applied sciences

MDPI

Article

How Do You Speak about Immigrants? Taxonomy and StereoImmigrants Dataset for Identifying Stereotypes about Immigrants

Javier Sánchez-Junquera [1,*], Berta Chulvi [1], Paolo Rosso [1] and Simone Paolo Ponzetto [2]

[1] PRHLT Research Center, Universitat Politècnica de València, 46022 València, Spain; machufe1@upvnet.upv.es (B.C.); prosso@dsic.upv.es (P.R.)
[2] Data and Web Science Group, University of Mannheim, 68131 Mannheim, Germany; simone@informatik.uni-mannheim.de
* Correspondence: juasanj3@doctor.upv.es

Abstract: Stereotype is a type of social bias massively present in texts that computational models use. There are stereotypes that present special difficulties because they do not rely on personal attributes. This is the case of stereotypes about immigrants, a social category that is a preferred target of hate speech and discrimination. We propose a new approach to detect stereotypes about immigrants in texts focusing not on the personal attributes assigned to the minority but in the frames, that is, the narrative scenarios, in which the group is placed in public speeches. We have proposed a fine-grained social psychology grounded taxonomy with six categories to capture the different dimensions of the stereotype (positive vs. negative) and annotated a novel *StereoImmigrants* dataset with sentences that Spanish politicians have stated in the Congress of Deputies. We aggregate these categories in two supracategories: one is *Victims* that expresses the positive stereotypes about immigrants and the other is *Threat* that expresses the negative stereotype. We carried out two preliminary experiments: first, to evaluate the automatic detection of stereotypes; and second, to distinguish between the two supracategories of immigrants' stereotypes. In these experiments, we employed state-of-the-art transformer models (monolingual and multilingual) and four classical machine learning classifiers. We achieve above 0.83 of accuracy with the BETO model in both experiments, showing that transformers can capture stereotypes about immigrants with a high level of accuracy.

Keywords: social bias; stereotypes about immigrants; social psychology based taxonomy; stereoimmigrants dataset; transformer models; Spanish

Citation: Sánchez-Junquera, J.; Chulvi, B.; Rosso, P.; Ponzetto, S.P. How Do You Speak about Immigrants? Taxonomy and StereoImmigrants Dataset for Identifying Stereotypes about Immigrants. *Appl. Sci.* **2021**, *11*, 3610. https://doi.org/10.3390/app11083610

Academic Editors: José Ignacio Abreu Salas and Yoan Gutiérrez Vázquez

Received: 26 February 2021
Accepted: 8 April 2021
Published: 16 April 2021

1. Introduction

Social bias in information is receiving more and more attention in computational science. The information on the web has a strong impact on how people perceive reality and consequently on the decision they can make, the attitude they develop, and the prejudice they hold [1]. Some general examples where we can find bias include political news [2,3], rumours [4], products reviews [5], among others. However, there is a kind of social bias which is massively present in everyday language, and of course on the web, which is the use of stereotypes. A recent work that measures stereotypical bias in pretrained language models has found that as the language model becomes stronger, so its stereotypical bias does too [6]. As the authors said, "this is unfortunate and perhaps unavoidable as long as we rely on real word distribution of corpora to train language models". The difficulty is clear but the need also: these stereotypes have a strong effect on the members of the stigmatised group, for instance, impacting the performance of individuals who face stereotype threats [7–9]. We have known from the beginning of social psychology that stereotypes are at the base of prejudice towards minorities and to spread prejudices is an efficient strategy for dogmatic groups and authoritarian ideologies [10,11].

A long tradition of research in social psychology defines stereotype as a set of widespread beliefs that are associated with a group category [12,13]. This set of beliefs facilitates the operation of prejudices and justifies them [1]. Research in prejudice has shown that this set of stereotyped beliefs may be both positive and negative [14]. The importance of positive beliefs in stereotyping social groups has been highlighted, especially in studies on gender stereotypes [15] but is less studied in relation to other stereotypes such as that of the social category of immigrants.

To understand how this social bias occurs in texts, we need to go beyond this common idea that a stereotype is a set of beliefs. A stereotype is a type of social bias that occurs when a message about a group disregards the great diversity existing within the members of this group and highlights a small set of features [12]. This process of homogenisation of a whole group of people is at the very heart of the stereotype concept [16]. We know from social science research that the main part of this definition process takes place in speeches from socially relevant actors [17]. Politicians, social movements, and mass media messages create and recreate a *frame* [18], a kind of scenario, where they speak about a group. Framing analysis [19] proposes that how citizens understand an issue—which features of it are central and which peripheral—is reflected in how the issue is framed.

Frame as a concept has a long tradition in psychology [20,21] and in sociology [22]. Gamson defines *frame* as *"a central organizing idea or story line that provides meaning to an unfolding strip of events, weaving a connection among them. The frame suggests what the controversy is about, the essence of the issue"* [23]. As Kinder [17] resumes "frames are rhetorical weapons created and sharpened by political elites to advance their interest and ideas. *Frames* also lived inside the mind; they are cognitive structures that help individual citizens make sense of the issues that animate political life". From the cognitive-linguistic area, George Lakoff [24] had used intensively this concept of *frame* to explain the use of language in US politics. Lakoff argues that politicians invoke *frames* to dominate debates because they know that it is crucial: to attack the opponents' *frame* has the unwanted effect of reinforcing their message.

We aim to address stereotypes about immigrants as a result of this activity of *framing* in political and media speeches proposing a taxonomy that focuses on the different *frames* that politicians use to speak about immigrants. The concept of *frame* allows us to consider social cognition more a narrative process than a conceptual one. If as Jerome Bruner [25] states, the principle that organises the cognitive system of commonsense thinking is narrative rather than conceptual, we would consider narrative scenarios more than attributes assigned to a group in our detection of stereotypes about immigrants.

With the *framing* approach to stereotypes, we could detect how politicians built a *frame* that tells a story about the group focusing only on some features of the collective. In these speeches, they shape a stereotype without using explicit attributes for the group. These *frames* are subtle but powerful mechanisms to associate a group with some characteristics that are different dimensions of the stereotype.

The repeated use of this *frames* about a collective is present in the texts that computational systems process. This replication of a stereotyped vision of certain groups has an undesirable impact on people's life when they interact with technology. If stereotyping is a common bias difficult to fight in social communication, the data on the web is more likely to suffer from this lack of diversity because most of the information is created on the web by a minority of active users. For instance, 7% from a total of 40,000 users provide 50% of the total amount of posts in Facebook: it is not difficult to assume that this minority of users does not represent the knowledge and opinion of the majority [26].

In addition, recent studies have shown the difficulty of detecting ideological bias manifested in, for example, hyperpartisan news, that is news that tends to provide strongly biased information or exaggeration ending in fake news. If hyperpartisan news is easy to accept by the public that sees in them a confirmation of their own beliefs [3,27] we can expect a great difficulty in mitigating the use of stereotypes.

In computational linguistics this problem has been addressed in some works where different techniques have been proposed to measure, represent, and reduce social bias, in particular, stereotypes and prejudice, concerning race, nationality, ethnic, and mostly gender and sex, among others [28,29]. Most of them use a word embeddings representation and rely on the association of attributes to a social group to approach the stereotype (or other social bias) detection. We aim at approaching the problem of identifying stereotypes from a narrative perspective where computational linguistics could play a major role in analysing the complex process in which social actors create a stereotype placing a group in specific *frames*. Approaching the problem of stereotypes from this new perspective could also help to develop more sensitive tools to detect social bias in texts and new strategies to mitigate it.

We observed in the literature of computational linguistics a lack of datasets annotated with stereotypes and also works addressing the stereotypes about immigrants. We found that [30] created a dataset in Italian and included a binary stereotype annotation, but this work is mainly focused on hate speech and only annotates the existence or not of a stereotype belief about the hate speech target. In [6] it is proposed a dataset that includes the domain of racism (additionally to gender, religion, and profession), and report *immigrate* as one of the most relevant keywords that characterise such domain of bias. However, the authors do not focus on the study of stereotypes about immigrants.

In order to detect how social bias about certain groups is present in everyday language, it is necessary to have a complex view of stereotypes taking into account both positive and negative beliefs and also how the different *frames* shape the stereotype. This more refined analysis of stereotypes would make it possible to detect social bias 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 a parliamentary debate.

In this paper we propose: (i) a social psychology grounded taxonomy (and an annotation guide) that considers the genesis of the stereotype taking into account the different *frames* in which the group is placed; (ii) *StereoImmigrants*, the first dataset annotated with dimensions of stereotypes about immigrants from political debates in a national parliamentary chamber; and (iii) a baseline for immigrant stereotype classification in the categories of the proposed taxonomy, using the state-of-the-art transformer models. For our experiments, we use some recent monolingual and multilingual transformer models (based on BERT) known for their effectiveness at the context-heavy understanding.

This paper aims to answer the following research questions:

RQ1: Is it possible to create a more fine grained taxonomy of stereotypes about immigrants from a social psychology perspective that focuses on *frames* and not on attributes defining the group?

RQ2: How feasible is to create a stereotype-annotated dataset relying on the new taxonomy?

RQ3: How effective classical machine learning and state-of-the-art transformers models are at distinguishing different categories of stereotypes about immigrants with this taxonomy?

The rest of the paper is structured as follows. Section 2 describes related work about stereotypes and social bias, both from social psychology and computational linguistics perspectives. Section 3 introduces the proposed social psychology grounded taxonomy and the annotation process that was employed to annotate the *StereoImmigrants* dataset. Sections 4 and 5 present the models that we use in the experiments and the experimental settings, respectively. In Section 6 we discuss the obtained results, and we conclude our work in Section 7 in which we also mention some directions for future work.

2. Related Work

From a computational perspective, there is a long list of works that address problems related to social bias like the detection of hate speech [31,32], aggressive language [33], abusive language [34], hostility [35], racism [36], and misogynistic language [37] among

others. In this paper, we focus our attention on studying the genesis of stereotypes, specifically about immigrants.

In computational linguistics, stereotypes have been studied in images and text as well. For instance, [38] offers a study on the fairness of the algorithms that detect the descriptions of people appearing in images and their inferred gender; while [39] also focused on gender stereotypes, the authors study how the description is affected by the context in the image. Other works show the linguistic biases that are present in the way that one uses language as a function of the social group of the person(s) being described in the descriptions of images depicting people [40].

In [40], the authors describe some of the evidences of linguistic biases: (i) category labels and (ii) descriptions of behaviours. The former consists of labels used to refer to social categories, for example, explicitly marking unexpected gender roles or occupation when this one is inconsistent with the stereotypically expected role for the person's gender (e.g., *female* surgeon, *male* nurse); labels for individuals showing behaviours that violate the general stereotype (e.g., a *nice* Moroccan, a *tough* woman); and the use of nouns compared to adjectives to describe a person (e.g., being *a Jew* vs. Jewish, or Paul is *a homosexual* vs. is homosexual). The latter includes the description of the subject instead of an observable action (e.g., Jack *is flirtatious*, vs. Jack talks to Sue); the use of relatively more concrete language to describe behaviour that is inconsistent with the stereotype (e.g., *he has tears in his eyes* vs. the female consistent stereotype *she is emotional*); and the tendency to provide relatively more explanations in descriptions of inconsistent stereotype to make sense of the incongruity, among others.

To sum up, people reveal their stereotype expectancies in many subtle ways in the words they use. This fact can explain the effectiveness of several computational works at measuring social bias (e.g., gender, racial, religion, and ethnic stereotypes among others) by using word representations [28,41,42]. In [41], social bias are quantified by using embeddings of representative words such as women, men, Asians living in the United States, and white people (i.e., non-Hispanic subpopulation from the United States). The authors computed the average Euclidean distance between each representative group vector and each vector in a neutral word list of interest, which could be occupations or adjectives (this association of adjectives/occupations to the social group is consistent with [40] regarding stereotypes). The difference of the average distances is the metric they used for capturing personality trait stereotypes that were contrasted with historical surveys, gender stereotypes from 1977 and 1990, and ethnic stereotypes from 1933, 1951, and 1969. The authors found a correlation between the embedding gender bias and quantifiable demographic trends in the occupation participation in that period. Similar experiments were carried out with ethnic occupation. The results showed that several adjectives (e.g., *delicate*, *artificial*, *emotional*, etc.) tend to be more associated to women than to men; also some occupations (e.g., *professor*, *scientist*, *engineer*, etc.) are more associated to Asians, and other occupations (e.g., *sheriff*, *clergy*, *photographer*, etc.) to white people.

In [28,42], word embeddings are used to measure (and reduce) the bias. In [28], a methodology based directly on word embeddings is proposed to differentiate gender bias associations (e.g., biased association between *receptionist* and *female*) from associations of related concepts (e.g., between *queen* and *female*); a neutralisation and equalisation debias process. The same debias process is used in [42], but in this work a new representation to detect the bias is proposed: the authors include a contextualisation step and create two subspaces of the attribute (e.g., gender), one for male and other for female. The contextualisation relies on a large and diverse set of sentences in which the bias attribute words (e.g., *he/she, man/woman*) appear; for example, if in one subspace it is included a sentence containing *he*, the same sentence is included in the subspace for female but replacing *he* by *she*. In addition, in this line, [29] proposes some metrics over word embeddings representations to measure the bias. In this work, they distinguish two different biases: (i) *implicit bias*, in which we only have sets of target terms with respect to which a bias is expected to exist in the embedding space (e.g., $T1 = \{physics; chemistry; experiment\}$ and

$T2 = \{poetry; dance; drama\}$ without any specification, one could expect in $T1$ and $T2$ a gender bias towards $\{man, father, woman, girl\}$); (ii) and *explicit bias*, in addition to sets $T1$ and $T2$, it is given one (e.g., $A = \{man, father, woman, girl\}$) or more (e.g., two opposite attribute sets $A1 = \{man, father\}$ and $A2 = \{woman, girl\}$) reference attribute sets.

Some other recent works on bias again face the gender stereotypes. A Gender Stereotype Reinforcement (GSR) measure was proposed in [43] to quantify the extent to which a search engine responds to stereotypically gendered queries with documents containing stereotypical language. Recently, in [44], the authors have compared the efficacy of lexicon-based approaches and end-to-end machine learning-based approaches (in particular BERT); the obtained results showed that the latter is significantly more robust and accurate, even when trained by moderately sized datasets. Differently, in [45] it is used Natural Language Inference (NLI) as the mechanism for measuring stereotypes [45]. The idea is that invalid inferences about sentences can expose underlying biases, and based on that, it is possible to see how gender biases affect inferences related to occupations. For example, a gender-biased representation of the word *accountant* may lead to a non-neutral prediction in which the sentence *The accountant ate a bagel* is an entailment or contradiction of the sentences *The man ate a bagel* and *The woman ate a bagel*; this could happen because of the gender bias concerning occupations. Therefore, the predictions of NLI on a set of entailment pairs that should be inherently neutral are used to compute the deviation from neutrality, which is assumed as the gender bias.

A similar idea of [45] has been used in [6] but with a different perspective. In [6], the authors propose two different level tests for measuring bias: (i) the intrasentence test, in which there are a sentence describing the target group and a set of three attributes which correspond to a stereotype, an antistereotype, and an unrelated option; and (ii) the intersentence test, consisting of a first sentence containing the target group, a second sentence containing a stereotypical attribute of the target group, a third one containing an antistereotypical attribute, and lastly an unrelated sentence. These tests remain the aforementioned idea of [45] to use NLI to measure entailment, contradiction, or neutral inferences to quantify the bias. To evaluate their proposal, the authors of [6] collected a dataset (StereoSet) for measuring bias related to four domains: gender, profession, race, and religion. For this purpose, they use specific words to represent each social group.

However, stereotypes are not always merely the association of *words* (seen as attributes or characteristics) to two opposite social groups (e.g., women vs. men), and it is not always clear to define the opposite groups by using specific keywords, for instance in the case of immigrants vs. nonimmigrants, the set of words to represent nonimmigrants is not clear. There are few works related to the detection of stereotypes about immigrants. In [46], a system is proposed that allows one to see what was said about Muslims and Dutch people. The authors use the collection of descriptions that a single text provides on a given entity or event (it was called *microportraits*: labels, descriptions, properties, roles). For example, the expression *the pious Muslim smiled* contains the label *Muslim*, the property *pious*, and the role *smiling*. This is an interesting study that helps explain how prejudice works according to social psychologists. In [30] an Italian dataset was created that focused on hate speech against immigrants, that included the annotation *{yes, no}* about whether a tweet is a (mostly untrue) stereotype or not. In the HaSpeeDe shared task at EVALITA 2020 [47], six teams submitted their results for the stereotype detection task in addition to their hate speech models, using the above dataset. Most of those teams only adapted the same hate speech model to stereotype identification, representing (and reducing) stereotypes to characteristics of hate speech. The authors of [47] observed that stereotype appears as a more subtle phenomenon that needs to be approached also as nonhurtful text.

From a psychosocial perspective, the better-established model to analyze the language that shapes a group of stereotypes is the Stereotype Content Model (SCM) developed by Fiske and colleagues [48–50]. Fiske has developed his model arguing that in encounters with conspecifics, social animals—i.e., humans—must determine, immediately, whether the "other" is friend or foe (i.e., intends good or ill) and, then, whether the "other" can

enact those intentions. Authors affirm that in answering these questions, humans use two universal dimensions of social cognition—warmth and competence—to judge individuals and groups. People perceived as warm and competent elicit uniformly positive emotions and behaviour, whereas those perceived as lacking warmth and competence elicit uniform negativity. People classified as high on one dimension and low on the other elicit predictable, ambivalent affective and behavioural reactions. This theoretical framework has been completed with the ambivalent stereotypes hypothesis: many groups are stereotyped as high in one dimension and low in the other [48].

Cuddy, Fiske and Glik [51] also have investigated how stereotypes and emotions shape behavioural tendencies toward different groups and have proposed the BIAS Map (Behaviors from Intergroup Affect and Stereotypes).They did a correlation study with a representative US sample and conclude that warmth stereotypes determine active behavioural tendencies—attenuating active harm (harassing) and eliciting active facilitation (helping). Competence stereotypes determine passive behavioural tendencies—attenuating passive harm (neglecting) and eliciting passive facilitation (associating). Admired groups (warm, competent) elicit both facilitation tendencies; hated groups (cold, incompetent) elicit both harm tendencies. Envied groups (competent, cold) elicit passive facilitation but active harm and pitied groups (warm, incompetent) elicit active facilitation but passive harm. In this research, the authors also find that emotions predict behavioural tendencies more strongly than stereotypes do and usually mediate stereotype-behavioural-tendency links. In this research [51] immigrants are placed between the set of groups that are seen as "low warmth and low competence", with other social groups seen as poor, homeless, including Latinos, Muslims, and Africans, in the particular US context. It is predicted that groups placed in this position evoked disgust and contempt in terms of emotions [52]. However, how does one explain the appeal to fear that right-wing politicians use intensively when they speak about immigrants? Why be afraid of a group we see as incompetent? The authors do not include fear among the emotions linked to the low competence factor.

Stereotypes about immigrants present specific difficulties that Fiske addressed in an early work [53]: the internal variability existing between the members of the social category "immigrants" led the authors to study a more fine-grained taxonomy of the stereotype, based on nationalities and socioeconomic status (documented, undocumented, farm-workers, the tech industry, first-generation, and third-generation). This research concludes that people conceptualise immigrants at three levels at least: the "generic immigrant", who is equally low in *competence* and *warmth*; clusters of immigrant groups uniquely defined by one attribute, such as low or high competence, or high warmth; and immigrants by specific origin.

One interesting remark arises from this study: the group that received the least favourable stereotype across both dimensions was "undocumented immigrants". In contrast, "documented immigrants" were perceived similarly to an American. Legal status alone determines whether an immigrant is perceived as a regular member of the mainstream society or as an outsider with the lowest status. The authors propose that one possible extension from this study could be the role of media framing of immigration status in perceived competition for finite amounts of societal resources. This idea of focusing on the framing activity is at the core of our research because we consider that it is this rhetorical activity of framing the one that enables the existence of what Lee and Fiske [53] define as "the generic immigrant".

Another interesting suggestion of these authors is that people's differing evaluations of documented and undocumented immigrants suggest that some dimensions (in this case, legal documentation) overwhelmingly bias judgement. The authors suggest for further research the question of which dimensions are most influential in perceiving immigrants when people receive information on multiple dimensions, for instance, if Asian immigrants are competent but undocumented immigrants are not, are undocumented Asian immigrants high or low in competence? They suspect that the more salient dimension would guide perception. In our research we will propose that the most salient dimension

will be the result of a *frame* that presents the group in a given scenario. Recently, Kervyn, Fiske, and Yzerbyt [54] introduce—in their experiments on stereotypes about immigrants—symbolic and realistic threats and found that they improve the SCM's prediction of warmth.

The realistic treat as origin of prejudice and stereotypes has a long tradition in the study of intergroup conflict [55,56] and also in the Integrated Threat Theory [57] that proposes two types of perceived threat from outgroups. The first type comes from research on Realistic Group Conflict Theory [55], which posits that groups compete for scarce resources and, therefore, one group's success threatens other groups' well-being, resulting in negative outgroup attitudes. The second type of intergroup threat originates from research on Symbolic Racism, which considers racism as coming from conflicting beliefs and values rather than conflicting goals [58,59]. Symbolic threat perceives the outgroup as threatening ingroup worldviews, assuming group differences in values, standards, beliefs, and attitudes.

Another line of research on prejudice towards minorities has been developed under the framework of Social Representations Theory by Moscovici [60–62]. For the study of prejudice, Moscovici [63] had the hypothesis that nature and culture constitute dimensions along which representations of human groups, that is to say, stereotypes, are organized in a sort of social ranking. Culture means "civilization" while Nature is "the primitive condition before human society". From this approach one of the key points to understand how people stereotype minorities is the role that these stigmatised groups play in the continuum between these two extremes of nature and culture. Perez, Moscovici and Chulvi [64,65] have shown in their experiments that the majority group see itself nearest to the culture extreme of this vector and place the minority group nearest to nature. Other research in the dehumanisation process has shown that it is present not only in the extreme manifestation of prejudice but it can also take subtle and everyday forms [66], for instance, differential attribution of uniquely human emotions to the ingroup versus an outgroup [67].

In the other extreme of the stereotyped continuum of minorities, we find the victimization bias [68,69]. Most of the groups that have been considered deviant or marginal gained the status of victims from a historical process that according to Barkan [70] reaches its peak in 1990. One consequence of this shift is a change in the way that minorities are named; for instance, persons formerly labelled "handicapped" are now categorised as "differently abled" or "immigrants" are named "migrants" or "non-nationals" [71] in a social effort to give restitution to this minorities. The status of victims confers a feeling of moral superiority but according to Steele [72] "binds the victim to its victimization by linking the power to his status of the victim". When this status of victims becomes salient in the narratives about a collective, the counterstereotyped cases are made invisible.

All of this social bias that fills the content of a stereotype of immigrants and other minority groups has two main features: first, it serves to maintain minorities' discrimination, and second, it occurs in everyday language. As stereotypes are present in everyday language, they are also in the texts that systems use to classify and retrieve information. Those minorities as immigrants that suffer from this stereotyped vision of their collectives are now, with the extension of the web and massive use of social media [26], in the face of a loudspeaker that amplifies their stigma to infinity.

At the beginning of this section we mentioned some works that reduce the gender bias: similar attempts are needed to attenuate the social bias about immigrants. Nevertheless, we think that in this case more knowledge is needed for developing automatic systems that could be effective in mitigating and detecting social bias. Computational linguistics could play a major role in understanding how the immigrant's stereotype are employed—in its general formulation—as Lee and Fiske said [53]. To do this, it would be necessary to move the annotation process from the words that are used to qualify the group, to the narratives—stories about what is going on—that social psychology defines as *"frames"* [18,23,73] in which the minority is placed, insistently, again and again, by a social actor in the public discourse. In Computational linguistics, the concept of *frame* is also used by [74], where

the authors propose Social Bias Frames, a novel conceptual formalism that aims to model pragmatic *frames* in which people project social biases and stereotypes on others.

To study more deeply this problem, the present work proposes an exhaustive taxonomy with different dimensions of the immigrants' stereotypes that have been used to annotate *StereoImmigrants*, a dataset of political speeches about immigration in Spain. Different from previous work, this work embraces the general picture about immigrants (in Spain) and not only the "negative" aspects of the stereotypes. Moreover, we do not rely on the attributes, characteristics, and roles that are played by the immigrants, but we use the *frames* (labelled by humans following the taxonomy shown in Appendix A) in which the immigrants are placed to detect the different dimensions of the stereotypes. More subtle examples of stereotypes are used to capture automatically more complex linguistic patterns in the way that stereotypes are present.

3. Social Psychology Grounded Taxonomy and StereoImmigrants Dataset

3.1. A Social Psychology Grounded Taxonomy

We have constructed a taxonomy trying to cover the whole attitudinal spectrum of stereotypes about immigrants, from the pro-immigrant attitude to the anti-immigrant attitude. Attitude is a theoretical concept that has been preeminent since the very beginnings of systematic research in social psychology, especially, in the study of prejudice [75]. If the stereotype is the cognitive aspect of the prejudice (a set of beliefs), the attitude expresses the effect (we could also say the feelings and emotions) that a group provokes. This taxonomy has six categories based on how the group is presented. We found that in public discourse immigrants could be presented as: (i) equals to the majority but the target of xenophobia (i.e., must have same rights and same duties but are discriminated), (ii) victims (e.g., people suffering from poverty or labour exploitation), (iii) an economic resource (i.e., workers that contribute to economic development), (iv) a threat for the group (i.e., cause of disorder because they are illegal and too many and introduce unbalances in our societies), or (v) a threat for the individual (i.e., a competitor for limited resources or a danger to personal welfare and safety). The sixth and last category presents immigrants as animals, excluding them—in whole or in part, explicitly or implicitly—from the supracategory "human beings".

The two first categories of the taxonomy (i.e., *Xenophobia's Victims* and *Suffering Victims*) hold a pro-immigrant attitude and we can aggregate them in a supracategory that we call *Victims*. Under this supracategory the goal of the speaker is to build a fair world. The speeches focus on xenophobic attitudes that are behind the problems of the minority and stress the causes of immigration. In the first category (*Xenophobia's Victims*) the speakers emphasise that the problem is not the minority but the racism and xenophobia from the majority. In this category, we include sentences such as *"We are ready to collaborate in all aspects that make life easier for our emigrants abroad, but at the same time we consider it important to work for the integration of immigrants in our country"* because they make a parallelism between the immigrant community in Spain and the Spaniards who emigrated focusing on the need to an integration strategy. In the second category (*Suffering Victims*), we include sentences such as *"You can say what you like, but the migratory movements affecting the planet are almost exclusively linked to the phenomenon of poverty and misery"*.

The third category of the taxonomy (immigrants as an economical resource) holds an ambivalent attitude [76] that presents immigrants as an instrument for achieving societal goals. The goal of the speaker is to manage with efficiency a phenomenon difficult to avoid. In this category, we include sentences like *"how can you say that there should be no more immigrants regularisation's if, after all, reports such as that of La Caixa or BBVA indicate that the Spanish labor market needs foreigners?"*.

The three last categories of the taxonomy—immigrants as collective threat (iv), individual threat (v), or as less humans (vi)—hold the anti-immigrant attitude. We can aggregate these three categories in a supracategory that we call *Threat*. The goal of the speaker is the protection of the "national" group in front of immigrants that are presented as a danger or

less human . Focused on the problems of the majority and critical about the immigrants, these speeches stress the negative effects of immigration. In the fourth category, we include sentences that consider immigration a source of problems such as *"it is clear that there is an increase in the number of people trying to enter Spain illegally"*. In the fifth category we include sentences that present immigrants as a threat not only for the collective but also for the health and security of the majority group in an explicit or implicit way: *"We need to tackle problems such as terrorism and immigration"*. In the last category of the taxonomy, which corresponds with the "dehumanization bias" we have not found examples in our dataset from the Spanish Parliament but some examples from statements made by Donald Trump could serve to illustrate the sense of this category. The former President of the United States said in a press conference at the White House: *"You wouldn't believe how bad these people are. These aren't people, these are animals, and we're taking them out of the country at a level and at a rate that's never happened before."* (NYT, 16 May 2018).

We have defined a finer level of granularity in each category to facilitate the annotation by humans (see Appendix A). Each category contains a subset of *frames* that politicians used to speak about immigrants. These *frames* do not describe the group but convey a homogeneous picture of the group placing it in a particular scenario. For example, in the fifth category, defined as "a personal threat", we have identified three *frames*: (i) immigrants compete with the country's population for resources such as jobs, health services, etc.; (ii) immigrants bring diseases; and (iii) immigrants are associated with crime.

3.2. Annotation of the StereoImmigrants Dataset

We annotated political speeches presented in the ParlSpeech V2 dataset [77], a dataset that has been already used in other tasks like Sentiment Analysis [78,79]. One of the peculiarities of ParlSpeech is that it is a transcription of a real debate between different and relevant social actors. Its dialogic nature makes it more difficult to approach from the perspective of computational linguistics, but it is also an opportunity to develop an interdisciplinary methodology that focuses on how social interaction takes place in language.

Specifically, we focused on the principal parliament in Spain, the Congress of Deputies (*Congreso de los Diputados*). This chamber is located in Madrid, has representatives from all regions, and elects the nation's prime minister. Using a list of 60 keywords (see Appendix B) we selected all the speeches that contained at least one keyword. We obtain 5910 speeches from different years (see Figure 1).

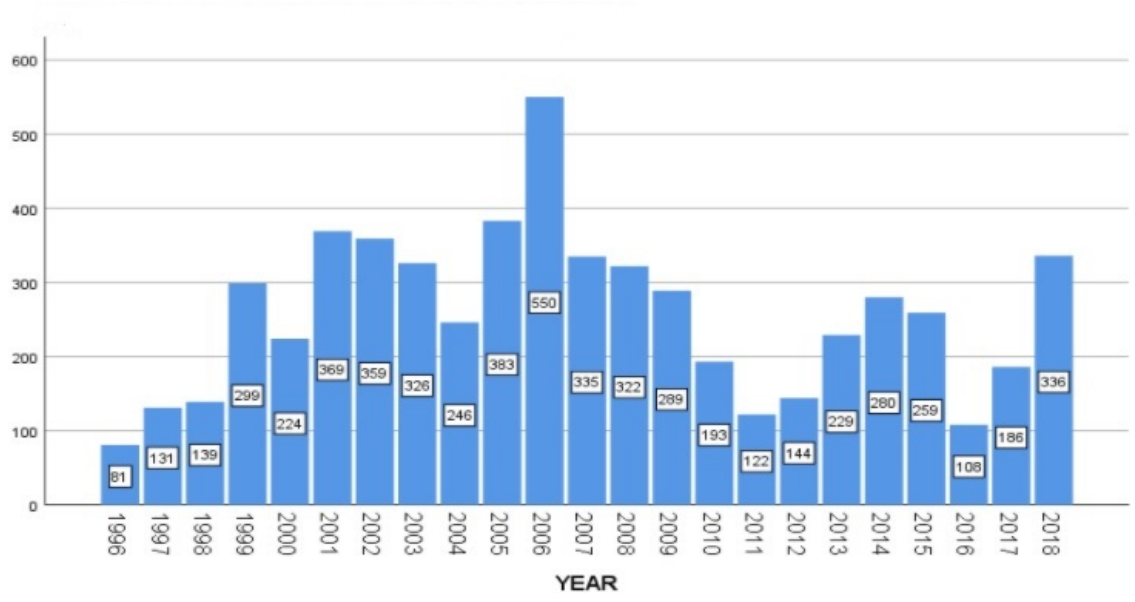

Figure 1. ParlSpeech V2 dataset: number of speeches with at least one immigrant-related keyword.

The year 2006 saw intense parliamentary activity on immigration. The Spanish media denominate this year the "Cayucos Crisis": the arrival of more boats than usual from different African countries to the coasts of the Canary Islands. The presence of these events in the media was very abundant, parliamentary activity on the issue was very intense, and Spanish public opinion was increasingly concerned about the issue of immigration (Based on the CIS Barometer: http://www.cis.es/cis/opencm/ES/11_barometros/index.jsp), accessed on 8 April 2021, (Figure 2).

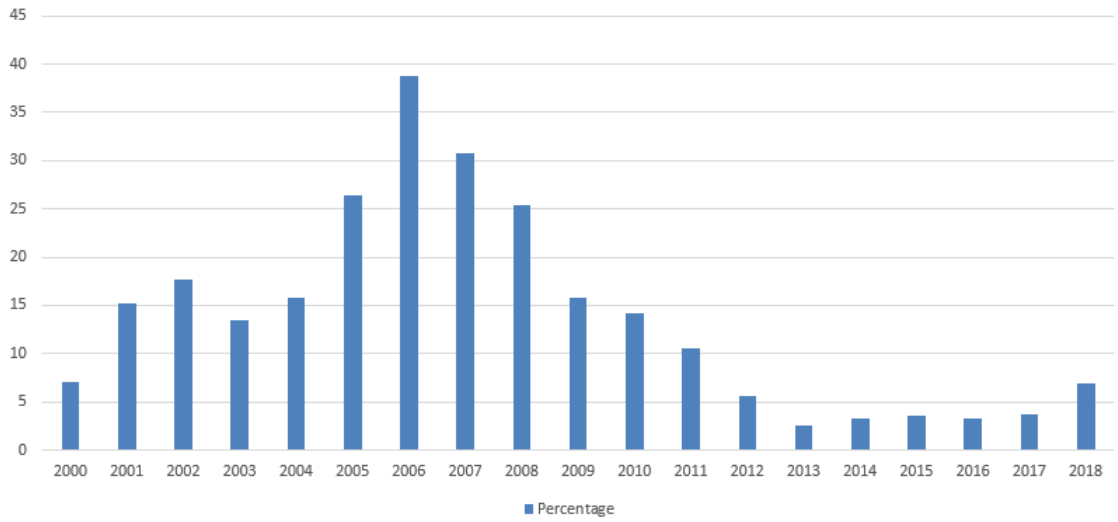

Figure 2. Percentage of people who consider immigration one of Spain's three main problems from a representative sample of the Spanish population. Source: CIS Barometer.

Official data [80] on immigration for 2007 contrast with the climax that we see in the Spanish parliament and in the mass media that covered the "Cayucos Crisis". The total immigrant population at the beginning of 2007 was 4.5 million. By origin, 40 percent came from Latin America and 33% from the European Union. Only 17% came from Africa despite the fact that the image of African people arriving in poor boats was the *frame* that was more used when the immigration phenomenon appeared in mass media and in political debates. The 2007 National Immigrant Survey indicates that only 10% of the immigrant's population was "undocumented" [80] but nearly 39% of the country's population was worried about illegal immigration how the CIS shows (see Figure 2). The 2007 National Immigrant Survey also shows that the immigrant population was quite similar to the Spaniards in some parameters: most parts of the immigrant population (59%) between 20 and 34 years old have completed lower and upper secondary education, 17% have higher education, and only 23% belong to the primary education or no education group. The different level of studies between the immigrant population and the Spanish population at this moment was not enormous: the group of Spaniards that has primary education or does not have any education in this age group was 13%. In fact, other studies, using other sources of data, highlight a situation in which the immigrant population has a very similar profile in terms of higher education to the Spanish population [81,82].

This gap between the variability of the real situation and the image that was conveyed through the mass media and the parliamentary interventions led us to build the dataset focusing on the speeches from 1996 to 1998, from 2006 to 2008, and from 2016 to 2018. In that way, we could contemplate speeches of consecutive years and also how these vary from one decade to another. We selected 582 speeches with more than one keyword and manually discarded the ones in which immigrants were mentioned alongside other groups but only tangentially. From these selected speeches, we manually extracted 3635 sentences for studying stereotypes about immigrants.

An expert in prejudice from the social psychology area annotated manually these sentences at the finest granularity of the taxonomy (i.e., identifying the different frames that fall into the same category, see Appendix A), and selected also negatives examples (i.e., *Nonstereotype* label) where politicians speak about immigration but do not use explicitly or implicitly any category of the stereotypes about immigrants. After this expert annotation we use a procedure with some similarities with [6]: five nonexperts 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 annotators where 77 undergraduate students from psychology, fine arts, and business. We only retained sentences where at least three validators agreed on the same category. In our dataset, each sentence is accompanied by the following information: politician's name, political party to which the speaker belongs, and date of the parliamentary session. This metainformation was hidden also for the expert annotator and for the nonexperts.

Table 1 depicts the distribution of instances per label. We include the mean of the length of the instance based on tokens to help us to define the hyperparameters of the models in Section 5. We can observe an imbalance across the dimensions of stereotype, where dimension 5 (i.e., *Personal threat*) is the smallest set of instances with only 81, and dimension 4 (i.e., *Collective threat*) is the biggest with 655 instances. In addition, the dataset has an imbalance regarding *Stereotype* (1673 instance) vs. *Nonstereotype* (1962 instances). In general, all the labels have a similar distribution according to the length of the instances, but the nonstereotyped instances are slightly more consistent in length, with a smaller standard deviation. We take into account this distribution in the experimental settings (Section 5).

Table 1. Statistics of the *StereoImmigrants* dataset. For each label, *Stereotype* and *Nonstereotype*, and for each category of the stereotype, we show the number of instances and the length based on tokens (ignoring punctuation marks).

			Length in Tokens		
		Instances	**Min**	**Mean ± Standard Deviation**	**Max**
Stereotype	**1. *Xenophobia's Victims***	186	6	50.55 ± 30.59	183
	2. *Suffering Victims*	557	7	47.32 ± 24.41	151
	3. *Economical Resources*	194	9	42.39 ± 22.31	128
	4. *Collective threat*	655	8	43.42 ± 23.28	162
	5. *Personal threat*	81	9	48.26 ± 25.56	149
	All dimensions joined	1673	6	45.62 ± 24.69	183
Nonstereotype	1962	3	36.00 ± 21.17	165	
Total	3635	3	40.43 ± 23.35	183	

3.3. Evaluation of the Taxonomy

We asked nonexpert annotators for their judgement about the attitude expressed in the text. Concretely, each annotator had to say if this text expressed a pro-immigrant, an anti-immigrant, or an ambivalent attitude that annotators qualified as neutral. The purpose of this second task was to test if the theoretical value assigned to each category in terms of positive or negative stereotype was justified. Our aim was to analyse if there were any significant relations among categories and attitudes. For instance, we expect that given a text labelled with the 1st category (i.e., *Xenophobia's Victims*), there will be a significant high probability of being judged to express a pro-immigrant attitude.

To test the relationship among the categories of the taxonomy and the attitudes towards immigrants, we performed a chi-square test and a residual analysis [83]. A residual is a difference between the observed and expected values for a cell. The larger the residual, the greater the contribution of the cell to the magnitude of the resulting chi-square obtained value. However, cells with the largest expected values also produce the largest raw

residuals. To overcome that redundancy, a standardised or Pearson residual is calculated by dividing the raw residual by the square root of the expected value as an estimate of the raw residual's standard deviation. If the standardised residual is beyond the range of ±2.58 that cell can be considered to be a major contributor to the chi-square significance.

Results confirm a significant relation (Pearson $\chi^2 = 3828.24$, $df = 8$, $p < 0.000$; see Table 2) of the taxonomy's categories and the positive, neutral, or negative evaluation. The residual analysis in Table 2 shows that category 1 (*Xenophobia's Victims*) and category 2 (*Suffering Victims*) are significantly associated with positive attitudes, category 3 (i.e., *Economical Resource*) is significantly associated with a neutral attitude, and categories 4 and 5 (i.e., *Collective Threat* and *Personal Threat*) are significantly associated with a negative attitude.

Table 2. The relation among categories of the taxonomy and attitudes expressed in the texts of the dataset. Chi-square test with adjusted standardised residuals.

		Attitudes towards Immigrants			
Taxonomy Categories		**Pro-Immigrant**	**Anti-Immigrant**	**Neutral**	**Total**
Xenophobia's Victims	Obs	916	90	143	1149
	Adj. Res	**25.4**	−23.6	−3.0	
Suffering Victims	Obs	2002	414	363	2779
	Adj. Res	**34.8**	−32.3	−4.2	
Economical resource	Obs	482	187	259	928
	Adj. Res	4.4	−12.7	**11.1**	
Collective threat	Obs	339	2406	508	3253
	Adj. Res	−50.5	**51.2**	0.3	
Personal threat	Obs	108	268	45	421
	Adj. Res	−8.2	**10.4**	−2.8	
	Total (Obs)	3847	3365	1318	8530

Taking into account these results we consider it appropriate to use categories 1 and 2 as a supracategory that we named *Victims* and categories 4 and 5 in a supracategory *Threat*. These supracategories will be used to evaluate the effectiveness of the models at the automatic classification of stereotypes about immigrants. category 3, that we qualify also as supracategory named *Resources*, evaluated as neutral, will be left out of the experiments because of the small number of instances. Figure 3 summarises the taxonomy at its different levels.

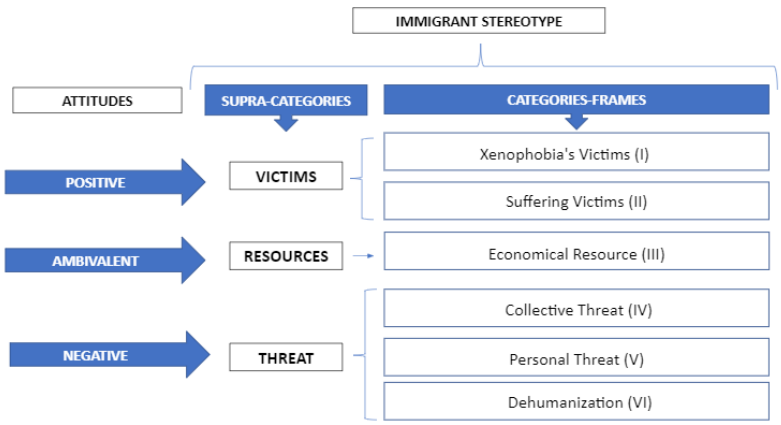

Figure 3. Explanatory taxonomy scheme.

4. Models

In this section, we briefly present the state-of-the-art models we have used for our experiments, which have been trained with huge general language datasets of pretrained systems based on Bidirectional Encoder Representations from Transformers (BERT).

BERT is a language representation model designed to pretrain deep bidirectional representations from an unlabeled text by jointly conditioning on both left and right context in all layers [84]. That is, BERT generates a representation of each word that is based on the other words in the sentence. This allows the model to capture complex patterns in the texts to study stereotypes, going beyond merely the use of words and capturing semantic and syntactic patterns in the same representation. BERT has also an attention mechanism that distinguishes if a word is attended by the model. These attention weights could be used also to give more insights about the results of the model and be used as a tool to support the work of human experts. Some important aspects of BERT include the pretraining, the fine-tuning, and the capability to be adapted to many types of Natural Language Processing (NLP) tasks like text classification.

To classify text in Spanish, we have used two monolingual models (BETO and SpanBERTa) and two multilingual models (MBERT and XLM-RoBERTa) briefly described below:

M-BERT: The Multilingual BERT is pretrained on the concatenation of monolingual Wikipedia datasets from 104 languages, showing good performance in many cross-lingual scenarios [84–86].

XLM-RoBERTa: It was trained on 2.5 TB of newly created clean CommonCrawl data in 100 languages. It provides strong gains over previously released multilingual models like M-BERT on downstream tasks like classification, sequence labelling, and question answering. In [87], it was reported with better results to the one obtained by fine-tuning with Spanish data only.

BETO: This is a recent BERT model trained on a big Spanish dataset [87]. It has been compared with multilingual models obtaining better or competitive results [88]. BETO was trained using 12 self-attention layers with 16 attention heads each and 1024 as hidden sizes. It was trained using the data from Wikipedia and all of the sources of the OPUS Project [89]. BETO also was ranked in a better place than Logistic Regression in the prediction of aggressive tweets [90].

SpanBERTa: SpanBERTa (https://skimai.com/roberta-language-model-for-spanish/), accessed on 8 April 2021, was trained on 18 GB of the OSCAR's Spanish corpus (https://oscar-corpus.com/), accessed on 8 April 2021, following the RoBERTa's training schema [91]. It is built on BERT and modifies key hyperparameters, removing the next-sentence pretraining objective and training with much larger mini-batches and learning rates.

Moreover, we use as baselines classical machine learning models such as Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), and Naïve Bayes (NB).

5. Experimental Settings

We evaluate each model on two tasks. First, given a text about immigration, to predict whether or not it contains a stereotype about immigrants; second, given a text already known that is reflecting a stereotype, to detect if the stereotype corresponds to see immigrants as *victims* or *threat*.

Experiment I: Stereotype vs. Nonstereotype This experiment is very relevant for us because we have annotated not attributes that speakers assign to a group but narratives about the group that in an implicit way convey a stereotyped vision of the collective. Due to negative examples also being sentences from members of the Parliament speaking about immigration, we want to see if the models detect the subtle difference that consists in approaching the issue without personifying the problem in one group, i.e., immigrants as a social category.

Experiment II: Victims vs. Threat With this experiment we tried to see if the model can detect which dimension of stereotype about immigrants has been used in the political discourse. One common rhetorical strategy used by politicians that present immigrants as a threat to the majority group is to dedicate a part of their speech to recognise the suffering of the minority. However, these claims of compassion are framed by a discourse that clearly presents immigration as a problem and migrants as a threat. We are interested to see if a model is able to distinguish the deeper meaning of that statement as the human annotators did.

We apply a 10-fold cross-validation procedure and report our results in terms of *accuracy*. For the execution of each model, we balance the two labelled classes by randomly removing examples from the more populated class. Therefore, in Experiment I we use 1673 examples per label and in Experiment II we use 736 examples per label.

We use two monolingual Spanish transformer models: BETO cased and SpanBERTa; and two multilingual transformer models: MBERT (*bert-base-multilingual-cased*) and XLM-RoBERTa (*xlm-roberta-base*). We search the following hyperparameter grid: *learning rate* $\in \{0.01, 3 \times 10^{-5}\}$; the *batch size* $\in \{16, 32\}$; and the optimizer $\in \{adam, rmsprop\}$. Moreover, we apply a dropout value of 0.3 to the last dense layer. Dropout is aimed at reducing overfitting by dropping a percentage of random units at each weight update.

Besides the length of the texts not being restricted to predicting stereotypes in general, we have to select a value for the *max_length* hyperparameter to use the transformer models. According to the characteristics of our data, the mean of the lengths of the instances is approximately 40 tokens (see the last row of Table 1), with a standard deviation of around 20 tokens. Taking this in mind and evaluating the number of instances with a length greater than 40 tokens, we finally select 60 as the value for the hyperparameter *max_length* in order not to lose too many instances. Accordingly, our transformer models expect an input text of around 60 tokens. In the case of longer texts, only the first 60 will be used while the rest is truncated.

Furthermore, we use the *sklearn* implementation of the four classical machine learning models. All the parameters were taken by default, except for LR in which we use the *newton-cg* optimization method. For SVM, we employ specifically the *LinearSVC* implementation, which uses a linear kernel and has more flexibility in the choice of penalties and loss functions. The number of trees in the forest of the RF classifier is 100 (the one by default). The four models were evaluated with the bag of words representation, using the *tfidf* term weighting. We tested with unigrams, bigrams, and trigrams of words, but unigrams allow for obtaining slightly better results. Stopwords and punctuation marks were removed in a preprocessing step.

6. Results and Discussion

In this section, we present the results of our two preliminary experiments. The optimal hyperparameter configuration for all the models is the following: $learning\_rate = 3 \times 10^{-5}$, $optimizer = adam$, and $batch\_size = 32$. In general terms, we observe that the best performances were consistently achieved by BETO and M-BERT models. It is not surprising that M-BERT obtained better results than SpanBERTa, being the latter pretrained specifically for Spanish: a similar comparison between a multilingual model and a monolingual model was reported in [87]. Another general observation is that either for *Stereotypes vs. Nonstereotypes* and *Victims vs. Threat*, BETO seems to capture more complex patterns than the classical machine learning models, which are based on the bag of words representation.

The next subsections discuss the results of Experiment I and Experiment II in more detail.

Table 3. Accuracy achieved by each model in Experiment I on Stereotype vs. Nonstereotype. We indicate the *p*-value of the Mann–Whitney U test regarding the alternative hypothesis that the results of BETO are the highest compared to the other transformers. With * we indicate when accuracy is significantly lower than the result of BETO. The hypothesis is accepted with $p < 0.05$ except for XLM-RoBERTa model.

BETO	SpanBERTa	M-BERT	XLM-RoBERTa
0.861 ± 0.016	$0.766 * \pm 0.021$ $_{p=0.00018}$	$0.829 * \pm 0.022$ $_{p=0.00736}$	0.780 ± 0.105 $_{p=0.06057}$
LR	**SVM**	**NB**	**RF**
0.82	0.81	0.73	0.81

6.1. Experiment I: Stereotype vs. Nonstereotype

In this section, we present the results concerning the identification of stereotype about immigrants. In Table 3 we can appreciate that all the models obtained an *accuracy* above 0.73. The highest result was obtained by BETO with 0.861 of *accuracy* and a standard deviation of 0.016. This performance was significantly better than the one of SpanBERTa and M-BERT for $p < 0.05$ using the Mann–Withney U Test. We use this test, also known as Wilcoxon–Mann–Whitney test (a nonparametric alternative to the Student's *t*-test), considering that we randomly removed instances from the most populated class each time. Therefore, we assume independence among the results of the models. We see that LR, SVM, and RF obtained a higher accuracy compared to SpanBERTa and XLM-RoBERTa. We believe that this could be associated with the size of the dataset since more data could have had an impact on deep learning models achieving better results [92].

Besides the fact that stereotypes involve more than the presence of specific words, word n-grams with the highest Pointwise Mutual Information (PMI) (PMI makes it possible to see the most relevant features (i.e., words, n-grams of words) for each topic and is computed as $PMI(L, w) = \log \frac{p(L,w)}{p(L)p(w)}$. Where $p(L, w)$ is the probability of a feature to appear in a text labeled as L, $p(L)$ is the probability of a label (we assume the label distribution to be uniform), and $p(w)$ is the probability of w.) to each label allow us to see that nonstereotypical texts talk more about *ayudas* (help) to refugees and Africa (the country some immigrants come from, and it is mostly mentioned in the speeches we are working with), *acuerdos* (agreement) between countries, etc. While in stereotypical texts we find more commonly bigrams such as *inmigración ilegal* (illegal immigration), *inmigrantes irregulares* (irregular immigrants), and *regularización masiva* (massive regularization) among others that indirectly reflect problems associated to immigration (see Table 4).

Interestingly, it is not evident from observing the relevant n-grams why some of them are more related to the stereotypes about immigrants and others are not. This confirms that for the study of stereotypes about immigrants we have to go beyond the representative keywords that could define the social group. In other words, we should not rely only on intuitive words to base the measuring of bias in the case of stereotypes. In this sense, automatic approaches can detect other patterns that escape human detection.

We also confirm that the detection of stereotypes in this work, concerning immigrants, is not about characterising two opposite social groups but the immigrant group only. The nonstereotypical texts are neutral in this sense referring only to the topic of immigration without stereotyping at all.

In Table 5, we present the confusion matrix of BETO when obtaining the results shown in the Table 3. The model was similarly effective at predicting stereotypes and nonstereotypes, with a bit more confusion with the label Nonstereotype. Table 6 shows some examples where BETO misclassified the texts and their predictions.

Table 4. Bigrams and trigrams with highest mutual information with respect to Stereotype and Nonstereotype labels.

	N-Grams with Highest Mutual Information
Nonstereotype	*inmigrantes irregulares; señor rajoy; señor zapatero; países africanos; abordar asunto; abordar problema; absolutamente acuerdo; acción exterior; acción política; acoger personas; acogida refugiados; acogida temporal; acorde derechos; acuerdo materia; acuerdos gobierno; acuerdos marruecos; acuerdos mauritania; acuerdos readmisión; adquisición nacionalidad; afecta unión europea; agencia europea fronteras; aguas canarias; aguas territoriales; asilo inmigración; asunto preocupa; atención inmigración; autoridad moral; ayuda refugiado; ayuda áfrica;...*
Stereotype	*efecto llamada; inmigrantes ilegales; política inmigración; inmigración irregular; unión europea; materia inmigración; derechos humanos; inmigrantes irregulares; consejo europeo; inmigración ilegal; drama humano; regularización masiva; islas canarias; seres humanos; política común; proceso regularización; inmigración delincuencia; llegada masiva; respeto derechos; situación irregular; economía sumergida; mujeres inmigrantes; centros acogida; orden expulsión; centros internamiento; costas canarias; miles personas; europea inmigración; política migratoria; política exterior; respeto derechos humanos; menores acompañados; acogida canarias; drama humanitario; empresarios sindicatos; menores inmigrantes; crecimiento económico; acogida inmigrantes;...*

Table 5. Confusion matrix of BETO in Experiment I on Stereotype vs. Nonstereotype.

	Predicted Labels	
	Stereotype	***Nonstereotype***
Stereotype	1451	222
Nonstereotype	240	1433

6.2. Experiment II: Victims vs. Threat

In Table 7, we can see all the performances are above the 0.70 of *accuracy*. The results are, in general, smaller than in Experiment I; this could be because the size of the training set for the current experiment is smaller or due to the difficulty that we have already mentioned in Section 5, describing the scenario of Experiment II: as some humans annotators reported, one of the common rhetorical strategies in political discourse is to precede any critical statement towards the immigrant collective with an expression of compassion towards the human drama that it represents in order not to be accused of xenophobia. For instance, the speaker is going to say that Spain could not admit more immigrants (threat) but she starts speaking about how many people died trying to arrive (*Victims*). Methodologically, we decided to assign only one label per sentence, and perhaps it would have been more effective to annotate different syntagmas within the same sentence, at least in those sentences in which this discursive strategy was developed.

Similar to the previous experiment, the highest accuracy was obtained by BETO, but this time with 0.834 of accuracy and a standard deviation of 0.034. This performance was significantly better than the one of SpanBERTa for $p < 0.05$ using the Mann–Whitney U Test.

In Table 8, we show some of the n-grams (without including stopwords) more relevant for each label, for example: *atención humanitaria* (humanitarian care) , *atención sanitaria* (healthcare), *acogida personas* (welcome of people). These relevant n-grams allow us to figure out that the phrases are more likely to reflect the needs and pain of immigrants when they are seen as *victims*.

In Table 9, we present the confusion matrix of BETO at obtaining its result from the Table 7. We can see that BETO is almost equally effective at detecting *Victims* and *Threat* dimensions. In Table 10 we show some texts that were classified correctly and wrongly, respectively.

Table 6. Examples of texts correctly classified and misclassified by BETO in the Experiment I on Stereotype vs. Nonstereotype. The true label is indicated in each example. For the stereotypes, we indicate the dimension to which each sentence belongs.

Classified Examples with the Right Label
1. *Nos gustaría que lo acompañara de política migratoria real también.* (Nonstereotype) (We would like that actual immigration policy was considered as well.) 2. *Señorías, la política de integración es la gran asignatura pendiente.* (Nonstereotype) (Ladies and gentlemen, integration policy is the great pending issue.) 3. *Es decir, se tiene que cambiar la política de inmigración del Gobierno.* (Nonstereotype) (In other words, the government's immigration policy has to be changed.) 4. *El secretario de empleo dijo: España seguirá necesitando inmigrantes.* (S: Economical Resource) (The employment secretary said: Spain will continue needing immigrants.) 5. *En lo que va de año han llegado a Canarias más de 3500 personas en pateras.* (S: Threat) (So far this year, more than 3500 people have arrived in the Canary Islands in boats.) 6. *El problema, señora vicepresidenta, está en que en el cuerno de África mueren todas las semanas 40,000 niños por falta de nutrición.* (S: Victims) (The problem, Vice President, is that 40,000 children die every week in the Horn of Africa due to lack of nutrition.)
Misclassified Examples
1. *Entendemos que España puede jugar un papel destacado en cuanto a este problema, pero Europa será más creíble si afronta problemas reales que los ciudadanos perciban.* (Nonstereotype) (We understand that Spain can play a leading role in this problem, but Europe will be more credible if it faces real problems that citizens perceive.) 2. *Evidentemente nos encontramos ante una situación compleja, la relativa a las remesas en un momento en el que la política migratoria ha adquirido una gran dimensión.* (Nonstereotype) (Obviously we face with a complex situation, relating to remittances at a time when migration policy has acquired a great dimension.) 3. *Desde que aprobamos en 1985 la Ley de extranjería, de los derechos y obligaciones de los extranjeros en España, ha mantenido una línea congruente.* (Nonstereotype) (Since we approved in 1985 the Law on foreigners, on the rights and obligations of foreigners in Spain, it has maintained a congruent line.) 4. *El asunto de la inmigración requiere medidas de control pero fundamentalmente -y lo apuntaba usted ayer- medidas de solidaridad y este y este es un reto europeo.* (S: Victims) (The issue of immigration requires control measures but fundamentally—and you pointed this out yesterday—solidarity measures and this is a European challenge.) 5. *Decían que lo que pasaba en España era un coladero para los distintos países de la Unión Europea y a ustedes no les importó lo más mínimo. En aquella época el ministro.* (S: Threat) (They said that what was happening in Spain was a drain for the different countries of the European Union and you did not care at all. At that time the minister) 6. *Por tanto, se abre un camino esperanzador, y yo solamente les deseo éxitos por el bien del conjunto de los trabajadores inmigrantes, por el bien de la política en el Estado.* (S: Economical Resource) (Therefore, a hopeful path opens, and I only wish you success for the good of all immigrant workers, for the good of politics in the State)

Table 7. Highest accuracy achieved by each model in Experiment II on Victims vs. Threat. We indicate the *p*-value of the Mann–Whitney U test regarding the alternative hypothesis that the results of BETO are the highest compared to the other transformers. With * we indicate when accuracy is significantly lower than the result of BETO. The hypothesis is accepted with $p < 0.05$ only for SpanBERTa model.

BETO	**SpanBERTa**	**M-BERT**	**XLM-RoBERTa**
0.834 ± 0.034	$0.704 * \pm 0.064$ $_{p=0.00024}$	0.809 ± 0.022 $_{p=0.4965}$	0.785 ± 0.070 $_{p=0.70394}$
LR	**SVM**	**NB**	**RF**
0.79	0.78	0.72	0.77

Table 8. Bigrams and trigrams with highest mutual information with respect to each label.

	N-Grams with More Mutual Information
Victims	*ley extranjería; ceuta melilla; guardia civil; política inmigración; presión migratoria; partido popular; acoger personas; acogida canarias; acogida inmigrantes; acogida integración; acogida personas; acogida temporal; acuerdos bilaterales; administración justicia; aeropuertos fronteras; fronteras terrestres; aflorar irregulares; afrontar problema; aguas canarias; aguas territoriales; amnistía internacional; apoyo; aquellas personas; aquellos inmigrantes; aquellos países; archipiélago canario; asuntos sociales; atención humanitaria; atención sanitaria;...*
Threat	*inmigración irregular; derechos humanos; regularización masiva; inmigración ilegal; inmigrantes ilegales; proceso regularización; efecto llamada; inmigrantes irregulares; señor caldera; asilo refugio; mujeres inmigrantes; inmigración delincuencia; personas muerto; control inmigración; derecho asilo refugio; mauritania senegal; canaria nueva; derecho asilo; inmigración problema; trafican seres humanos; principal problema; inmigración clandestina;...*

Table 9. Confusion matrix of BETO in Experiment II on Victims vs. Threat.

	Predicted Labels	
	Victim	***Threat***
Victim	611	125
Threat	119	617

Table 10. Examples of texts correctly classified and misclassified by BETO in Experiment II on Victims vs. Threat. The annotators label is indicated in each example.

Classified examples with the right label
1. *¿Por qué ha muerto una persona joven?* (*Victims*) (Why did a young person die?) 2. *Derechos de ciudadanía para los inmigrantes.*(*Victims*) (Citizenship rights for immigrants.) 3. *Esta no es la forma de enfrentarse con un problema que requiere, sobre todo, grandes dosis de solidaridad.* (*Victims*) (This is not the way to deal with a problem that requires, above all, large doses of solidarity.) 4. *Hay en España más ciudadanos irregulares que nunca.* (*Threat*) (There are more irregular citizens in Spain than ever.) 5. *España hoy está desbordada con la inmigración ilegal.* (*Threat*) (Spain today is overwhelmed with illegal immigration.) 6. *Hay más llegadas de inmigrantes irregulares que nunca.* (*Threat*) (There are more arrivals of irregular immigrants than ever.)
Misclassified examples
1. *Por cierto, el Gobierno debería explicarnos cuántos inmigrantes se fugaron este fin de semana del centro de Las Raíces, si fueron veinte, como dice el delegado del Gobierno, o si fueron cien, como afirman fuentes policiales.* (Threat) (By the way, the Government should explain to us how many immigrants escaped this weekend from the center of Las Raices, if there were twenty, as the Government delegate says, or if there were a hundred, as stated by police sources.) 2. *No queremos olvidar la operación Melilla, la expulsión de los 103 ciudadanos.* (*Victims*) (We do not want to forget the Melilla operation, the expulsion of the 103 citizens.) 3. *Por tanto, apostamos por una política de retorno, de repatriación humanitaria.* (*Threat*) (Therefore, we are committed to a policy of return, of humanitarian repatriation.) 4. *Esto es un escándalo, esto son más trabas a los migrantes cuando ya se encuentran dentro.* (*Victims*) (This is a scandal, these are more obstacles to migrants when they are already inside.) 5. *Ya son 25,000 los inmigrantes llegados a Canarias en lo que va de año y se cuentan por miles los que han dejado su vida en el intento.* (*Threat*) (There are already 25,000 immigrants who have arrived in the Canary Islands so far this year and there are thousands who have lost their lives in the attempt.) 6. *¿Por qué en tres meses no han tomado ninguna de las medidas propuestas para evitar las avalanchas que han generado tanto sufrimiento?* (Threat) (Why in three months have they not taken any of the measures proposed to avoid the avalanches that have generated so much suffering?)

As we see in the misclassified examples 5 and 6 in Table 10, speakers mention that thousands of people have arrived, or specifically the term *avalanchas* (avalanches of people) but also refer, in the same sentence, to some words or phrases that are in the semantic field of compassion and victims such as *sufrimiento* (suffering) and *han dejado su vida en el intento* (have lost their lives in the attempt). We could think that fear is an emotion stronger than compassion, so humans give more weight to the part of the sentence that generates fear than to the part that generates compassion and consider that in this *frame* migrants are presented as a *threat*.

To understand better the confusion matrix shown in Tables 9 and 10, we explore the hypothesis that some parties perform a rhetorical strategy to avoid being tagged as xenophobic that consists in mentioning some expression of compassion just before or after they present immigrants as a threat, a strategy that does not convince humans' annotators but confuses the performance of transformers. As we have seen in Table 8, the presence of the n-grams *personas muertas* (death people) and *trafican seres humanos* (traffic human beings) in the dimension of *Threat* could be indicative of this rhetorical strategy that we have mentioned above: "appealing to pity and misfortune just before presenting the immigrant collective as a threat".

To explore this hypothesis, we did an analysis using the word "party" in the dataset, which indicates the party that uttered the sentence. We have 12 different parties, but we keep only those parties that have more than one hundred sentences in the dataset and we created a new category for the rest (*Other Parties*). We performed a chi-square test and a residual analysis to identify if the disagreement between human annotators and the model has any relation with the rhetorical strategy of a party. We found a significant relation (Pearson X2 = 36,979, df = 8, $p < 0.000$, see Table 10) between parties and the type of confusion, see Table 11. As it was mentioned in Section 5, for this experiment we balanced the classes to have 736 examples per class and, therefore, in this analysis we used a total of 1472 labelled examples.

Table 11. Relation between the type of confusion and political parties. Chi-square test with adjusted standardised residual.

			Annotators Say *Victims* but BETO Predicts *Threat*	**Annotators Say *Threat* but BETO Predicts *Victims***	**Annotators and BETO Agreement**	**Total**
Parties	**Coalición Canaria**	**Obs.**	14	10	127	151
		Adj. Res.	0.4	−0.7	0.2	
	IU	**Obs.**	22	2	130	154
		Adj. Res.	2.7	−3.3	0.3	
	PP	**Obs.**	11	40	330	381
		Adj. Res.	−4.6	2.0	1.9	
	PSOE	**Obs.**	42	29	324	395
		Adj. Res.	1.8	−0.6	−0.9	
	Other Parties	**Obs.**	36	38	317	391
		Adj. Res.	0.6	1.4	−1.5	
Total (Obs.)			125	119	1228	1472

With the general category, *Other Parties* and with "Coalición Canaria", a regional party from the Canary Islands, very concerned about immigration, we did not find any significant differences in the type (One type of confusion is when the annotators label is *Victims* and BETO predicted it as *Threat*, and the other type is the opposite, the annotators label is *Threat*, but BETO predicted it as *Victims*.) of confusion between humans and the model (sometimes humans label *Victim* and the model predicts *Threat* and the opposite) but in case of the right win party Partido Popular (PP) humans label more often *Threat* when the model labels is *Victim*, whereas with the left win party Izquierda Unida (IU) the type

of confusion is the opposite: humans label is *Victim* and the model labels predicts *Threat*. With PSOE, the socialist center-left party, the type of confusion goes in the same direction as with IU but is not statistically significant.

This result leads us to think that the model's confusion is based on the fact that politics use the same words for different purposes trying to avoid the label of xenophobic. This rhetorical strategy could be detected by humans that make an inference about the intentionality of the speaker, but computational models have more difficulties to detect it.

There is a great deal of research about how the human communication process occurs and which role the inference of the speaker's intentionality plays [93–95]. We think that one interesting approach to a further exploration of the computational linguistic difficulties is the one that Watzlawick and colleague suggest in their Pragmatic of Human Communication Theory [96]. They suggest that the study of human communication can be subdivided into three areas: (i) *syntactic*—problems of transmitting information (a matter of mathematical logic), (ii) *semantic*—meaning of communication (a matter of philosophy of science), and (iii) *pragmatic*—communication affecting behaviour (a matter of psychology). While a clear conceptual separation of the three areas is possible, they are nevertheless interdependent. The same act of communication can express a content (then the question is "what" is being said and, therefore, we would be in the area of syntactic or semantic) but also the same act of communication can express a personal or a social relationship (the question is "how" is being said and, therefore, we would be in the psychology area).

Following this reasoning, Watzlawick said that humans communicate both digitally and analogically. *Digital concept* refers to humans conveying meanings by using words (syntactic and semantics) and *analogical concept* refers to when humans convey relational content. For the analogical level, Watzlawick mentions nonverbal communication, posture, gesture, facial expression, voice inflection, rhythm, and cadence of words, etc. From a psychosocial point of view, we can reinterpret this definition of the analogical level of communication that Watzlawick identifies [96], in the broadest sense of the ability that allows humans to capture the level of the social relations that is behind the words. Using this human ability, people infer the speaker's intentionality in a sentence. For instance, when one speaker say "It is a humanitarian drama, more than 300 hundred boats had arrived at the Canarian Island this summer", the reader infers that for this speaker, the important part of the message is the second one, that means: "Spain can not accept more immigrants". This kind of inferences about the intention of the speakers is a natural cognition activity for humans, but it is more difficult for computational models.

7. Conclusions and Future Work

In order to advance in the study of stereotypes, for instance about immigrants, social sciences need to complement the classical paradigm that focuses on how a group is defined in terms of personal attributes with a new paradigm that emphasises the *frames*, that is, the narrative scenarios, in which a group (e.g., immigrants) is mentioned. In this work, we developed a taxonomy to identify stereotypes about immigrants not from the attributes that are assigned to the group but from the narrative scenarios in which the speaker places the group (**RQ1**). This fine grained taxonomy allows for the study of stereotypes about immigrants and covers the whole spectrum of the stereotype: from positive images of the group (as equals to the nationals or victims of xenophobia) to more negative images (a threat to the nationals or a less human people).

Based on psychosocial research on prejudice and stereotypes, this language-independent taxonomy is a new conceptual instrument with two objectives: (i) to provide computational linguistics with a new conceptual tool to detect and mitigate social bias in datasets, specifically, stereotypes about immigrants; and (ii) to strengthen the collaboration between social sciences and computational linguistics to understand better how stereotypes are generated in the context of public discourse.

We have validated our taxonomy considering how each category is related with some attitude (pro-immigrant, anti-immigrant, or neutral). We have identified two opposite

supracategories of the stereotypes about immigrants: one that presents the minority as *victim* and the other that presents the minority as a *threat*. We annotated political speeches of the ParlSpeech V2 dataset, focusing on the speeches from 1996 to 1998, from 2006 to 2008, and from 2016 to 2018. The resulting *StereoImmigrants* stereotype-annotated dataset was created relying on the new taxonomy (**RQ2**). The dataset will be made public to the research community to foster the research on stereotypes about immigrants. *StereoImmigrants* was used to carry out some preliminary experiments using state-of-the-art transformer models and classical machine learning models. We obtained results between 0.70 and 0.86 of accuracy in the two experiments we performed: Stereotype vs. Nonstereotype and Victims vs. Threat. The best performance was obtained by BETO, a monolingual Spanish transformer model, suggesting that this model could be capturing a richer representation of stereotypes and their dimensions, than the classical machine learning models do (**RQ3**). We also point up that with these preliminary experiments we prove the existence of social bias, in particular stereotypes about immigrants, in political speeches, and the effectiveness of automatically detecting them.

From the most relevant n-grams from the examples labelled as *Stereotype*, we confirmed that they are not trivially associated with the immigrant group (they are not always biased attributes) and, therefore, we should not rely only on a set of keywords to represent it. Additionally, we confirmed that in the study of stereotypes about immigrants, we have to consider not to define two opposite social groups, since the nonstereotypical texts are neutral phrases talking about immigration in general without stereotyping at all.

We have analysed the confusion matrices considering metadata from our *StereoImmigrants* dataset, in particular, the political party the speaker belongs to. This analysis has shown that the confusion could be explained because certain rhetorical strategies are particularly difficult to infer for transformers and not for humans. When speakers use the same words for different purposes, humans elaborate the different meanings of the sentence making an inference about the intention of the speaker, but transformers have difficulties inferring the rhetorical strategies. More work is necessary in this direction.

The taxonomy, although used to label text in Spanish, is applicable to other languages because it classifies into categories the different *frames* in which immigrants are placed. These *frames* are common in Western cultures. These categories express the dimensions of the stereotype about immigrants on the north area of a north-south axis of economical inequality. Moreover, we aim to apply this taxonomy on the Europarl corpus, as well as on texts of different genres like newspaper datasets.

As future work, we could test if these two big dimensions of immigrant stereotype (*Victims* and *Threat*) could be applied to other minorities' stereotypes as feminist or LGTBI people. In fact, these minorities are presented also as a *threat*, for instance, when feminist women are presented in scenarios that emphasise conflict and then are defined as *feminazis*. In addition, feminist women are presented as "victims" when the narrative context emphasises gender violence. A general idea for future work will be that the study of minorities' stereotype needs these two dimensions (*Victims* vs. *Threat*) to complement the well established Stereotype Content Model [49] that proposes *warmth* and *competence* as two fundamental dimensions of stereotypes.

Furthermore, we plan to analyse how social bias (in particular stereotypes) is reflected in the attention values of the transformer layers, in order to facilitate the explainability of the results and a further debiasing process. Moreover, we think that it will be interesting to enrich the dataset with more examples of each stereotype category for evaluating the multiclass classification by using the transformer models (e.g., BETO).

Author Contributions: Conceptualisation: B.C., P.R., S.P.P.; Methodology: B.C., P.R., S.P.P.; Implementation: J.S.-J.; Software: J.S.-J.; Validation: J.S.-J.; Investigation: J.S.-J., B.C.; Formal Analysis: J.S.-J., B.C., S.P.P.; Resources: J.S.-J., B.C.; Data Curation: B.C.; Writing—Original Draft Preparation: J.S.-J., B.C.; Writing—Review and Editing: P.R., S.P.P.; Visualisation: J.S.-J.; Supervision: P.R., S.P.P.; Project Administration: P.R.; Funding Acquisition: P.R. All authors have read and agreed to the published version of the manuscript.

Funding: The work of the authors from the Universitat Politècnica de València was funded by the Spanish MICINN under the research project MISMIS-FAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31). The work of Paolo Rosso was done also in the framework of the research project PROMETEO/2019/121 (DeepPattern) funded by the Generalitat Valenciana.

Data Availability Statement: The dataset is available in github.com/jjsjunquera/StereoImmigrants, accessed on 8 April 2021.The *StereoImmigrants* dataset was created by filtering and processing the ParlSpeech dataset (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/L4OAKN), accessed on 8 April 2021, focusing on the speeches from 1996 to 1998, from 2006 to 2008, and from 2016 to 2018. We include in our data the speech, speaker, speaker's party, the speech's date, and the annotation given by the five annotators regarding the taxonomy and the stereotype's attitude (i.e., pro-immigrant, anti-immigrant, and ambivalent).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BERT	Bidirectional Encoder Representations from Transformers
GSR	Gender Stereotype Reinforcement
LR	Logistic Regression
M-BERT	Multilingual BERT
ML	Machine Learning
NB	Naïve Bayes
NER	Name Entity Recognition
NLI	Natural Language Inference
NLP	Natural Language Processing
RF	Random Forest
SCM	Stereotype Content Model
SVM	Support Vector Machine

Appendix A. Taxonomy: Categories and Frames

Annotators should use the *frames* (labels in two digits) to label the texts, if it is possible; in other case, the label should be the category (label in one digit). For example, if the annotator recognises that a target text is saying that immigrants have the same rights as the nationals, she should label the text using the *frame 1.1* (from category 1). However, if the annotator identifies that the text belongs to category 1, but she cannot specify the *frame*, then the label to put should be *1*.

If one text contains fragments that correspond to different *frames* from the same category, the label should be at one digit level: the one corresponding to the category.

If one text contains fragments that correspond to different *frames* from different categories, we ask the annotators to choose the most important or to discard the sentence.

Appendix A.1. Category 1: Xenophobia's Victims

1.1 With the same rights and with the same obligations. They are named as *ciudadanos*(citizens), *nueva ciudadanía* (new citizenship), etc.

1.2 They are presented doing a simile with the Spaniards who emigrated.

1.3 It is suggested that the problem is the racist or xenophobic attitudes of people.

1.4 It is claimed that immigration topic is used as an electoral or partisan weapon and that this is not right. Some party is accused of being racist and/or xenophobic. The rise of racist or xenophobic parties is seen as a problem.

1.5 It is stated that immigration is not a problem for coexistence. The population is not concerned about the presence of immigrants and the problem is xenophobia or racism.

1.6 Immigration is considered to bring cultural diversity, pluralism, etc. and that is positive for the country.

Appendix A.2. Category 2: Suffering Victims

2.1 Victims of suffering and poverty. It is argued that poverty and suffering in their countries of origin is the cause of the immigration. In addition, they are victims of suffering once they are here.

2.2 Victims of injustice and discrimination. Victims of labour exploitation and mafias. It is reported that they do not respect human rights in the treatment of immigrants or it is stated that they have to be respected.

2.3 Solidarity is required or manifested in the face of immigrant problems.

2.4 It is suggested that they die trying to get there (Spain in this work), for example, there is talk about the rescues.

Appendix A.3. Category 3: Economical Resource

3.1 They do the jobs that the Spanish do not want to do. They support the black economy. They are seen as workers in a situation of vulnerability, with special difficulties.

3.2 They bring economic prosperity: they pay taxes, send remittances abroad, etc.

3.3 They solve the problem of lack of population.

3.4 They propose measures to hire immigrant workers in their countries of origin: they must be approved, etc.

3.5 The entry of immigrants must be regulated according to the needs of the labour market.

Appendix A.4. Category 4: Collective Threat

4.1 They come in droves and create a situation of chaos. It could be mentioned *avalancha* (avalanche), *falta de control* (lack of control), *llegada a las costas* (arrival at the coast), and so on.

4.2 The problem is that they are illegal. They refer to them as *ilegales* (illegal) or *irregulares* (irregular) or using the category *inmigrantes* (immigrants) or *inmigracion* (immigration). It could be mentioned *repatriaciones* (repatriations), *devoluciones* (returns), or *expulsiones* (expulsions).

4.3 It is stated that immigration is a problem for the host society, causing imbalances in coexistence of the group.

Appendix A.5. Category 5: Personal Threat

5.1 It is argued that immigrants compete with the country's population for resources such as work, health services, and education. Immigration remains as a problem with regard to the use of these resources.

5.2 Immigrants are reported to bring diseases or are referred to as carriers of new diseases.

5.3 Immigration is associated with crime.

Appendix A.6. Category 6: Dehumanisation

6.1 They do not know how to live as human beings do.

6.2 They behave like animals.

6.3 Their deaths are not our problem: they come because they want to.

Appendix B. Keywords Used to Filter Immigration-Related Speeches

The keywords shown in Table A1 were used to discard those speeches that were not talking about immigration as a central topic. These keywords were carefully defined by a social psychologist who payed attention to the important historical events that occurred during the periods of the speeches.

Table A1. Keywords used to filter relevant speeches.

anti-inmigrante	*deportado*	*inmigración*	*pateras*
anti-inmigrantes	*deportados*	*inmigrante*	*permisos de residencia*
asilada	*deportar*	*inmigrantes*	*polizones*
asiladas	*desheredados de la tierra*	*islamofobia*	*racismo*
asilado	*devolución*	*migrantes*	*racista*
asilados	*efecto llamada*	*migratoria*	*refugiada*
centro de acogida	*efecto salida*	*migratorias*	*refugiadas*
centros de acogida	*emigrantes*	*migratorio*	*refugiado*
ciudadanía inmigrada	*etnocentrismo*	*multiculturalismo*	*refugiados*
ciudadano emergente	*expatriada*	*nativismo*	*repatriación*
ciudadanos emergentes	*expatriadas*	*nuevas ciudadanas*	*schengen*
colonialismo	*expatriado*	*nuevos ciudadanos*	*sociedad de acogida*
deportación	*expatriados*	*países de recepción*	*xenófoba*
deportada	*extranjería*	*países emisores*	*xenofobia*
deportadas	*indocumentados*	*países en tránsito*	*xenófobo*

References

1. Brown, R. *Prejudice Its Social Psychology*; Wiley-Blackwell: New York, NY, USA, 2010.
2. Shu, K.; Sliva, A.; Wang, S.; Tang, J.; Liu, H. Fake news detection on social media: A data mining perspective. *ACM Sigkdd Explor. Newsl.* **2017**, *19*, 22–36. [CrossRef]
3. Sánchez-Junquera, J.; Rosso, P.; Montes-y Gómez, M.; Ponzetto, S.P. Unmasking Bias in News. *arXiv* **2019**, arXiv:1906.04836.
4. Zubiaga, A.; Aker, A.; Bontcheva, K.; Liakata, M.; Procter, R. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.* **2018**, *51*, 1–36. [CrossRef]
5. Sánchez-Junquera, J.; Villaseñor-Pineda, L.; Montes-y Gómez, M.; Rosso, P.; Stamatatos, E. Masking domain-specific information for cross-domain deception detection. *Pattern Recognit. Lett.* **2020**, *135*, 122–130. [CrossRef]
6. Nadeem, M.; Bethke, A.; Reddy, S. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv* **2020**, arXiv:2004.09456.
7. Steele, C.M.; Aronson, J. Stereotype threat and the intellectual test performance of African Americans. *J. Personal. Soc. Psychol.* **1995**, *69*, 797–811. [CrossRef]
8. Desombre, C.; Jury, M.; Bagès, C.; Brasselet, C. The Distinct Effect of Multiple Sources of Stereotype Threat. *J. Soc. Psychol.* **2018**, *159*, 1–14. [CrossRef]
9. Desombre, C.; Jury, M.; Renard, M.; Brasselet, C. Validation factorielle d'une mesure des menaces du stéréotype en langue française. *L'Année Psychol.* **2020**, *120*, 251–269. [CrossRef]
10. Smith, M.B. The Open and Closed Mind. Investigations into the nature of belief systems and personality systems. *Science* **1960**, *132*, 142–143. [CrossRef]
11. Kossowska, M.; Czernatowicz-Kukuczka, A.; Sekerdej, M. Many faces of dogmatism: Prejudice as a way of protecting certainty against value violators among dogmatic believers and atheists. *Br. J. Psychol.* **2017**, *108*, 127–147. [CrossRef]
12. Tajfel, H.; Sheikh, A.A.; Gardner, R.C. Content of stereotypes and the inference of similarity between members of stereotyped groups. *Acta Psychol.* **1964**, *22*, 191–201. [CrossRef]
13. Dovidio, J.F.; Hewstone, M.; Glick, P.; Esses, V.M. Prejudice, stereotyping and discrimination: Theoretical and empirical overview. *SAGE Handb. Prejud. Stereotyping Discrim.* **2010**, *80*, 3–28.
14. Bergsieker, H.B.; Leslie, L.; Constantine, V.S.; Fiske, S. Stereotyping by omission: Eliminate the negative, accentuate the positive. *J. Personal. Soc. Psychol.* **2012**, *102*, 1214–1238. [CrossRef] [PubMed]
15. Glick, P.; Fiske, S. The Ambivalent Sexism Inventory: Differentiating Hostile and Benevolent Sexism. *J. Personal. Soc. Psychol.* **1996**, *70*, 491–512. [CrossRef]
16. Lipmann, W. *Public Opinion*; Harcourt Brace: New York, NY, USA, 1922.
17. Kinder, D. Opinion and action in the realm of politics. In *The Handbook of Social Psychology*; Gilbert, D.T., Fiske, S., Lindzey, G., Eds.; Wiley: New York, NY, USA, 1998; Volume 2, pp. 778–867.
18. Scheufele, D.A. Framing as a Theory of Media Effects. *J. Commun.* **2006**, *49*, 103–122. [CrossRef]
19. Pan, Z.; Kosicki, G. Framing Analysis: An Approach to News Discourse. *Political Commun.* **1993**, *10*, 55–75. [CrossRef]
20. Bateson, G. Ecology of Mind. *Psychyatric Res. Rep.* **1955**, *2*.
21. Tversky, A.; Kahneman, D. The framing of decisions and the psychology of choice. *Science* **1981**, *211*, 453–458. [CrossRef] [PubMed]
22. Goffman, E. *Frame Analysis*; Harper & Row: New York, NY, USA, 1974.
23. Gamson, W.; Modigliani, A. The Changing Culture of Affirmative Action. In *Research in Political Sociology*; Braungart, R., Ed.; JAI Press Inc.: Greenwich, CT, USA, 1987; Volume 3, pp. 137–177.
24. Lakoff, G. *Don't Think of an Elephant! Know Your Values and Frame the Debate*; Chelsea Green Publishing: Vermont, VT, USA, 2004.

25. Bruner, J. *Acts of Meaning*; Harvard University Press: Cambridge, MA, USA, 1990.
26. Baeza-Yates, R. Bias on the web. *Commun. ACM* **2018**, *61*, 54–61. [CrossRef]
27. Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; Stein, B. A stylometric inquiry into hyperpartisan and fake news. *arXiv* **2017**, arXiv:1702.05638.
28. Bolukbasi, T.; Chang, K.W.; Zou, J.Y.; Saligrama, V.; Kalai, A.T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 4349–4357.
29. Lauscher, A.; Glavaš, G.; Ponzetto, S.P.; Vulić, I. A general framework for implicit and explicit debiasing of distributional word vector spaces. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 8131–8138.
30. Sanguinetti, M.; Poletto, F.; Bosco, C.; Patti, V.; Stranisci, M. An italian twitter corpus of hate speech against immigrants. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
31. Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; Smith, N.A. The Risk of Racial Bias in Hate Speech Detection. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Stroudsburg, PA, USA; pp. 1668–1678.
32. Saleem, H.M.; Dillon, K.P.; Benesch, S.; Ruths, D. A web of hate: Tackling hateful speech in online social spaces. *arXiv* **2017**, arXiv:1709.10159.
33. la Peña Sarracén, G.L.D.; Rosso, P. Aggressive Analysis in Twitter using a Combination of Models. In Proceedings of the Iberian Languages Evaluation Forum Co-Located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, 24 September 2019.
34. Glavaš, G.; Karan, M.; Vulić, I. XHate-999: Analyzing and Detecting Abusive Language Across Domains and Languages. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 6350–6365.
35. Shekhar, C.; Bagla, B.; Maurya, K.K.; Desarkar, M.S. Walk in Wild: An Ensemble Approach for Hostility Detection in Hindi Posts. *arXiv* **2021**, arXiv:2101.06004.
36. Manzini, T.; Lim, Y.C.; Black, A.W.; Tsvetkov, Y. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. In *Long and Short Papers, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019*; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; Volume 1, pp. 615–621.
37. Anzovino, M.; Fersini, E.; Rosso, P. Automatic Identification and Classification of Misogynistic Language on Twitter. In *Natural Language Processing and Information Systems*; Silberztein, M., Atigui, F., Kornyshova, E., Métais, E., Meziane, F., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 57–64.
38. Kyriakou, K.; Barlas, P.; Kleanthous, S.; Otterbacher, J. Fairness in proprietary image tagging algorithms: A cross-platform audit on people images. In Proceedings of the International AAAI Conference on Web and Social Media, Munich, Germany, 11–14 June 2019; pp. 313–322.
39. Barlas, P.; Kyriakou, K.; Guest, O.; Kleanthous, S.; Otterbacher, J. To "See" is to Stereotype: Image Tagging Algorithms, Gender Recognition, and the Accuracy-Fairness Trade-off. *Proc. ACM Hum. Comput. Interact.* **2021**, *4*, 1–31. [CrossRef]
40. Beukeboom, C.J.; Forgas, J.; Vincze, O.; Laszlo, J. Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. *Soc. Cogn. Commun.* **2014**, *31*, 313–330.
41. Garg, N.; Schiebinger, L.; Jurafsky, D.; Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E3635–E3644. [CrossRef] [PubMed]
42. Liang, P.P.; Li, I.M.; Zheng, E.; Lim, Y.C.; Salakhutdinov, R.; Morency, L.P. Towards debiasing sentence representations. *arXiv* **2020**, arXiv:2007.08100.
43. Fabris, A.; Purpura, A.; Silvello, G.; Susto, G.A. Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Inf. Process. Manag.* **2020**, *57*, 102377. [CrossRef]
44. Cryan, J.; Tang, S.; Zhang, X.; Metzger, M.; Zheng, H.; Zhao, B.Y. Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–11.
45. Dev, S.; Li, T.; Phillips, J.M.; Srikumar, V. On Measuring and Mitigating Biased Inferences of Word Embeddings. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 7659–7666.
46. Fokkens, A.; Ruigrok, N.; Beukeboom, C.; Sarah, G.; Van Atteveldt, W. Studying muslim stereotyping through microportrait extraction. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
47. Manuela, S.; Gloria, C.; Di Nuovo, E.; Frenda, S.; Stranisci, M.A.; Bosco, C.; Tommaso, C.; Patti, V.; Irene, R. HaSpeeDe 2@ EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task. In Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Turin, Italy, 12–13 December 2020; pp. 1–9.
48. Fiske, S.; Xu, J.; Cuddy, A.; Glick, P. (Dis)respecting versus (Dis)liking: Status and Interdependence Predict Ambivalent Stereotypes of Competence and Warmth. *J. Soc. Issues* **1999**, *55*, 473–489. [CrossRef]

49. Fiske, S.T.; Cuddy, A.J.C.; Glick, P.; Xu, J. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *J. Personal. Soc. Psychol.* **2002**, *82*, 878–902. [CrossRef]
50. Fiske, S.T.; Cuddy, A.J.C.; Glick, P. Universal dimensions of social perception: Warmth and competence. *Trends Cogn. Sci.* **2007**, *11*, 77–83. [CrossRef]
51. Cuddy, A.; Fiske, S.; Glick, P. The BIAS Map: Behaviors from Intergroup Affect and Stereotypes. *J. Personal. Soc. Psychol.* **2007**, *92*, 631–648. [CrossRef] [PubMed]
52. Fiske, S.T. Stereotype content: Warmth and Competence Endure. *Curr. Dir. Psychol. Sci.* **2018**, *27*, 67–73. [CrossRef] [PubMed]
53. Lee, T.L.; Fiske, S.T. Not an outgroup, not yet an ingroup: Immigrants in the Stereotype Content Model. *Int. J. Intercult. Relations* **2006**, *30*, 751–768. [CrossRef]
54. Kervyn, N.; Fiske, S.; Yzerbyt, V. Forecasting the primary dimension of social perception: Symbolic and realistic threats together predict warmth in the Stereotype Content Model. *Soc. Psychol.* **2015**, *46*, 36–45. [CrossRef] [PubMed]
55. University of Oklahoma; Institute of Group Relations; Sherif, M. *Intergroup Conflict and Cooperation: The Robbers Cave Experiment*; University Book Exchange: Norman, OK, USA, 1961; Volume 10.
56. Zanna, M. On the nature of prejudice. *Can. Psychol.* **1994**, *35*, 11–23. [CrossRef]
57. Stephan, W.G.; Diaz-Loving, R.; Duran, A. Integrated Threat Theory and Intercultural Attitudes: Mexico and the United States. *J. Cross Cult. Psychol.* **2000**, *31*, 240–249. [CrossRef]
58. McConahay, J.B.; Hough, J.C. Symbolic Racism. *J. Soc. Issues* **1976**, *32*, 23–45. [CrossRef]
59. Kinder D.R.; Sears, D.O. Prejudice and politics: Symbolic racism versus racial threats to the good life. *J. Personal. Soc. Psychol.* **1981**, *40*, 414–431. [CrossRef]
60. Moscovici, S. The coming era of representations. In *Cognitive Analysis of Social Behaviour*; Codol, J., Leyens, J., Eds.; Nijhoff: The Hague, The Netherlands, 1982.
61. Moscovici, S. The phenomenon of social representation. In *Social Representations*; Farr, R., Moscovici, S., Eds.; Cambridge University Press: Cambridge, NY, USA, 1984; pp. 3–69.
62. Breakwell, G.; Canter, D. *Empirical Approaches to Social Representations*; Oxford Science Publications; Clarendom Press: Oxford, UK, 1993.
63. Moscovici, S. *Essai Sur l'histoire Humaine de la Nature*; Flammarion: Paris, France, 1968.
64. Pérez, J.; Moscovici, S.; Chulvi, B. Nature and culture as principles for social classification. Anchorage of social representations on ethnical minorities. *Int. J. Soc. Psychol.* **2002**, *17*, 51–67. [CrossRef]
65. Perez, J.; Moscovici, S.; Chulvi, B. The taboo against group contact: Hypothesis of Gypsy ontologization. *Br. J. Soc. Psychol.* **2007**, *46*, 249–272. [CrossRef] [PubMed]
66. Haslam, N.; Loughnan, S.; Reynolds, C.; Wilson, S. Dehumanization: A New Perspective. *Soc. Personal. Psychol. Compass* **2007**, *1*, 409–422. [CrossRef]
67. Leyens, J.P.; Paladino, P.M.; Rodriguez-Torres, R.; Vaes, J.; Demoulin, S.; Rodriguez-Perez, A.; Gaunt, R. The emotional side of prejudice: The attribution of secondary emotions to ingroups and outgroups. *Personal. Soc. Psychol. Rev.* **2000**, *4*, 186–197. [CrossRef]
68. Moscovici, S.; Pérez, J. A study of minorities as victims. *Eur. J. Soc. Psychol.* **2007**, *37*, 725–746. [CrossRef]
69. Moscovici, S.; Pérez, J.A. A new representation of minorities as victims. In *Coping with Minority Status: Responses to Exclusion and Inclusion*; Butera, F., Levine, J.M., Eds.; Cambridge University Press: Cambridge, NY, USA, 2009; pp. 82–103.
70. Barkan, E. *The Guilt of Nations: Restitution and Negotiating Historical Injustices*; Norton: New York, NY, USA, 2000.
71. Sironi, A.; Bauloz, C.; Emmanuel, M. *Glossary on Migration*; International Organization for Migration (IOM): Geneva, Switzerland, 2019.
72. Steele, S. *The Content of our Character*; St. Martin's Press: New York, NY, USA, 1990.
73. Gamson, W. *Talking Politics*; Cambridge University Press: New York, NY, USA, 1992.
74. Sap, M.; Gabriel, S.; Qin, L.; Jurafsky, D.; Smith, N.A.; Choi, Y. Social Bias Frames: Reasoning about Social and Power Implications of Language. *arXiv* **2020**, arXiv:1911.03891.
75. Eagly, A.H.; Chaiken, S. *The Psychology of Attitudes*; Harcourt Brace Jovanovich College Publishers: Brussels, Belgium, 1993.
76. Kaplan, K.J. On the ambivalence-indifference problem in attitude theory and measurement: A suggested modification of the semantic differential technique. *Psychol. Bull.* **1972**, *77*, 361–372. [CrossRef]
77. Rauh, C.; Schwalbach, J. The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies. *Harv. Dataverse* **2020**. [CrossRef]
78. Rudkowsky, E.; Haselmayer, M.; Wastian, M.; Jenny, M.; Emrich, Š.; Sedlmair, M. More than bags of words: Sentiment analysis with word embeddings. *Commun. Methods Meas.* **2018**, *12*, 140–157. [CrossRef]
79. Proksch, S.O.; Lowe, W.; Wäckerle, J.; Soroka, S. Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legis. Stud. Q.* **2019**, *44*, 97–131. [CrossRef]
80. Reher, D.S.; Cortés, L.; González, F.; Requena, M.; Sánchez, M.I.; Sanz, A.; Stanek, M. *Informe Encuesta Nacional de Inmigrantes (ENI—2007)*; Instituto Nacional de Estadística: Buenos Aires, Argentina, 2008.
81. CES. *La Inmigración y el Mercado de Trabajo en España*; Colección Informes; Consejo Económico y Social: Madrid, Spain, 2004.
82. Izquierdo Escribano, A. *Inmigración: Mercado de Trabajo y Protección Social en España*; Consejo Económico y Social: Madrid, Spain, 2003.

83. Sharpe, D. Chi-Square Test is Statistically Significant: Now What? *Pract. Assess. Res. Eval.* **2015**, *20*. [CrossRef]
84. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
85. Wang, Z.; Mayhew, S.; Roth, D. Cross-lingual ability of multilingual bert: An empirical study. *arXiv* **2019**, arXiv:1912.07840.
86. Pires, T.; Schlinger, E.; Garrette, D. How multilingual is multilingual bert? *arXiv* **2019**, arXiv:1906.01502.
87. Cañete, J.; Chaperon, G.; Fuentes, R.; Ho, J.H.; Kang, H.; Pérez, J. Spanish Pre-Trained BERT Model and Evaluation Data. In Proceedings of the ICLR 2020, Addis Ababa, Ethiopia, 30 April 2020.
88. Úbeda, P.L.; Díaz-Galiano, M.C.; Lopez, L.A.U.; Martín-Valdivia, M.T.; Martín-Noguerol, T.; Luna, A. Transfer learning applied to text classification in Spanish radiological reports. In Proceedings of the LREC 2020 Workshop on Multilingual Biomedical Text Processing (MultiligualBIO 2020), Marseille, France, 16 May 2020; pp. 29–32.
89. Tiedemann, J. Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, 23–25 May 2012; pp. 2214–2218.
90. Casavantes, M.; López, R.; González, L. UACh at MEX-A3T 2020: Detecting Aggressive Tweets by Incorporating Author and Message Context. In Proceedings of the 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain, 23 September 2020.
91. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
92. Menger, V.; Scheepers, F.; Spruit, M. Comparing Deep Learning and Classical Machine Learning Approaches for Predicting Inpatient Violence Incidents from Clinical Text. *Appl. Sci.* **2018**, *8*, 981. [CrossRef]
93. Jones, E.E.; Davis, K.E. From acts to dispositions: The attribution process in social perception. In *Advances in Experimental Social Psychology*; Academic Press: New York, NY, USA, 1965; Volume 2, pp. 219–266.
94. Jones, E.E.; Harris, V.A. The attribution of attitudes. *J. Exp. Soc. Psychol.* **1967**, *3*, 1–24. [CrossRef]
95. Jones, E.E. The rocky road from acts to dispositions. *Am. Psychol.* **1979**, *34*, 107–117. [CrossRef]
96. Watzlawick, P.; Beavin, J.; Jackson, D. *Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies, and Paradoxes*; W. W. Norton & Company Incorporated: New York, NY, USA, 1967.