 2018). For the pre-training, the collected data included Wikipedia and other Spanish sources such as United Nations and Government journals, TED Talks, Subtitles, News Stories among others. The model has 12 self-attention layers with 16 attention-heads each, and uses 1024 as hidden size, with a total of 110M parameters. The vocabulary contains 32K tokens.

For fine-tuning, the model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the downstream task, which in our case is a stereotype-annotated dataset (see Section 4). The first token of every sequence is always a special classification token ([CLS]), which is used as the aggregate sequence representation for classification tasks. In our work, we add to the [CLS] representation two dense layers and a Softmax function to obtain the binary classification.

The masking technique: it consists of transforming the original texts to a distorted form where the textual structure is maintained while irrelevant words are masked, i.e., replaced by a neutral symbol. The irrelevant terms are task-dependent and have to be defined in advance, following some frequency criteria or the expert's intuition.

The masking technique replaces each term t of the original text by a sequence of "*". The length of the sequence is determined by the number of characters that t contains. One example of this is shown in Figure 1 considering the Spanish stopwords. In Section 5.1, we explain which are the relevant words that we considered better to mask.

After all texts are distorted by the masking technique, we use a traditional classifier to be compared with BETO. In our experiments, we use Logistic Regression (LR) classifier which has been used before to be compared with BERT (Alaparthi and Mishra, 2021).

4 Dataset

We use the StereoImmigrants dataset[2] for identifying stereotypes about immigrants (Sánchez-Junquera et al., 2021). In this previous work we collected texts on immigrant stereotypes from the political speeches of the ParlSpeech V2 dataset (Rauh and Schwalbach, 2020); from where we also extracted the negative examples (labeled as Non-stereotypes). These texts are extracted from the speeches of the Spanish Congress of Deputies (*Congreso de los Diputados*), and are written in Spanish.

In the construction of StereoImmigrants (Sánchez-Junquera et al., 2021), we proposed a new approach to the study of immigrant stereotyping elaborating a taxonomy to annotate the corpus that covers the whole spectrum of beliefs that make up the immigrant stereotype. The novelty of this taxonomy and this annotation process is that the work has not focused on the characteristics attributed to the group but on the narrative contexts in

[2]https://github.com/jjsjunquera/StereoImmigrants.

Original text
la inmigración sigue siendo hoy - lo confirman los últimos sondeos del CIS - el principal problema que preocupa a los ciudadanos del estado
(Immigration is still today - confirmed by the latest CIS polls - the main problem that worries the citizens of the state)
Masking stopwords
*** inmigración sigue siendo hoy ** confirman *** últimos sondeos *** cis ** principal problema *** preocupa * *** ciudadanos ** *** ******
Masking the non-stopwords
*la ********** ***** ****** *** lo ********* los ******* ******* del *** el ********* ******** que ******** los ********** del estado*

Figure 1: Example of masking the stopwords, or keeping only the stopwords unmasked.

which the immigrant group is repetitively situated in the public discourses of politicians. To do this, the authors applied the frame theory –a social psychology theory– to the study of stereotypes. The frame theory allows us to show that politicians in their speeches create and recreate different *frames* (Scheufele, 2006), i.e. different scenarios, where they place the group. The result of this rhetorical activity of framing ends with the creation of a stereotype: a diverse group is seen only with the characteristics of the main actor in a particular scenario.

In (Sánchez-Junquera et al., 2021), we identify different frames used to speak about immigrants that could be classified in one of the following categories: (i) present the immigrants as equals to the majority but the target of xenophobia (i.e., they must have the same rights and same duties but are discriminated), (ii) as victims (e.g., they are people suffering from poverty or labor exploitation), (iii) as an economic resource (i.e., they are workers that contribute to economic development), (iv) as a threat for the group (i.e., they are the cause of disorder because they are illegal, too many, and introduce unbalances in societies), or (v) as a threat for the individual (i.e., they are competitors for limited resources or a danger to personal welfare and safety). In the construction of the StereoImmigrants dataset, an expert in prejudice from the social psychology area annotated manually the sentences at the finest granularity of the taxonomy and selected also negatives examples where politicians speak about immigration but do not refer, explicitly or implicitly, to the people that integrates the group "immigrants". After this expert annotation, five non-experts annotators read the label assigned by the expert to each sentence and decided if they agreed with it or considered that another label from the taxonomy was better suited for this sentence. The dataset only contains sentences where at least three annotators agreed on the same category.

Label	**Length**	**Texts**	
Stereotype	45.62 ± 24.69	1673	3635
Non-stereotype	36.00 ± 21.17	1962	
Victims	48.93 ± 27.5	743	1479
Threat	45.84 ± 24.42	736	

Table 1: Distribution of texts per label and the average length (with standard deviation) of their instances. The texts labeled as *Victims* or *Threat* are a subset of the texts labeled as *Stereotype*.

In (Sánchez-Junquera et al., 2021), attending to a second annotation of the attitudes that each sentence expresses, we proposed two supra-categories of the stereotypes annotated as *Victims* or *Threat*, where the categories (i) and (ii) belong to the *Victims* supra-category, and (iv) and (v) belong to the *Threat* supra-category. Table 1 shows the distribution per label of the dataset.

Table 2 shows examples of Non-stereotypes and Stereotypes labels. The Stereotypes examples specify if they were labeled as *Victims* or *Threat*. From these examples, it is possible to see that the dataset contains stereotypes that are not merely the association of attributes or characteristics to the group, but texts which reflect biased representations of the group (i.e., how the immigrants are indirectly perceived or associated with specific situations and social issues).

5 Experimental Settings

We applied a 10-fold cross-validation procedure and reported our results in terms of *F-measure*.

For BETO, we searched the following hy-

Non-stereotype
No nos vale que se contabilice todo lo que se dedica a inmigración porque no estamos hablando de lo mismo. (We are not worth accounting for everything that is dedicated to immigration because we are not talking about the same thing.)
El Gobierno está desbordado por la inmigración, por su política improvisada, irresponsable, descoordinada y unilateral. (The Government is overwhelmed by immigration, by its improvised, irresponsible, uncoordinated and unilateral policy.)
Stereotype: Victims
¿Por qué ha muerto una persona joven? (Why did a young person die?)
Hay una situación de desamparo en muchas personas a la que necesitamos dar una solución. (There is a helplessness situation in many people to which we need to provide a solution.)
Stereotype: Threat
España hoy está desbordada con la inmigración ilegal. (Spain today is overwhelmed with illegal immigration.)
Esta alarmante situación, agravada por la incapacidad del Gobierno socialista, ha producido el colapso, el desbordamiento de los servicios humanitarios, judiciales y policiales que han generado una gran alarma social. (This alarming situation, aggravated by the incapacity of the socialist government, has produced the collapse, the overflow of humanitarian, judicial and police services that have generated great social alarm.)

Table 2: Examples from each label of the dataset.

perparameter grid to obtain the results of BETO: *learning rate* $\in \{0.01, \mathbf{3e\text{-}5}\}$; the *batch size* $\in \{16, \mathbf{34}\}$; and the optimizer$\in$ $\{\mathbf{adam}, rmsprop\}$ (in bold we highlighted the optimal hyperparameter values). Moreover, we applied a dropout value of 0.3 to the last dense layer. We have selected a value of 180 for the *max_length* hyperparameter according to the maximum length of all the texts in the dataset. The model was fine-tuned for 10 epochs on the training data for each task.

For the masking-based approach, we used the *sklearn* implementation of the LR classifier. All the parameters were taken by default, except for the optimization method: we selected *newton-cg*. The model used the bag of words representation, using the *tfidf* term weighting. We tested with unigrams, bigrams, and trigrams of words; and with characters n-grams (n$\in \{3, 4, 5, 6\}$) obtaining the better results with **character 4-grams**. When we used LR with the original texts, unigrams of words achieved better results than character n-grams.

5.1 Unmasking Stereotypes

Related works on social bias detection have found a list of words that tend to be associated with one of two opposite social groups (e.g., female vs. male, Asian vs Hispanic people) (Bolukbasi et al., 2016; Garg et al., 2018). In the immigrant stereotypes case, it is particularly difficult to define two opposite groups and consequently to find such biased words. In this paper, we use the dataset described in Section 4 to find which could be the most relevant terms to be used in the masking process.

Intuitively, in the immigrant stereotypes' context, the relevant words could be content-related, although style-related terms like function words could play also an interesting role. After preliminary experiments, we found higher results by masking the words out of the following lists: (i) the words with higher relative frequency ($RelFreq$), i.e., the k words with a frequency in one class remarkably higher than its frequency in the opposite class; and (ii) the k words with the highest absolute frequency ($AbsFreq$) in all the collection, excluding stopwords (i.e., stopwords were masked). In our experiments we achieved better results with $k = 1000$.

Each list was computed using the corresponding set of texts depending on the classification task: Stereotype vs. Non-stereotype, or Victims vs. Threat. The information that is kept unmasked corresponds to the content-related words.

6 Results and Discussion

We report the results of the models in Table 3. It is possible to see high results of LR with the original texts. However, we observe that masking the terms out of the list $RelFreq$ is slightly better than using the original text. These results suggest that the masking technique improves the quality of the stereotype detection and its dimensions.

In comparison with $AbsFreq$, maintain-

	S/N	V/T
Original text	0.82	0.79
Masking Technique with *AbsFreq*	0.79	0.75
Masking Technique with *RelFreq*	0.84	0.81
BETO	**0.86**	**0.83**

Table 3: F-measure in both classification tasks: Stereotype vs. Non-stereotype (S/N), and Victims vs. Threat (V/T).

ing unmasked the *RelFreq* words helps to ignore more words that are less discriminative for classification tasks. This could be explained because *AbsFreq* includes words similarly frequent in both classes, which could not help at predicting immigrant stereotypes: *países* (countries), *gobierno* (government), *señor* (mister), *partido* (party); or at identifying the immigrant-stereotype dimension: *fronteras* (frontiers), *política* (politic), *seguridad* (security), *grupo* (group). Table 4 shows examples of words included in *RelFreq* that are indeed reflecting some bias according to the category. For example, it is not surprising to find words like *derechos* (rights), *humanos* (human[3] or humans), *pobreza* (poverty), *muerto* (dead), and *hambre* (hunger) more associated to immigrants seen as victims; and words like *irregular* (irregular), *ilegal* (illegal), *regularización* (regularization), *masiva*(massive), and *problema* (problem), more used in speeches where immigrants are seen as collective or personal threat.

BETO achieves the highest results in both classification tasks (**RQ1**). This is not surprising because the transformer-based models are known for their properties at capturing semantic and syntactic information, and richer patterns in which the context of the words are taken into account. However, we do not observe a significant difference between the results of such a resource-hungry model, and the combination of the masking technique with the traditional LR classifier. Considering the computational capabilities that BETO demands, and the less complexity of the masking technique, the latter shows a better trade-off between effectiveness and efficiency than the latter.

6.1 Discriminating Words

Motivated by the similar results of BETO and the masking technique, we wanted to observe and compare what portions of the texts they could be focusing on. For this purpose, we looked at the last layer of BETO and computed the average of the attention heads. Therefore, for each text, we had an attention matrix from which we could compute the attention that the transformer gave to each word in that texts. Figure 2 shows examples of texts where the two models agreed on the right label.

[3]It refers to the adjective: *human* rights.

Stereotype	Non-stereotype	Victims	Threat
personas	política	derechos	inmigración
canarias	europea	personas	canarias
derechos	unión	humanos	gobierno
problema	materia	derecho	irregular
país	grupo	mujeres	ilegales
irregular	políticas	países	irregulares
situación	consejo	pobreza	regularización
regularización	cooperación	integración	ilegal
ilegales	gobierno	mundo	españa
humanos	moción	vida	problema
ciudadanos	europeo	solidaridad	proceso
irregulares	ley	asilo	masiva
efecto	parlamentario	condiciones	llegado
origen	comisión	millones	aeropuertos
centros	cámara	muerto	ministro
ilegal	desarrollo	refugiados	control
drama	consenso	social	efecto
acogida	subcomisión	miseria	pateras
llamada	socialista	internacional	llamada
menores	común	ciudadanos	medidas
vida	temas	xenofobia	llegada
mafias	tema	hambre	inmigrantes
llegada	asuntos	viven	marruecos
masiva	grupos	emigrantes	cayucos
extranjeros	emigrantes	muerte	presión

Table 4: Examples of the relevant words that were not masked, considering the list *RelFreq* in each classification task.

From the figure, it is possible to see what words were relevant for both approaches. Although some of the relevant words are function words (e.g., *para*, *muy*) and are not too informative at first glance for human interpretation, we can observe that some content-related words can be helpful for expert's analysis. For instance, the text labeled as Stereotype has as relevant words *fenómeno* (phenomenon), *inmigración* (immigration), *problema* (problem), *terrorismo* (terrorism), *paro* (unemployment), among others. The text labeled as Victims contains *desamparo* (abandonment), *personas* (people), *necesitamos dar una solución* (we need to give a solution), reflecting how immigrants were seen as people more than their illegal status (e.g., see Tables 4), and the target of problems that need solutions. Moreover, in the example of Threat, some of the words and phrases receiving more importance (such as, *problema muy serio*, *problema muy importante*) reflect how immigrants were seen as a problem to

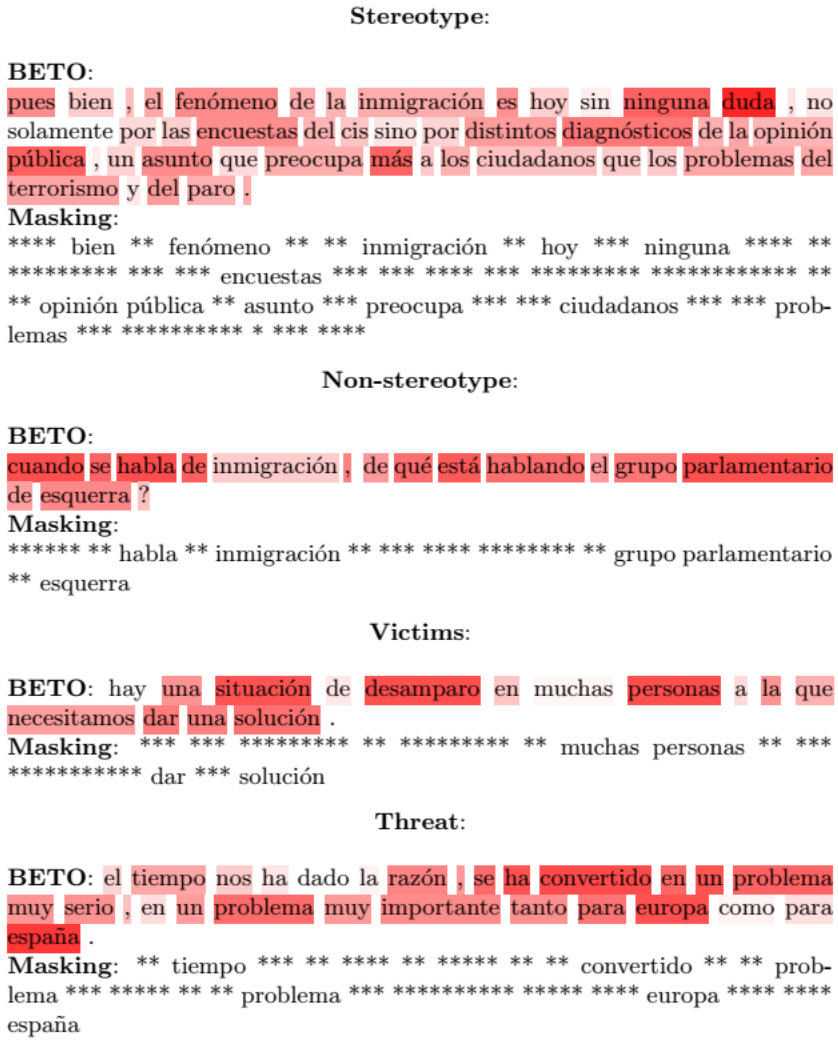

Figure 2: Examples of attention visualization and masking transformation over the same texts. These examples were correctly classified by both models. The more intense the color, the greater is the weight of attention given by the model.

the continent and the country, but not the country where immigrants come from.

Table 5 presents some of the words with the highest attention scores in only the true positive predictions of each class. Therefore, these words could be among the most discriminative for stereotype identification.

We contrasted the list of words with more attention on the BETO true positive predictions, with the *RelFreq* words used by the masking technique as more discriminative for each class. Table 6 shows the percentage of *RelFreq* words (which were not masked) that were present in the top of the ranking as *more discriminative* from BETO. In the top 30 of the ranking, we found the vast majority of the not masked words. This suggests that the two approaches have seen similar cues.

For now, we have seen that BETO and the masking technique achieved similar results and have an intersection in the discriminative words they focused on in the texts (which in fact answer **RQ2**), despite one of them is a resource-hungry model and the other requires less computational resources. We do not think that BETO should not be used because of its complexity: one of the differences we should highlight is that for the masking technique the list of words should be predefined with some limitations and algorithm bias that this could imply. However, BETO learns by itself to score the words gradually, instead of giving a binary score like in the masking technique (to mask or keep unmasked). Therefore, we can apply from the transformer a comparison of the importance of the different words (like it was visualized in Figure 2), which was automatically learned

Stereotype	Non-stereotype	Victims	Threat
drama	*consejo*	*derecho*	*masiva*
llegada	*asuntos*	*esclavitud*	*pateras*
ilegales	*temas*	*mujeres*	*llamada*
efecto	*comparecen*	*asilo*	*avalancha*
irregulares	*producen*	*refugiados*	*aeropuertos*
llamada	*recibir*	*pobreza*	*trasladar*
costas	*acuerdos*	*xenofobia*	*llegada*
expulsiones	*esfuerzo*	*muerto*	*zapatero*
trabajadores	*diálogo*	*sistema*	*caldera*
pateras	*pacto*	*devoluciones*	*alarma*
xenofobia	*congreso*	*miseria*	*ayudas*
mafia	*necesidad*	*desgracia*	*ilegalmente*
condiciones	*cumbre*	*difícil*	*afrontar*
legalidad	*proyecto*	*grupos*	*judiciales*
dinero	*zapatero*	*racismo*	*capacidad*
avalancha	*enmiendas*	*hambre*	*archipiélago*
vienen	*miembros*	*refugio*	*delincuencia*
peninsula	*colabora*	*persona*	*grave*
muertes	*conferencia*	*situaciones*	*congreso*
miles	*gobiernos*	*explotación*	*tropicales*
humanitaria	*exterior*	*denuncia*	*fallecido*
coladero	*importantes*	*muerte*	*coladero*
preocupa	*acción*	*democracia*	*oleada*

Table 5: Words whose attention scores are the highest only on the true positive predictions of BETO in each class.

Attention Ranking	**% from a total of unmasked words**	
	S/N	V/T
top 10	48.96%	35.89%
top 20	78.92%	65.11%
top 30	95.84%	82.98%

Table 6: The percentage of *RelFreq* words that are in the top of words with the highest attention.

from the context of the words in the texts. In the next sections, we confirm the advantages of both models by analyzing the results of an ideal ensemble and other utilities of the attention mechanisms.

6.2 An Ideal Ensemble

We have seen the results achieved by the proposed models and the intersection set of words they focus on in the texts. Therefore, one could think that these models are classifying correctly the same texts. In this section, we report that the models are misclassifying different instances in general.

Table 7 shows the misclassified instances of LR with the masked texts and BETO in each classification task. The models have good performances, so it is licit to think in an ideal ensemble that could wisely combine their predictions. The resultant ensemble will miss only the texts where both models are wrong: 272 texts at distinguishing Stereotype vs. Non-stereotype, which means that

	S/N	V/T
Total of instances	3630	1477
misclassified by LR	624	263
misclassified by BETO	518	259
misclassified by both	272	115
Well predicted by an ideal ensemble	92.5%	92.2%

Table 7: Misclassified instances and the performance of an ideal ensemble for Stereotype vs. Non-stereotype and Victims vs. Threat tasks.

92.5% of the 3630 texts will be correctly classified. A similar analysis can be done in the Victims vs. Threat classification task, which will result in 92.2% of the 1477 texts that will be potentially correctly classified.

6.3 Relations with the Highest Attention Scores

Another advantage of the attention mechanism is the relations between the non-discriminative words and other words from each class. We could find noisy features similarly present in the opposite classes. One of the words with the highest attention scores in our dataset is *inmigración* (immigration); since we found its scores high in the two opposite classes, we did not count it as *discriminative* by BETO. However, we think that as the heads have the attention that each word gives to the others in the texts, we can observe how the "noisy" words are used in the opposite classes, by looking at the relations with their context. We hypothesize that the immigration-related words are used in different contexts in the opposite classes.

Table 8 shows an example of the words whose relation with *inmigración* are the most scored in each class. We omitted the ones in the Non-stereotype class due to they are not informative. Interestingly, the words associated with *inmigración* are also describing differently the Stereotype, Victims, and Threat classes. For example, with this strategy we observe that words like *criminal*, and *enfermedades* (diseases) are now in the top of discriminating words of the *Threat* category (in contrast to Table 5). We conclude that the attention mechanism should be exploited in the future in this sense. Probably the attention scores could be a source of interesting cues not only in terms of biased words from *RelFreq* list or the ones shown in Table 5,

Stereotype	Victim	Threat
muertes	discriminación	llega
saturado	colectivos	nuevo
miseria	mujeres	delincuencia
pobres	consenso	aeropuertos
policiales	dentro	procedente
internamiento	refugiados	zapatero
dramaticos	familias	saturado
descontrol	educativo	policiales
empresarios	planteamos	retención
humanitario	miseria	madrid
costas	refugiados	enfermedades
avalancha	pobreza	congreso
garantías	reto	evolución
delincuencia	mujeres	entran
tráfico	voto	francia
devueltos	voluntad	aeropuertos
explotación	especificas	tropicales
llamada	pobreza	criminal
alarman	iniciativa	aeropuerto
ilegales	enmienda	intentos
pateras	podían	llegaron
expulsión	saben	coladero

Table 8: Words with the highest attention scores in relation to *inmigración* (immigration).

but also concerning the forms in which neutral terms are contextualized.

7 Conclusion and Future Work

This work is a contribution to the immigrant stereotype identification problem. The particularities of the immigration phenomenon make this bias detection task differs from other kinds of bias that have received much more attention (e.g., gender bias). We addressed two classification tasks, the Stereotype vs. Non-stereotype detection, and Victims vs. Threat dimensions identification using an annotated dataset in Spanish. We proposed two different models: BETO, a resource-hungry model which demands strong computational capabilities; and a masking technique, a less complex approach that transforms the texts to be used by a traditional classifier. We demonstrate that both approaches are suitable for immigrant stereotype identification; and interestingly, the masking technique achieves almost the same results of BETO, despite its simplicity (RQ1).

We developed a comparison between the attention mechanism of BETO, and the list of relevant terms that the masking technique uses. These two different approaches focused on similar portions of the texts. Specifically, the majority of the relevant words maintained unmasked are at the top of the words that BETO gave the highest attention. Furthermore, with these models it is possible to highlight some stereotype cues that could be considered as *local explanations* for further studies about immigrant stereotypes (RQ2).

On the basis of the reported results, we conclude that both models are effective at identifying the immigrant stereotypes, and could be combined to build an ideal ensemble that overcomes the results of each one. We also point out that BETO can help to investigate with more detail the bias towards immigrants with the attention mechanisms. For these reasons, we think we cannot rule out the use of either model.

To our knowledge, this is the first work on immigrant stereotypes identification that compares deep learning with traditional machine learning approaches paying special attention to the explicability of the models in this task. However, more work is necessary to explore more deeply the advantages of the attention mechanisms in this sense. In future work, we plan to combine the two approaches to increase the performance; and to use discriminative words to find debiasing strategies to mitigate the immigrant stereotypes in social media and political speeches.

Agradecimientos

The work of the authors from the Universitat Politècnica of València was funded by the Spanish Ministry of Science and Innovation under the research project MISMIS-FAKEnHATE on MISinformation and MIScommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31). Experiments were carried out on the GPU cluster at PRHLT thanks to the PROMETEO/2019/121 (DeepPattern) research project funded by the Generalitat Valenciana.

References

Alaparthi, S. and M. Mishra. 2021. Bert: a sentiment analysis odyssey. *Journal of Marketing Analytics*, pages 1–9.

Bodria, F., A. Panisson, A. Perotti, and S. Piaggesi. 2020. Explainability methods for natural language processing: Applications to sentiment analysis. In *SEBD*.

Bolukbasi, T., K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Advances in neural information processing systems, 29:4349–4357.

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

Clark, K., U. Khandelwal, O. Levy, and C. D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy, August. Association for Computational Linguistics.

Croce, D., D. Rossini, and R. Basili. 2019. Auditing deep learning processes through kernel-based explanatory models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4037–4046, Hong Kong, China, November. Association for Computational Linguistics.

Danilevsky, M., K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen. 2020. A survey of the state of explainable ai for natural language processing. *arXiv preprint arXiv:2010.00711*.

Dennison, J. and A. Geddes. 2019. A rising tide? the salience of immigration and the rise of anti-immigration political parties in western europe. *The political quarterly*, 90(1):107–116.

Dev, S., T. Li, J. M. Phillips, and V. Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *AAAI*, pages 7659–7666.

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

Fokkens, A., N. Ruigrok, C. Beukeboom, G. Sarah, and W. Van Atteveldt. 2018. Studying muslim stereotyping through microportrait extraction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Garg, N., L. Schiebinger, D. Jurafsky, and J. Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Granados, A., M. Cebrián, D. Camacho, and F. De Borja Rodríguez. 2011. Reducing the loss of information through annealing text distortion. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):1090–1102. cited By 19.

Islam, S. R., W. Eberle, S. K. Ghafoor, and M. Ahmed. 2021. Explainable artificial intelligence approaches: A survey. *arXiv preprint arXiv:2101.09429*.

Jarquín-Vásquez, H. J., M. Montes-y-Gómez, and L. Villaseñor-Pineda. 2020. Not all swear words are used equal: Attention over word n-grams for abusive language identification. In K. M. Figueroa Mora, J. Anzurez Marín, J. Cerda, J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, and J. A. Olvera-López, editors, *Pattern Recognition*, pages 282–292, Cham. Springer International Publishing.

Liang, P. P., I. M. Li, E. Zheng, Y. C. Lim, R. Salakhutdinov, and L.-P. Morency. 2020. Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100*.

Lipmann, W. 1922. *Public Opinion*. New York:Harcourt Brace.

Mathew, B., P. Saha, S. Muhie Yimam, C. Biemann, P. Goyal, and A. Mukherjee. 2020. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *arXiv e-prints*, page arXiv:2012.10289, December.

Mullenbach, J., S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana, June. Association for Computational Linguistics.

Nadeem, M., A. Bethke, and S. Reddy. 2020. Stereoset: Measuring stereotypical bias

in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Rauh, C. and J. Schwalbach. 2020. The parlspeech v2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies. *Harvard Dataverse*.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Sanguinetti, M., G. Comandini, E. Di Nuovo, S. Frenda, M. A. Stranisci, C. Bosco, C. Tommaso, V. Patti, R. Irene, et al. 2020. Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task. In *EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, pages 1–9. CEUR.

Sanguinetti, M., F. Poletto, C. Bosco, V. Patti, and M. Stranisci. 2018. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Scheufele, D. A. 2006. Framing as a Theory of Media Effects. *Journal of Communication*, 49(1):103–122, 02.

Stamatatos, E. 2017. Authorship attribution using text distortion. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, 1:1138–1149. cited By 31.

Sánchez-Junquera, J., B. Chulvi, P. Rosso, and S. P. Ponzetto. 2021. How do you speak about immigrants? taxonomy and stereoimmigrants dataset for identifying stereotypes about immigrants. *Applied Sciences*, 11(8).

Sánchez-Junquera, J., L. Villaseñor-Pineda, M. Montes-y-Gómez, P. Rosso, and E. Stamatatos. 2020. Masking domain-specific information for cross-domain deception detection. *Pattern Recognition Letters*, 135:122–130.

Tajfel, H., A. A. Sheikh, and R. C. Gardner. 1964. Content of stereotypes and the inference of similarity between members of stereotyped groups. *Acta Psychologica,*, 22(3):191–201.

Tessler, H., M. Choi, and G. Kao. 2020. The anxiety of being asian american: Hate crimes and negative biases during the covid-19 pandemic. *American Journal of Criminal Justice*, 45(4):636–646.